Análisis y predicción de datos abiertos MIBICI Guadalajara

Guillermo Segura Gómez

Avance del proyecto

Introducción

Para construir un modelo predictivo eficaz sobre cualquier fenómeno, es necesario comenzar con una base sólida de datos. Un "dataset" se puede definir como una colección de observaciones sobre un fenómeno específico [1]. Sin embargo, un dataset por sí solo no es suficiente para realizar un correcto análisis sobre los datos. Una buena visualización y un correcto análisis de los datos son cruciales para comprender y extraer inferencias significativas del conjunto de datos. Las herramientas gráficas juegan un papel vital en este proceso, ya que permiten resumir y destacar características clave de los datos. Mientras tanto, los análisis de inferencia se centran en formular y explorar preguntas específicas, lo que conduce a la construcción de modelos estadísticos.

En este trabajo, nos enfocamos en un análisis exploratorio del sistema Mibici, un servicio de bicicletas compartidas en Guadalajara. Este análisis no solo proporciona una visión general del perfil y los patrones de uso de los usuarios del sistema, sino que también incluye un análisis de regresión detallado. El objetivo es responder a preguntas específicas planteadas sobre los patrones de uso y las preferencias de los usuarios, utilizando tanto técnicas de visualización de datos como métodos estadísticos avanzados para obtener insights valiosos y orientados a la acción.

Carga de datos

Para comenzar con el análisis primero es necesario cargar los datos. Se utiliza la codificación UTF-8 para poder leer los datos desde R. Los datos escogidos son del mes de mayo de 2023, esto es por elección personal ya que el mes de mayo es el mes de mi cumpleaños. El análisis hasta ahora solo contempla este año, pero sería algo sumamente interesante explorar comparativas por año. En las preguntas se abordará un poco mas acerca de esto.

```
datos <- read.csv("datos_abiertos_2023_05.csv", fileEncoding = "UTF-8")
```

Se realiza una inspección incial.

```
str(datos)
```

```
## 'data.frame':    364158 obs. of  8 variables:
##  $ Viaje_Id        : int  28467098 28467099 28467100 28467101 28467102 28467103 28467104 28467105 284
##  $ Usuario_Id      : int  70123 2237235 2051727 2246225 324247 1723717 440033 515507 607361 1617437
##  $ Genero          : chr  "M" "M" "F" "M" ...
##  $ Año_nacimiento  : int  1967 1980 2002 1969 1975 1995 1988 1984 1988 1996 ...
##  $ Inicio_del_viaje: chr  "2023-05-01 00:00:03" "2023-05-01 00:00:23" "2023-05-01 00:01:05" "2023-05
##  $ Fin_del_viaje   : chr  "2023-05-01 00:22:19" "2023-05-01 00:04:26" "2023-05-01 00:10:21" "2023-05
##  $ Origen_Id       : int  64 36 96 33 226 257 261 25 50 8 ...
##  $ Destino_Id      : int  141 172 296 255 231 54 265 279 259 271 ...
```

```r
summary(datos)
```

```
##     Viaje_Id          Usuario_Id        Genero          Año_nacimiento
##  Min.   :28467098   Min.   :    102   Length:364158      Min.   :1920
##  1st Qu.:28574350   1st Qu.: 442464   Class :character   1st Qu.:1984
##  Median :28684164   Median :1140731   Mode  :character   Median :1992
##  Mean   :28683190   Mean   :1132535                      Mean   :1989
##  3rd Qu.:28791049   3rd Qu.:1737235                      3rd Qu.:1997
##  Max.   :28897246   Max.   :2371510                      Max.   :2022
##                                                          NA's   :557
##  Inicio_del_viaje   Fin_del_viaje        Origen_Id     Destino_Id
##  Length:364158      Length:364158      Min.   :  2   Min.   :  2
##  Class :character   Class :character   1st Qu.: 51   1st Qu.: 51
##  Mode  :character   Mode  :character   Median :132   Median :120
##                                        Mean   :138   Mean   :138
##                                        3rd Qu.:224   3rd Qu.:232
##                                        Max.   :327   Max.   :327
##
```

Existen muchos valores NaN, en año de nacimiento. Es necesario hacer una limpieza de los datos.

Limpieza de los datos

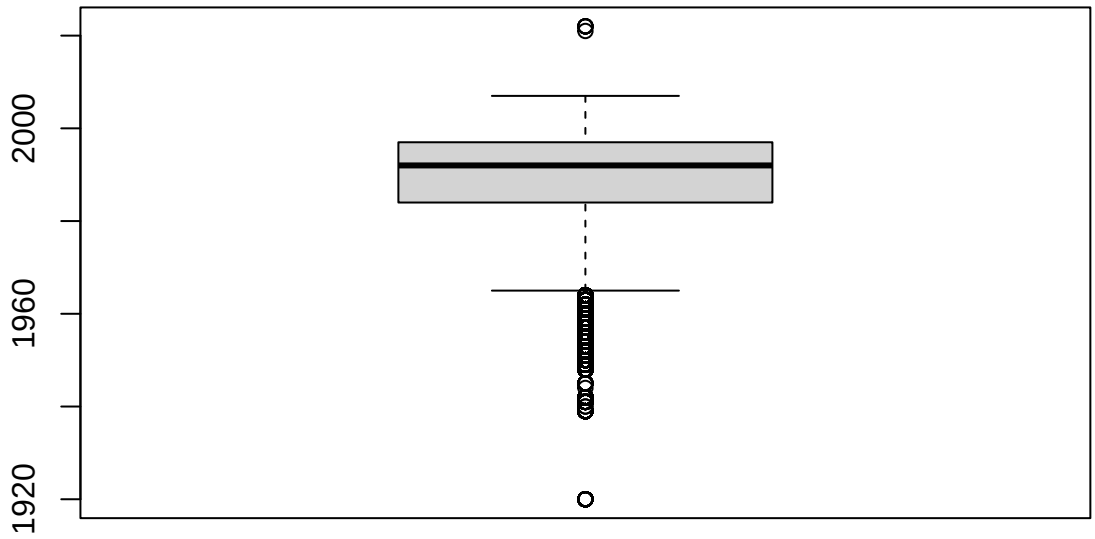
```r
datos <- na.omit(datos)
summary(datos)
```

```
##     Viaje_Id          Usuario_Id        Genero          Año_nacimiento
##  Min.   :28467098   Min.   :    102   Length:363601      Min.   :1920
##  1st Qu.:28574357   1st Qu.: 441551   Class :character   1st Qu.:1984
##  Median :28684195   Median :1140370   Mode  :character   Median :1992
##  Mean   :28683196   Mean   :1131974                      Mean   :1989
##  3rd Qu.:28791073   3rd Qu.:1740054                      3rd Qu.:1997
##  Max.   :28897246   Max.   :2371510                      Max.   :2022
##  Inicio_del_viaje   Fin_del_viaje        Origen_Id     Destino_Id
##  Length:363601      Length:363601      Min.   :  2   Min.   :  2
##  Class :character   Class :character   1st Qu.: 51   1st Qu.: 51
##  Mode  :character   Mode  :character   Median :132   Median :125
##                                        Mean   :138   Mean   :138
##                                        3rd Qu.:224   3rd Qu.:232
##                                        Max.   :327   Max.   :327
```

Además de limpiar los valores NaN, se tiene que trabajar con los valores atipicos, puesto que por ejemplo en el campo "Año de nacimiento" el valor máximo es el año 2022. Es evidente que una persona nacida en el 2022 no puede utilizar una bicicleta.

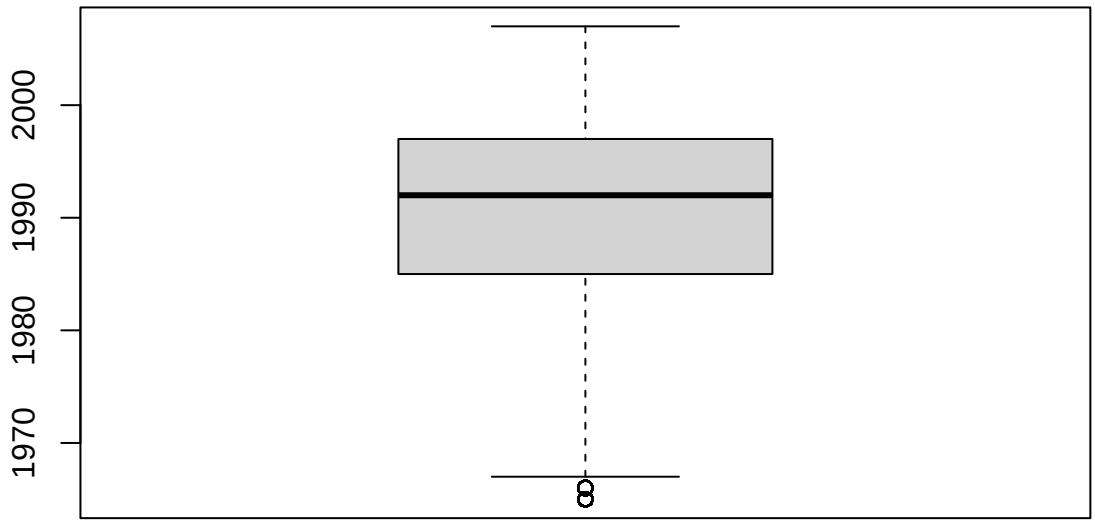
Para la correcta visualización de los outliers se utiliza un boxplot. Se tienen que eliminar los valores atipicos.

```r
boxplot(datos$Año_nacimiento)
```




```
# Identificar valores atípicos y eliminarlos
outliers <- boxplot.stats(datos$Año_nacimiento)$out
datos <- datos[!datos$Año_nacimiento %in% outliers,]
```

```
boxplot(datos$Año_nacimiento)
```




```
head(datos)
```

```
##   Viaje_Id Usuario_Id Genero Año_nacimiento    Inicio_del_viaje
## 1 28467098      70123      M           1967 2023-05-01 00:00:03
## 2 28467099    2237235      M           1980 2023-05-01 00:00:23
## 3 28467100    2051727      F           2002 2023-05-01 00:01:05
## 4 28467101    2246225      M           1969 2023-05-01 00:01:07
## 5 28467102     324247      M           1975 2023-05-01 00:01:26
## 6 28467103    1723717      M           1995 2023-05-01 00:01:37
##         Fin_del_viaje Origen_Id Destino_Id
## 1 2023-05-01 00:22:19        64        141
## 2 2023-05-01 00:04:26        36        172
## 3 2023-05-01 00:10:21        96        296
## 4 2023-05-01 00:04:14        33        255
## 5 2023-05-01 00:13:18       226        231
## 6 2023-05-01 00:17:22       257         54
```

```
str(datos)
```

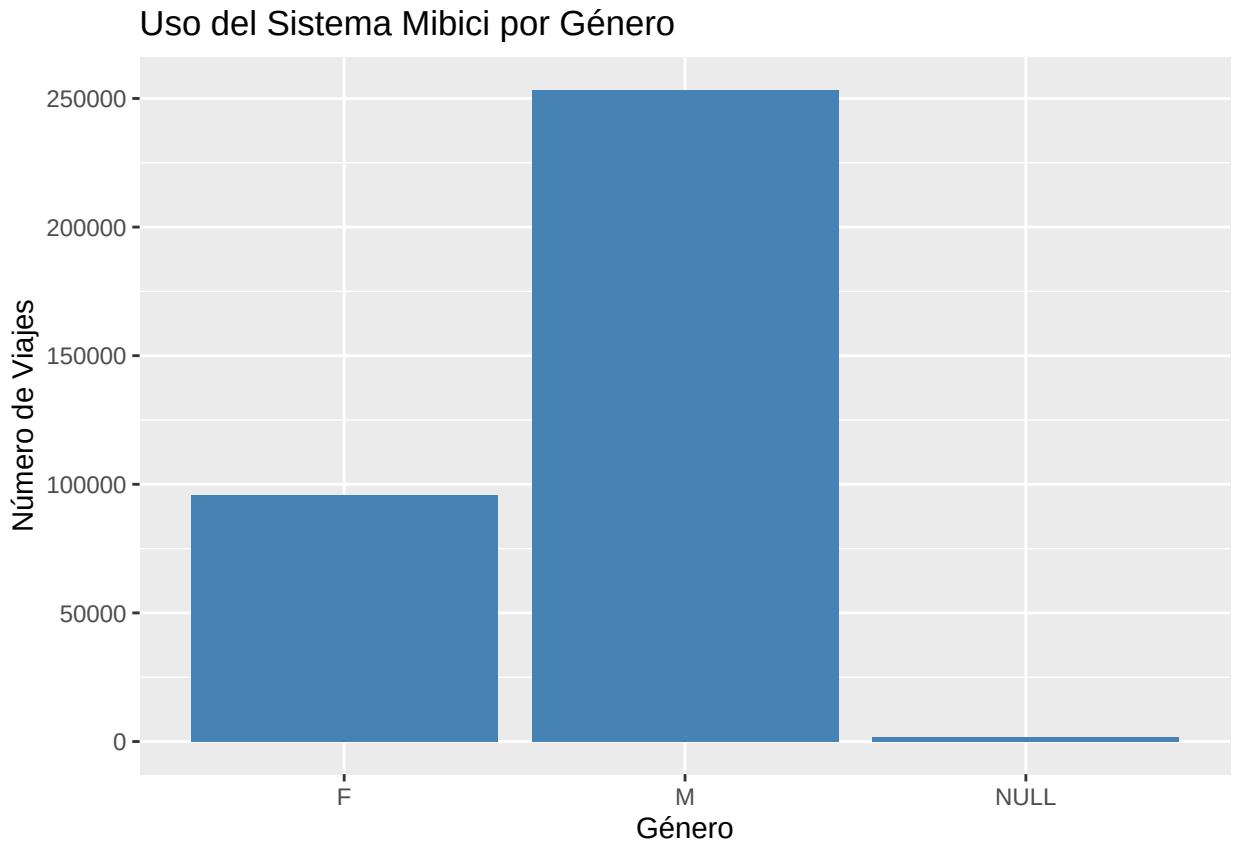
```
## 'data.frame':    351007 obs. of  8 variables:
##  $ Viaje_Id        : int  28467098 28467099 28467100 28467101 28467102 28467103 28467104 28467105 284
##  $ Usuario_Id      : int  70123 2237235 2051727 2246225 324247 1723717 440033 515507 607361 1617437
##  $ Genero          : chr  "M" "M" "F" "M" ...
##  $ Año_nacimiento  : int  1967 1980 2002 1969 1975 1995 1988 1984 1988 1996 ...
##  $ Inicio_del_viaje: chr  "2023-05-01 00:00:03" "2023-05-01 00:00:23" "2023-05-01 00:01:05" "2023-05-
```

```
##  $ Fin_del_viaje   : chr  "2023-05-01 00:22:19" "2023-05-01 00:04:26" "2023-05-01 00:10:21" "2023-05-
##  $ Origen_Id       : int  64 36 96 33 226 257 261 25 50 8 ...
##  $ Destino_Id      : int  141 172 296 255 231 54 265 279 259 271 ...
##  - attr(*, "na.action")= 'omit' Named int [1:557] 1022 1778 2633 4773 4881 4932 5202 5252 6056 6269
##   ..- attr(*, "names")= chr [1:557] "1022" "1778" "2633" "4773" ...
```

Exploración gráfica

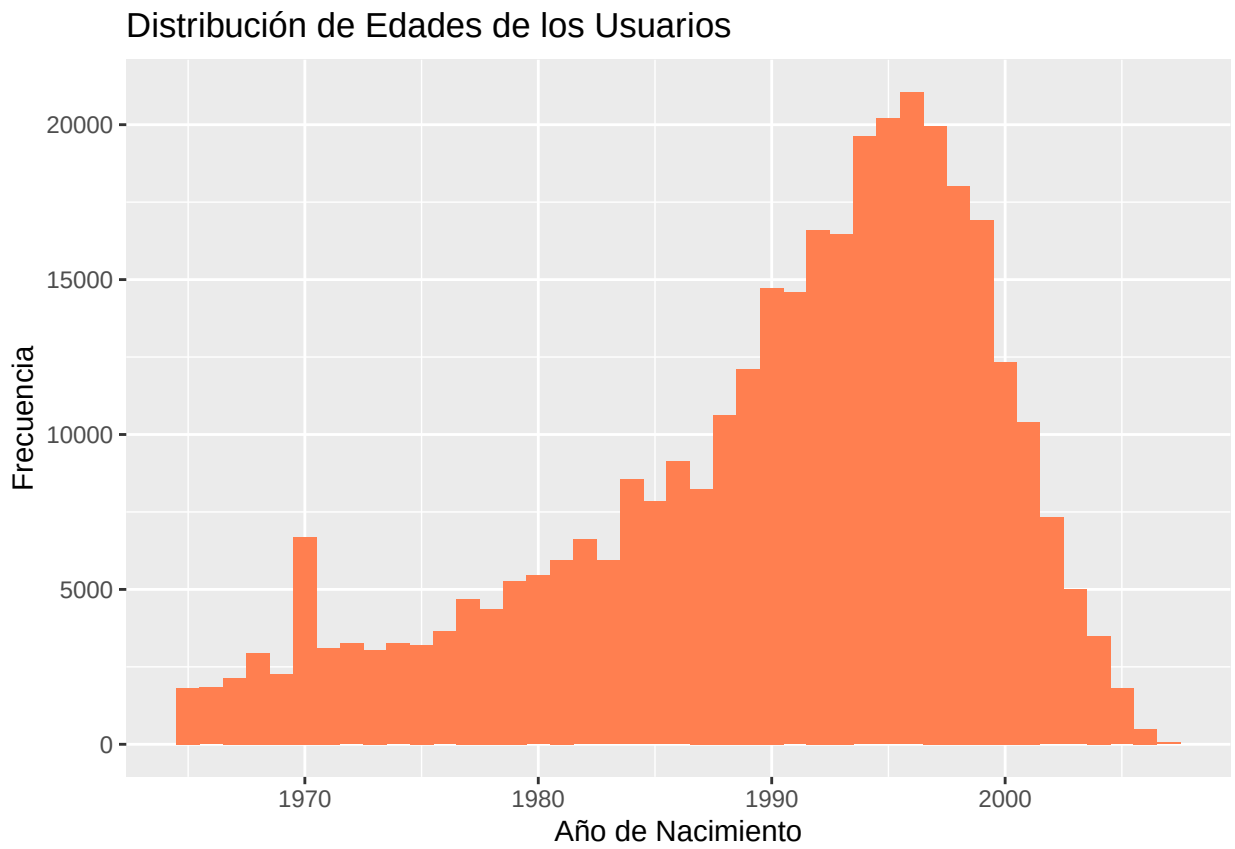
Antes de plantear las preguntas es necesario hacer una exploración gráfica para identificar comportamientos interesantes.

```
library(ggplot2)
ggplot(datos, aes(x=Genero)) +
  geom_bar(fill="steelblue") +
  labs(title="Uso del Sistema Mibici por Género", x="Género", y="Número de Viajes")
```



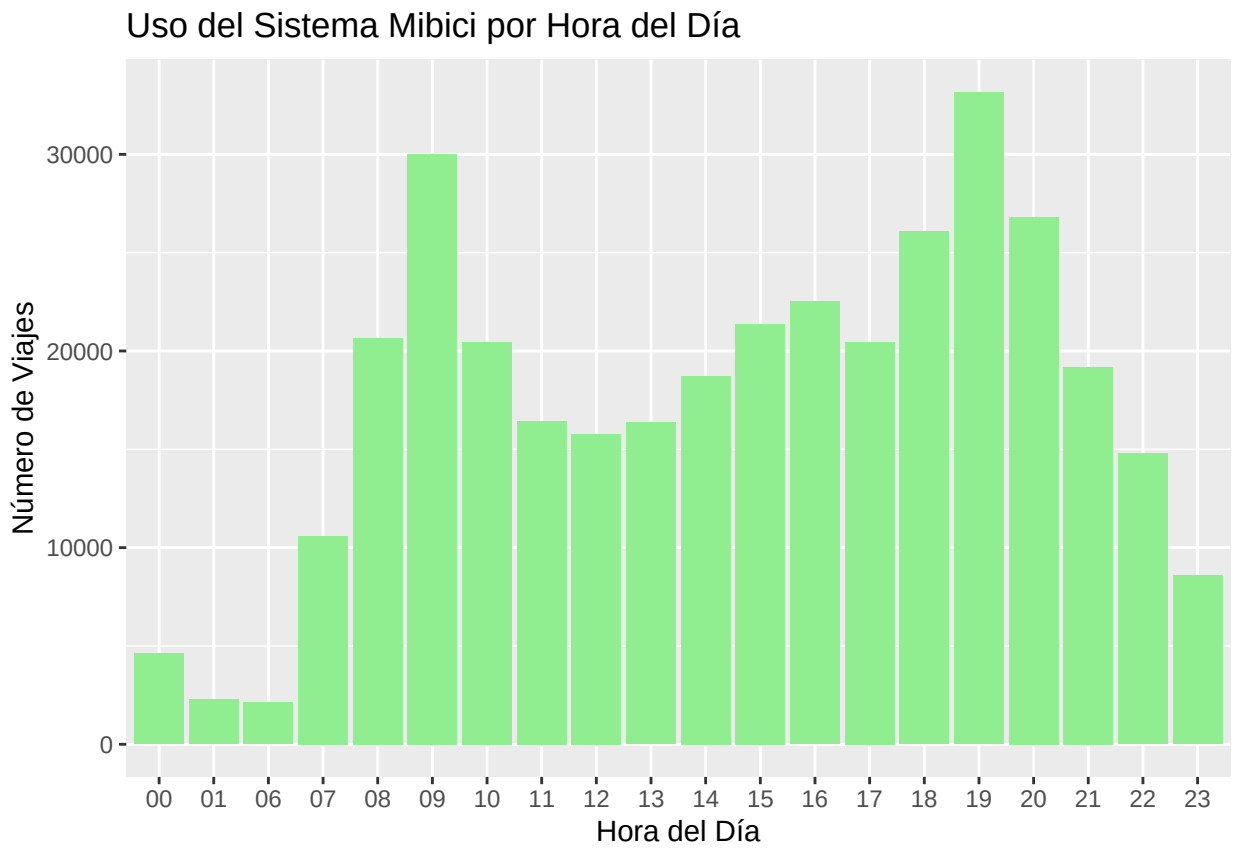
Hay áreas de oportunidad a mejorar en la gráfica anterior. Parece que aún existen algunos datos nulos.

```
ggplot(datos, aes(x=Año_nacimiento)) +
  geom_histogram(binwidth=1, fill="coral") +
  labs(title="Distribución de Edades de los Usuarios", x="Año de Nacimiento", y="Frecuencia")
```



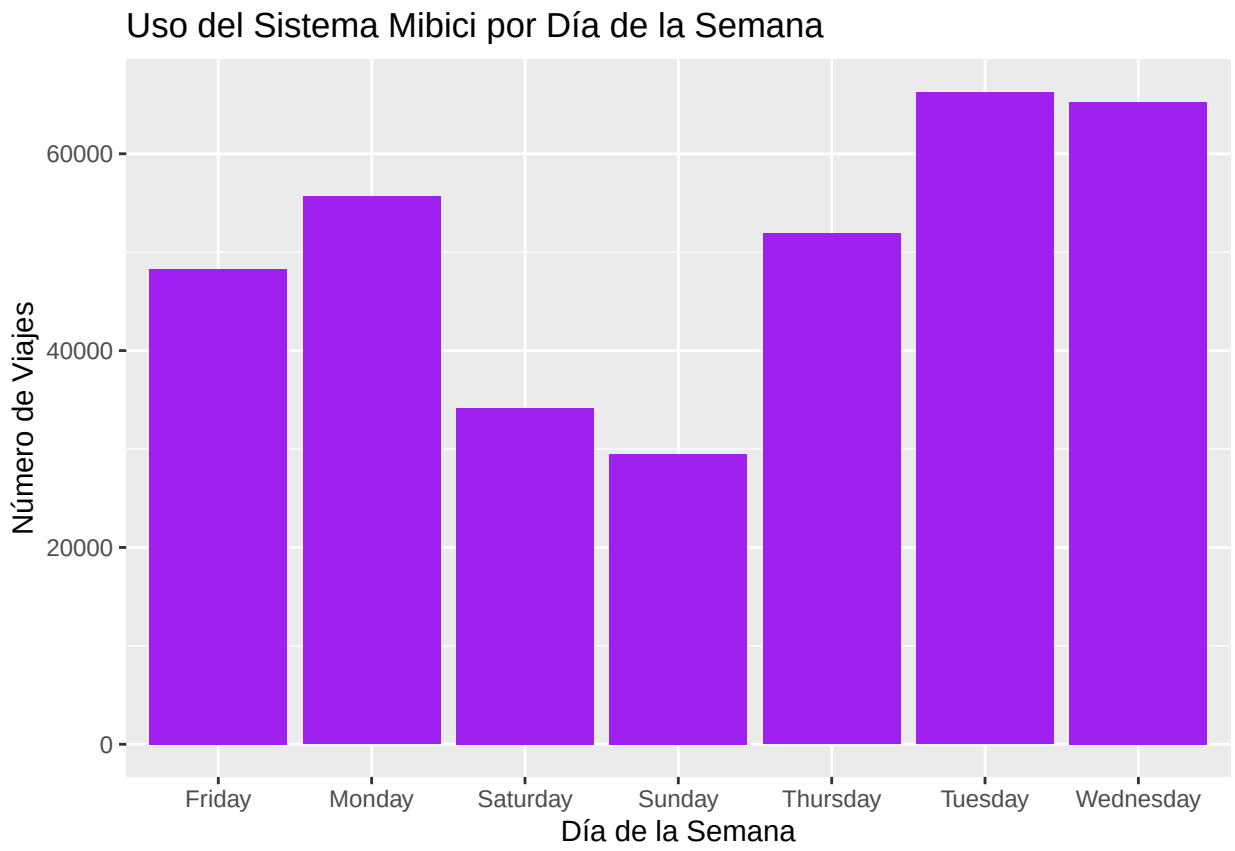

```r
# Extraer la hora del día de la columna 'Inicio_del_viaje'
datos$Hora_del_dia <- format(as.POSIXct(datos$Inicio_del_viaje, format="%Y-%m-%d %H:%M:%S"), "%H")

ggplot(datos, aes(x=Hora_del_dia)) +
  geom_bar(stat="count", fill="lightgreen") +
  labs(title="Uso del Sistema Mibici por Hora del Día", x="Hora del Día", y="Número de Viajes")
```



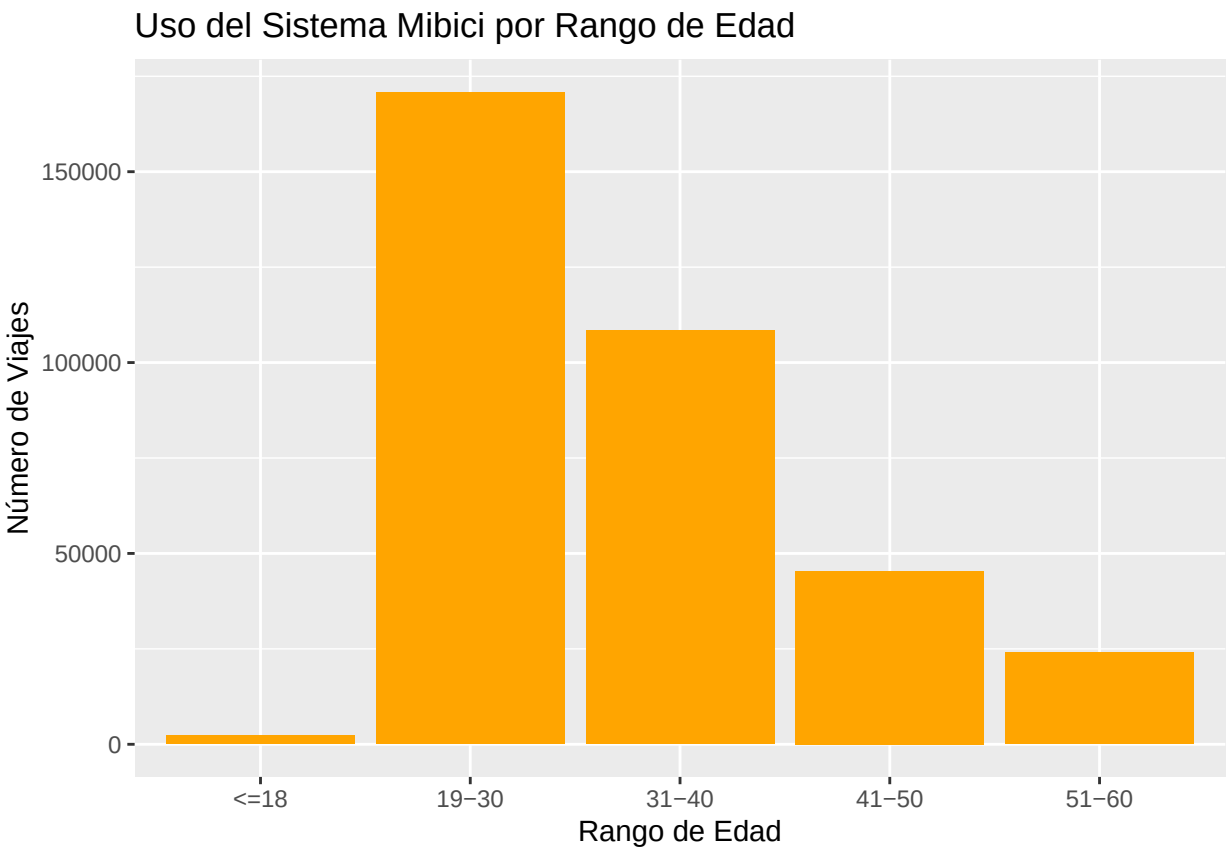

```r
# Convertir 'Inicio_del_viaje' a formato de fecha y extraer el día de la semana
datos$Dia_de_la_semana <- weekdays(as.Date(datos$Inicio_del_viaje))

ggplot(datos, aes(x=Dia_de_la_semana)) +
  geom_bar(fill="purple") +
  labs(title="Uso del Sistema Mibici por Día de la Semana", x="Día de la Semana", y="Número de Viajes")
```




```r
# Crear una nueva columna para rangos de edad
datos$Rango_edad <- cut(as.numeric(format(Sys.Date(), "%Y")) - datos$Año_nacimiento,
                        breaks=c(0, 18, 30, 40, 50, 60, Inf),
                        labels=c("<=18", "19-30", "31-40", "41-50", "51-60", "60+"))

ggplot(datos, aes(x=Rango_edad)) +
  geom_bar(fill="orange") +
  labs(title="Uso del Sistema Mibici por Rango de Edad", x="Rango de Edad", y="Número de Viajes")
```



De las gráficas podemos extraer algunas conclusiones a simple vista. Por ejemplo la gráfica del número de viajes en función de la hora del día presenta un comportamiento bimodal. Lo cual puede ser explicado debido a la hora de entrada y salida de las personas del trabajo. Habría que comparar con el horario laboral de las personas en Guadalajara. Esto puede ser reforzado por la gráfica del uso del sistema por semana, ya que los fines de semana es cuando menos se utiliza el sistema. Sugiriendo que la mayoría de los usuarios utilizan mibici para transportarse al trabajo.

Con base a estas gráficas y la exploración de datos se plantean las preguntas que se buscará responder utilizando modelos predictivos, asi como el uso de mas y mejores gráficas específicas para cada caso.

Preguntas

¿Cómo influyen las horas pico laborales en el uso del sistema Mibici?

Esto puede tener la ventaja de poder configurar la disponibilidad del sistema para asi mejorar la calidad el servicio.

Modelo a utilizar: Regresión lineal para analizar la relación entre las horas del día y el número de viajes, ajustando por día de la semana.

¿Cuál es el perfil demográfico predominante entre los usuarios de Mibici?

Conocer el perfil demográfico predominante puede ayudar en la planificación de campañas de marketing y en la adaptación del servicio a las necesidades de los usuarios más frecuentes.

Modelo a utilizar: Análisis descriptivo seguido de regresión logística para explorar la relación entre características demográficas (edad, género) y la frecuencia de uso.

¿Se puede predecir la demanda de bicicletas en diferentes estaciones basándose en factores como la hora del día, el día de la semana o condiciones climáticas?

Predecir la demanda puede mejorar la gestión de inventario y la distribución de bicicletas, asegurando que haya suficientes bicicletas disponibles donde y cuando se necesiten.

Modelo a utilizar: Revisar si se puede implentar algún tipo de modelos predictivos multivariados variables predictoras y relaciones no lineales.

¿Existe una correlación entre la duración del viaje y factores como la edad o el género del usuario?

Entender cómo diferentes factores demográficos afectan la duración del viaje puede ayudar a personalizar y mejorar la experiencia del usuario, así como a planificar mejor las rutas y la disponibilidad de bicicletas.

Modelo a utilizar: Regresión múltiple para evaluar cómo diferentes variables (edad, género, hora del día) influyen en la duración del viaje.

Referencias

[1] Dekking, Frederik Michel, et al. A Modern Introduction to Probability and Statistics: Understanding why and how. Vol. 488. London: springer, 2005.