INSTITUTO TECNOLÓGICO DE BUENOS AIRES – ITBA

ESCUELA DE (INGENIERÍA Y TECNOLOGÍA – INGENIERÍA Y GESTIÓN - POSTGRADO)

Análisis del Sistema Ecobici en la Ciudad de Buenos Aires

AUTOR/ES: Calvo, Mario Daniel

TUTOR/ES: Aizemberg, Ariel

TRABAJO FINAL PRESENTADO PARA LA OBTENCIÓN DEL TÍTULO DE:

ESPECIALISTA EN CIENCIAS DE DATOS

BUENOS AIRES

PRIMER CUATRIMESTRE, 2018

Abstract

Los sistemas públicos de bicicletas se han extendido en todas las grandes ciudades el mundo. Buenos Aires ha desarrollado uno y en este estudio buscamos entender cuáles podrían ser los factores espaciales y ambientales, y el perfil de los usuarios qué expliquen el uso del sistema. Para ello se utilizaron datos provistos por el gobierno de CABA, respecto de los viajes realizados, y datos del clima. Se aplicaron métodos de minería de datos tales como: regresiones para identificar relaciones entra las variables y Kmeans, para clasificar datos.

Los hallazgos más importantes fueron el efecto de la temperatura sobre el uso del sistema, y la identificación del grupo etario que más viajes realiza. La diferencia significativa entre el uso de los días de semana y los fines de semana, y su correspondencia con la hora del día. Estos hallazgos permitieron inferir que el uso en los días de semana se debe al traslado hacia el trabajo y/o centros de estudio. En tanto que el uso en los fines de semana se orienta a un uso recreativo.

Contenido

Introducción

Los sistemas de públicos de bicicletas desarrollados en las grandes ciudades, similares al Ecobici, se proponen resolver la necesidad de cubrir el recorrido entre puntos del sistema de transporte y destinos requeridos por quienes transitan por las áreas más congestionadas de la ciudad.
EL sistema Ecobici se suma a la red de transportes públicos de la ciudad. Y a diferencia del resto de los transportes públicos de la ciudad, no hay rutas prefijadas, ni tablas de horarios. Esto ofrece más libertad a los usuarios, y se ajusta mejor a sus necesidades. El comportamiento de los usuarios del sistema es diferente para los distintos grupos de usuarios.

Aunque las estaciones de Ecobici se distribuyen aleatoriamente a través de la red, una razón para explicar su uso podría ser los factores espaciales y ambientales que se superponen en estas estaciones. Es por ello que se analizaron diferentes aspectos

- Análisis basado en el clima (temperatura)
- Análisis basado en los viajes (hora del día)
- Análisis basado en las estaciones (población y distancia al metro)
- Análisis basado en el usuario (edad y sexo)

Los hallazgos presentados en el análisis muestran que los motivos del uso de las estaciones de bicicletas son complejas y diversas. La presunción de que los patrones de actividad y la ubicación de la estación se correlacionan parecía ser prometedor. Sin embargo los hallazgos más significativos fueron en relación con atributos propios de los usuarios del sistema.

Antecedentes

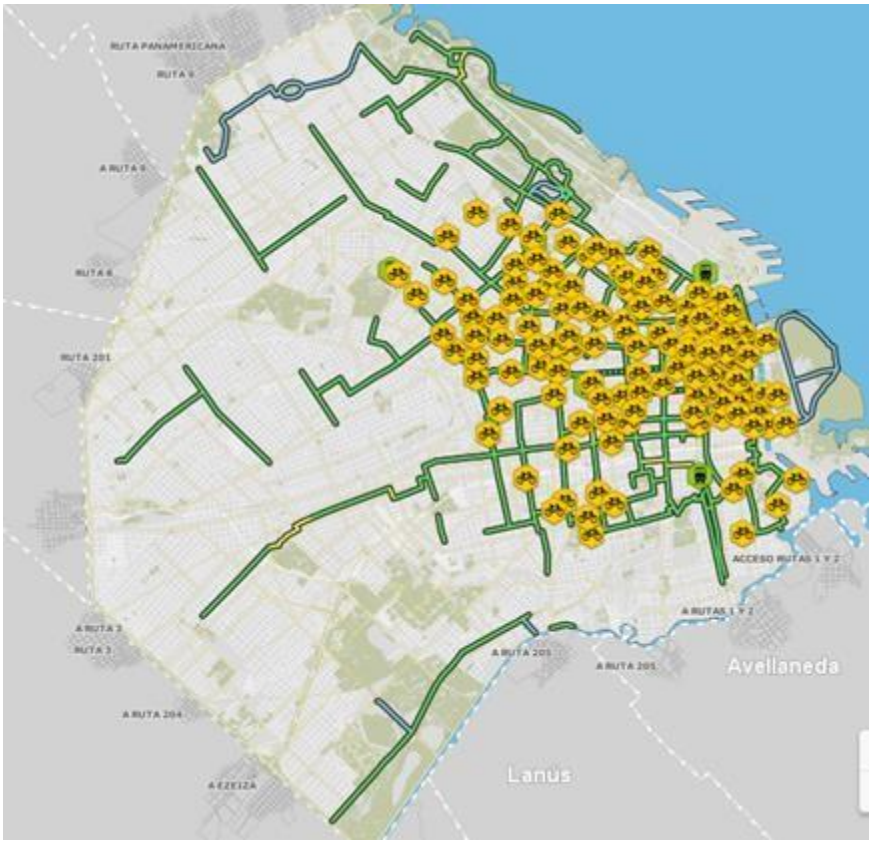
La Ciudad de Buenos Aires desarrolló un Sistema de Transporte Público en Bicicletas, el cual se denomina "**Ecobici**". El sistema consta de 2 componentes principales:

- Una red de ciclo vías
- Una red de estaciones de bicicletas.

La red de ciclo vías se desarrolló a partir de 2009. En la actualidad dicha red alcanza los 165 km. La red fue especialmente diseñada para integrar distintos puntos estratégicos de la ciudad como centros de transbordo, universidades, escuelas y hospitales permitiendo también la interconexión con otros medios de transporte.

Red de estaciones de bicicletas: las estaciones comenzaron siendo manuales, con la atención de una persona dentro de un rango horario acotado de 8 a 20. Luego se agregaron estaciones semiautomáticas con apoyo, y finalmente con la incorporación de tecnología se agregaron estaciones automáticas. Esto permite que las estaciones funcionen las 24 horas del día, todos los días del año. Se estima finalizar el año 2017 con 200 estaciones.

En la actualidad el sistema permite hacer un registro online o presencial del usuario. Luego de completar el registro, y obtener su tarjeta, el usuario podrá de manera gratuita retirar una bici en una de las estaciones, usarla y devolver la bici en cualquiera de las estaciones del sistema. Se dispone de una aplicación para teléfonos celulares que permite consultar de manera online la disponibilidad de bicicletas en cada estación. En el siguiente mapa podemos observar la red de ciclo vías existentes.



Definición del problema

Determinar qué factores inciden en el uso del sistema “Ecobici”

Justificación del estudio

Aplicar técnicas de ciencia de datos sobre datos públicos que permitan comparar con resultados producidos por otras publicaciones obtenidos con otros métodos.

Alcances del trabajo y limitaciones

Limitaciones

El gobierno de la ciudad lleva un registro del uso del sistema desde su inicio. Estos datos están disponibles en la página del gobierno de la ciudad. Se dispone de los datos de uso del sistema desde 2010 hasta el 2017. Además se dispone de otro conjunto de datos con el registro de la temperatura con registro horario de cada día del año, para los años 2011 y 2012.
Tenemos algunas limitaciones que surgen de los datos. La primera es que solo tenemos correspondencia de 2 años entre los conjuntos de datos de viajes y clima. La segunda limitación es la no completitud del conjunto de datos del clima. La tercera limitación es que el formato de registro de los viajes fue cambiando en los diferentes años. Solo en el año 2017 se incorporó algo de información respecto del perfil de los usuarios, como ser: edad y sexo.

Alcance:
Generar un análisis del sistema Ecobici, similar a los análisis existentes sobre otros sistemas de bicicletas públicas en otras ciudades del mundo.

Objetivos

Objetivo general

Determinar si los factores climáticos, horarios, geográficos y el grupo etario inciden en el uso del sistema “Ecobici”.

Objetivos específicos

1. Hacer profiling de los datos, para detectar problemas de calidad de los datos en los conjuntos de datos.
2. Aplicar técnicas de limpieza y/o enriquecimiento de los datos
3. Verificar si existe relación entre la variable de uso del sistema y la temperatura.
4. Verificar si existe relación entre la variable de uso del sistema y la hora del día.
5. Verificar si existe relación entre la variable uso del sistema y la cantidad de población cercana a la estación de Ecobici
6. Verificar si existe relación entre la variable uso del sistema y la edad de los usuarios.

Hipótesis

1. El uso de las bicicletas es mayor cuando la temperatura está entre los 15 y los 25 grados.
2. El uso de las bicicletas es mayor en el horario pico.
3. Las estaciones de Ecobici más usadas dependen de la densidad de población del barrio.
4. Las estaciones de Ecobici más usadas están cercanas a las estaciones del metro.
5. El sistema es más utilizado por los usuarios menores de 30 años.

Para verificar la hipótesis se utilizarán las siguientes variables independientes:

Temperatura (contenidas en el conjunto de datos con datos del clima).
Día y hora (contenidas en el conjunto de datos del uso del sistema).
Distancia entre las estaciones de subte y Ecobici. (google maps)
Población alrededor de las estaciones de Ecobici (censo 2010)
Edad de los usuarios.

Analizaremos las siguientes variables dependientes:

Cantidad de viajes.

Tiempo de viaje

Descripción de datos

Origen de los datos:

Los conjuntos de datos fueron obtenidos de los siguientes sitios:

Sistema de Ecobici

https://data.buenosaires.gob.ar/dataset/bicicletas-publicas

Registro de temperatura de Bs As.

https://data.buenosaires.gob.ar/dataset/informacion-meteorologica

Censo

https://www.indec.gov.ar/ftp/cuadros/territorio/codgeo/Codgeo_CABA_con_datos.zip

Estaciones de subte

https://data.buenosaires.gob.ar/dataset/subte-estaciones

Detalle de los conjuntos de datos Sistema Ecobici:

Este conjunto de datos contiene la información de los viajes realizados, y contiene los siguientes atributos:

Campos	Años
Fecha-hora de inicio del viaje	2011 - 2012 - 2013 - 2014 - 2015 -2016 - 2017
Fecha-hora fin del viaje	2011 - 2012 - 2013 - 2014
Estación de inicio del viaje	2011 - 2012 - 2013 - 2014 - 2015 - 2016 -2017
Estación destino del viaje	2011 - 2012 - 2013 - 2014 - 2015 - 2016 -2017
Tiempo de uso formato 1	2011 - 2012
Tiempo de uso formato 2	2015 - 2016 - 2017
ID	2013 – 2014
ID_USUARIO	2016 – 2017
Fecha_creacion	2016 – 2017

Edad	2017
Sexo	2017

Registro del clima de Bs As.

https://data.buenosaires.gob.ar/dataset/informacion-meteorologica

Este conjunto de datos contiene la información de las mediciones de temperatura en diferentes zonas de la ciudad de Buenos Aires. El nivel de detalle es por día y por hora, y contiene los siguientes atributos:

Campos	**Años**
Fecha	2011- 2012
hora	2011- 2012
ID estación meteorológica	2011- 2012
temperatura	2011- 2012
presión atmosférica	2011- 2012
lluvia(en mm)	2011- 2012
radiación solar	2011- 2012
radiación ultravioleta	2011- 2012

Censo

Este conjunto de Datos contiene los datos del último censo del año 2010, y contiene los siguientes atributos:

Campos
población total por sexo
total de hogares
total de viviendas particulares
total de viviendas particulares habitadas

Estaciones de subte

Este conjunto de Datos contiene la información de geolocalización de las estaciones de subte de la ciudad de Buenos aires, y contiene los siguientes atributos:

Campos
Latitud
Longitud
ID
Nombre de la estación
Línea

Metodología de investigación

Las tareas para la realización del análisis fueron:

- Análisis de los datos utilizando los metadatos
- Análisis de los datos haciendo profiling
- Limpieza de datos
- Pre procesamiento de datos
- Procesamiento de datos con los métodos de minería de datos
- Análisis de resultados
- Elaboración del informe

Análisis de los datos utilizando los metadatos

Sistema Ecobici

Se realizó un análisis de los diferentes conjuntos de datos. Para ello se usó el metadato provista por el GCBA. El resultado de este primer análisis mostró que los grupos de datos de

los viajes fueron variando en los diferentes años. De acuerdo a los atributos que los conforman estos conjuntos quedaron agrupados en 4 grupos:

- Grupo de datos 1

  Años 2011 - 2012

- Grupo de datos 2

  Años 2013 - 2014

- Grupo de datos 3

  Años 2015 - 2016

- Grupo de datos 4

  Año 2017

En particular el año 2017 fue modificado a los largo del año. A mitad de año 2017 se le agregaron 2 campos que enriquecieron de manera significativa el valor de los datos. Estos campos son: Edad y sexo del usuario del viaje. Pero este agregado al conjunto de datos no es homogéneo, ya que contiene columnas con valores nulos.

Análisis de datos haciendo profiling

Se realizó el análisis de la metadato, esto permitió la definición de las tablas a crear en la base de datos. Luego se procedió a cargar los conjuntos de datos en las tablas de una base de datos postgres, creando una tabla para cada año del conjunto de datos del sistema Ecobici. También se generaron y cargaros tablas para los conjuntos de datos del registro de temperatura, del censo y de las estaciones de subte. Se generaron tablas de agregación para facilitar la construcción de vistas.

Durante el análisis de los datos usando comandos SQL, fue posible identificar algunos temas de calidad de datos en los siguientes conjuntos de datos:

Registro del clima de Bs As.: En este caso se detectó una falta de completitud de los registros del año 2012. Por lo que se trabajó con el conjunto de datos del año 2011,

descartando el año 2012. Se creó una vista donde se realizó una agregación de los datos obteniendo un valor único por día y hora. Ya que el conjunto de datos tenia los registros de todas las estaciones de clima de la ciudad.

Sistema Ecobici: se verificaron temas de inconsistencia funcional del dato, en aquellos casos en que los viajes tienen la misma estación del origen y destino, y el tiempo de uso es menor a dos minutos, se aplicó el criterio de considerar que ese viaje no se realizó. Ya que pueden considerarse como una devolución de la bicicleta por problemas técnicos Y que no merecen ser considerados para el análisis. De ese análisis se rechazan la siguiente cantidad de registros:

Año	Viajes < 2 minutos	Total de viajes	Porcentaje viajes no considerados
2011	7140	408450	1.75
2012	12990	644640	2.02
2013	22718	1066310	2.13
2014	22564	1073800	2.10
2015	1872	494139	0.38
2016	849	718138	0.12
2017	779	1356668	0.06

En el conjunto de datos de los recorridos del año 2017 fue necesario completar datos nulos

Limpieza de datos

Durante el proceso de profiling se detectaron valores nulos, y valores inconsistentes en los conjuntos de datos de los recorridos. Para el caso de los valores nulos del año 2017 se generó una nueva tabla de usuarios con su edad y sexo, lo que permitió completar los datos de sexo y edad de los registros del primer semestre.

Hubo algunos casos donde no fue posible inferir el valor correcto, por lo que se eliminaron algunos registros. Este último caso tuvo poca significancia, ya no dentro de un volumen de millones de registros, solo alcanzo algunas decenas.

Pre procesamiento de los datos para ser usado por los métodos del Minería de datos

Se trabajó sobre la Base de datos postgres, creando vistas con agregaciones y filtros de los datos. Luego se usaron los resultados de las consultas de las vistas mencionadas para generar archivos delimitados, dejando estos archivos listos para ser procesados con las herramientas mencionadas.

Basado en el análisis del meta dato, y del profiling de datos, se verificó que el conjunto de datos del registro de temperatura del año 2012 tiene faltantes de datos. Por lo que solo se usó el conjunto de datos del registro de temperatura del año 2011 y su correspondiente conjunto de datos de los viajes del año 2011, para buscar la relación entre las variables.

Procesamiento de datos con los métodos de minería de datos

En el trabajo se utilizaron métodos de minería de datos para el análisis de los datos y verificación de la hipótesis. Las hipótesis plantean la existencia de variables independientes y variables dependientes. Los métodos que se utilizaron para demostrar que existe una relación fueron:

Regresión lineal simple
Regresión polinómica
Kmeans

Regresión

El análisis de regresión engloba a un conjunto de métodos estadísticos que usamos cuando tanto la variable de respuesta como la(s) variable(s) predictiva(s) son continuas y queremos predecir valores de la primera en función de valores observados de las segundas. En esencia, el análisis de regresión consiste en ajustar un modelo a los datos, estimando coeficientes a partir de las observaciones, con el fin de predecir valores de la variable de respuesta a partir de una (regresión simple) o más variables (regresión múltiple) predictivas o explicativas.
El análisis de regresión juega un papel central en la estadística moderna y se usa para:

- identificar a las variables predictivas relacionadas con una variable de respuesta
- describir la forma de la relación entre estas variables y para derivar una función matemática óptima que modele esta relación

- predecir la variable de respuesta a partir de la(s) explicativas o productoras

Tipos de regresión

Tipo de regresión	Uso típico
lineal simple	Predicción de una variable de respuesta cuantitativa a partir de una variable predictora cuantitativa
polinómica	Predicción de una variable de respuesta cuantitativa a partir de una variable predictora cuantitativa, donde la relación se modela como una función polinómica de orden *n*

Regresión lineal simple

La regresión lineal simple se basa en estudiar los cambios en una variable, no aleatoria, que afectan a una variable aleatoria, en el caso de existir una relación funcional entre ambas variables, que puede ser establecida por una expresión lineal, es decir, su representación gráfica es una línea recta. Es decir, se está en presencia de una regresión lineal simple cuando una variable independiente ejerce influencia sobre otra variable dependiente. Ejemplo: Y = f(x)

Regresión lineal polinómica

Es una forma de regresión lineal en el que la relación entre la variable independiente x y la variable dependiente Y se modela como un polinomio de orden n. Regresión polinómica se ajusta a una relación no lineal entre el valor de x y la media condicional correspondiente de y, denotado E, y se ha utilizado para describir los fenómenos no lineales

¿Cómo saber si un Modelo de regresión es Válido?

Tenemos tres tipos de validaciones

Test de Hipótesis:

Test de Fisher

Test t-student

Test Numérico:

Coeficiente de Determinación R^2

Gráficos:

Gráfico de la dispersión

Gráfico de la recta

Test de Fisher of-test

- $H_0 : \forall i, \beta_i = 0$
- $H_1 : \exists i, \beta_i \neq 0$

Analizamos el $p - valor$:
Si el $p - valor < 0{,}05$ rechazamos H_0 y entonces el modelo es válido.

Student-test

Se Testea el efecto de β_1

- $H_0 : \beta_1 = 0$
- $H_1 : \beta_1 \neq 0$

Student-test

Se Testea el efecto de β_1

- $H_0 : \beta_1 = 0$
- $H_1 : \beta_1 \neq 0$

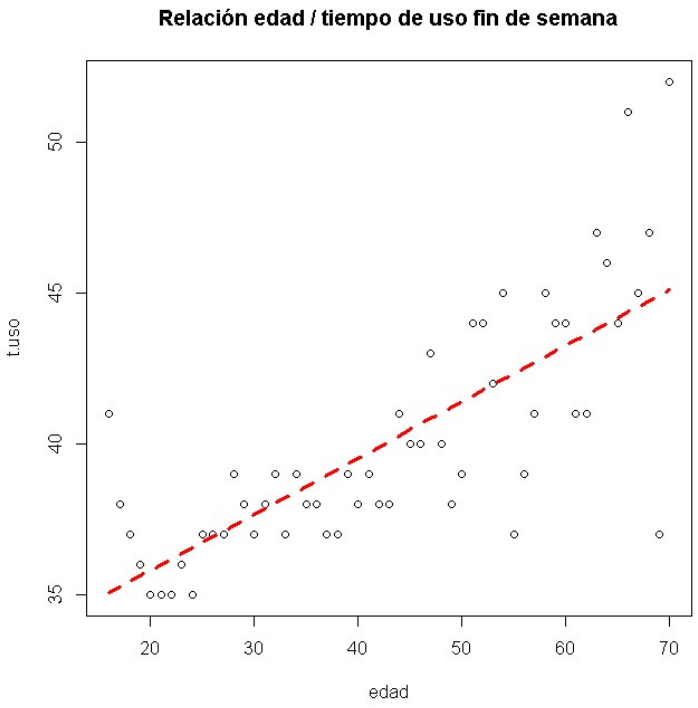
Analizamos el p-valor $p - valor < 0{,}05$ efecto significativo de β_1 (rechazamos H_0)
$p - valor > 0{,}05$ no hay efecto significativo de β_1 (aceptamos H_0)

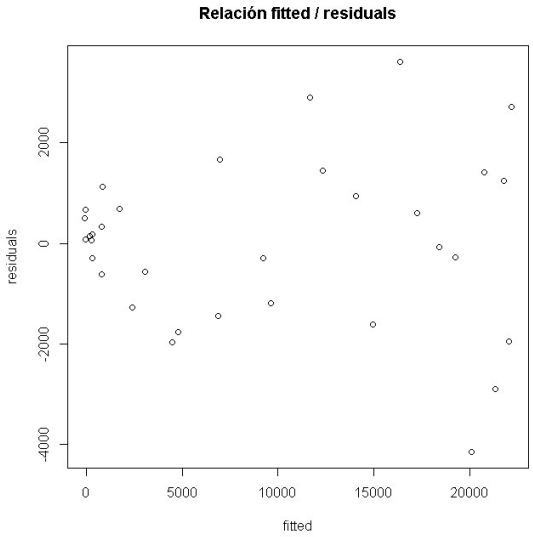
Coeficiente de Determinación o Porcentaje de Varianza Explicada

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Además $0 \leq R^2 \leq 1$.
Valores cercanos a 1 indica un mejor ajuste.

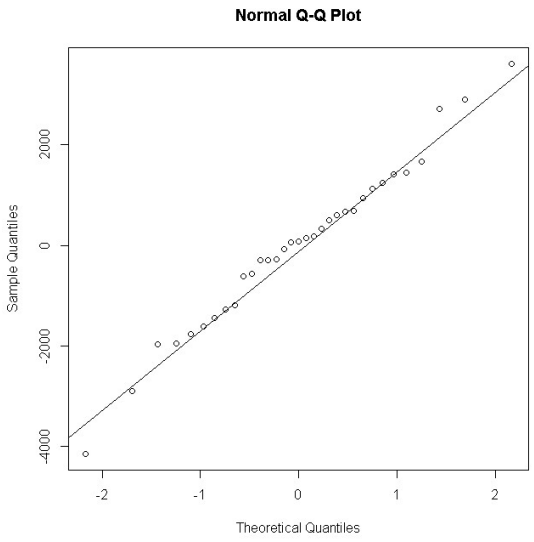
Gráfico de recta de ajuste





La gráfica nos ayuda a decidir si las variables están linealmente relacionadas. Si es así, no debería de existir una relación sistemática entre los residuos (errores) y los valores predichos (ajustados). Es decir, el modelo debería de capturar toda la varianza sistemática de los datos, dejando sólo ruido estocástico sin explicar. Por tanto, esta gráfica debe de verse como "las estrellas en el firmamento", sin un patrón claro de asociación. Si es así, sugiere además que se cumple el supuesto de homocedasticidad.

Test de normalidad de los residuos

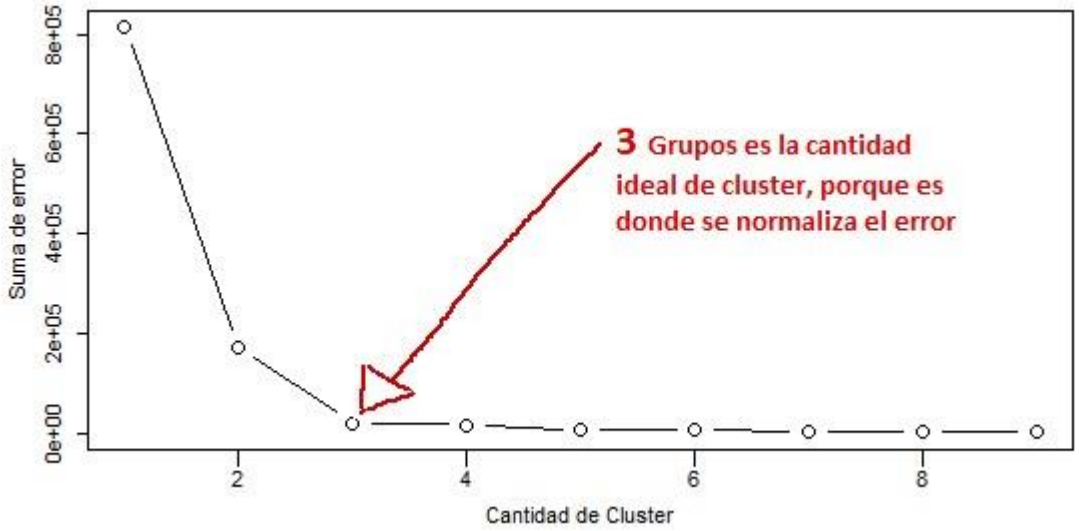


La **gráfica de quantil-quantil normal**. Los puntos deberían seguir la diagonal si los residuos están normalmente distribuidos. Si aparecen patrones tipo "S" o "banana", posiblemente necesitemos ajustar otro modelo.

Kmeans

Kmeans es un método de agrupamiento, que tiene como objetivo la partición de un conjunto de n observaciones en k grupos en el que cada observación pertenece al grupo cuyo valor medio es más cercano.

Kmeans (MacQueen, 1967) es uno de los algoritmos de aprendizaje no supervisado más simples que resuelven el conocido problema de agrupamiento. El procedimiento sigue una forma simple y fácil de clasificar un conjunto de datos dado a través de un cierto número de clusters (supongamos k clusters) fijados a priori. La idea principal es definir k centroides, uno para cada grupo. Estos centroides deben colocarse de una manera astuta debido a que la ubicación diferente causa un resultado diferente. Entonces, la mejor opción es ubicarlos lo más lejos posible el uno del otro. El siguiente paso es tomar cada punto perteneciente a un conjunto de datos dado y asociarlo al centroide más cercano. Cuando no hay ningún punto pendiente, se completa el primer paso y se realiza un grupaje anticipado. En este punto, necesitamos volver a calcular k nuevos centroides como baricentros de los grupos resultantes del paso anterior. Después de tener estos k nuevos centroides, se debe realizar un nuevo enlace entre los mismos puntos de ajuste de datos y el centroide nuevo más cercano. Se ha generado un ciclo. Como resultado de este ciclo, podemos observar que los k centroides cambian su ubicación paso a paso hasta que no se realizan más cambios. En otras palabras, los centroides no se mueven más. Para poder realizar la segmentación con Kmeans, es necesario identificar previamente la cantidad ideal de grupos. Una técnica es usar el algoritmo de Kmeans de forma repetida (iterada), aumentando la cantidad de grupos (clúster) en cada ejecución y obteniendo el error intracluster en cada iteración, donde este error calcula la distancia entre cada observación y el clúster al cual pertenece. Con esta técnica (conocida como SSE o Sum of Squares Error) se puede obtener la siguiente gráfica que permite identificar la cantidad ideal de clúster o grupos:



Herramientas utilizadas

Considerando la hipótesis, y los análisis estadísticos a realizar, se seleccionaron las siguientes herramientas informáticas:

Base de datos Postgres-SQL

Como repositorio de datos y motor de consulta. Y Pre procesamiento

Rstudio

Para el proceso de los datos.

QGis

Para geo localizar las estaciones, recorridos y generar mapas.

Tableau Public

Para gráficos, y análisis.

Excel

Para gráficos, y análisis.

Resultados

Hipótesis 1 - El uso de las bicicletas es mayor cuando la temperatura está entre los 15 y los 25 grados

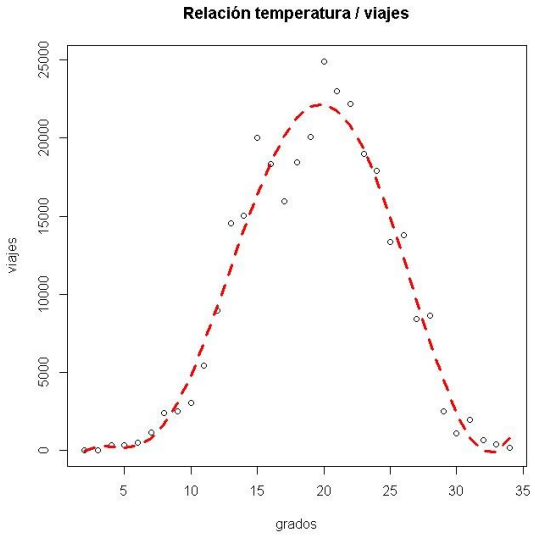
Se realizó una agregación de los viajes del año 2011 por hora, de manera que la granularidad de los datos de los viajes fuera la misma que el conjunto de datos del registro de temperatura.

Para el análisis calculamos la correlación y la regresión polinómica de grado 6
Resultado de R-squared 0.9813096

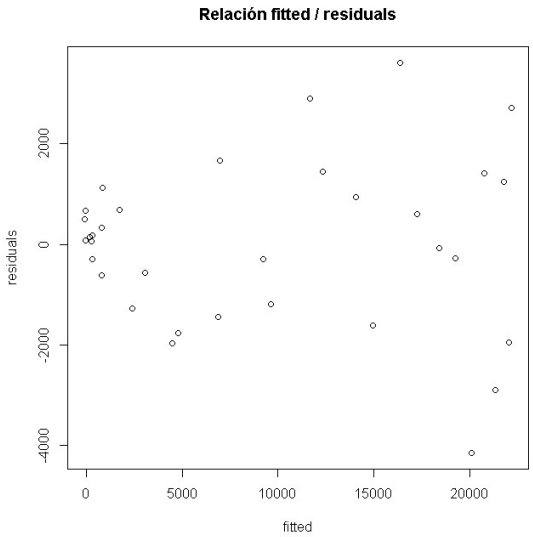
EL R-squared: 0.963, indica el modelo se ajusta bien.

EL p-value: < 2.2e-16, indica que los coeficientes son significativos, con una puntuación de "***" ya que su valor es muy pequeño, y por lo tanto la cola de probabilidad es de solo 2.2e-16
Por lo tanto puedo rechazar la hipótesis nula β = 0. Indica que existe una relación entre las variables.

Verifico el ajuste del modelo, utilizando el modelo con nuevos datos generados

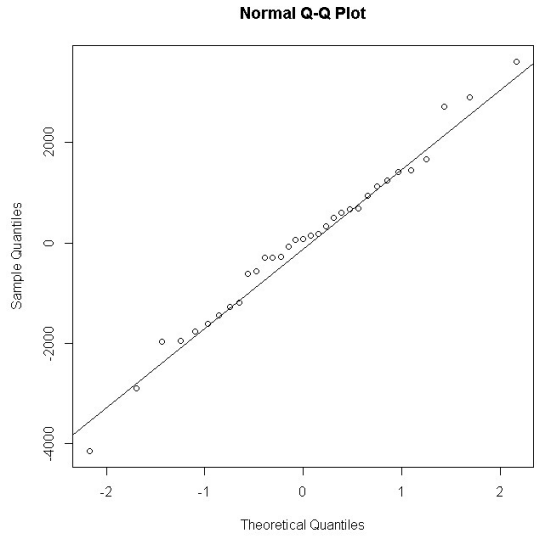


Verifico el ajuste utilizando los residuos



Se puede considerar razonable la distribución de los residuos

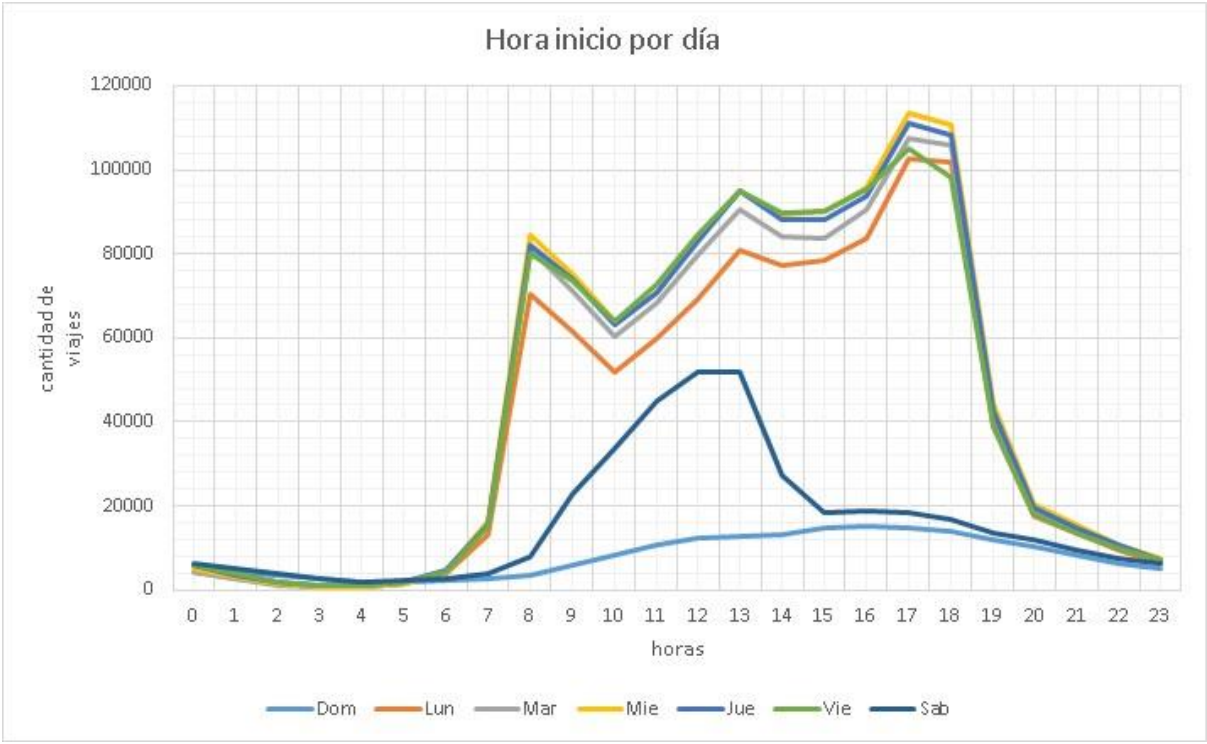
Test de normalidad de los residuos



Los puntos se ajustan bastante a la recta, por lo los residuos están normalmente distribuidos

Hipótesis 2 – El uso de las bicicletas es mayor en el horario pico.

Para este análisis se utilizaron los conjuntos de datos de los años 2011-2012-2013-2014-0152016-2017. Se realizó una unión de los conjuntos de datos, y se hizo una agregación de datos por hora del día. Al analizar el conjunto de datos resultante se observa que el uso permite dividir el análisis en dos, uso diurno y uso nocturno, como se observa en los gráficos.



Basados en el análisis del grafico se consideró solo el uso diurno del sistema Ecobici. También se realizó una segmentación entre los días de semana y los fines de semana, ya que como se observa en el grafico el uso tiene patrones diferentes.

Por lo descripto se realizaron dos análisis:

1- Días de semana de 8 a 19
2- Fines de semana de 8 a 20

Se seleccionó el año 2017, ya que en este año el total de estacione fue de 199, y son todas automáticas. Se realizó una agregación de datos sobre el conjunto de datos, utilizando para ello comandos de agregación de SQL. Luego se utilizaron los siguientes métodos de análisis:

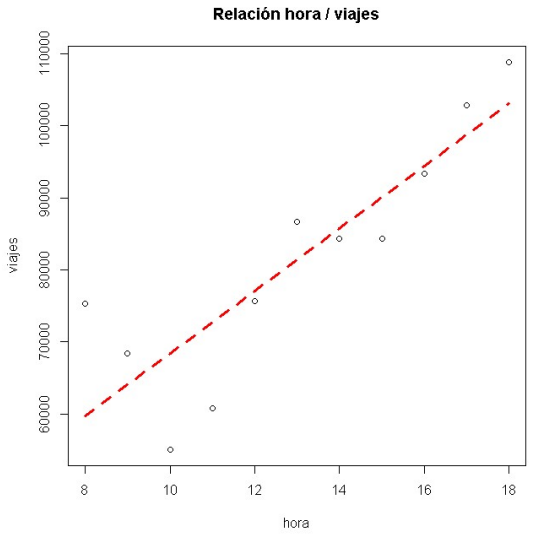
1- Regresión lineal (días de semana)
2- Regresión polinómica de grado 2 (fin de semana)

Regresión lineal (días de semana)

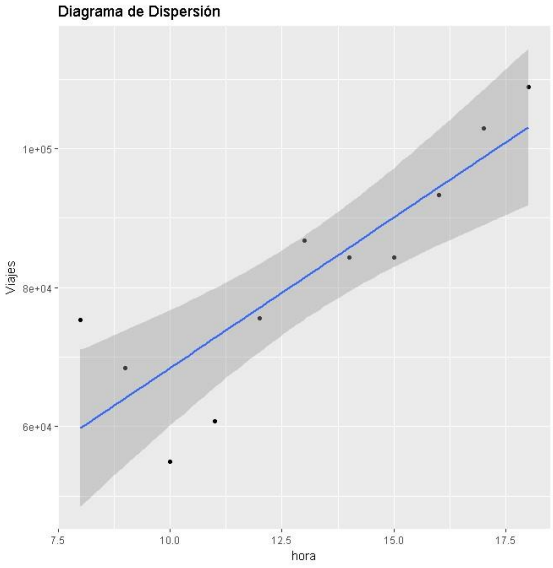
Sumario del cálculo de la regresión:

Multiple R-squared 0.7461 indica que el modelo se ajusta bien.

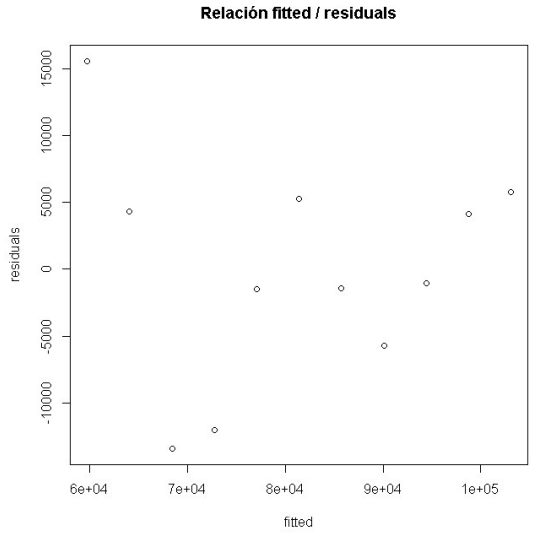
P-value 0.0006089, la cola de probabilidad es de solo 2.2e-16 , que es menos a 0.05. Por lo tanto puedo rechazar la hipótesis nula β = 0. Indica que existe una relación entre las variables.



Validación del modelo

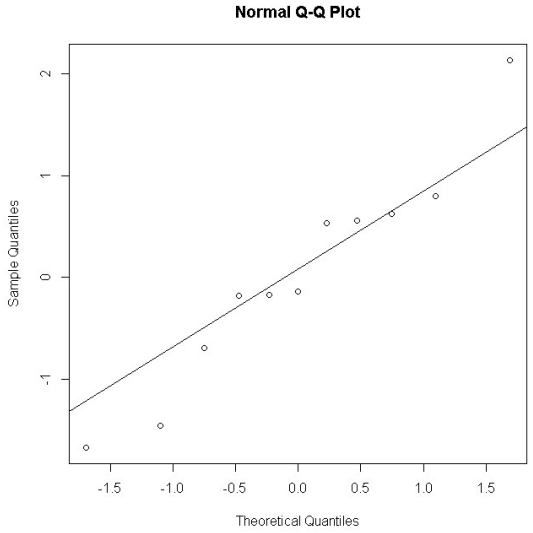


Verifico el ajuste utilizando los residuos



Resulta aceptable la distribución de puntos

Test de normalidad de los residuos



Considero aceptable la proximidad de los puntos a la recta.

Regresión polinómica de grado 2 (fin de semana)

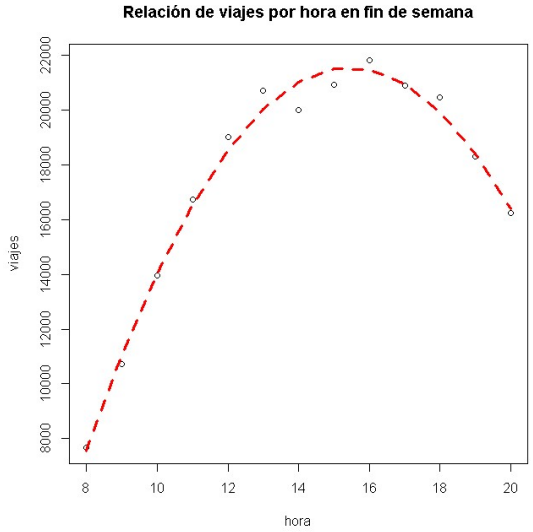
Sumario del cálculo de la regresión:

Multiple R-squared 0.9884 indica que el modelo se ajusta bien.

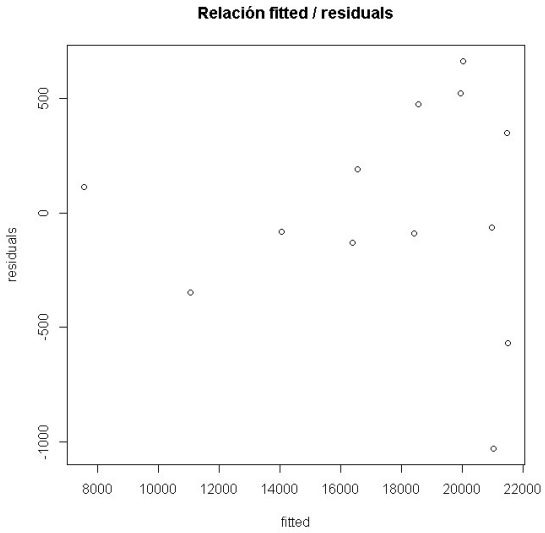
P-value 2.138e-10 la cola de probabilidad es de mayor a 0.05. Por lo tanto puedo aceptar

La hipótesis nula $\beta = 0$. Indica que existe una relación entre las variables.

Validación del modelo usando las predicciones

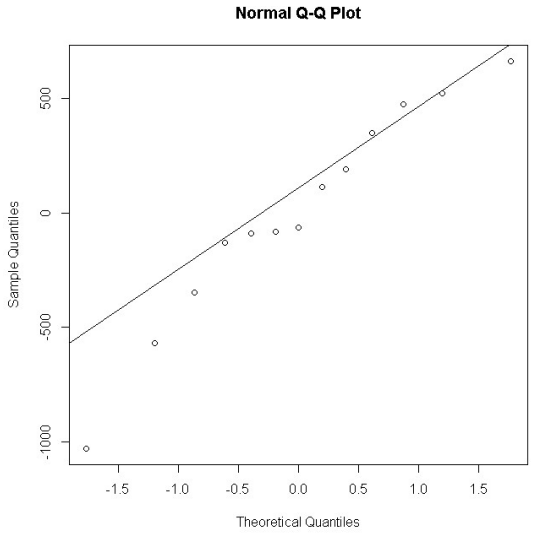


Verifico el ajuste utilizando los residuos



Resulta aceptable la distribución de puntos

Test de normalidad de los residuos



Se considera aceptable la proximidad de los puntos a la recta.

Hipótesis 3 - Las estaciones de Ecobici más usadas dependen de la densidad de población del barrio.

En este análisis se utilizaron los conjuntos de datos de los viajes realizados entre los años 2011 y 2016, en conjunto con los datos del censo de población del año 2010. Se buscó verificar la relación entre la cantidad de población cercana a las estaciones y la cantidad de viajes iniciada en las estaciones. Para ello se realizó una agregación considerando la estación origen de cada viaje. Para el análisis calculamos la correlación y la regresión lineal.

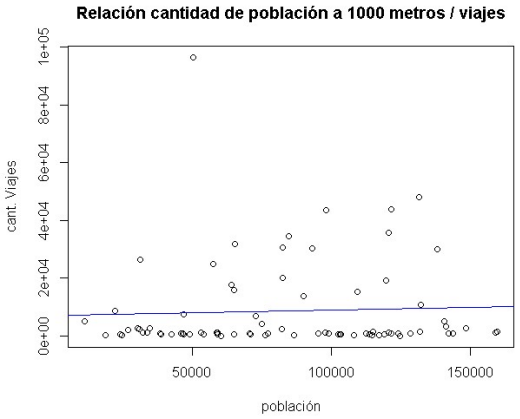
Sumario del cálculo de la regresión

Multiple R-squared: 0.002185. El modelo no se ajusta

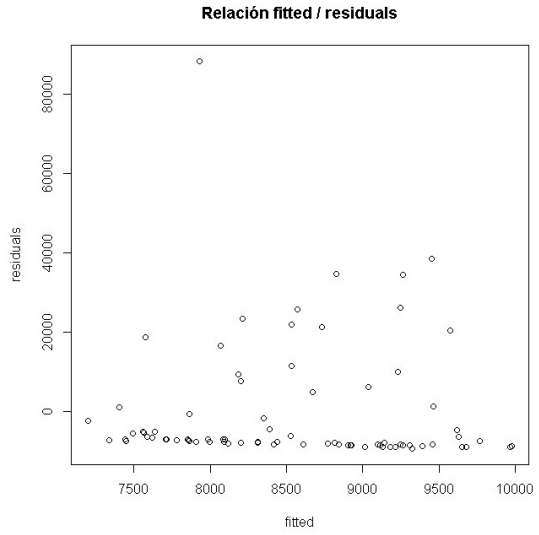
P-value: 0.6825 la cola de la probabilidad es mayor a 0.05. Por lo tanto puedo aceptar

La hipótesis nula $\beta = 0$. Indica que No existe una relación entre las variables

Validación del modelo usando las predicciones

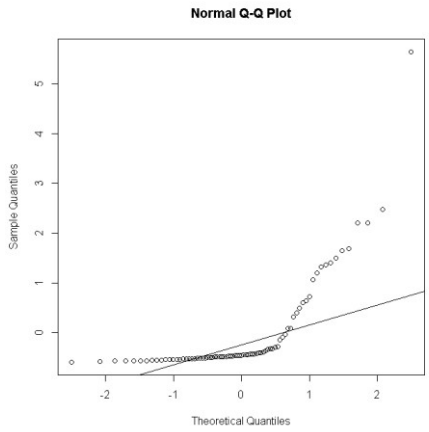


Verifico el ajuste utilizando los residuos



No cumple con la distribución esperada

Test de normalidad de los residuos

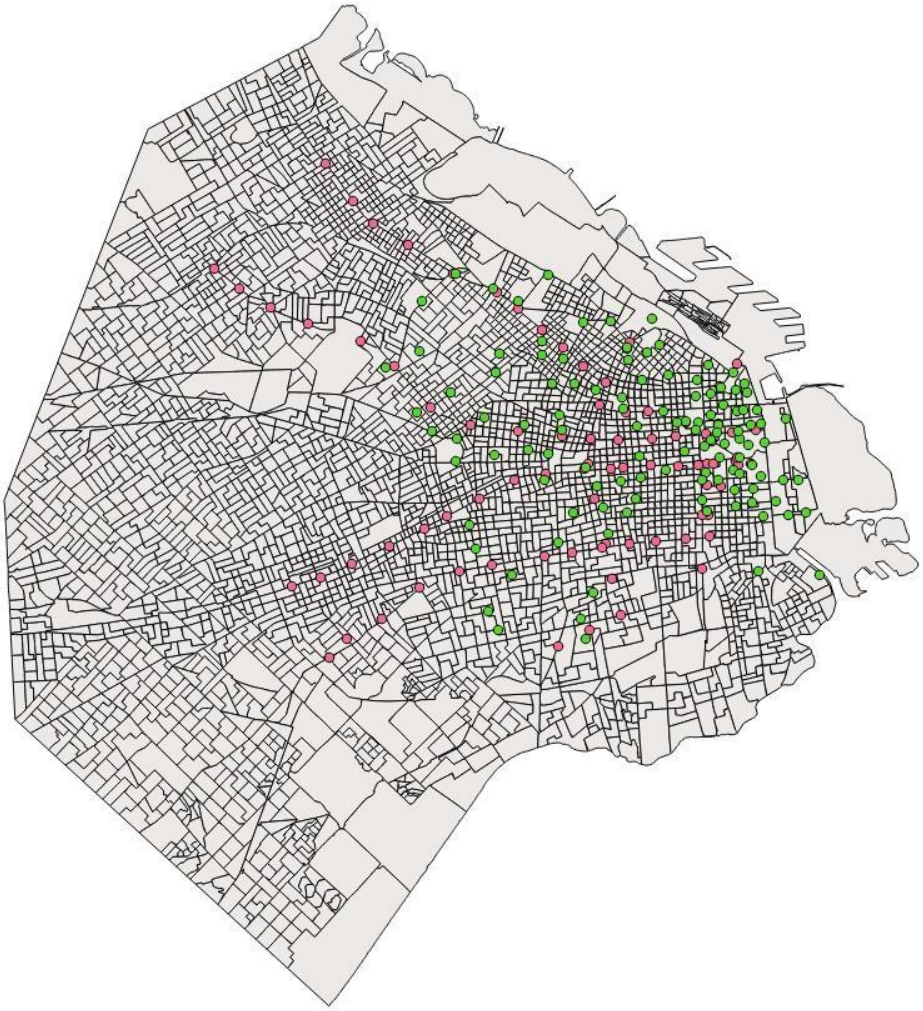


Los puntos No se ajustan a la recta.

Hipótesis 4 - Las estaciones de Ecobici más usadas están cercanas a las estaciones del metro.

En este caso se buscó verificar si la proximidad de las estaciones de metro incidía en el uso del sistema Ecobici. Se puede observar en el mapa la relación geográfica entre las estaciones de Ecobici (verde) y las estaciones de metro (rosa). Para el estudio se obtuvieron las distancias entre las estaciones de Ecobici y las estaciones de metro, utilizando mapas de google.

En el siguiente mapa se observan en verde las estaciones de subte y en rojo las estaciones de Ecobici.



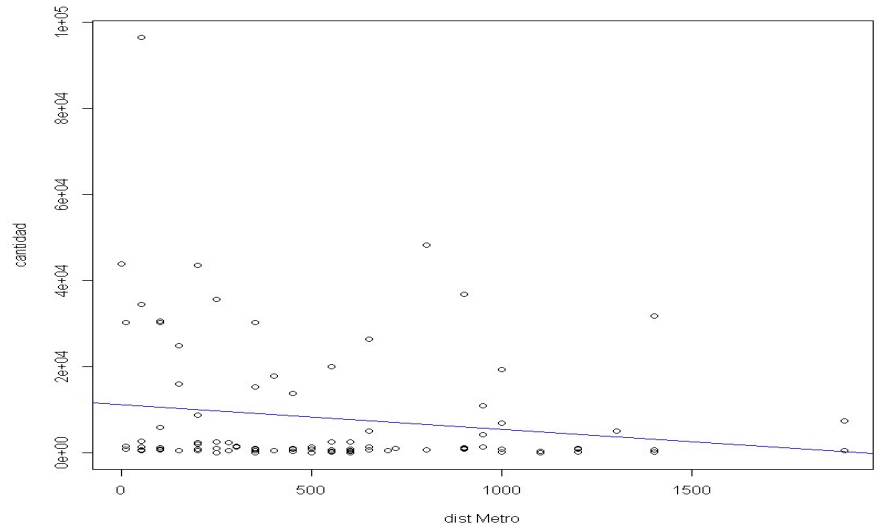
Para el análisis calculamos la correlación y la regresión lineal.

Sumario del cálculo de la regresión

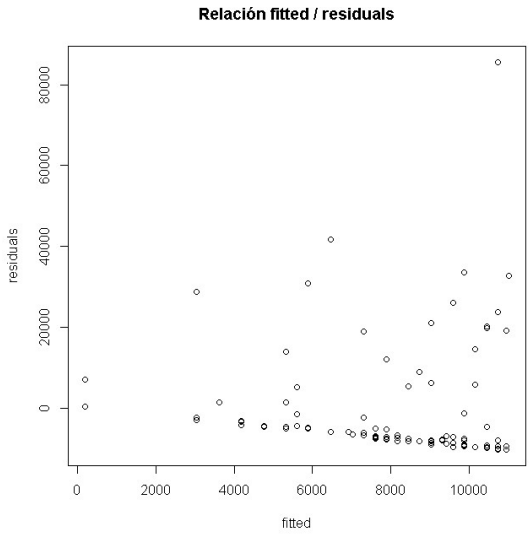
EL R-squared: 0.02567, indica el modelo no se ajusta.

EL p-value: 0.1229, que su valor es muy alto mayor a 0.05, y por lo tanto puedo aceptar que la hipótesis nula β = 0. Indica que No existe una relación entre las variables.

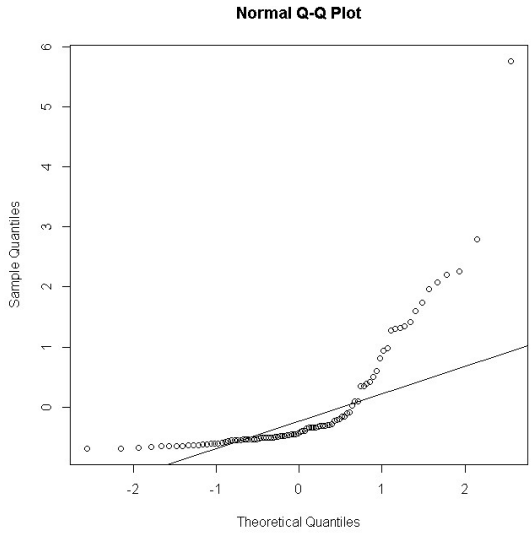
Validación del modelo usando las predicciones



Verifico el ajuste utilizando los residuos



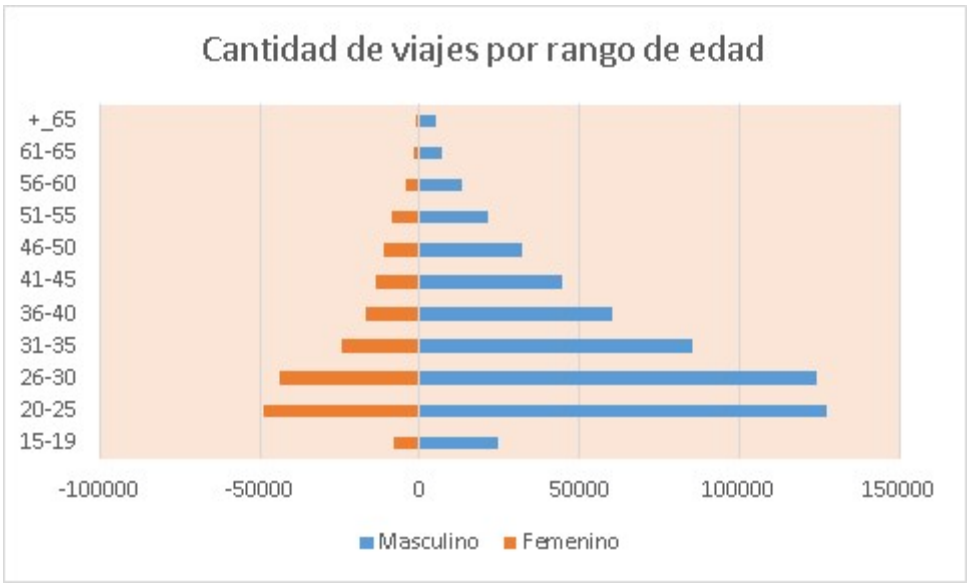
No cumple con la distribución esperada



No se ajustan a la recta

Hipótesis 5 - El sistema es más utilizado por los usuarios menores de 30 años.

En este análisis se utilizaron los datos del conjunto de datos de los viajes realizados en el año 2017. Ya que este es el primer conjunto de datos con información de los usuarios del sistema. Se hicieron análisis de los datos, realizando agregaciones con comando SQL, y luego se graficaron los resultados. En el grafico se observa la utilización del sistema Ecobici por rango etario.



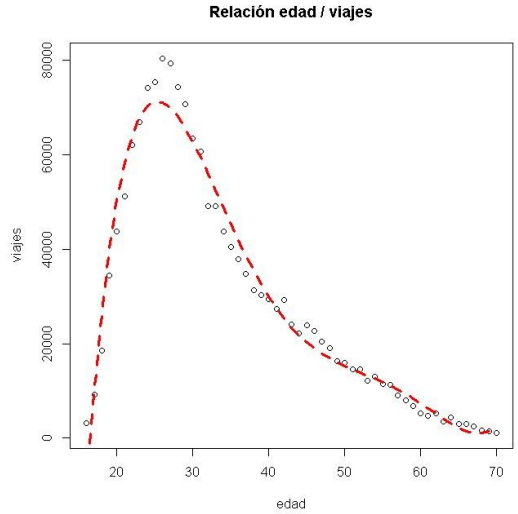
Luego se hicieron diferentes pruebas para identificar qué modelo ajustaba mejor a conjunto de datos. De ello resultó la selección de una regresión no lineal de un polinomio de grado 6. También se procedió a eliminar outliers, que solo representan el 0.3 % de los viajes.

A partir de allí se calculó la regresión polinómica.

Sumario de la regresión

R-squared: 0.99775

P-value: 2.2e-16



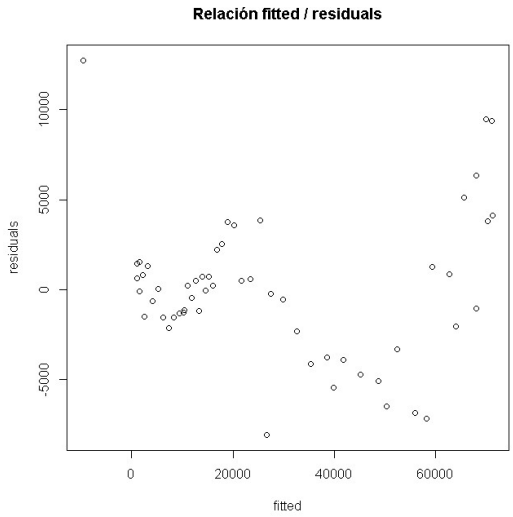
Realizo un test de Shapiro, para validar la muestra.

```
        Shapiro-Wilk normality test

data:  residuals(m.nls2) W
= 0.98714, p-value = 0.78
```

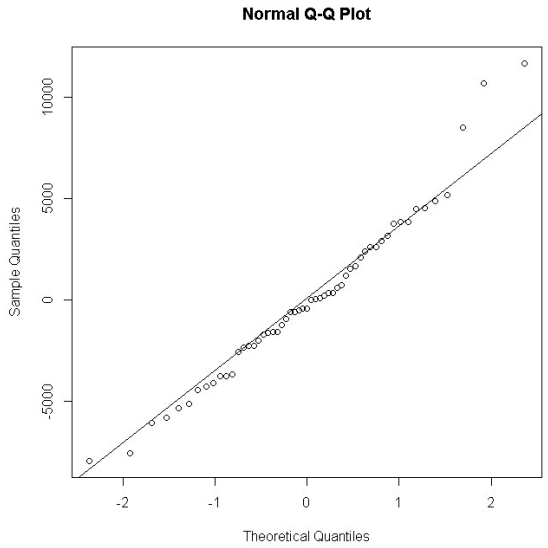
Como el p-value obtenido es 0.78, y es mayor que 0.05 aceptamos la hipótesis nula (H0), por lo que podemos afirmar que nuestros datos se distribuyen siguiendo una normal.

Verifico el ajuste utilizando los residuos



Resulta aceptable la distribución de puntos

Test de normalidad de los residuos



Considero aceptable la proximidad de los puntos a la recta **Análisis surgidos durante la realización del trabajo**

a) El tiempo de uso está relacionado con el día de la semana

b) Tipo de uso del sistema por parte de los usuarios individuales

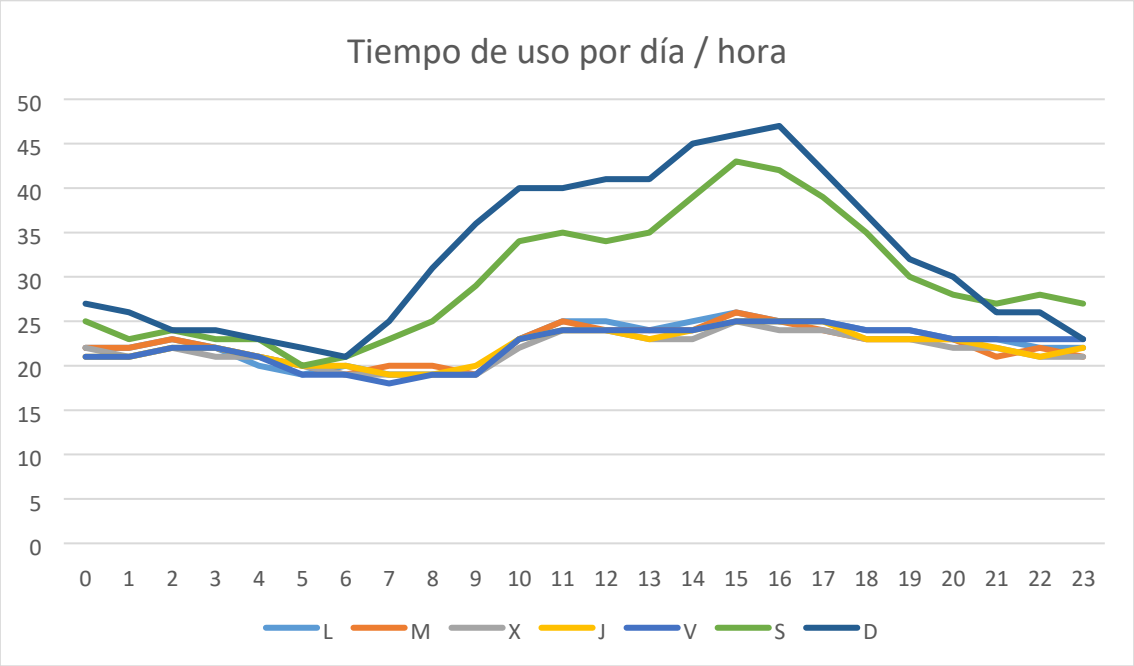
a) El tiempo de uso está relacionado con el día de la semana

En este análisis se utilizaron los datos del conjunto de datos de los viajes realizados en el año 2017. Ya que este el año con mayor cantidad de estaciones, y todas son automáticas (desatendidas).

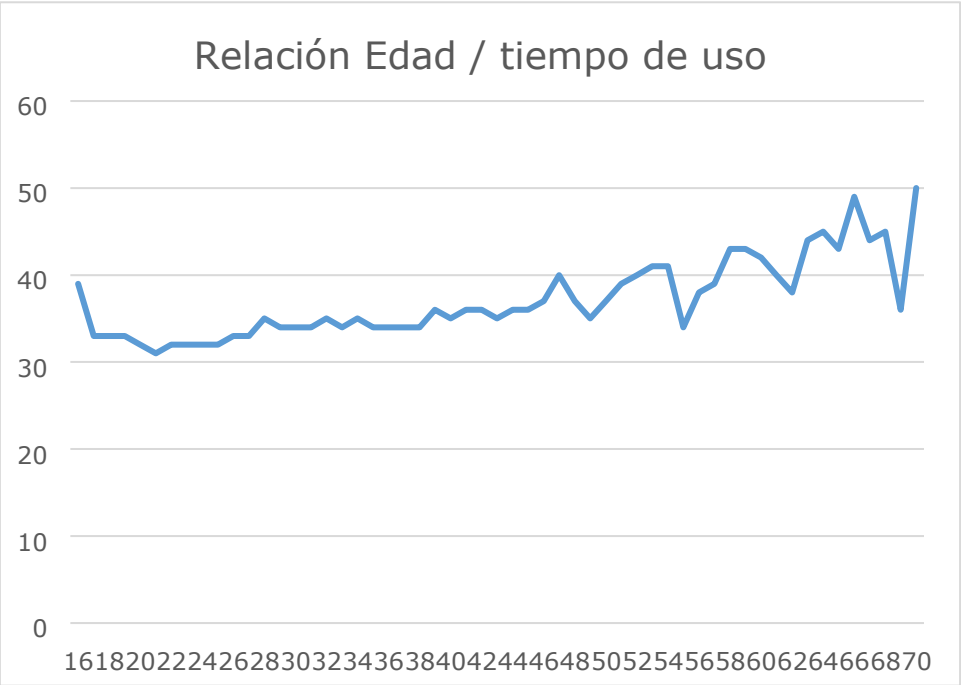
Este análisis de realizo utilizando agregaciones da datos con comandos SQL, y haciendo un gráfico del tiempo promedio de duración de los viajes por día / hora.

En él grafico se identifican dos patrones de uso muy diferentes. Por lo tanto se plantearon dos análisis:

1. Días de semana
2. Fin de semana



Luego de separar el conjunto de datos en dos. Se realizaron pruebas para identificar que variables podían incidir en el tiempo del viaje. Se identificó como candidata de la edad los usuarios. Para ello se realizó un gráfico que permitiera verificar visualmente la hipótesis.

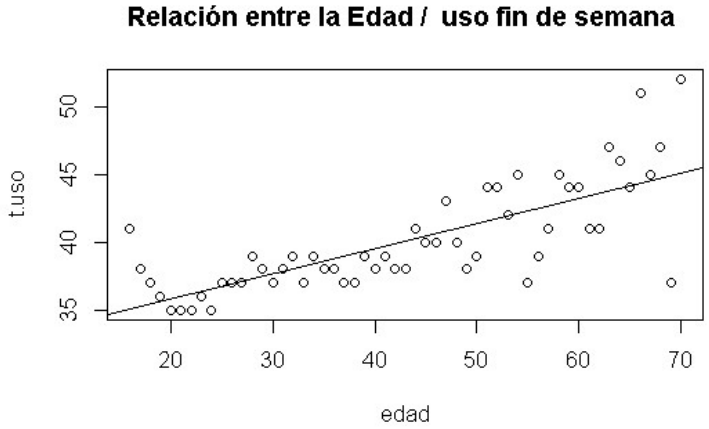


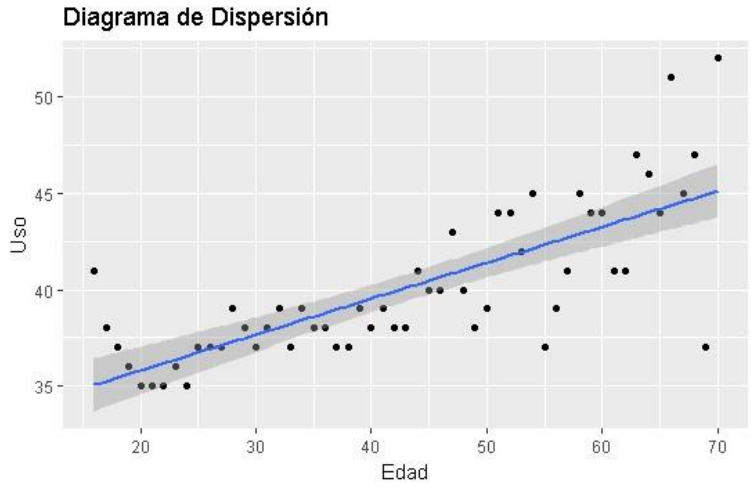
A partir de esto se planteó se planteó como un modelo de regresión lineal, entre la edad y la duración del viaje, para los fines de semana.

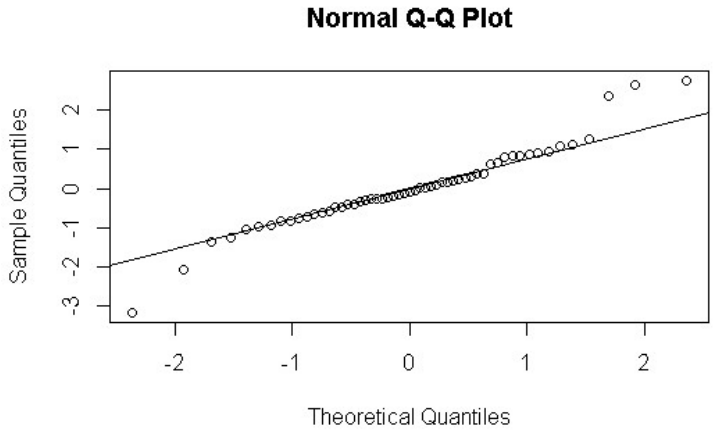
Sumario de la regresión

R-squared: 0.5745 Indica que existe una relación entre las variables, pero de un grado medio.

P-value: 2.075e-11 la cola de probabilidad es de mayor a 0.05. Por lo tanto puedo aceptar La hipótesis nula $\beta = 0$. Indica que existe una relación entre las variables







Considero aceptable la proximidad de los puntos a la recta.

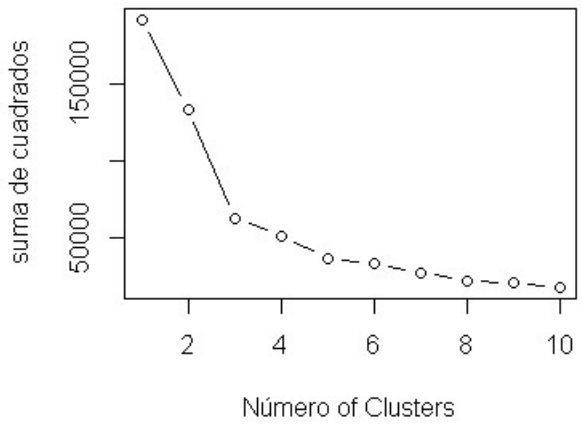
b) El tipo de uso del sistema por parte de los usuarios individuales

Durante la realización del trabajo surgió un nuevo análisis que fue ver el tipo de uso de los usuarios individuales en relación con la cantidad de viajes.

El sistema Ecobici tenía 95.975 usuarios individuales registrados en el año 2017.

Se definió el uso de Kmeans como técnica para agrupar a los usuarios basados en la cantidad de viajes por año. Para ello se determinó el número óptimo de clúster para el uso de Kmeans.

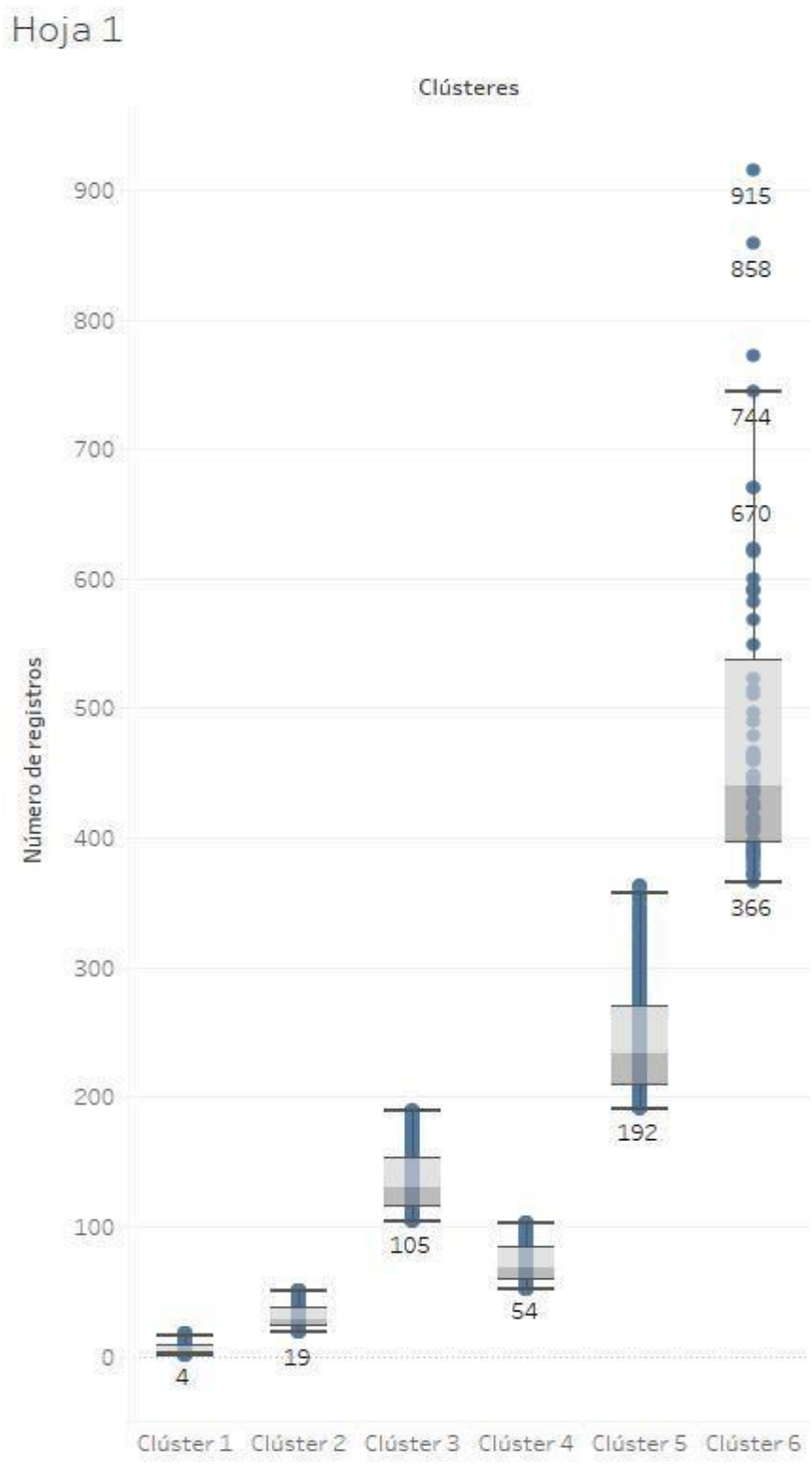
Determinación del número de Clúster a utilizar



Se identificó como 6 el valor a utilizar

A partir de tener el número de cluster , se aplicó Kmeans para obtener la clasificación de usuarios basada en la cantidad de viajes por año.

En el grafico se observan los grupos resultantes.



En la siguiente tabla se muestran los usuarios agrupados en cluster, la cantidad total de viajes de cada grupo de usuarios individuales, y su peso relativo en relación con el total.

Clúster	viajes	Porcentaje de viajes	Usuarios	Porcentaje de usuarios	Promedio
Clúster1	387,310	23.48	72,991	76.05	5
Clúster2	465,863	28.24	15,242	15.88	31
Clúster3	258,802	15.69	1,922	2.00	135
Clúster4	369,343	22.39	5,193	5.41	71
Clúster5	137,490	8.33	563	0.59	244
Clúster6	30,794	1.87	64	0.07	481
Total	1649602	100	95975	100	

Del análisis de la tabla anterior surgió que el 8 % de los usuarios (7.742) realizaron el 48.28 % de los viajes. Como se puede observar en la tabla siguiente:

Clúster	viajes	Porcentaje de viajes	Usuarios	Porcentaje de usuarios	Promedio
Clúster1	387,310	23.48	72,991	76.05	5
Clúster2	465,863	28.24	15,242	15.88	31
Clúster3	258,802	15.69	1,922	2	135
Clúster4	369,343	22.39	5,193	5.41	71
Clúster5	137,490	8.33	563	0.59	244
Clúster6	30,794	1.87	64	0.07	481
		48.28	7,742	8.07	

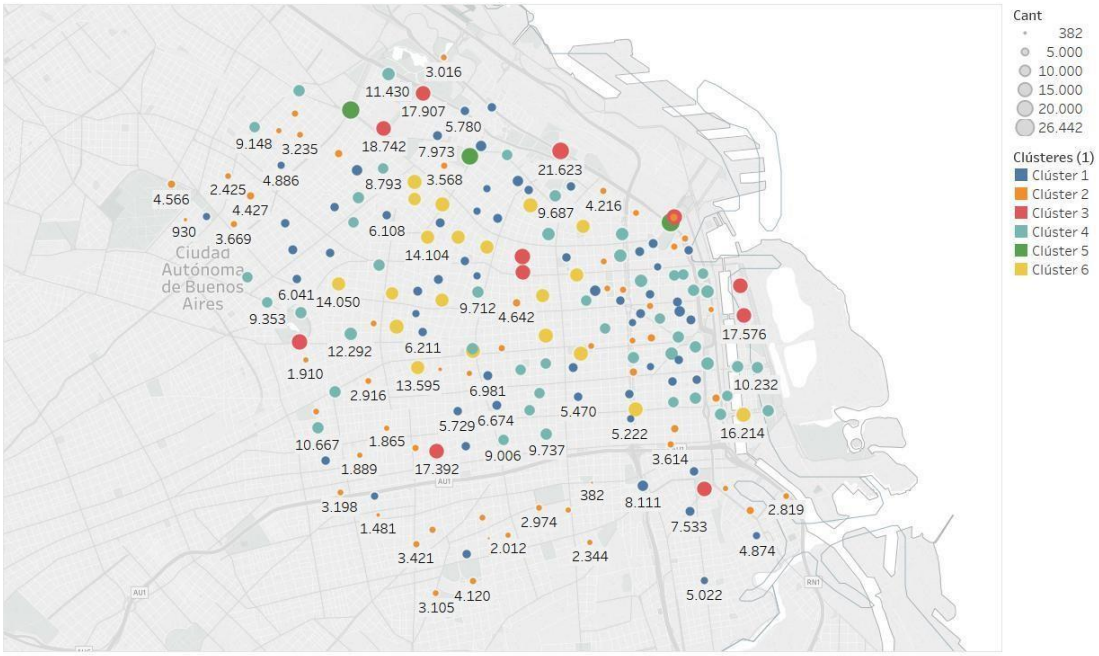
Se destaca el grupo de usuarios del Clúster 6, en el que el promedio de uso del sistema es muy alto.

Análisis visuales

Se realizó un análisis visual de las estaciones por cantidad de total de viajes (incluye inicio y fin), y su ubicación en el mapa. Para ello se usó la técnica de Clúster. En este caso se usó 6 como el número de Clusters.

Aquí se puede observar cual es la distribución de las estaciones más usadas, basados en el grupo del clúster en el que fueron clasificadas.

Hoja 1



Mapa basado en LON y LAT. El color muestra detalles acerca de Clústeres (1). El tamaño muestra suma de Cant. Las marcas se etiquetan por suma de Cant.
Se muestran detalles para Nombre de medidas.

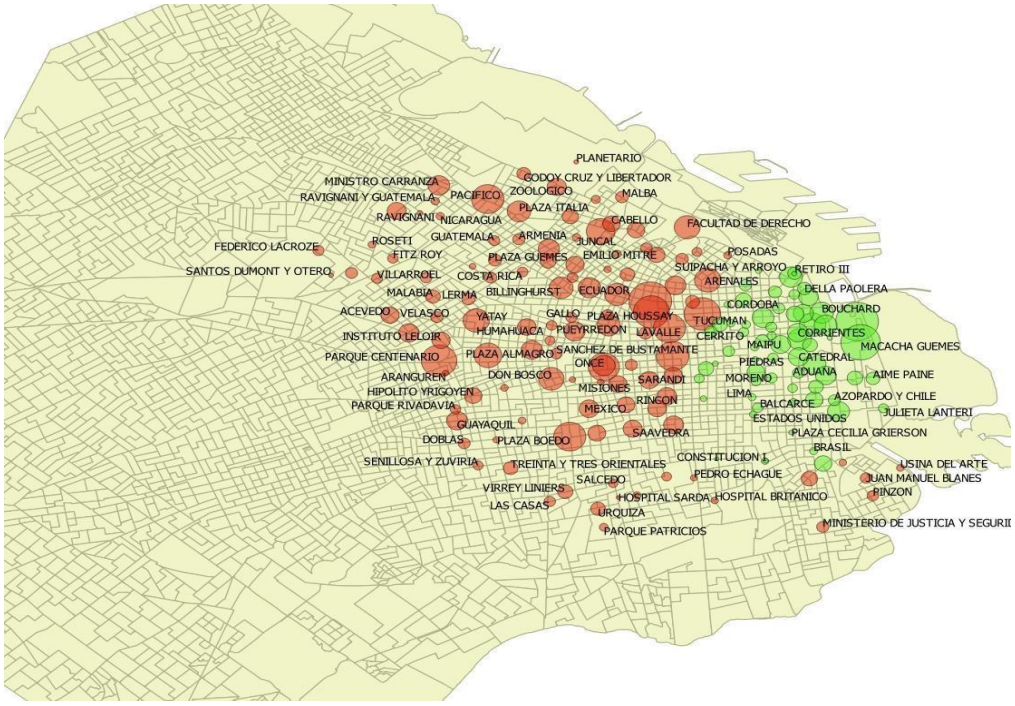
En esta tabla se observa que las estaciones del Clúster5 tuvieron un uso muy intenso. Dichas estaciones son: Pacifico – Retiro – Pza. Las Heras

Clúster nro.	Cantidad de	Viajes e Totales	Promedio
Clúster 5	3	73,810	24,603
Clúster 3	11	206,811	18,801
Clúster 6	20	293,503	14,675
Clúster 4	47	467,073	9,938
Clúster 2	56	374,074	6,680
Clúster 1	59	172,548	2,925

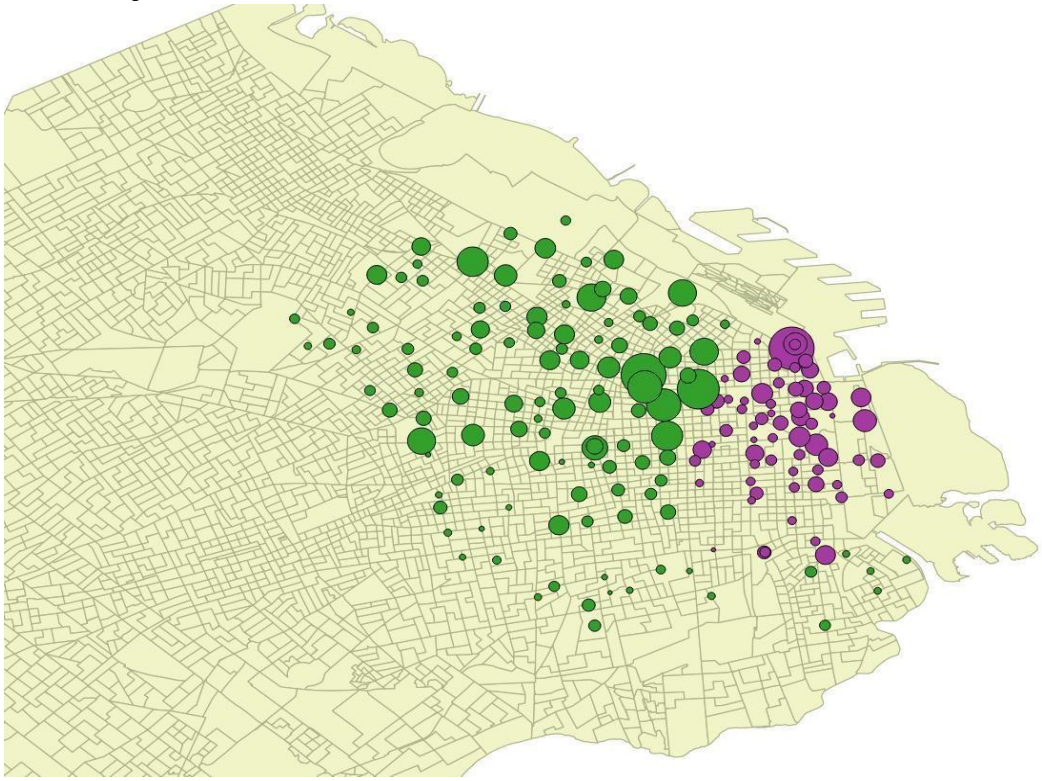
Distribución de los diferentes Clúster de estaciones en el mapa. Donde vemos para cada grupo de estaciones cuál es su ubicación.

Análisis visual de las estaciones origen y destino en los horarios de 7 a 10 horas, y de 16 a 19 horas.

Viajes de 7 a 10 horas. En rojo el origen de los viajes. En verde el destino de los viajes.



Viajes de 16 a 19 horas. En violeta el origen de los viajes. En verde el destino de los viajes



Discusión de resultados

Se propusieron cinco objetivos para analizar el uso del sistema:

1. Relación del uso con la temperatura: esta hipótesis quedó demostrada, dentro del rango de los 15 a los 25 grados ocurren la mayor parte de los viajes. Se observa también que en las temperaturas extremas (altas o bajas) la cantidad de viajes disminuye mucho.
2. Relación del uso con la hora del día: en el análisis de esta hipótesis se encontraron 2 patrones de uso, días de semana y fines de semana. Y dentro de cada uno de ellos también quedan en evidencias 2 tipos de uso, por lo que el día debe ser analizado en dos partes, diurno y nocturno. Para el caso de los días de semana el uso diurno aumenta entre los 08 y las 19 horas. de manera lineal En cuanto el uso los fines de semana , la utilización también muestra una relación con la hora, pero en una relación de tipo polinómica de grado 2
3. Relación del uso con cantidad de población: esta hipótesis no pudo demostrar una dependencia entre las variables analizadas.
4. Relación entre el uso y la proximidad al metro: esta hipótesis no pudo demostrar una dependencia entre las variables analizadas.
5. Relación entre el uso y la edad de los usuarios: esta hipótesis permitió encontrar una dependencia entre las variables analizadas. Se observó que el grupo etario que más utiliza el sistema es el de entre los 20 y los 30 años.
6. Durante el análisis de la relación edad / cantidad de viajes, surgió un nuevo análisis que es la relación del tiempo de uso / edad del usuario. Aquí se observó que el tiempo de uso en los días de semana es similar para los diferentes días laborables. Pero que en el fin de semana el tiempo de uso llega a duplicarse. Para este caso se realizó un análisis de los fines de semana. Se pudo verificar que a medida que la edad del usuario aumenta, también aumento el tiempo de uso. Si bien el grado de correlación del modelo de 0.57, se lo puede considerar valido para indicar una relación entre las variables.
7. Otro análisis que surgió durante el trabajo fue el análisis de los usuarios en base a la cantidad de viajes realizados por año. Esto permite diferenciar a los usuarios frecuentes de los usuarios ocasionales. Del análisis surge que el 8 % de usuarios frecuentes realizaron el 48 % del total de viajes.
8. Se realizaron diferentes visualizaciones, las cuales permitieron identificar cual es la ubicación de las estaciones más utilizadas. También se puede observar un patrón de

uso del sistema, donde durante la hora pico de la mañana el movimiento es desde la periferia hacia el macro y micro centro de la ciudad. En contrapartida al final del día se observa que el movimiento es desde el centro de la ciudad hacia la periferia.

El resumen de la verificación de algunas de las hipótesis, como la relación entre algunas de ellas podemos concluir lo siguiente:

El uso del sistema es muy influenciado por la temperatura en la ciudad. El uso del sistema en los días de semana se infiere que es utilizado para ir a regresar del trabajo o de los centros de estudio, en especial terminales ferroviarias y las universidades más importantes. En los días de semana la duración media del viaje refleja el tiempo necesario para la realización del trayecto. También destaca que el grupo etario que más usa el sistema está entre los 20 y los 30 años. En cuanto al uso de los fines de semana, se observa que el uso es del tipo recreativo, esto se pone de manifiesto por el tiempo promedio de uso, que duplica al día de semana. También aparece que el uso más intenso se da en grupos de mayor edad.

Otro dato importante es que existe un grupo de usuarios frecuentes cercanos al 8%, qué representan el 48% de los viajes.

Conclusiones y Trabajo Futuro

Se puede concluir que en investigaciones futuras tener un mejor perfil de los usuarios, con la incorporación de más datos personales, permitiría entender el comportamiento de los usuarios y cómo afecta esto al uso del sistema de Ecobici.

También será muy importante poder geolocalizar instituciones educativas y otras instituciones gubernamentales, qué permitan estudiar su relación con el origen y destino de los viajes. La aparición de usuarios frecuentes requiere poder analizar si estos usuarios están utilizando al sistema como herramienta de trabajo.

Esto se alinea con las tendencias actuales, donde el uso de los servicios tiene una alta relación con quienes son los usuarios, y que es lo que ellos buscan en cada servicio.

Otro dato necesario para entender si existe demanda insatisfecha por falta de bicicletas es la cantidad de puestos de cada estación. Esto permitirá relacionar la cantidad de puestos con las salidas y llegadas de bicicletas a cada estación. Y detectar posibles problemas de logística de

distribución de bicicletas. Ya que podría existir una demanda insatisfecha, y que la posible falta de bicicletas afecte al inicio de viajes en algunas estaciones.

Queda abierta la posibilidad de obtener los datos adicionales mencionados, que permitan mejorar la profundidad del análisis, y generar una versión más enriquecida del presente documento.

Agradecimientos

Lic. Ariel Aizemberg (tutor)
Lic. Juliana Gambini (consultas)

Bibliografía

A Little Book of R For Multivariate Analysis Release 0.1
Avril Coghlan

R para Principiantes
Emmanuel Paradis

Practical Regression and Anova using R
Julian J. Faraway

Data Analysis and Graphics Using R: An Example-based Approach (Cambridge Series in Statistical and Probabilistic Mathematics)

Métodos estadísticos con R
Jose Saez

Referencias a trabajos similares

Bicycle sharing systems demand
Inês Fradea, Anabela Ribeiroa

Understanding Bike-Sharing Systems using Data Mining: Exploring Activity Pattern Patrick Vogela, Torsten Greisera, Dirk Christian Mattfelda

ANALYSING PUBLIC BIKE SYSTEM USE
CASE STUDY ZHONGSHAN, CHINA PROF. DR. IR. MARTIN VAN MAARSEVEEN

Bicycle-Sharing System Analysis and Trip Prediction
Jiawei Zhang, Xiao Pan†, Moyin Li, Philip S. Yu, University of Illinois at Chicago, Chicago, IL, USA
†Shijiazhuang Tiedao University, China

Apéndice

Código R

```
#  Hipótesis 1 - Analizar si existe relación del uso con la temperatura
#
# leo el archivo
setwd("F:/Postgrado/Taller/datos_tp")
dfgv <- read.table("grados_viajes_2011.csv", header = TRUE, sep =";")
summary(dfgv) names(dfgv) attach(dfgv)
# ========================
# paso a numero
dfgvn = as.matrix(as.data.frame(lapply(dfgv, as.numeric)))
xv<-c (dfgv[,1]) yv<-
c (dfgv[,2])
hist(dfgv$viajes)
pairs(dfgv) cor(dfgv)
# ======================== #
Plot los valores
plot(xv,yv,main="Relación temperatura / viajes",xlab="grados",ylab="viajes")
# ========================
# genero modelo
#m<- nls(yv ~  b1*xv^6   + b2*xv^5  - b3*xv^4  + b4*xv^3  - b5*xv^2  +  b6*xv   - b7,
#        start = list(b1=-0.0014,b2=0.1779 ,b3=7.9595 ,b4=154.13 ,b5=1234.7 ,b6=
4203.4 ,b7=4726.7)  ) m<-
lm(  yv  ~  poly(xv,6)  )
summary(m)
# ======================== #
1- estimación del ajuste
cor(yv,predict(m))
plot(xv,yv,main="Relación temperatura / viajes",xlab="grados",ylab="viajes")
lines(xv,predict(m),lty=2,col="red",lwd=3)
# No aplican en este modelo
#confint(m, level = 0.9)
#abline(m, col="red")
# ======================== #
2 - fitted - residuals
plot(fitted(m),residuals(m),main="Relación fitted /
residuals",xlab="fitted",ylab="residuals")
# ========================
# 3 -Plot con nuevos datos y pruebo la predicción
plot(xv,yv,main="Relación temperatura / viajes",xlab="grados",ylab="viajes") new.dt<-
data.frame(xv = seq(min(xv),max(xv),len = 33))
lines(new.dt$xv,predict(m,newdatab = new.dt),lty=2,col="red",lwd=3)
# ========================
# 4 - suma de los cuadrados de los residuos
print(sum(resid(m)^2)) #
========================
# 5 - Test de normalidad de los residuos
residuosm <- rstandard(m)
qqnorm(residuals(m))
qqline(residuals(m))
shapiro.test(residuals(m)) #
========================
# 6 -  validar modelo
# NA en este modelo
#par(mfrow = c(2,2))
#plot(m)
```

```
#par(mfrow = c(1,1))
# ========================
# 7 - Disgrama de dispersión
# NA en este modelo
# ========================
# 8 -  distribucion de los residuos
# NA en este modelo
#x<-rstandard(m)
#hist(x,probability=TRUE)
#xxh <- seq(min(x), max(x),length=100)
#lines(xxh, dnorm(xxh, mean=0,sd=1),col= 'blue')



#  Hipótesis 2 -Analizar si existe relación del uso con la hora
# Relación hora cantidad para los días de semana
#
# leo el archivo
setwd("F:/Postgrado/Taller/datos_tp")
dfrl <- read.table("hora_viajes.csv", header = TRUE, sep =";")
summary(dfrl) names(dfrl) attach(dfrl) # paso a numero
dfrln = as.matrix(as.data.frame(lapply(dfrl, as.numeric)))
hist(dfrln) xv<-c (dfrl[,1]) yv<-c (dfrl[,2]) pairs(dfrl)
cor(dfrl)
# ======================== #
Plot los valores
plot(dfrl$hora,dfrl$viajes,xlab="hora", ylab="viajes" ,main='Relación de viajes por
hora')
# ========================
# genero modelo
regrel <- lm( dfrl$viajes ~ dfrl$hora   , data = dfrl) summary(regrel)
# ======================== # 1-
estimación del ajuste
cor(dfrl$viajes,predict(regrel))
plot(dfrl$hora,dfrl$viajes,main="Relación hora / viajes",xlab="hora",ylab="viajes")
lines(dfrl$hora,predict(regrel),lty=2,col="red",lwd=3) confint(regrel, level = 0.9)
abline(regrel, col="red") # ======================== # 2 - fitted - residuals
plot(fitted(regrel),residuals(regrel),main="Relación fitted /
residuals",xlab="fitted",ylab="residuals")
# ========================
# 3 - Plot con nuevos datos y pruebo la predicción
# na en este modelo
plot(dfrl$hora,dfrl$viajes,main="Relación hora / viajes",xlab="hora",ylab="viajes")
new.dt<-data.frame(xv = seq(min(xv),max(xv),len = 11))
lines(new.dt$xv,predict(regrel,newdatab = new.dt),lty=2,col="red",lwd=3)
# ========================
# 4 - suma de los cuadrados de los residuos print(sum(resid(regrel)^2))
# ========================
# 5 Test de normalidad de los residuos
residuosrl <- rstandard(regrel)
qqnorm(residuosrl) qqline(residuosrl)
shapiro.test(residuals(regrel)) #
========================
# 6 validar modelo par(mfrow
= c(2,2)) plot(regrel)
par(mfrow = c(1,1)) # 7 -
Disgrama de dispersión
library(ggplot2)
ggplot(dfrl, aes(x=hora, y=viajes)) + geom_point()  + ggtitle("Diagrama de
Dispersión") +
  xlab("hora") + ylab("Viajes") + geom_smooth(method=lm)
```

```
# ======================== # 8 -  distribucion
de los residuos x<-rstandard(regrel)
hist(x,probability=TRUE) xxh <- seq(min(x),
max(x),length=100) lines(xxh, dnorm(xxh,
mean=0,sd=1),col= 'blue')
# ========================


#  Hipótesis 2 - Analizar  si existe relación del uso con la hora el fin de semana #
valido con esto http://blog.minitab.com/blog/adventures-in-statistics-2/why-youneed-
to-check-your-residual-plots-for-regression-analysis
#
# leo el archivo
setwd("F:/Postgrado/Taller/datos_tp")
dfrlpf <- read.table("hora_viajes_F.csv", header = TRUE, sep =";")
# veo el contenido
summary(dfrlpf)
names(dfrlpf)
attach(dfrlpf) #
paso a numero
dfrlpfn = as.matrix(as.data.frame(lapply(dfrlpf, as.numeric)))
hist(dfrlpfn) xv<-c (dfrlpf[,1]) yv<-c (dfrlpf[,2]) # Plot los
valores plot(dfrlpf)
#=======================================
# Genero modelo model<- lm(
yv ~ poly(xv,2) )
summary(model)
# 1 - estimación del ajuste cor(yv,predict(model))
plot(dfrlpf,xlab="hora", ylab="viajes" ,main='Relación de viajes por hora en fin de
semana')
lines(dfrlpf$hora,predict(model),lty=2,col="red",lwd=3)
#=======================================
# 2 - fitted - residuals
plot(fitted(model),residuals(model),main="Relación fitted /
residuals",xlab="fitted",ylab="residuals")
#======================================= # 3 - Plot con nuevos
datos y pruebo la predicción plot(xv,yv,main="viaje x hora
finde",xlab="hora",ylab="viajes") new.dt<-data.frame(xv =
seq(min(xv),max(xv),len = 13))
lines(new.dt$xv,predict(model,newdatab = new.dt),lty=2,col="red",lwd=3)
#======================================= #
4 - suma de los cuadrados de los residuos
print(sum(resid(m)^2))
#======================================= #
5 Test de normalidad de los residuos
residuosrl <- rstandard(model)
qqnorm(residuals(model))
qqline(residuals(model))
shapiro.test(residuals(modeld))
#=======================================
# 6 validar modelo
par(mfrow = c(2,2))
plot(modeld) par(mfrow
= c(1,1))
#=======================================
# 7 - Disgrama de dispersión library(ggplot2)
ggplot(dfrlpf, aes(x=hora, y=viajes)) + geom_point()  + ggtitle("Diagrama de
Dispersión") +
  xlab("hora") + ylab("Viajes") + geom_smooth(method=lm)
#=======================================
```

```
---------------------------------------------------------------------------------------------------------------------------
# Hipótesis 3 -Analizar  si existe relación del uso respecto a la cantidad de
población a 1000
# Relación plblocaión a 1000 metros de la estación y su uso
# seteo el wd
#
# leo el archivo
setwd("F:/Postgrado/Taller/datos_tp")
dfrlc <- read.table("Poblacion_recorridos2.csv", header = TRUE, sep =";")
summary(dfrlc) names(dfrlc) attach(dfrlc)
# ======================== #
paso a numero
# ======================== #
paso a numero
dfrlcn = as.matrix(as.data.frame(lapply(dfrlc, as.numeric)))
hist(dfrlc$cantidad) pairs(dfrlc) cor(dfrlc)
# ======================== #
Plot los valores
plot(dfrlc$poblacion1000m/1000,dfrlc$cantidad/100)
# ======================== #
genero modelo
regrec <- lm( dfrlc$cantidad ~ dfrlc$poblacion1000m , data = dfrlc)
summary(regrec) # ======================== # 1- estimación del ajuste
confint(regrec, level = 0.9) plot(dfrlc$poblacion1000m,dfrlc$cantidad
,xlab="población", ylab="cant. Viajes",,main='Relación cantidad de
población a 1000 metros / viajes' ) abline(lm(dfrlc[,2]~dfrlc[,3]),
col="blue")
# ======================== #
2 - fitted - residuals
plot(fitted(regrec),residuals(regrec),main="Relación fitted /
residuals",xlab="fitted",ylab="residuals")
# ========================
# 3 - Plot con nuevos datos y pruebo la predicción
# na en este modelo
#plot(dfrlc$poblacion1000m,dfrlc$cantidad,main="Relación población /
viajes",xlab="Población",ylab="viajes")
#new.dt<-data.frame(dfrlc =
seq(min(dfrlc$poblacion1000m),max(dfrlc$poblacion1000m),len = 33))
#lines(new.dt$xv,predict(regrec,newdatab = new.dt),lty=2,col="red",lwd=3)
# ========================
# 4 - suma de los cuadrados de los residuos print(sum(resid(regrec)^2))
# ========================
# 5 Test de normalidad de los residuos residuos
<- rstandard(regrec)
qqnorm(residuos) qqline(residuos)
shapiro.test(residuals(regrec)) #
======================== # 6
validar modelo par(mfrow =
c(2,2)) plot(regrec) par(mfrow =
c(1,1)) #
======================== # 7 -
Disgrama de dispersión
library(ggplot2)
ggplot(dfrlc, aes(x=poblacion1000m, y=cantidad)) + geom_point()  + ggtitle("Diagrama
de Dispersión") +
  xlab("Poblacion") + ylab("Cant. Viajes") + geom_smooth(method=lm)
# ======================== # 8 -  distribucion
de los residuos x<-rstandard(regrec)
hist(x,probability=TRUE) xxh <- seq(min(x),
max(x),length=100) lines(xxh, dnorm(xxh,
mean=0,sd=1),col= 'blue')
# ========================
```

```
# Hipótesis 4 - Analizar si existe relación basada en la distancia  entre las
estaciones de subte  #  y las estaciones de ecobici
# Relación cantidad de viajes y distancias de estaciones de metro
#
# leo el archivo
setwd("F:/Postgrado/Taller/datos_tp")
df4 <- read.table("viajes_distancia.csv", header = TRUE, sep =";")
summary(df4)
names(df4) attach(df4)
# ========================
# paso a numero
df4n = as.matrix(as.data.frame(lapply(df4, as.numeric)))
hist(df4$viajes) pairs(df4) cor(df4)
# ======================== #
Plot los valores
plot(df4$distancia,df4$viajes)
# ========================
# genero modelo
regre4 <- lm( df4[,1] ~ df4[,2] , data = df4) summary(regre4)
# ======================== #
1- estimación del ajuste
plot(df4$distancia,df4$viajes,main="Relación ditancia metro /
viajes",xlab="distancia",ylab="viajes")
lines(df4$distancia,predict(regre4),lty=2,col="red",lwd=3)

plot(df4[,2] ,df4[,1],xlab="distancia al Metro", ylab="cant. viajes",,main='Uso /
distancia del metro ' ) confint(regre4, level = 0.9)
abline(lm(df4[,1]~df4[,2]), col="blue")
#abline(m, col="red")
# ======================== #
2 - fitted - residuals
plot(fitted(regre4),residuals(regre4),main="Relación fitted /
residuals",xlab="fitted",ylab="residuals")
# ========================
# 3 -Plot con nuevos datos y pruebo la predicción
#plot(df4$distancia,df4$viajes,main="Relación distancia metro /
viajes",xlab="distancia",ylab="viajes")
#new.dt<-data.frame(xv = seq(min(xv),max(xv),len = 33))
#lines(new.dt$xv,predict(m,newdatab = new.dt),lty=2,col="red",lwd=3)
# ========================
# 4 - suma de los cuadrados de los residuos print(sum(resid(regre4)^2))
# ========================
# 5 - Test de normalidad de los residuos
# NA en este modelo residuos
<- rstandard(regre4)
qqnorm(residuos) qqline(residuos)
shapiro.test(residuals(regre4))
# ========================
# 6 -  validar modelo # NA
en este modelo par(mfrow =
c(2,2)) plot(regre4)
par(mfrow = c(1,1)) #
======================== # 7
- Disgrama de dispersión
library(ggplot2)
ggplot(df4, aes(x=distancia, y=viajes)) + geom_point()  + ggtitle("Diagrama de
Dispersión") +
  xlab("distancia al metro ") + ylab("cant. viajes") + geom_smooth(method=lm)
# ========================
```

```
-----------------------------------------------------------------------------------------------------------------------------
# 8 -  distribucion de los residuos
# NA en este modelo x<-rstandard(regre4)
hist(x,probability=TRUE) xxh <- seq(min(x),
max(x),length=100) lines(xxh, dnorm(xxh,
mean=0,sd=1),col= 'blue')
# ========================
# verlo

# H_5 Analizar si existe la relación entre la edad y el uso
# shapiro test https://rpro.wikispaces.com/Test+de+Shapiro-Wilk
#
# leo el archivo
setwd("F:/Postgrado/Taller/datos_tp")
dfec <- read.table("edad_cantidad.csv", header = TRUE, sep =";")
#dfec <- read.table("edad_cantidad2.csv", header = TRUE, sep =";")
dfec2 <- read.table("edad_cantidad3.csv", header = TRUE, sep =";")
summary(dfec) names(dfec) attach(dfec)
# ========================
# paso a numero
dfecn = as.matrix(as.data.frame(lapply(dfec, as.numeric)))
dfec2n= as.matrix(as.data.frame(lapply(dfec, as.numeric)))
hist(dfec$viajes) xv<-c (dfec[,1]) yv<-c (dfec[,2]) zx<-c
(dfec2[,1]) pairs(dfec) cor(dfec) # Plot valores
plot(dfec)
# ========================
# genero modelo
modeld<- lm(dfec$viajes ~ poly(dfec$edad,6) ) summary(modeld)
# ========================
# 1- estimación del ajuste

plot(dfec$edad,dfec$viajes,main="Relación edad / viajes",xlab="edad",ylab="viajes")
lines(dfec$edad,predict(modeld),lty=2,col="red",lwd=3)
plot(dfec$edad,dfec$viajes,xlab="edad", ylab="cant. viajes",main='Edad / viajes ' )
confint(modeld, level=0.95)
lines(dfec$edad,predict(modeld),lty=2,col="red",lwd=3)
# ======================== #
2 - fitted - residuals
plot(fitted(modeld),residuals(modeld),main="Relación fitted /
residuals",xlab="fitted",ylab="residuals")
# ========================
# 3 -Plot con nuevos datos y pruebo la predicción
plot(dfec$edad,dfec$viajes,main="Relación edad / viajes",xlab="edad",ylab="viajes")
new.dt<-data.frame(zx = seq(min(zx),max(zx),len = 28))
lines(new.dt$zx,predict(modeld,newdatab = new.dt),lty=2,col="red",lwd=3)
# ========================
# 4 - suma de los cuadrados de los residuos print(sum(resid(modeld)^2))
# ========================
# 5 - Test de normalidad de los residuos
# NA en este modelo residuos <-
rstandard(modeld)
qqnorm(residuals(modeld))
qqline(residuals(modeld))
shapiro.test(residuals(modeld))
# ========================
# 6 -  validar modelo # NA
en este modelo par(mfrow =
c(2,2)) plot(modeld)
par(mfrow = c(1,1)) #
======================== # 7
```

```
---------------------------------------------------------------------------------------------------------------------------
- Disgrama de dispersión
library(ggplot2)
ggplot(dfec, aes(x=edad, y=viajes)) + geom_point()  + ggtitle("Diagrama de
Dispersión") +
  xlab("distancia al metro ") + ylab("cant. viajes") + geom_smooth(method=lm)
# ========================
# 8 -  distribucion de los residuos
# NA en este modelo x<-rstandard(modeld)
hist(x,probability=TRUE) xxh <- seq(min(x),
max(x),length=100) lines(xxh, dnorm(xxh,
mean=0,sd=1),col= 'blue')
# ========================


#H H6 # H_5 Analizar si existe relación entre el  Tiempo de uso  con la edad fin de
semana
# cumple parcialmente con 0.57 paro con algunos indicadores que muestran una falta de
ajuste del modelo
#
# leo el archivo
setwd("F:/Postgrado/Taller/datos_tp")
dfeu <- read.table("edad_uso_f.csv", header = TRUE, sep =";")
summary(dfeu) names(dfeu) attach(dfeu)
# ========================
# paso a numero
dfeun = as.matrix(as.data.frame(lapply(dfec, as.numeric)))
hist(dfeu$uso) xv<-c (dfeu[,1]) yv<-c (dfeu[,2])
pairs(dfeu) cor(dfeu)
# ========================
# Plot valores
plot(dfeu$edad,dfeu$uso,xlab='Edad',ylab='tiempo de uso',main="Tiempo de uso en
relación con la edad ") # ========================
# genero modelo
regreu <- lm( dfeu$uso ~ dfeu$edad , data = dfeu) summary(regreu)
# ======================== #
1- estimación del ajuste
plot(dfeu$edad,dfeu$uso ,main="Relación edad / tiempo de uso fin de
semana",xlab="edad",ylab="t.uso")
lines(dfeu$edad,predict(regreu),lty=2,col="red",lwd=3)
plot(dfeu$edad,dfeu$uso ,main="Relación edad / tiempo de uso fin de
semana",xlab="edad",ylab="t.uso") confint(regreu, level=0.95)
lines(dfeu$edad,predict(regreu),lty=2,col="red",lwd=3) abline(regreu)
# ======================== #
2 - fitted - residuals
plot(fitted(regreu),residuals(regreu),main="Relación fitted /
residuals",xlab="fitted",ylab="residuals")
# ========================
# 3 -Plot con nuevos datos y pruebo la predicción
plot(dfeu$edad,dfeu$uso,main="Relación edad / tiempo de uso",xlab="edad",ylab="t.uso")
new.dt<-data.frame(xv = seq(min(xv),max(xv),len = 55))
lines(new.dt$xv,predict(regreu,newdatab = new.dt),lty=2,col="red",lwd=3)
# ========================
# 4 - suma de los cuadrados de los residuos print(sum(resid(regreu)^2))
# ========================
# 5 - Test de normalidad de los residuos
# NA en este modelo residuos
<- rstandard(regreu)
qqnorm(residuos) qqline(residuos)
shapiro.test(residuals(regreu))
# ========================
```

```
---------------------------------------------------------------------------------------------------------------------------
# 6 -  validar modelo # NA
en este modelo par(mfrow =
c(2,2)) plot(regreu)
par(mfrow = c(1,1)) #
======================== # 7
- Diagrama de dispersión
library(ggplot2)
ggplot(dfeu, aes(x=edad, y=uso)) + geom_point()  + ggtitle("Diagrama de Dispersión")
+
  xlab("Edad") + ylab("Uso") + geom_smooth(method=lm)
# ========================
# 8 -  distribucion de los residuos
# NA en este modelo x<-rstandard(regreu)
hist(x,probability=TRUE) xx <- seq(min(x),
max(x),length=100) lines(xx, dnorm(xx,
mean=0,sd=1),col= 'blue')
# ========================
## Dado que los puntos No están bastante alineados, la normalidad también parece
aceptable
stdres = rstandard(regreu) plot(dfeu$edad,stdres)
# H7 determinar la cantidad de centroides en el clustrer - Clasificación de usuarios
en cluster
#  cantidad de cluster es ==> 6
setwd("F:/Postgrado/Taller/datos_tp")
mydata <- read.table("viajes_x_usuario.csv", header = TRUE, sep =";")
head(mydata) # Preparo
mydata <- na.omit(mydata) # listwise deletion of missing mydata
<- scale(mydata) # standardize variables
# Determine number of clusters wss <- (nrow(mydata)-
1)*sum(apply(mydata,2,var)) for (i in 2:10) wss[i] <-
sum(kmeans(mydata, centers=i)$withinss) plot(1:10, wss, type="b",
xlab="Número of Clusters",      ylab="suma de cuadrados")
```