Universidad Andrés Bello®

# Actividad Evaluativa #1
# Grupo 5
# Universidad Andrés Bello
# Facultad de Ingeniería
# Ingeniería Civil Informática
# Ingeniería Civil Industrial

Martín Fernández 1[1], Tomás Moya 2, Wesly Ocampo 3[1], Alan Tovar 4[1]

**Abstract**

Describir brevemente en qué consiste este primer análisis de los datos, e incluir el objetivo del estudio. (máximo 150 palabras)

*Corresponding author

*Email addresses:* nombre@uandresbello.edu (Martín Fernández 1), nombre@uandresbello.edu (Tomás Moya 2), nombre@uandresbello.edu (Wesly Ocampo 3), nombre@uandresbello.edu (Alan Tovar 4)

**Introducción**

El preprocesamiento de datos y la administración de los mismos nos permiten la recolección de datos de distintas fuentes, el tratamiento de filas y cabeceras (headers o columnas). Con esto podemos hacer una limpieza adecuada, eliminar errores, corregir inconsistencias y aumentar la calidad de la minería de datos, una correcta gestión de datos también nos sirve para acceder a diferentes datos de una forma más fácil, así obtenemos información estadística y comparativa que permite una correcta toma de decisiones y abarcar de mejor manera los distintos problemas empresariales.

En este caso nuestra base de datos es de una renta de bicicletas que cuenta con 5856 observaciones clasificadas en las siguientes 14 variables con su tipo de variable correspondiente:

1. Fecha (Categórica)
2. Recuento de bicicletas alquiladas (Numérica)
3. Hora (Numérica)
4. Temperatura (Numérica)
5. Humedad (Numérica)
6. Velocidad del viento (Numérica)
7. Visibilidad (Numérica)
8. Temperatura de punto de rocío (Numérica)
9. Radiación solar (Numérica)
10. Lluvia (Numérica)
11. Nevada (Numérica)
12. Estaciones (Categórica)
13. Vacaciones (Categórica)
14. Día de funcionamiento (Categórica)

Nuestro objetivo es analizar esta base de datos, variables categóricas y numéricas para ver su distribución, determinar si hay datos atípicos, datos faltantes y ver como se relacionan las distintas variables entre sí.

**Desarrollo**

*Resumen de medidas estadísticas*

1

Table 1: Resumen de medidas estadísticas.

| variables | min | Q1 | mean | median | Q3 | max | zero | minus | outlie |
|---|---|---|---|---|---|---|---|---|---|
| Rented_Bike_Count | 0 | 327.75 | 905.83 | 813.00 | 1298.50 | 3556.00 | 295 | 0 | 5 |
| Hour | 0 | 5.75 | 11.50 | 11.50 | 17.25 | 23.00 | 244 | 0 | |
| Temperature | -3 | 12.90 | 19.19 | 19.90 | 25.20 | 39.40 | 1 | 21 | |
| Humidity | 0 | 46.00 | 61.22 | 61.00 | 77.00 | 98.00 | 17 | 0 | |
| Windspeed | 0 | 0.90 | 1.63 | 1.50 | 2.20 | 7.40 | 48 | 0 | 10 |
| Visibility | 27 | 1003.00 | 1470.78 | 1731.00 | 2000.00 | 2000.00 | 0 | 0 | |
| Dew_point_temperature | -19 | 4.10 | 10.71 | 11.40 | 18.70 | 27.20 | 43 | 726 | |
| Solar_Radiation | 0 | 0.00 | 0.67 | 0.05 | 1.17 | 3.52 | 2664 | 0 | 19 |
| Rainfall | 0 | 0.00 | 0.20 | 0.00 | 0.00 | 35.00 | 5392 | 0 | 46 |
| Snowfall | 0 | 0.00 | 0.02 | 0.00 | 0.00 | 8.80 | 5805 | 0 | 5 |

*Análisis de la variable Seasons*

En la figura 1 solo tiene 3 estaciones, podemos asumir que no existe recolección de datos durante el invierno debido a las características climáticas, presentando un clima demasiado frió y con posible nieve que son razones por las cuales los clientes no quieran rentan bicicletas y que la empresa no opere durante ese tiempo por las nulas ganancias.

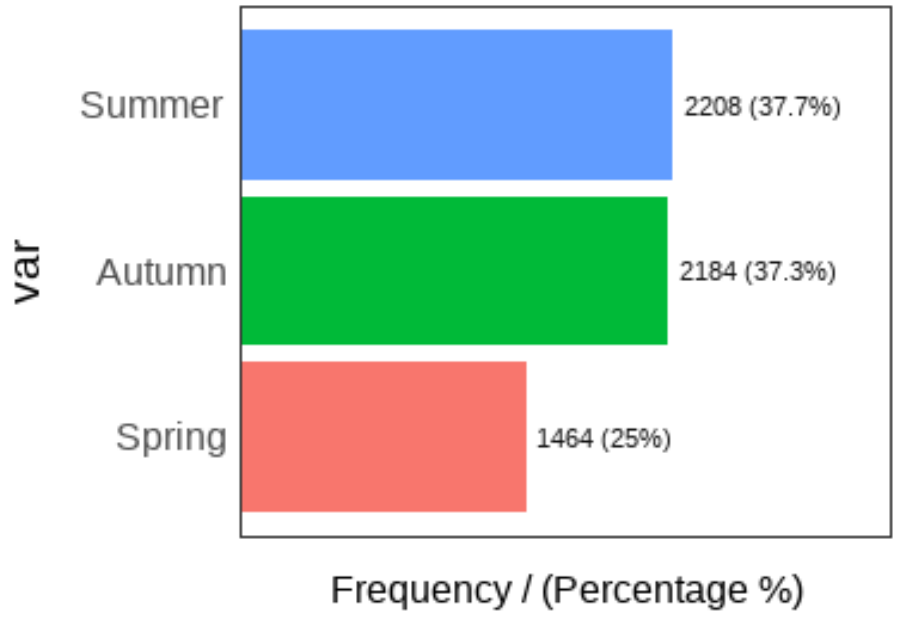


Figure 1: Histograma de Seasons.

*Análisis de la variable Bicicletas Rentadas Rented Bike Count*

Table 2: Tabla de frecuencias de Bicicletas Rentadas Rented Bike Count.

| variables | levels | N | freq | ratio | rank |
|---|---|---|---|---|---|
| Rented_Bike_Count | (140,160] | 641 | 91 | 14.20 | 1 |
| Rented_Bike_Count | (20,40] | 641 | 87 | 13.57 | 2 |
| Rented_Bike_Count | (160,180] | 641 | 82 | 12.79 | 3 |
| Rented_Bike_Count | (0,20] | 641 | 80 | 12.48 | 4 |
| Rented_Bike_Count | (120,140] | 641 | 75 | 11.70 | 5 |
| Rented_Bike_Count | (40,60] | 641 | 72 | 11.23 | 6 |
| Rented_Bike_Count | (60,80] | 641 | 56 | 8.74 | 7 |
| Rented_Bike_Count | (100,120] | 641 | 54 | 8.42 | 8 |
| Rented_Bike_Count | (80,100] | 641 | 44 | 6.86 | 9 |

El gráfico del histograma con curva de densidad correspondiente a la figura 2 podemos analizar que la mayoría de las frecuencias está concentrada en los primeros intervalos, es decir la variable presenta una asimetría positiva y una variación heterogénea, presentando una tendencia a una mayor densidad mientras se esté más cerca del cero mientras menos bicicletas hayan sido alquiladas

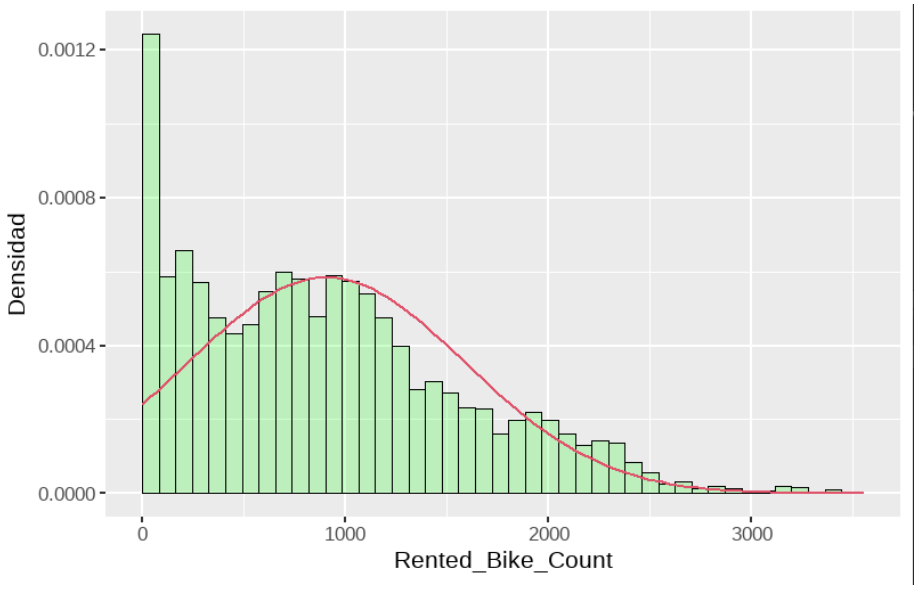


Figure 2: Histograma con curva de densidad de Bicicletas Rentadas Rented Bike Count.

Según el gráfico Q-Q obtenido correspondiente a la figura 3 se puede concluir que los datos en su mayoría mantienen una distribución asimétrica positiva.

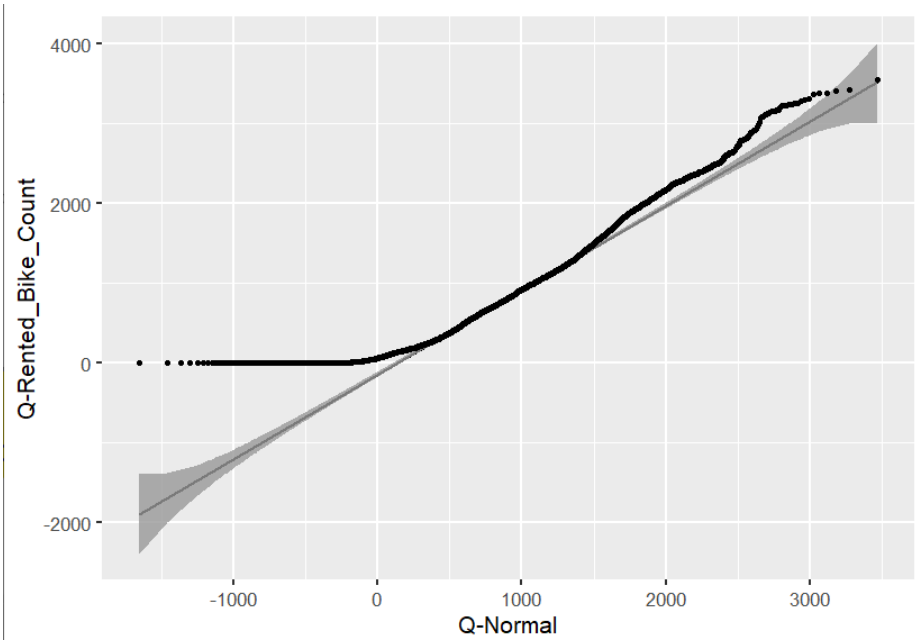


Figure 3: Q-Q plot de Bicicletas Rentadas Rented Bike Count.

Segun el test de kolmogorov-smirnov de la variable bicicletas rentadas el valor de probabilidad es menor a 2.2e-16. Por ende rechazamos la hipótesis nula al ser nuestro valor no mayor a 0.05, es decir la distribución no es normal.

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  datos$Rented_Bike_Count
## D = 0.092264, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

*Valores Atípicos*

En este gráfico de boxplot correspondiente a la figura 4 observamos los datos atípicos, en esta variable previamente vimos que teníamos 55 datos atípicos, todos son superiores, no tenemos datos atípicos inferiores y se encuentran en un rango entre 2500 y 4000 aproximadamente.

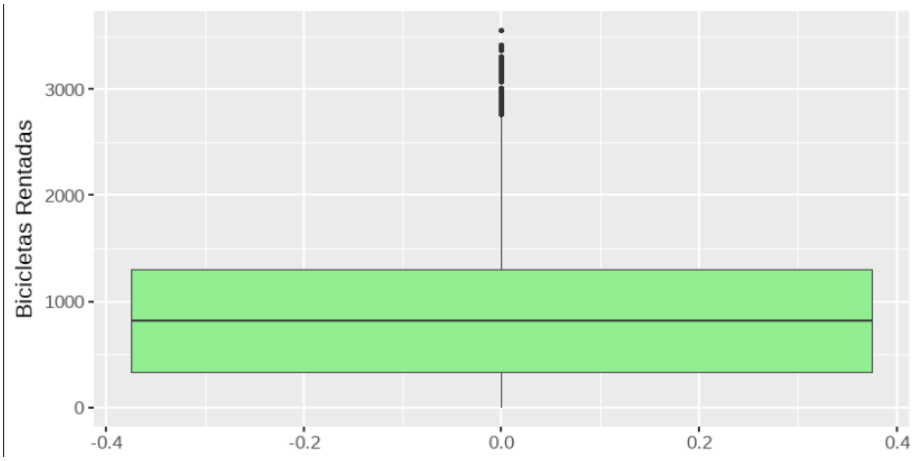


Figure 4: Datos Atípicos de Bicicletas Rentadas Rented Bike Count.

En este caso el boxplot de la figura 5 relacionamos las variables categórica y numérica estudiadas anteriormente, aquí podemos observar las bicicletas rentadas según la estación Podemos concluir que todas las estaciones tienen valores atípicos y si vemos las medianas sabremos que en verano se rentan más bicicletas.

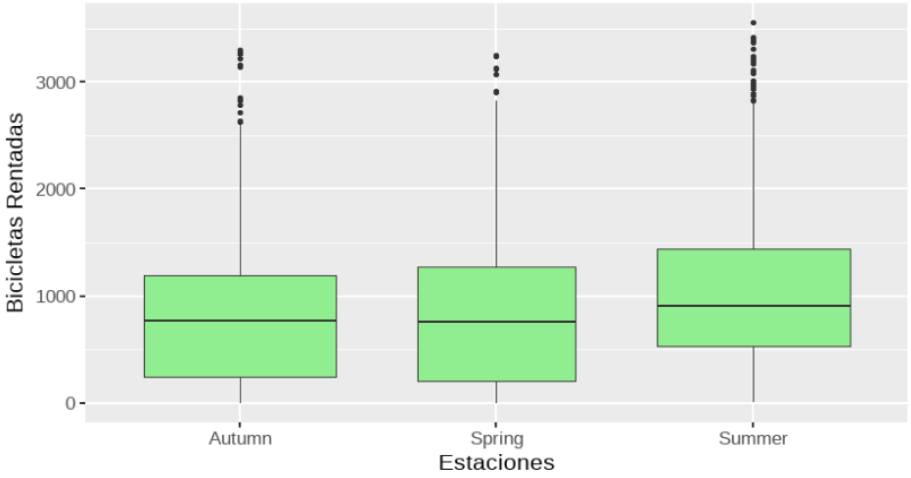


Figure 5: Boxplot de la variable Seasons.

*Datos Faltantes*

Determinar proporción de datos faltantes

En este gráfico de proporcion de datos faltantes, figura 6podemos observar que ninguna variable de nuestra base de datos tiene datos perdidos. Por ende no es necesario realizar ninguna imputación

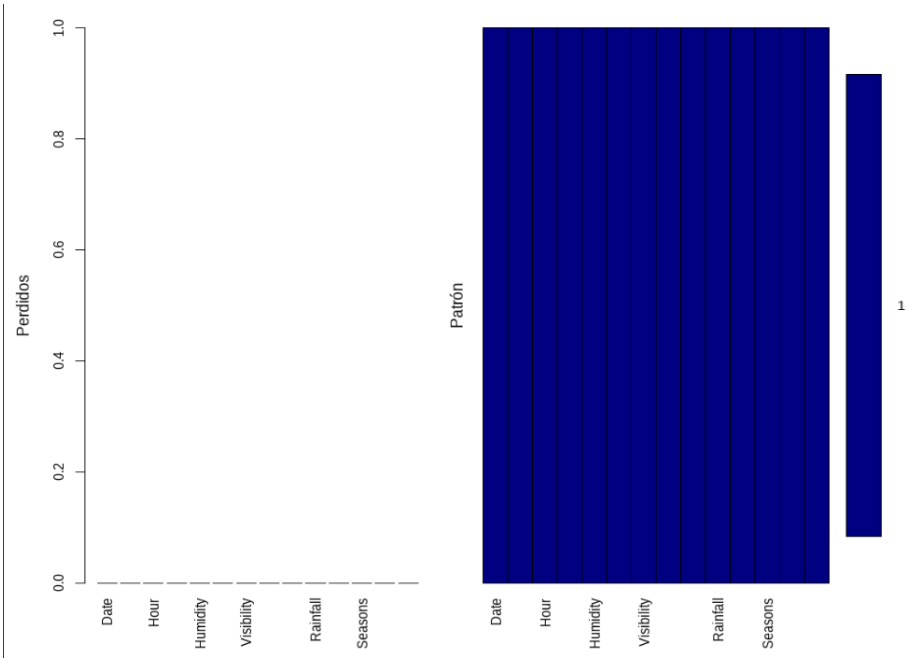


Figure 6: Gráfico de Datos faltantes.

*Análisis de Correlación*

En este grafico correspondiente a la figura 7 podemos observar cómo se relacionan entre si las variables de nuestra base de datos, es decir, que tanto afecta una a la otra. En el caso de nuestra base de datos podemos observar Correlación nula o despreciable entre Temperatura de rocío y bicicletas alquiladas, hora, velocidad del viento, radiación solar, también entre Lluvia y hora, temperatura, velocidad del viento y entre Radiación solar y visibilidad Por otro lado, con Correlación fuerte positiva tenemos a Temperatura de rocío y temperatura, humedad o a Bicicletas alquiladas y hora Por último, tenemos Correlación fuerte negativa entre Humedad y radiación solar, visibilidad, bicicletas rentadas.

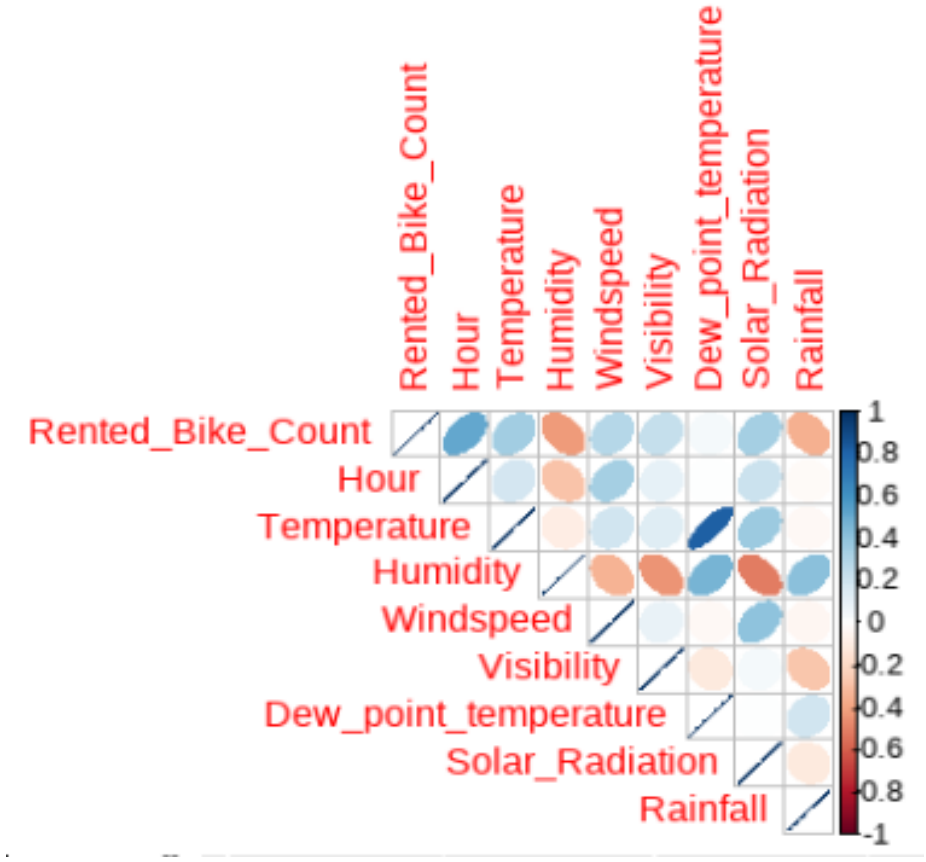


Figure 7: Matriz de Correlación.

**Conclusiones**

Resuma las principales conclusiones de cada análisis realizado como parte del desarrollo.

Luego de este análisis podemos concluir que el análisis exploratorio de datos es indispensable para organizar los datos, entender su contenido, visualizar y extraer información relevante del set de datos para poder decidir cuál será la técnica más adecuada para procesarlos posteriormente. En el caso de nuestra base de datos definimos sus variables, sus tipos, luego del análisis de la variable bicicletas rentadas concluimos que su distribución es asimétrica positiva, como se relacionaban las variables entre sí, no teníamos datos faltantes, contábamos con datos atípicos en algunas variables, no contábamos con datos en invierno por lo que creemos que la empresa no trabajaba en esta estación ya que es muy difícil transportarse en bicicleta en climas extremadamente bajos y también pudimos ver que en el verano se rentaba una mayor cantidad de bicicletas.