**PRUEBA DE CURSO**

N° de Libreta:

Nombre Participante:

Instructor: Andrés Andruskevicius – Ana Cecilia Plavan

Fecha:

Resultado práctico:

**Total**:

**EJERCICIOS – Supervivencia en el Titanic**

Pon todas tus habilidades a prueba para realizar un análisis exploratorio de datos en un data set de machine learning: Titanic survival.

**Un poco de historia**

El TITANIC era un transoceánico que chocó contra un iceberg realizando el recorrido desde Reino Unido a Nueva York y se hundió en 1912. Más de 1.500 de los 2.224 pasajeros estimados murieron en el accidente, haciendo este desastre el más grande que ha pasado fuera de la guerra.

El barco llevaba un amplio rango de pasajeros de todas las edades y géneros, desde viajeros pertenecientes a clase alta viajando en primera clase a inmigrantes en las clases más bajas. Pero no todos los pasajeros tuvieron iguales chances de sobrevivir al accidente. Utilizaremos datos reales de una selección de 891 pasajeros para obtener quienes estaban a bordo del Titanic y que chande de supervivencia tuvieron.

**Librerías, opciones y datos**

Antes de continuar asegúrate de tener instalado el paquete **titanic**.

Define el dataset titanic de la librería como se muestra en el código a continuación:

options(digits = 3) # report 3 significant digits

library(tidyverse)

library(titanic)

titanic <- titanic\_train %>%  
 select(Survived, Pclass, Sex, Age, SibSp, Parch, Fare) %>%  
 mutate(Survived = factor(Survived),  
 Pclass = factor(Pclass),  
 Sex = factor(Sex))

### **Pregunta 1 – Tipo de variables**

0.0/3.0 puntos (calificable)

Inspecciona los datos y utiliza ?titanic\_train para aprender más de las variables del data set.

*Relaciona las variables a su tipo. Hay al menos una variable de cada tipo (categórica ordinal, categórica no ordinal, continua, discreta)*

*Ingresa una ‘X’ en el tipo correcto de la variable*

|  |  |
| --- | --- |
| Variable **Survived** | Marca con una X la correcta |
| Categórica ordinal |  |
| Categórica no-ordinal |  |
| Continua |  |
| Discreta |  |

|  |  |
| --- | --- |
| Variable **Pclass** | Marca con una X la correcta |
| Categórica ordinal |  |
| Categórica no-ordinal |  |
| Continua |  |
| Discreta |  |

|  |  |
| --- | --- |
| Variable **Sex** | Marca con una X la correcta |
| Categórica ordinal |  |
| Categórica no-ordinal |  |
| Continua |  |
| Discreta |  |

|  |  |
| --- | --- |
| Variable **SibSp** | Marca con una X la correcta |
| Categórica ordinal |  |
| Categórica no-ordinal |  |
| Continua |  |
| Discreta |  |

|  |  |
| --- | --- |
| Variable **Parch** | Marca con una X la correcta |
| Categórica ordinal |  |
| Categórica no-ordinal |  |
| Continua |  |
| Discreta |  |

|  |  |
| --- | --- |
| Variable **Fare** | Marca con una X la correcta |
| Categórica ordinal |  |
| Categórica no-ordinal |  |
| Continua |  |
| Discreta |  |

### **Pregunta 2 – Demografía de los pasajeros**

0.0/3.5 puntos (calificable)

Realiza un gráfico de densidad de las edades de los pasajeros agrupados por sexo. Experimenta con las herramientas ***faceting, alpha blending, stacking*** y utilizando counts en el eje y para responder las siguientes preguntas.

*¿Cuál de las siguientes afirmaciones es verdadera?*

*Ingresa una ‘X’ en la/s correcta/s*

|  |  |
| --- | --- |
| Respuesta | Marca con una X la/s correcta/s |
| Las mujeres y hombres tienen en general la misma curva de distribución por edad |  |
| La distribución de edad era bimodal, con un modo alrededor de los 25 años de edad y otro modo más pequeño cercano a los 5 años de edad |  |
| Habían más mujeres que hombres |  |
| La cuenta de hombres de la edad de 40 era más alta que la cuenta de mujeres de 40 años |  |
| La proporción de hombres de edades entre 18-35 era mayor a la proporción de mueres entre 18-35 anos |  |
| La proporción de mueres menores a 17 años era mayor a la proporción de hombres menores a 17 |  |
| La persona de más edad en el barco era una mujer |  |

### **Pregunta 3 – Grafico QQ para la distribución de edades**

0.0/1 punto (calificable)

Utiliza *geom\_qq()* para realizar un gráfico QQ de las edades de los pasajeros y agrega una línea de identidad con *geom\_abline().* Filtra los individuos que tengan edad igual a NA. Utiliza el siguiente objeto como el argumento *dparams* en *geom\_qq()*

params <- titanic %>%

filter(!is.na(Age)) %>%

summarize(mean = mean(Age), sd = sd(Age))

*¿Cuál de los siguientes gráficos obtienes siguiendo las instrucciones de arriba?*

*Ingresa una ‘X’ en la/s correcta/s*

|  |  |
| --- | --- |
| Respuesta | Marca con una X el correcto |
|  |  |
|  |  |
|  |  |
|  |  |

### **Pregunta 4 – Supervivencia por género**

0.0/2.0 puntos (calificable)

Para responder las siguientes preguntas realiza un gráfico de barras utilizando geom\_bar() de los supervivientes vs género, variables Survived y Sex respectivamente. Intenta graficar por una variable y rellenar por la otra variable. Inicialmente intenta con el gráfico por defecto y luego agrega los parámetros position = position\_dodge() y geom\_bar()para realizar barras separadas para cada grupo.

Puedes encontrar más información sobre cómo realizar gráficos de barras en [github.](https://rafalab.github.io/dsbook/distributions.html#other-geometries)

*¿Cuál de las siguientes afirmaciones es verdadera?*

*Ingresa una ‘X’ en la/s correcta/s*

|  |  |
| --- | --- |
| Respuesta | Marca con una X la/s correcta/s |
| Menos de la mitad de los pasajeros sobrevivieron |  |
| La mayoría de los supervivientes fueron mujeres |  |
| La mayoría de hombre sobrevivió |  |
| La mayoría de mujeres sobrevivió |  |

### **Pregunta 5 – Supervivencia por edad**

0.0/3.0 puntos (calificable)

Realiza un gráfico de densidad de supervivencia survival status. Cambia el eje y para que muestre la cantidad counts y setea alpha=0.2.

*Ingresa una ‘X’ en la/s correcta/s*

*¿Cuál de los rangos de edades fue más propenso a sobrevivir que a morir?*

|  |  |
| --- | --- |
| Respuesta | Marca con una X la/s correcta/s |
| 0 - 8 |  |
| 10 - 18 |  |
| 18 - 30 |  |
| 30 - 50 |  |
| 50 – 70 |  |
| 70 - 80 |  |

*¿Cuál de los rangos de edades tuvo mayor cantidad de muertos?*

|  |  |
| --- | --- |
| Respuesta | Marca con una X la/s correcta/s |
| 0 – 8 |  |
| 10 – 18 |  |
| 18 – 30 |  |
| 30 – 50 |  |
| 50 – 70 |  |
| 70 – 80 |  |

*¿Cuál de los rangos de edades tuvo la mayor proporción de muertes?*

|  |  |
| --- | --- |
| Respuesta | Marca con una X la/s correcta/s |
| 0 – 8 |  |
| 10 – 18 |  |
| 18 – 30 |  |
| 30 – 50 |  |
| 50 – 70 |  |
| 70 – 80 |  |

### **Pregunta 6 – Supervivencia por tarifa**

0.0/2.5 puntos (calificable)

Filtra los datos para quitar a los individuos que no hayan pagado la tarifa (fare=0).

Realiza un gráfico de cajas de las tarifas (fare) agrupadas por supervivencia (survival status). Intenta una transformación log2 de las tarifas. Agrega los puntos de datos al gráfico con jitter y alfa blending

*¿Cuál de las siguientes afirmaciones es/son verdadera/s?*

*Ingresa una ‘X’ en la/s correcta/s*

|  |  |
| --- | --- |
| Respuesta | Marca con una X la/s correcta/s |
| Generalmente los pasajeros que sobrevivieron pagaron tarifas más altas que los que no |  |
| El rango inter cuartil para las tarifas es más pequeño que para los pasajeros que sobrevivieron |  |
| El valor medio de las tafias era menor para pasajeros que no sobrevivieron |  |
| Solo un individuo pago una tarifa de alrededor de $500 y este individuo sobrevivió |  |

### **Pregunta 7 – Supervivencia por clase de pasajero**

0.0/3.0 puntos (calificable)

La variable Pclass corresponde a la clase de cada pasajero. Para este ejercicio realiza 3 gráficos de barras:

Para el primero, un gráfico clásico de la clase del pasajero relleno (filled) por Survival.

Para el segundo, realiza el mismo gráfico que el primero pero utilizando el argumento position = position\_fill() para mostrar la proporción relativa de cada grupo en lugar de la cuenta.

Para el tercero, realiza un gráfico de barras de la supervivencia Survival filled por la clase utilizando position = position\_fill() .

Para recordar la geometría de los gráficos de barras puedes acceder a [github](https://rafalab.github.io/dsbook/distributions.html#other-geometries)*.*

*¿Cuál de las siguientes afirmaciones es/son verdadera/s?*

*Ingresa una ‘X’ en la/s correcta/s*

|  |  |
| --- | --- |
| Respuesta | Marca con una X la/s correcta/s |
| Hay más pasajeros en la tercera clase que pasajeros en la primera y segunda combinadas. |  |
| La clase con menos cantidad de pasajeros fue la primera, luego la segunda y la clase con más pasajeros fue la tercera. |  |
| La proporción de supervivencia mayor fue para los pasajeros en primera clase seguida de los de segunda clase. La tercera clase tuvo la menor proporción de supervivencia. |  |
| Casi todos los pasajeros de la primera clase sobrevivieron. Casi todos los pasajeros de las clases restantes no sobrevivieron. |  |
| La mayoría de los sobrevivientes fueron de la primera clase (por mayoría nos referimos a más del 50%) |  |
| La mayoría de los pasajeros que no sobrevivieron fueron de la tercera clase |  |

### **Pregunta 8 – Supervivencia por edad, género y clase de pasajero**

0.0/3.0 puntos (calificable)

Crea una grilla de gráficos de densidad para edad, **filled** por supervivencia con cuenta en el eje Y, **faceted** por sex y clase de pasajero

*¿Cuál de las siguientes afirmaciones es/son verdadera/s?*

*Ingresa una ‘X’ en la/s correcta/s*

|  |  |
| --- | --- |
| Respuesta | Marca con una X la/s correcta/s |
| El grupo más grande de pasajeros son los hombres de tercera clase |  |
| La distribución de edad es la misma para todas las clases de pasajeros |  |
| La distribución de género es la misma para todas las clases de pasajeros |  |
| La mayoría de las mujeres de primera clase y segunda clase sobrevivieron |  |
| La mayoría de los hombres de segunda clase no sobrevivieron, con la excepción de los niños |  |