Para el desarrollo del análisis se propusieron las siguientes preguntas:

1. ¿Cuál era el porcentaje de hombres y mujeres del total de población de los pasajeros?
2. ¿Cuál era la clase más poblada? Es decir ¿Cuál clase tenía mayor número de pasajeros?
3. ¿Cuál era el promedio de edad de los pasajeros?
4. ¿Dónde embarcó la mayoría de los pasajeros?
5. ¿Cuál es el porcentaje de sobrevivientes?
6. ¿De cuál clase eran la mayoría de los sobrevivientes?
7. ¿Cuál es el porcentaje de hombres/mujeres sobrevivientes?
8. ¿Cuál es el promedio de edad de los sobrevivientes?

Mediante la implementación de los algoritmos y comandos de R que vimos a lo largo de las sesiones se planteó el brindar solución a estas preguntas. A continuación se presentan las soluciones obtenidas con el código que se desarrolló para su obtención.

Para efectos prácticos, en este documento se utilizará el término *Población* para referirnos a los registros de la tabla.

1. Para obtener el porcentaje de hombres y mujeres de la población de pasajeros, se utilizó la función *Count,* de la siguiente forma:

```
#Porcentaje de hombres y mujeres en la poblacion
df %>% count(Sex) %>% mutate(porcentaje = (n/sum(n))*100)
```

Obtenemos estos números en formato de porcentaje al dividir cada uno de las cuentas de registros por sexo entre la suma del total de registros. De esta sentencia de código obtuvimos que:

- El total de hombres en la población es de 266, que es equivalente al 63.63 % de la población.
- Las mujeres son el 36.36% restante, con un total de 152 registros.

2. Para la clasificación de la población por su registro de *Clase* se desarrolló la siguiente sentencia de código:

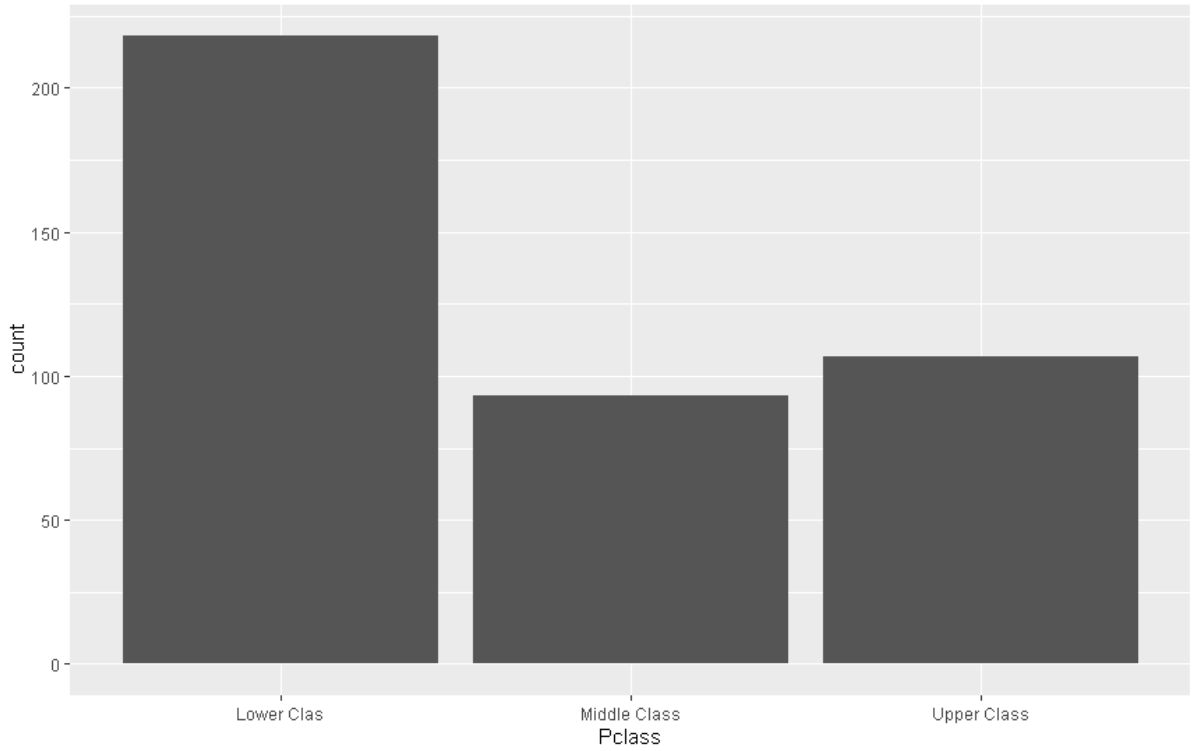
```
#Cual es el porcentaje de problacion por clase?
df %>% count(Pclass) %>% mutate(porcentaje = (n/sum(n))*100)
```

El porcentaje se calcula de la misma forma que en la pregunta anterior. Se obtuvo la siguiente tabla con el código:

Pclass	n	Porcentaje
Lower Class	218	52.15311
Middle Class	93	22.24880
Upper Class	107	25.59809

La clase más poblada era la clase baja (Lower Class) con 218 pasajeros, le seguía la clase alta (Upper Class) con 107 pasajeros y el último lugar era la clase media (Middle Class) con 93 pasajeros, siendo el 52, 26 y 22 por ciento de la población respectivamente.

Para ayudar al análisis de esta información, se construyó la siguiente imágen-



El código implementado para generar el histograma se presenta a continuación.

```
df %>%   ggplot(aes(x=Pclass)) +
  geom_bar()
```

En el histograma se puede apreciar que la clase predominante es Lower Class, que pasa de la barrera de los 200 pasajeros, la siguiente es Upper Class, que sobrepasa los 100 y la última es Middle Class, que no llega a tocar esa barrera. Esto nos confirma la data mostrada por el código anterior.

3. El promedio de edad se obtuvo con la función *Mean().* La sentencia escrita para este apartado es la siguiente:

```
mean(na.omit(df$Age))
```

Dando como resultado 30.27259 como edad promedio de la población. El siguiente diagrama de dispersión nos muestra la distribución de edades de los pasajeros en relación con la clase a la que pertenecían sus boletos.

Podemos apreciar que, las clases bajas tenían una población mayormente joven, pues la mayoría de los datos se encuentran entre las líneas de los 10 a los 40 años. Las clases medias tienen una dispersión más amplia, lo cual nos indica que su población estaba constituida por varios rangos de edad, teniendo una pequeña concentración en las líneas de los 20 a los 30 años, por lo cual se puede considerar que esta clase estaba constituida por población mayormente joven, pero con un amplio abanico de edades. Las clases altas tienen una concentración de personas mayores de edad, con registros que van desde los 30 años hasta los 80 años, por lo cual se estima que el pasajero de más edad en el barco pertenecía a esta clase.

El máximo y mínimo de edad de los pasajeros se calculó con las funciones *max()* y *min()*, de la siguiente forma:

```
max(na.omit(df$Age))
min(na.omit(df$Age))
```

Obteniendo la edad mínima de 0.17 años y 76 años como máximo.

4. El porcentaje de sobrevivientes se obtuvo de forma similar a los apartados 1 y 2. Para facilitar el análisis de los datos de supervivientes se construyó otro *dataframe.*

```
#Crear el dataframe de supervivientes
df_survivors <- df %>% filter(Survived =="Yes")
```

```
#Porcentaje de supervivientes
survivors <- df %>% count(Survived) %>% mutate(porcentaje = (n/sum(n))*100)
survivors
```

De acuerdo con el resultado del código anterior, sobrevivieron 152 personas, representando el 36.36% de la población total.

El siguiente diagrama

5. Para obtener el porcentaje de pasajeros embarcados en cada puerto, se desarrollo un código similar al de los apartados 1 y 2.

```
#Cual fue el lugar donde embarcaron mas pasajeros
df %>% group_by(Embarked) %>% tally() %>% mutate(porcentaje = (n/sum(n))*100)
```

Embarked	n	porcentaje
Cherbourg	102	24.4
Queenstown	46	11.0
Southampton	270	64.6

En la tabla se puede observar que la mayoría de los pasajeros embarcaron en Southampton, que según los registros, fue el puerto del que zarpó el 10 de abril de 1912, siendo 270 pasajeros, equivalentes al 64.6% de la población. El segundo lugar con más pasajeros embarcados fue Cherbourg, dónde embarcaron 102 personas, equivalentes al 24% , y por último, Queenstown con 46 pasajeros abordados, equivalentes al 11%.

6. Para obtener el número de supervivientes por clase se hizo una consulta similar al apartado 2, utilizando el dataframe de supervivientes.

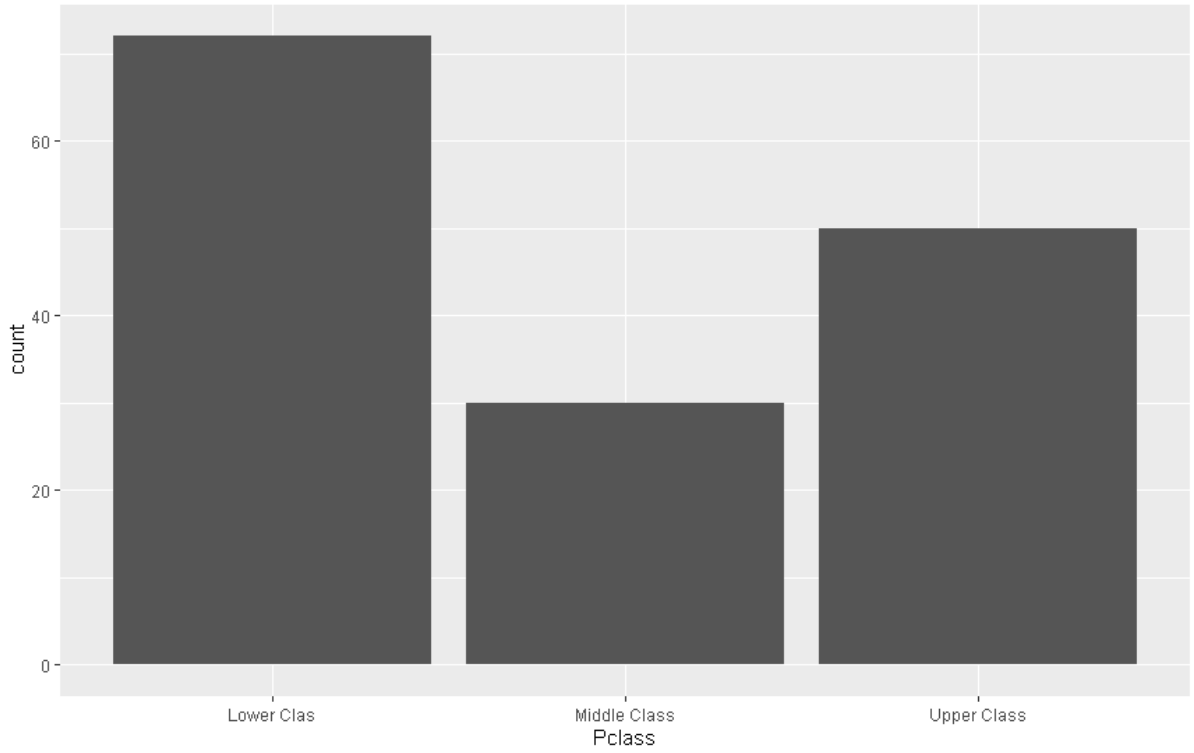
```
#clases supervivientes
df_survivors %>% count(Pclass) %>% mutate(porcentaje = (n/sum(n))*100)
```

Dando como resultado la siguiente tabla:

Pclass	n	Porcentaje
Lower Class	72	47.36

Middle Class	30	19.73
Upper Class	50	32.89

El siguiente histograma presenta los resultados contenidos en la tabla, en un formato visual.



7. Para el cálculo de la distribución de sexo de los supervivientes, se desarrolló el siguiente código

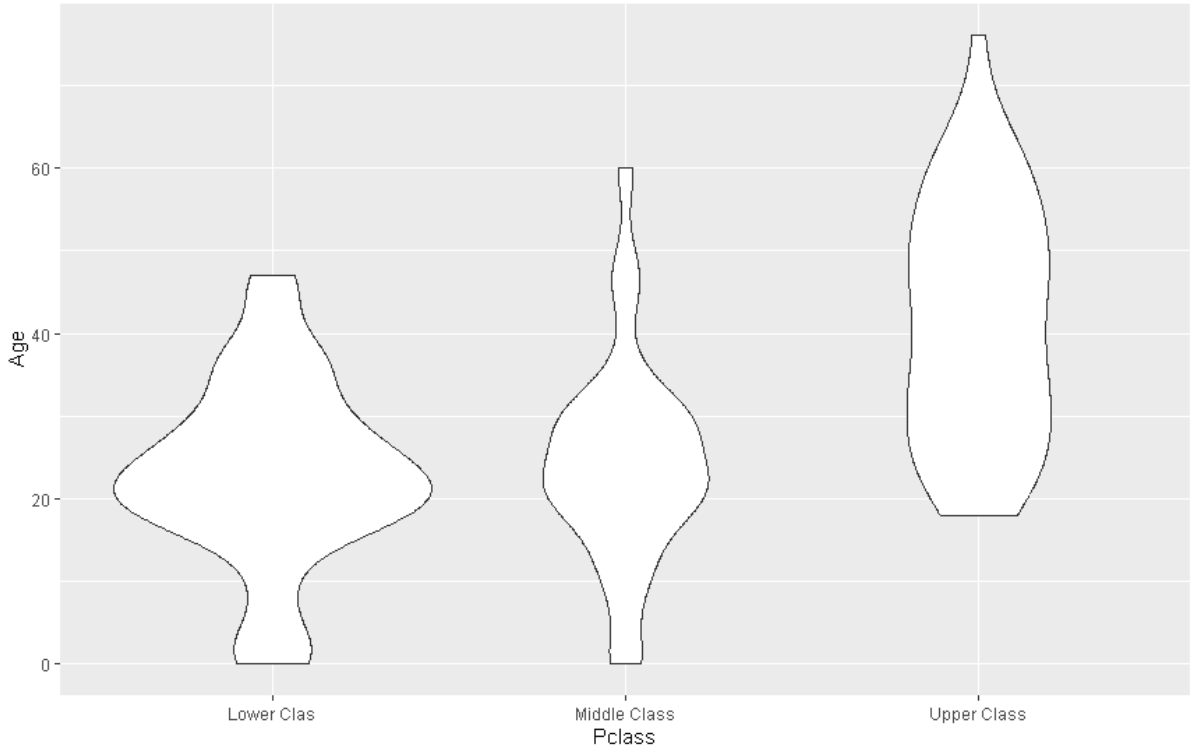
```
#sexo de los supervivientes
df_survivors %>% count(Sex)
```

Resultante de ésto, los 152 supervivientes fueron únicamente mujeres.

8. El promedio de edad, así como los máximos y mínimos se calcularon con las funciones dedicadas de R, como en la sección tres de este documento.

```
#Medidas de tendencia central para los supervivientes
mean(na.omit(df_survivors$Age))
median(na.omit(df_survivors$Age))
max(na.omit(df_survivors$Age))
min(na.omit(df_survivors$Age))
```

Los resultados obtenidos fueron 30.25 años como promedio, 0 como la edad mínima y 76 años como la máxima. El diagrama de violín presentado a continuación nos muestra la distribución de edades de los supervivientes al incidente, de acuerdo a la clase a la que pertenecía su boleto del barco.



Es apreciable que la clase alta era la que concentraba a las supervivientes de mayor edad, pues los registros van desde poco menos de los 20 años, hasta poco antes de la línea de los 80 años. La clase baja concentra la mayor cantidad de supervivientes jóvenes, pues se aprecia que su tendencia es más ancha alrededor de la línea de los 20 años. Las clases medias tienen una distribución más madura, sin llegar a la tercera edad, ya que su distribución se ensancha en el intervalo correspondiente a los 30 a 40 años.