Universidad Nacional Autónoma de México

Facultad de Ciencias, 2024 - 1

Almacenes y Minería de Datos

Tarea 04. Análisis de Datos

PostgresandoesoSQLazos

Integrantes

Adrian Aguilera Moreno	421005200
Marco Antonio Rivera Silva	318183583
Sebastían Alejandro Gutiérrez Medina	318287021
Israel Hernández Dorantes	318206604
Alejandra Ortega García	420002495

1. Lectura de Artículo

1.1. Resumen

Muchas veces cuando empezamos a trabajar con los datos, la idea es aclarar la estructura general de los datos, obtener información simple resúmenes descriptivos para que en un futuro tal vez contemos con ideas para un análisis más sofisticado. A esto se le llama EDA (Exploratory Data Analysis). Las virtudes que tiene la EDA en la Investigación Operativa se pueden ver con ejemplos de varias áreas, incluida la teoría de colas, la integridad y el pronóstico de series de tiempo. Por desgracia es difícil definir el EDA de forma precisa, pues dependiendo del autor cambia esta definición, sin embargo, algo que si podemos definir es el objetivo del EDA. Podríamos decir que tiene 2 objetivos principales, los cuales son la **descripción de los datos** y la **formulación de modelos** .Para que se cumplan estos objetivos, obviamente primero debemos de resumir los datos. Podemos decir que la formulación de modelos es (de los dos objetivos) el más importante ya que hay muchas situaciones para las cuales el EDA es vital para generar hipótesis, sugerir procedimientos o construir modelos adecuados.

Después de aclarar los objetivos de la investigación y obtener suficiente información básica, el analista debe comenzar por evaluar la estructura de los datos, ya que el análisis dependerá crucialmente, no sólo del número de observaciones, sino también del número de observaciones. variables y si son continuas, discretas, cualitativas, binarias, etc. Si el analista no fue cuidadoso de recopilar los datos, entonces es importante averiguar cómo lo hizo. La calidad de los datos debe comprobarse especialmente en lo que respecta a errores, valores atípicos y observaciones faltantes. Los valores atípicos son observaciones que no parecen ser consistentes con el resto de los datos y pueden crear problemas graves.

Después de examinar los datos, el análisis generalmente continúa con lo que a menudo se denomina "Estadísticas Descriptivas". Se deben calcular estadísticas resumidas para los datos en su conjunto y para subgrupos importantes. Generalmente incluyen la media y la desviación estándar para cada variable, aunque a veces se prefiere el rango a la desviación estándar como medida descriptiva para comparar la variación en muestras de tamaño aproximadamente igual.

Una parte importante del análisis de los datos es la representación visual que podemos hacer de ello, en este caso el articulo nos explica la importancia del histograma y la gráfica de caja, las cuales han sido muy útiles a lo largo del tiempo que se lleva haciendo EDA.

Uno de los tipos de problemas más simples se relaciona con la evaluación de una forma distributiva adecuada para una sola variable de interés. Esto generalmente se hace observando un histograma de datos relevantes, aunque también puede ser útil un diagrama de tallo y hojas o un diagrama de caja. Por ejemplo, el estudio de los sistemas de colas requiere que se hagan suposiciones sobre la distribución de los tiempos de servicio y de los tiempos entre llegadas, y los histogramas apropiados deben indicar qué suposiciones son razonables.

Muchas técnicas aritméticas incluyen métodos para resumir un solo grupo de observaciones, para evaluar la relación entre variables, para analizar tablas de dos y tres entradas y para resumir datos de conteo en forma distributiva.
En teoría, la mayoría de los estadísticos aceptan la necesidad de algún tipo de EDA. Sin embargo, en la práctica, la literatura sugiere que el EDA a menudo se infravalora. Cuando se produce esta omisión, se puede implementar una técnica inferencial sofisticada que resulta realmente inapropiada. Por suerte, hay señales de que el uso del EDA está aumentando.

La idea de que la mayoría de la gente "hace" EDA pero no se molesta en hablar de ello puede ser cierta hasta cierto punto, pero no es una situación deseable ni cierta en general. A muchos analistas no les gusta la EDA por su falta de teoría. Sin embargo, esto no hace que el tema seatrivial, contrariamente a algunas expectativas. De hecho, EDA puede ser más exigente que los métodos inferenciales más avanzados debido a la necesidad de un juicio subjetivo cuidadoso y porque el analista no puede confiar completamente en un paquete informático. Otro argumento en contra del uso de EDA es que no se basa en modelos y que dicho análisis corre el riesgo de arrojar conclusiones inválidas. La implicación aquí es que el analista puede verse tentado a pensar que una EDA es suficiente. Si bien esto conlleva ciertos riesgos, hay ocasiones en las que los resultados son tan claros que no es necesario realizar más análisis, como se señaló anteriormente. Sin embargo, quisiera enfatizar que la EDA está pensada principalmente como un ejercicio preliminar y normalmente irá seguida de un análisis formal basado en modelos. Si bien los análisis que no se basan en ningún modelo pueden ser peligrosos.

La construcción de modelos en quirófano suele depender, al menos en parte, de un examen exploratorio de un conjunto de datos relevante. Esto se describe como Análisis Exploratorio de Datos (EDA).

1.2. Adrian Aguilera Moreno

El Análisis Exploratorio de Datos (EDA, por sus siglas en inglés) es una técnica fundamental en la estadística que se utiliza para examinar y comprender los datos antes de aplicar cualquier modelo estadístico. El EDA es una herramienta poderosa que permite a los analistas de datos descubrir patrones, tendencias y relaciones en los datos que no son evidentes a simple vista. En este ensayo, discutiremos los objetivos del EDA, sus ingredientes y cómo se utiliza en la construcción de modelos.

El objetivo principal del EDA es proporcionar una descripción completa y detallada de los datos. Esto incluye la identificación de valores atípicos, la distribución de los datos, la presencia de patrones y tendencias, y la relación entre las variables. El segundo objetivo del EDA es la formulación de modelos. El EDA puede ayudar a los analistas de datos a identificar qué modelo es el más adecuado para los datos y cómo se deben ajustar los parámetros del modelo.

Los ingredientes del EDA incluyen gráficos, tablas y estadísticas descriptivas. Los gráficos son una herramienta poderosa para visualizar los datos y descubrir patrones y tendencias. Los gráficos más comunes utilizados en el EDA incluyen histogramas, diagramas de caja y bigotes, gráficos de dispersión y gráficos de series de tiempo. Las tablas y las estadísticas descriptivas se utilizan para resumir los datos y proporcionar una descripción numérica de los mismos.

El EDA se utiliza en la construcción de modelos en una variedad de temas, como el control de inventario, la fiabilidad, la regresión y la predicción. Por ejemplo, en el control de inventario, el EDA se utiliza para identificar patrones en la demanda de productos y para determinar cuánto inventario se debe mantener en stock. En la fiabilidad, el EDA se utiliza para identificar patrones en los fallos de los equipos y para determinar cuánto tiempo se puede esperar que un equipo funcione antes de que falle.

En conclusión, el Análisis Exploratorio de Datos es una técnica esencial en la estadística que permite a los analistas de datos comprender mejor los datos antes de aplicar cualquier modelo estadístico. El EDA proporciona una descripción completa y detallada de los datos, lo que ayuda a los analistas de datos a identificar patrones, tendencias y relaciones en los datos que no son evidentes a simple vista. Además, el EDA es una herramienta poderosa para la formulación de modelos, ya que ayuda a los analistas de datos a identificar qué modelo es el más adecuado para los datos y cómo se deben ajustar los parámetros del modelo.

1.3. Marco Antonio Rivera Silva

La idea principal del articulo es darnos a entender que el Análisis Exploratorio de Datos (EDA) constituye una técnica fundamental en la estadística que se emplea para examinar y comprender los datos antes de aplicar cualquier modelo estadístico. El EDA nació como una valiosa herramienta que permite a los analistas de datos descubrir patrones, tendencias y relaciones en los datos que no resultan evidentes de manera inmediata. En el presente ensayo, exploraremos los objetivos del EDA, sus componentes y cómo se aplica en la creación de modelos.

El autor también menciona que el propósito primordial del EDA radica en suministrar una descripción exhaustiva y minuciosa de los datos. Esto abarca la detección de valores atípicos, la distribución de los datos, la identificación de patrones y tendencias, así como la comprensión de las relaciones entre las variables. El segundo objetivo del EDA es el desarrollo de modelos. De hecho, el EDA puede asistir a los analistas de datos en la identificación del modelo más adecuado para los datos y la configuración de los parámetros del modelo.

Se destaca que los elementos del EDA incluyen representaciones gráficas, tablas y medidas descriptivas. Las representaciones gráficas representan una herramienta efectiva para visualizar los datos y descubrir patrones y tendencias. Entre los gráficos más comúnmente utilizados en el EDA se encuentran los histogramas, los diagramas de caja y bigotes, los gráficos de dispersión y los gráficos de series de tiempo. Las tablas y medidas descriptivas se utilizan para resumir los datos y brindar una descripción numérica de los mismos.

Por último se concluye que el EDA se emplea en la creación de modelos en diversas áreas, tales como el control de inventario, la fiabilidad, la regresión y la predicción. Por ejemplo, en el control de inventario, el EDA se utiliza para detectar patrones en la demanda de productos y determinar la cantidad de inventario que debe mantenerse en stock. En la fiabilidad, el EDA se aplica para identificar patrones en las fallas de equipos y estimar el tiempo de funcionamiento esperado de un equipo antes de que ocurra una falla.

1.4. Sebastían Alejandro Gutiérrez Medina

En el articulo se discute la importancia de realizar un examen exploratorio de los datos, o Exploratory Data Analysis (EDA), antes de realizar cualquier análisis estadístico formal, los dos objetivos principales de EDA son la descripción de los datos y la formulación de modelos.

Además, el autor sugiere que es importante ver EDA como una parte integral de la inferencia estadística en general, y no como un tema separado, después se presentan varios ejemplos para mostrar cómo EDA se utiliza en la construcción de modelos en relación con temas como el control de colas, la confiabilidad, el control de inventario, la regresión y la predicción.

Algunas de las ventajas de EDA que se destacan en el articulo son la importancia de realizar un examen exploratorio de los datos antes de realizar cualquier análisis estadístico formal, ya que esto puede ayudar a identificar patrones y tendencias en los datos, detectar cualquier desviación del proceso y seleccionar un modelo adecuado. Además, se sugiere que es importante ver EDA como una parte integral de la inferencia estadística en general, y no como un tema separado.

La conclusión principal del artículo es que EDA es una parte integral de la inferencia estadística en general, y no como un tema separado y que el autor, Chris Chatfield, destaca la importancia de realizar EDA antes de realizar cualquier análisis estadístico formal, ya que esto puede ayudar a identificar patrones y tendencias

en los datos, detectar cualquier desviación del proceso y seleccionar un modelo adecuado, además, sugiere que es importante ver EDA como una parte integral de la inferencia estadística en general, y no como un tema separado.

1.5. Israel Hernández Dorantes

En el artículo ”Exploratory data analysis” podemos notar que la idea principal es definir el **Análisis Exploratorio de Datos** como una herramienta fundamental en el análisis de datos y la formulación de modelos. El autor destaca la importancia de realizar una exploración informal y descriptiva de los datos antes de llevar a cabo un análisis más *sofisticado*.

Y podemos ver que el **Análisis Exploratorio de Datos** se relaciona con los conceptos de la materia de Almacenes y Minería de Datos en el sentido de que busca comprender la estructura general de los datos, obtener resúmenes descriptivos simples y generar ideas para un análisis más detallado. El **Análisis Exploratorio de Datos** permite obtener una visión general de los datos y formular hipótesis antes de aplicar técnicas más avanzadas.

La temática central del artículo es la importancia y los objetivos del **Análisis Exploratorio de Datos**. El autor menciona que el Análisis Exploratorio de Datos tiene dos objetivos principales: la descripción de los datos y la formulación de modelos. La descripción de los datos implica resumir y destacar las características más relevantes de los datos, mientras que la formulación de modelos se refiere a la generación de hipótesis y la selección de técnicas estadísticas adecuadas para analizar los datos.

Como conclusión, tenemos que reconocer la importancia del Análisis Exploratorio de Datos como una herramienta valiosa en el proceso de análisis de datos, realizar una exploración inicial de los datos puede ayudar a comprender su estructura y proporcionar ideas para un análisis más detallado. Además, el autor destaca la necesidad de integrar el Análisis Exploratorio de Datos en el *análisis estadístico* y la *construcción de modelos*, en lugar de considerarlos por separado; y poder aprovechar el potencial de este Análisis Exploratorio de Datos en nuestras prácticas analíticas.

1.6. Alejandra Ortega García

En el articulo **Exploratory Data Analysis**, **Chris Chatfield**, expone razones por las cuales EDA debe considerarse como una etapa dentro de la estadística inferencial, "es simple ... pero no tanto como una estadística descriptiva". Algunos autores consideran que EDA carece de teoría y que al no basarse de modelos se corre el riesgo de llegar a conclusiones erróneas. ¿Cómo debe considerarse a EDA?

La idea del Análisis Exploratorio de Datos (EDA), es **comprender** la estructura general de los datos, obtener resúmenes descriptivos simples y tal vez obtener ideas para un análisis más complejo.

Cuando un analista utiliza EDA tiene como propósito, verificar la calidad de los datos, calcular estadísticas, hacer uso de gráficas adecuadas para representar los datos. Pero también debe cumplirse con los dos objetivos principales que tiene EDA, la descripción de los datos y la creación de modelos. El segundo es mas importante, pues de acuerdo a las hipótesis que se generaron se construye el modelo adecuado y en base a él se proponen soluciones.

Para realizar un Análisis Exploratorio de Datos, primero el analista debe de evaluar la estructura de los datos; debe comprobar la calidad de los datos, esto es, identificar si hay valores atípicos o observaciones perdidas.

Después, se debe obtener las estadísticas resumidas para el conjunto de datos y finalmente representar los datos de la forma que se considere adecuada.

En conclusión, el Análisis Exploratorio de Datos (EDA) es una etapa esencial en la estadística que va más allá de la mera estadística descriptiva. Aunque EDA puede parecer simple en su enfoque, radica en su capacidad (los ingredientes) para ayudar a los analistas a comprender la estructura general de los datos.

2. Análisis Exploratorio de Datos

1. En caso de existir **variables categóricas**, guarda dichas variables como tal con la función `as.factor()`.

   Al realizar un head sobre los datos, descubrimos que había 2 variables categoricas :

   - **Star.color**
   - **Spectral Class**

```
# Guardamos las 2 variables categoricas como factores

datos$Star.color <- as.factor(datos$Star.color)
datos$Spectral.Class <- as.factor(datos$Spectral.Class)
```

2. Realiza un **resumen del conjunto de datos**. ¿Qué puedes decir de las **variables numéricas** y de las **categóricas**? ¿Qué sucede con la variable **Star type**? ¿Existen valores ausentes?

```
> # Hacemos un resumen de los datos
> summary(datos)
 Temperature..K. Luminosity.L.Lo.    Radius.R.Ro.       Absolute.magnitude.Mv.   Star.type
 Min.   : 1939   Min.   :     0.0   Min.   :   0.0084   Min.   :-11.920        Min.   :0.0
 1st Qu.: 3344   1st Qu.:     0.0   1st Qu.:   0.1027   1st Qu.: -6.232        1st Qu.:1.0
 Median : 5776   Median :     0.1   Median :   0.7625   Median :  8.313        Median :2.5
 Mean   :10497   Mean   :107188.4   Mean   : 237.1578   Mean   :  4.382        Mean   :2.5
 3rd Qu.:15056   3rd Qu.:198050.0   3rd Qu.:  42.7500   3rd Qu.: 13.697        3rd Qu.:4.0
 Max.   :40000   Max.   :849420.0   Max.   :1948.5000   Max.   : 20.060        Max.   :5.0

  Star.color        Spectral.Class
 Length:240         A: 19
 Class :character   B: 46
 Mode  :character   F: 17
                    G:  1
                    K:  6
                    M:111
                    O: 40
```

Figure 1: Resumen de los datos

   De nuestras variables numéricas podemos decir que todas trabajan con valores positivos (excepto Absolute.magnitude.Mv que también tiene negativos), algunas solo tienen números enteros. Por otro lado nuestras variables categóricas **Star.color y Star.Class**, en caso de **Star.color** los datos no están limpios pues no es consistente la forma de los datos, mientras que en Star.Class simplemente son letras mayúsculas.

   La variable **start type** es una variable numérica, que inicia en 0 curiosamente cada 10 filas aumenta en 1 su cantidad hasta llegar a 5 y después vuelve al 0.

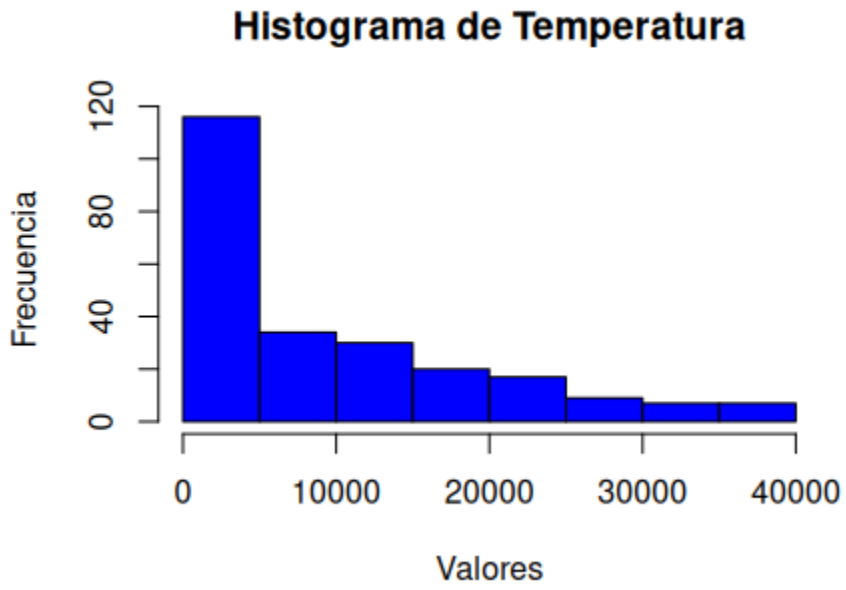
   Por útlimo, no existen valores ausentes en nuestro conjunto de datos.

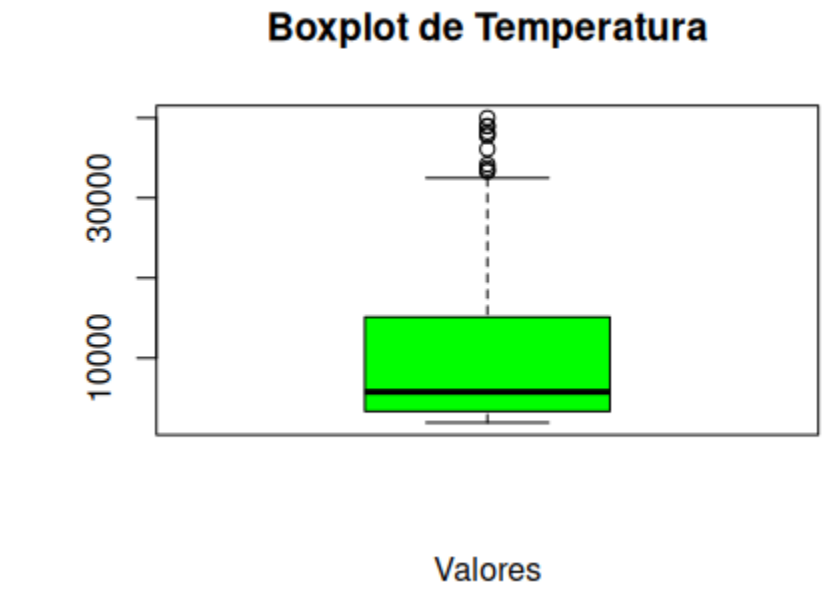
3. Realiza los **histogramas** y **boxplots** de las **variables numéricas**. ¿Qué puedes decir de la **distribución** de cada una de las variables? ¿Dirías que se puede **ajustar una distribución de probabilidad** a alguna variable? ¿Existen **datos atípicos**?

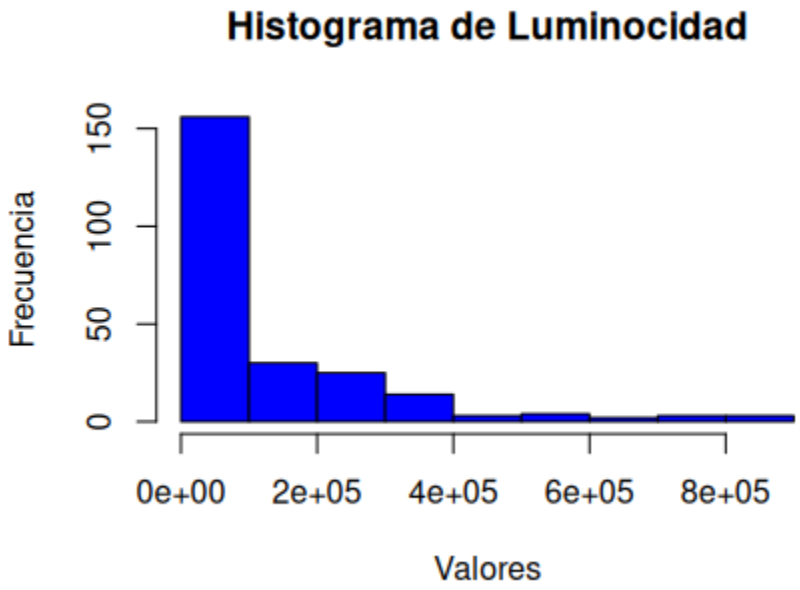
```
# Histogramas
# Temperatura K
```

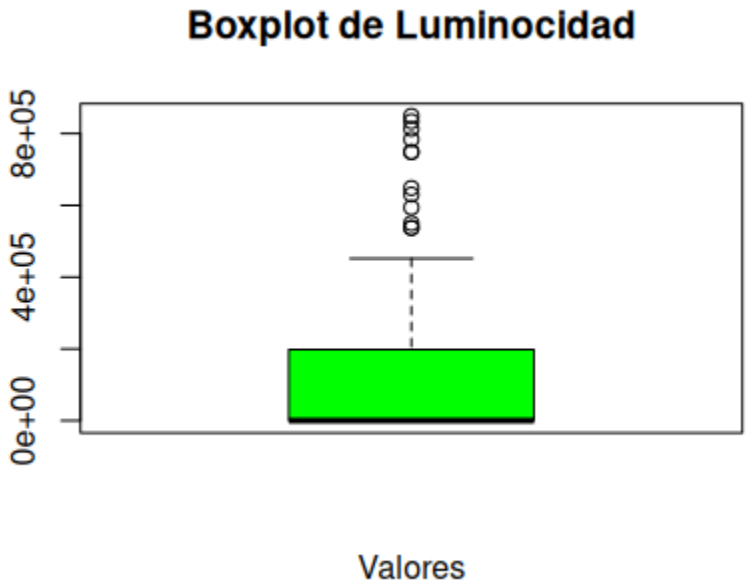
```r
hist(datos$Temperature..K.,
     main = "Histograma de Temperatura",
     xlab = "Valores",
     ylab = "Frecuencia",
     col = "blue",
     border = "black")

boxplot(datos$Temperature..K.,
        main = "Boxplot de Temperatura",
        xlab = "Valores",
        col = "green")
# Luminocidad
hist(datos$Luminosity.L.Lo.,
     main = "Histograma de Luminocidad",
     xlab = "Valores",
     ylab = "Frecuencia",
     col = "blue",
     border = "black")

boxplot(datos$Luminosity.L.Lo.,
        main = "Boxplot de Luminocidad",
        xlab = "Valores",
        col = "green")
# Radio
hist(datos$Radius.R.Ro.,
     main = "Histograma de Radio",
     xlab = "Valores",
     ylab = "Frecuencia",
     col = "blue",
     border = "black")

boxplot(datos$Radius.R.Ro.,
        main = "Boxplot de Radio",
        xlab = "Valores",
        col = "green")
# Magnitud Absoluta
hist(datos$Absolute.magnitude.Mv.,
     main = "Histograma de Magnitud Absoluta",
     xlab = "Valores",
     ylab = "Frecuencia",
     col = "blue",
     border = "black")

boxplot(datos$Absolute.magnitude.Mv.,
        main = "Boxplot de Magnitud Absoluta",
        xlab = "Valores",
        col = "green")
```

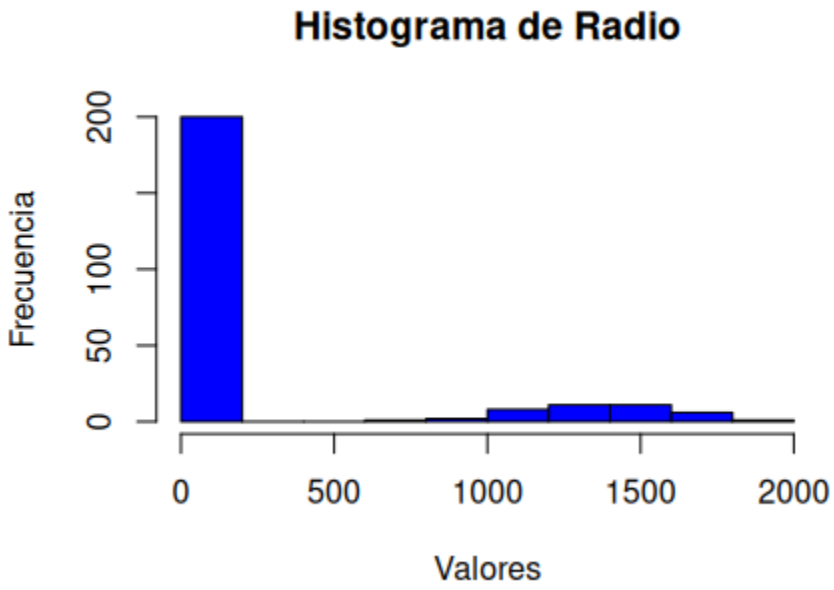
```r
# Tipo de estrella
hist(datos$Star.type,
     main = "Histograma de Tipo de estrella",
     xlab = "Valores",
     ylab = "Frecuencia",
     col = "blue",
     border = "black")

boxplot(datos$Star.type,
        main = "Boxplot de Tipo de estrella",
        xlab = "Valores",
        col = "green")
```

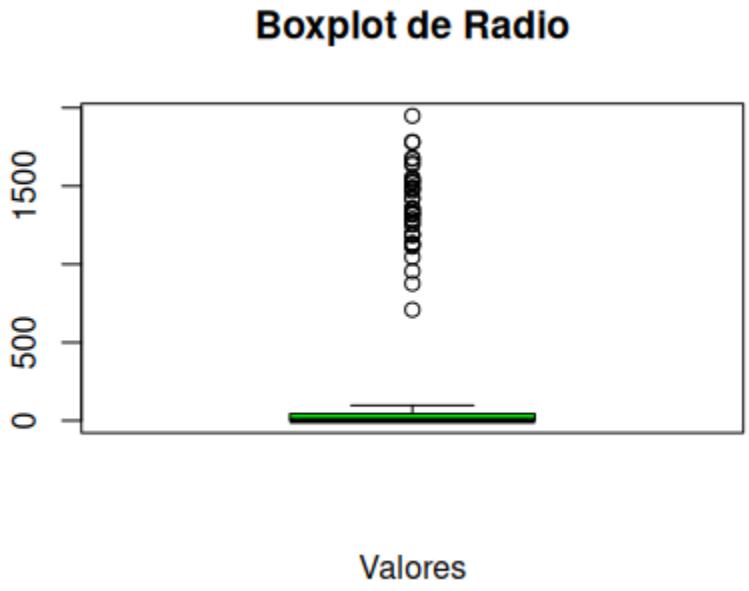


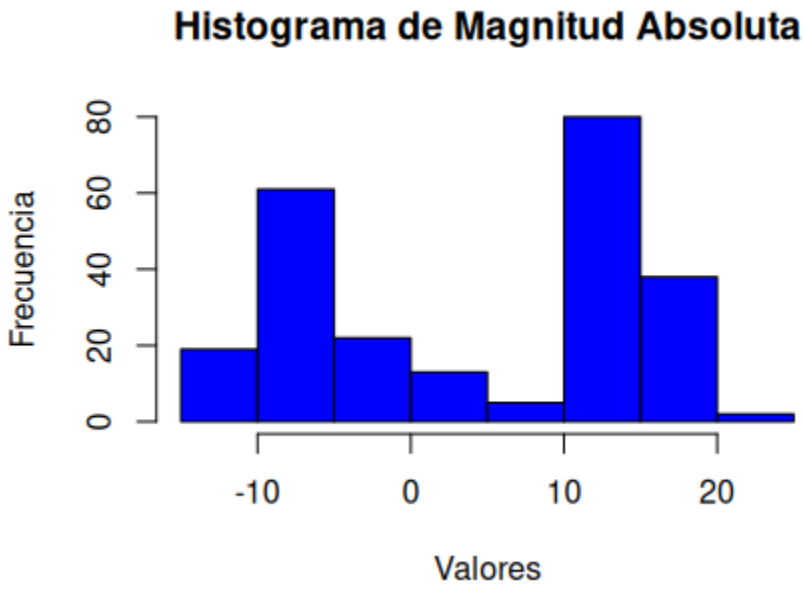

Boxplot de Temperatura
10000
30000
Valores

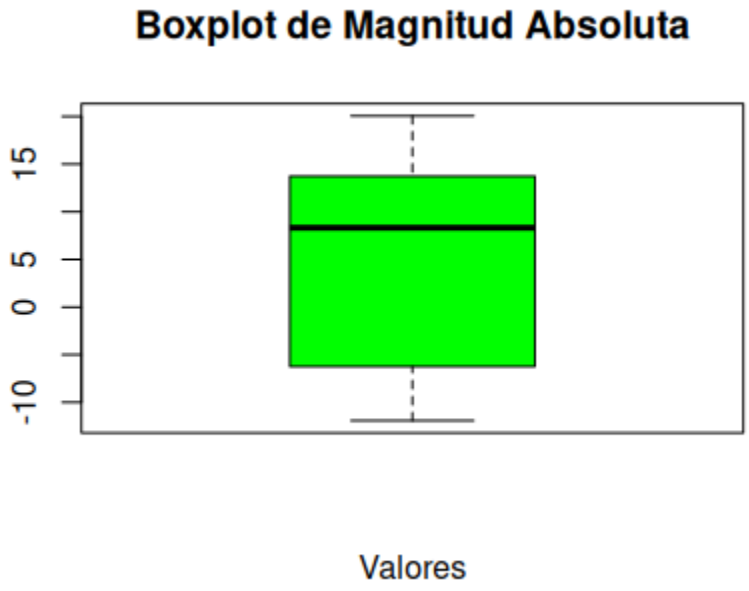

Histograma de Luminocidad
Frecuencia
0
50
100
150
0e+00
2e+05
4e+05
6e+05
8e+05
Valores

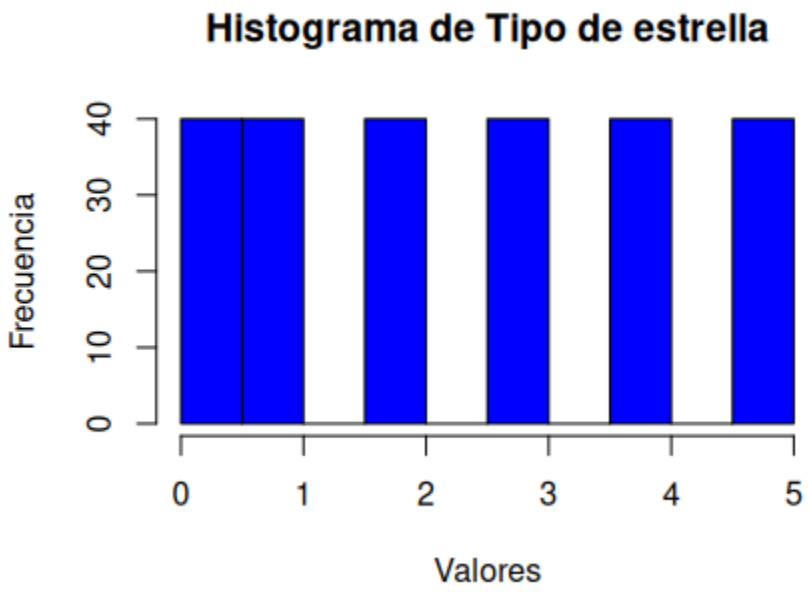

Boxplot de Luminocidad
0e+00
4e+05
8e+05
Valores

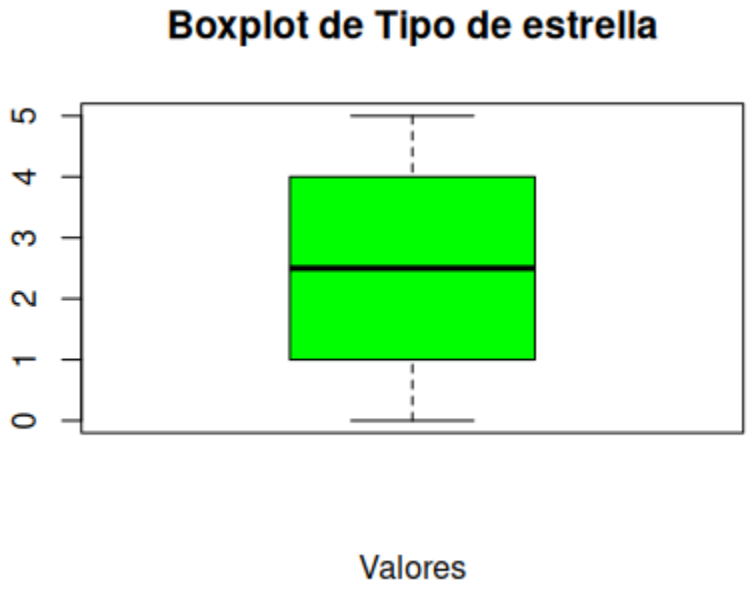

Histograma de Radio
Frecuencia
0
50
100
200
0
500
1000
1500
2000
Valores


Boxplot de Radio
0
500
1500
Valores


Histograma de Magnitud Absoluta
Frecuencia
0
20
40
60
80
-10
0
10
20
Valores


Boxplot de Magnitud Absoluta
-10
0
5
15
Valores


Histograma de Tipo de estrella
Frecuencia
0
10
20
30
40
0
1
2
3
4
5
Valores

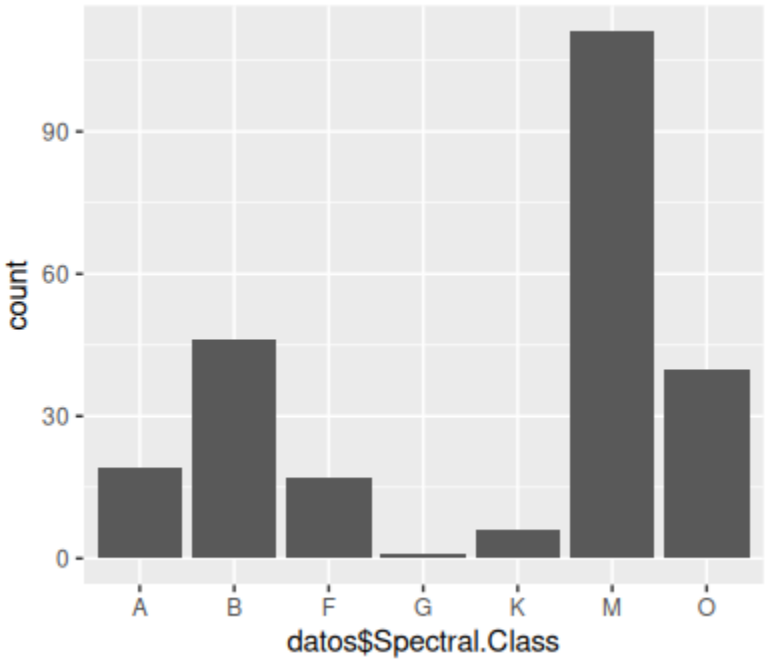


- ¿Qué puedes decir de la distribución de cada una de las variables?
  La mayoría de las variables tienen una distribución sesgada hacia la derecha, lo que significa que la mayoría de los valores se encuentran en el extremo inferior del rango y hay algunos valores atípicos en el extremo superior.

- ¿Dirías que se puede ajustar una distribución de probabilidad a alguna variable?
  Tal vez podríamos sobre-ajustar nuestras muestras en busca de un mejor comportamiento. Una alternativa es normalizar nuestras muestras.

- ¿Existen datos atípicos?
  Se pueden observar algunos valores atípicos en los boxplots de algunas variables, como la variable Absolute, magnitud.

4. Realiza una **gráfica de barras** de la variable `Spectral class` utilizando la biblioteca `ggplot2`. ¿Cuál es la clase con más datos?

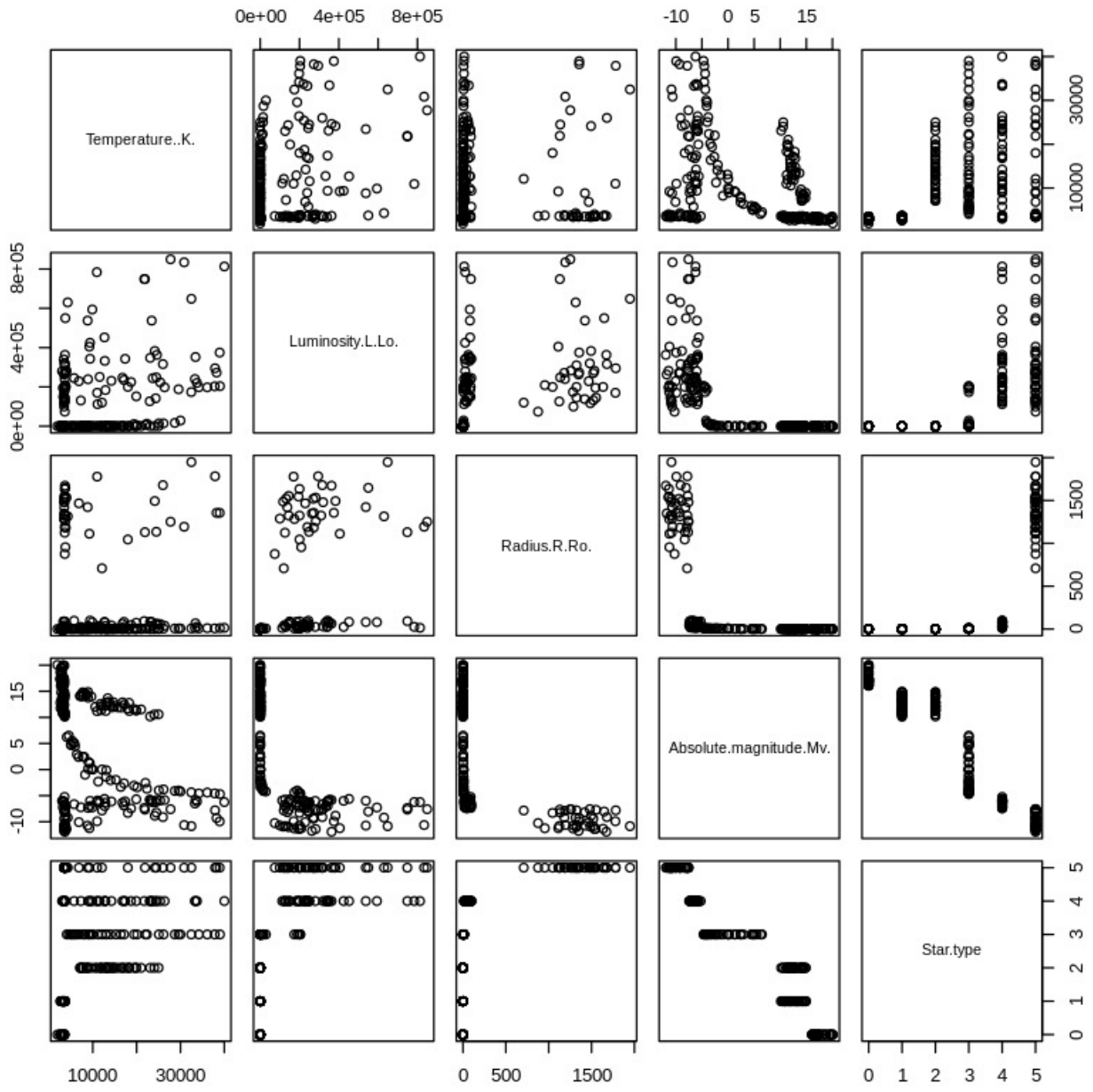


Cómo podemos ver, la clase con más datos es *M*. Para obtener esta gráfica usamos el siguiente comando:

```
library("ggplot2")
ggplot(datos, aes(x = datos$Spectral.Class)) + geom_bar()
```

5. Gráfica el **diagrama de dispersión** de las .**variables numéricas**. ¿Existe alguna **relación** entre algún **par de variables**?

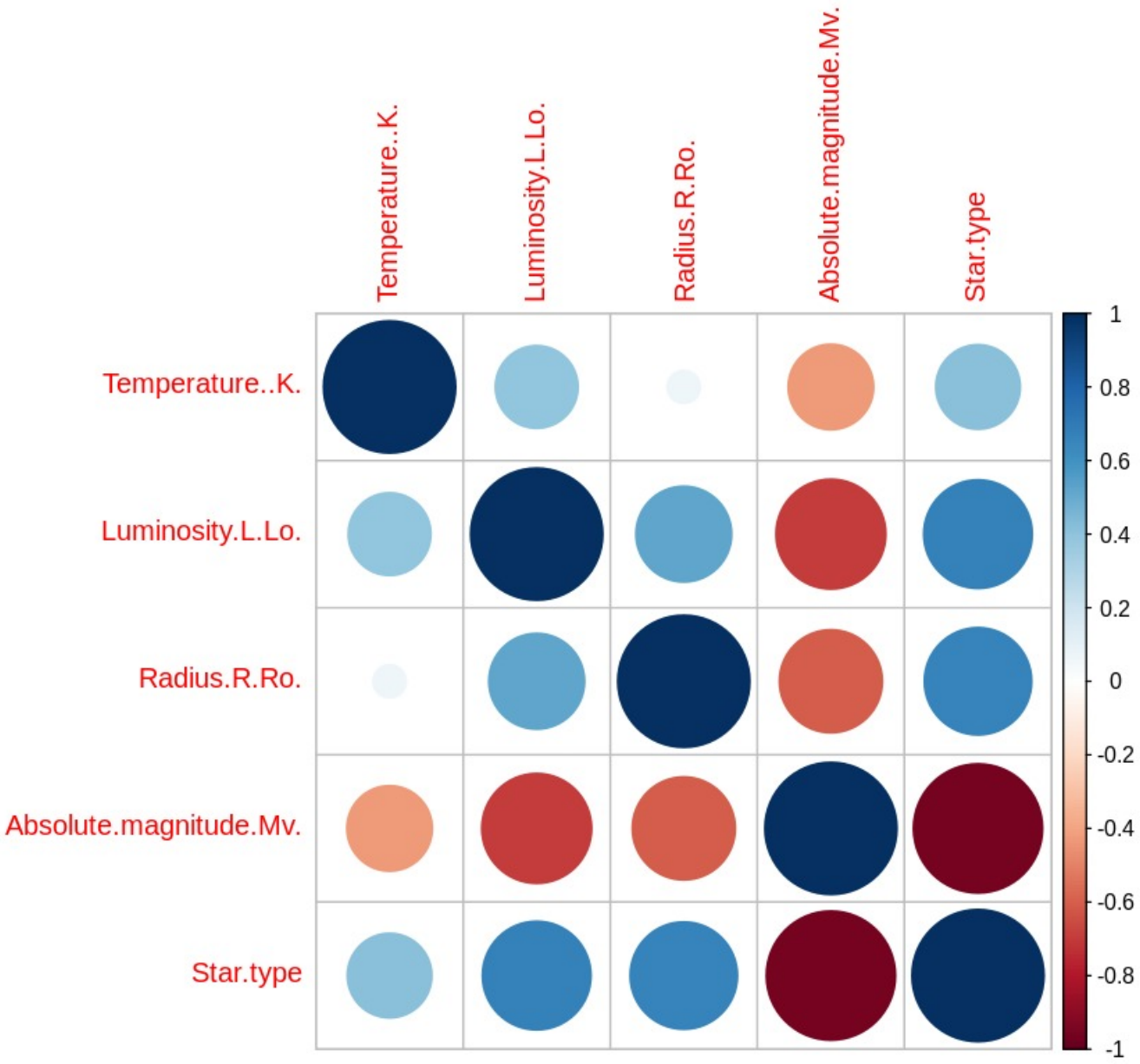
```
# Grafica la tabla de dispersión sin las variables categoricas
library(car)
pairs(datos[, 1:(ncol(datos) - 2)])
```



La relación que encontramos a simple vista fué entre las variables **Radius.R.Ro.** y **Star.type**, ya que consideramos que hay una relación entre ambas. De igual forma con la Luminosidad y la temperatura.

6. Finalmente gráfica la **correlación entre las variables numéricas**. Describe la correlación entre los pares de variables.

```
# Grafica la tabla de dispersión sin las variables categoricas
library(corrplot)
matriz_correlacion <- cor(datos[, 1:(ncol(datos) - 2)])
corrplot(matriz_correlacion)
```



Las correlaciones son las siguientes:

(a) Las variables consigo mismas tienen alta correlación ya que son iguales.

(b) Las variables con baja correlación son:

- Absolute.magnitude.Mv. y Star.type
- Absolute.magnitude.Mv. y Radius.R.Ro.
- Absolute.magnitude.Mv. y Luminosity.L.Lo.
- Absolute.magnitude.Mv. y Temperature..K.

(c) Las variables con alta correlación son:

- Star.type y Temperature..K.
- Star.type y Luminosity.L.Lo.
- Star.type y Radius.R.Ro.
- Radius.R.Ro. y Luminosity.L.Lo.
- Luminosity.L.Lo. y Temperature..K.

(d) Las variables con correlación mediana son:

- Radius.R.Ro. y Temperature..K.