Práctica 2 (35% nota final)

Presentación

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar las herramientas de integración, limpieza, validación y análisis de las mismas. Para hacer esta práctica tendréis que trabajar en grupos de 2 personas. Tendréis que entregar un solo archivo con el enlace Github (https://github.com) donde se encuentren las soluciones incluyendo los nombres de los componentes del equipo. Podéis utilizar la Wiki de Github para describir vuestro equipo y los diferentes archivos que corresponden a vuestra entrega. Cada miembro del equipo tendrá que contribuir con su usuario Github. Aunque no se trata del mismo enunciado, los siguientes ejemplos de ediciones anteriores os pueden servir como guía:

- Ejemplo: https://github.com/Bengis/nba-gap-cleaning
- Ejemplo complejo (archivo adjunto).

Competencias

En esta práctica se desarrollan las siguientes competencias del Máster de Data Science:

- Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.
- Capacidad para aplicar las técnicas específicas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis.

Objetivos

Los objetivos concretos de esta práctica son:

- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.
- Saber identificar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyecto analítico.
- Aprender a analizar los datos adecuadamente para abordar la información contenida en los datos.
- Identificar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado en el proceso analítico.
- Actuar con los principios éticos y legales relacionados con la manipulación de datos en

función del ámbito de aplicación.
- Desarrollar las habilidades de aprendizaje que les permitan continuar estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo.
- Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

Descripción de la Práctica a realizar

El objetivo de esta actividad será el tratamiento de un dataset, que puede ser el creado en la práctica 1 o bien cualquier dataset libre disponible en Kaggle (https://www.kaggle.com).
Algunos ejemplos de dataset con los que podéis trabajar son:

- Red Wine Quality (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009)
- Titanic: Machine Learning from Disaster (https://www.kaggle.com/c/titanic)

El último ejemplo corresponde a una competición activa de Kaggle de manera que, opcionalmente, podéis aprovechar el trabajo realizado durante la práctica para entrar en esta competición.

Siguiendo las principales etapas de un proyecto analítico, las diferentes tareas a realizar (y **justificar**) son las siguientes:

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

   El siguiente dataset que se va a utilizar extraído de Kaggle es “Titanic: Machine Learning from Disaster” , este dataset me parece bastante interesante ya que nos otorga datos sobre cuantas personas separadas por genero sobrevivieron al titanic, esta información es importante debido a que tomando tal catástrofe es necesario saber la mayor cantidad de información para tomar distintas conclusiones, por ejemplo utilizaremos este ejercicio y este dataset para responder a la pregunta si al momento de evacuar fueron primero las mujeres y niños o evacuaron por la clase a la que pertenecían las personas.

2. Integración y selección de los datos de interés a analizar.

   En este apartado realizaremos el ingreso de los datos, y nos enfocaremos en los que van a ser más necesarios analizar:

   Primero vamos a cargar los datos:

```
data<-read.csv("./titanic.csv",header=T,sep=",")
attach(data)
```

   utilizamos en la funcion read el sep=“,” debido a que es el separador por el cual está dividido el data set, así que empezaremos haciendo un breve análisis de los datos ya que

nos interesa tener una idea general de los datos que disponemos. Por ello, primero calcularemos las dimensiones de nuestra base de datos y analizaremos qué tipos de atributos tenemos.

```
dim(data)
```

```
## [1] 2201    4
```

Para lograr calcular las dimensiones de la base de datos utilizamos la función dim(). Obtenemos que disponemos de 2201 registros o pasajeros (filas) y 4 variables (columnas).

Ahora vamos a utilizar la función str() y podemos observar que tenemos cuatro variables categóricas o discretas, es decir, toman valores en un conjunto finito. La variable CLASS hace referencia a la clase en la que viajaban los pasajeros (1ª, 2ª, 3ª o crew), AGE determina si era adulto o niño (Adulto o Menor), la variable SEX si era hombre o mujer (Hombre o Mujer) y la última variable (SURVIVED) informa si el pasajero murió o sobrevivió en el accidente (Muere o Sobrevive).

```
str(data)
```

```
## 'data.frame':    2201 obs. of  4 variables:
##  $ CLASS   : Factor w/ 4 levels "1a","2a","3a",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ AGE     : Factor w/ 2 levels "Adulto","Menor": 1 1 1 1 1 1 1 1 1 1 ...
##  $ SEX     : Factor w/ 5 levels "","H","Hombre",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ SURVIVED: Factor w/ 2 levels "Muere","Sobrevive": 2 2 2 2 2 2 2 2 2 2 ...
```

3. Limpieza de los datos.

3.1. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

ya que nos interesa saber si existen valores nulos (elementos vacíos) es recomendable empezar el análisis con una visión general de las variables. Mostraremos para cada atributo la cantidad de valores perdidos mediante la función summary.

```
##   CLASS         AGE           SEX            SURVIVED
##  1a  :325   Adulto:2092         :   3   Muere    :1490
##  2a  :285   Menor : 109   H     :   4   Sobrevive: 711
##  3a  :706                 Hombre:1724
##  crew:885                 M     :   7
##                           Mujer : 463
```

Como podemos observar tenemos valores vacios, y en el caso de SEX tenemos datos que necesitan ser normalizados tambien, asi que primero realizaremos un summary solo de SEX, y luego vamos a normalizar estos datos, debido a que la cantidad de hombres es mucho mayor al de mujeres, los datos vacios que tenemos los ubicaremos con el valor de Hombre de la siguiente manera:

```
##             H Hombre      M  Mujer
##      3      4   1724      7    463
```

```
# Al tener muchos valores distintos vamos a normalizar reemplazando el valor erroneo con el correcto
data$SEX <- as.character(data$SEX)
data$SEX[data$SEX == 'H'] <- "Hombre"
data$SEX[data$SEX == ''] <- "Hombre"
data$SEX[data$SEX == 'M'] <- "Mujer"
data$SEX <- as.factor(data$SEX)
summary(data$SEX)
```

```
## Hombre  Mujer
##   1731    470
```

De esta manera podemos observar que los datos ahora estan correctos sin datos vacios o sin normalizar.

3.2. Identificación y tratamiento de valores extremos.

Para visualizar los valores extremos podemos utilizar el comando summary y obtenemos los siguientes datos:

```
##   CLASS          AGE           SEX            SURVIVED
##  1a  :325   Adulto:2092   Hombre:1731   Muere    :1490
##  2a  :285   Menor : 109   Mujer : 470   Sobrevive: 711
##  3a  :706
##  crew:885
```

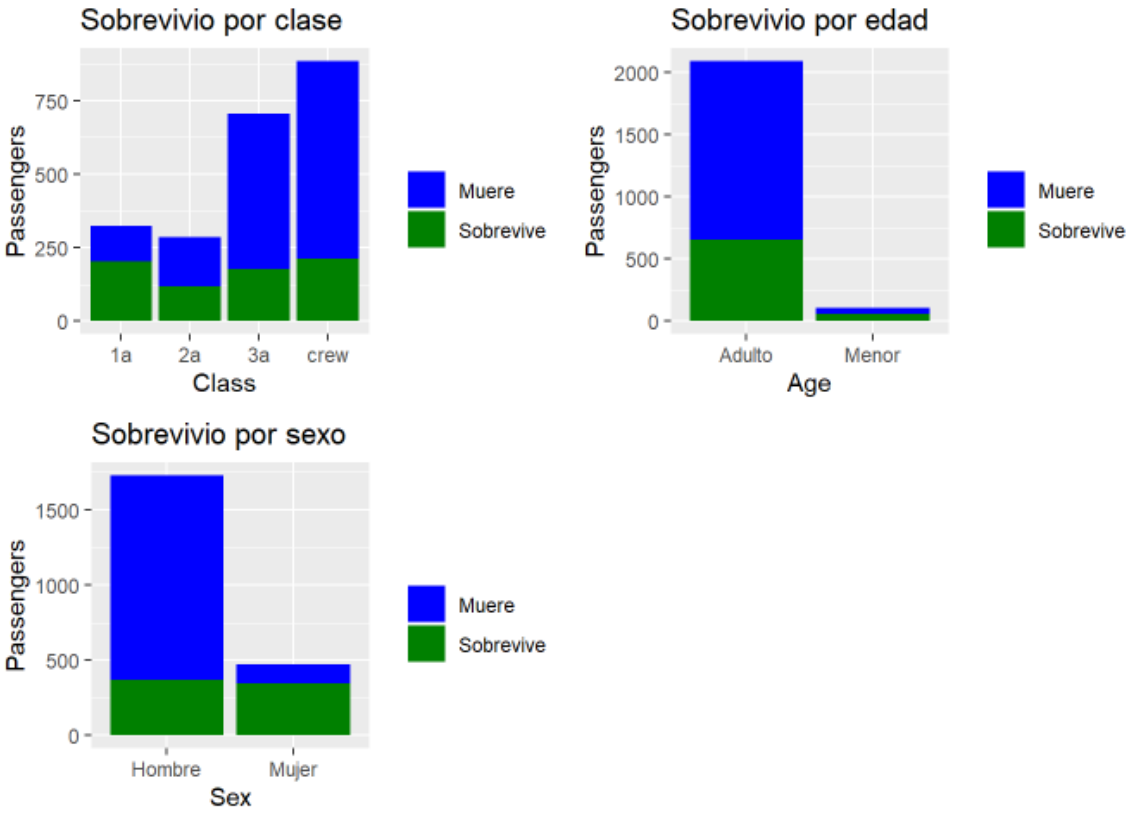
Al no tener valores numéricos o distintos valores que nos permite calcular extremos, no podemos realizar un tratamiento de los mismos.

4. Análisis de los datos.

4.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

Nos interesa describir la relación entre la supervivencia y cada una de las variables mencionadas anteriormente, entonces graficaremos mediante diagramas de barras la cantidad de muertos y supervivientes según la clase en la que viajaban, la edad o el sexo.

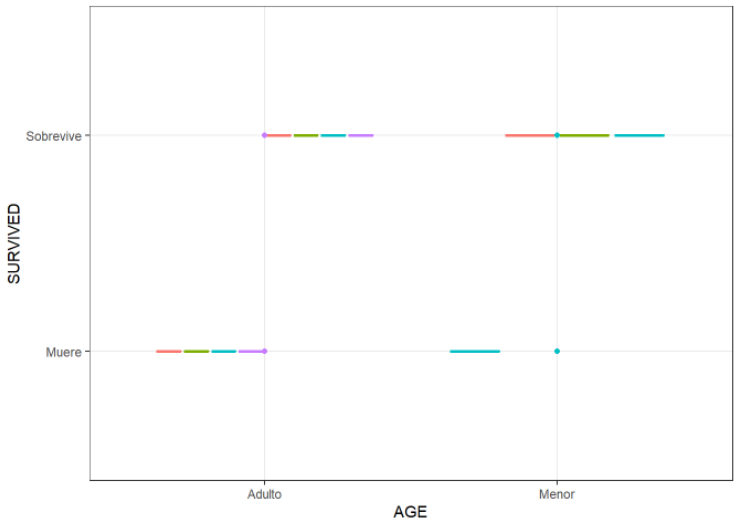
Para obtener los datos que estamos graficando utilizaremos el comando table para dos variables que nos proporciona una tabla de contingencia.



Con los graficos podemos sacar distintas conclusiones, por ejemplo, la cantidad que sobrevivieron es similar en hombres y mujeres (hombres: 367 y mujeres 344), teniendo en cuenta que viajaban muchos más hombres que mujeres, en cuanto a la clase, los que viajaron en primera clase fueron los que más sobrevivieron, teniendo en cuenta que un porcentaje muy pequeño de niños sobrevivió.

4.2. Comprobación de la normalidad y homogeneidad de la varianza.

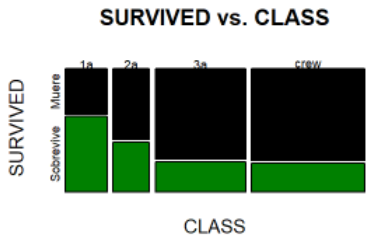
para la homogeneidad de varianza o homocedasticidad tenemos que tener en cuenta que, si los grupos con tamaños muestrales pequeños son los que tienen mayor varianza, la probabilidad real de cometer errores de hipotesis sera menor, para realizar un gráfico sobre este test de varianza utilizaremos el ggplot

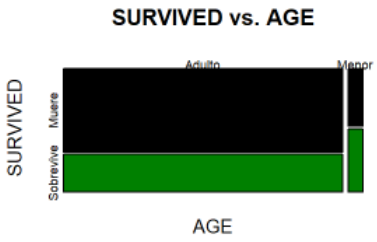


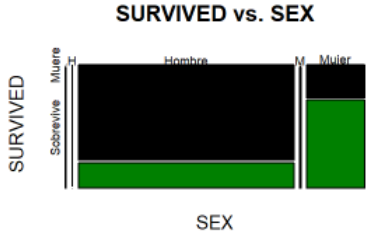
podemos tener en cuenta la homogeneidad de la varianza es mucho menor ya que estamos calculando entre valores discretos.

4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

En función de los datos vamos a realizar una tabla de contingencia, es una manera realizar una prueba estadística sobre los datos que tenemos de una manera más visual:







gracias a esto podemos observar de manera más visual la cantidad de personas por su clase, edad y sexo que sobrevivieron.

para responder a la pregunta inicial y tener valores estadísticos reales utilizando variables discretas la mejor manera es mediante un árbol de decisiones, el cual toma en cuenta cada hipó tesis para poder analizar qué tipo de pasajero del Titanic tenía probabilidades de sobrevivir o no. Por lo tanto, la variable por la que clasificaremos es el campo de si el pasajero sobrevivió o no. De todas maneras, al imprimir las primeras (con head) y últimas 10 (con tail) filas nos damos cuenta de que los datos están ordenados.

```
head(data,10)
```

```
##    CLASS    AGE    SEX  SURVIVED
## 1     1a Adulto Hombre Sobrevive
## 2     1a Adulto Hombre Sobrevive
## 3     1a Adulto Hombre Sobrevive
## 4     1a Adulto Hombre Sobrevive
## 5     1a Adulto Hombre Sobrevive
## 6     1a Adulto Hombre Sobrevive
## 7     1a Adulto Hombre Sobrevive
## 8     1a Adulto Hombre Sobrevive
## 9     1a Adulto Hombre Sobrevive
## 10    1a Adulto Hombre Sobrevive
```

```
tail(data,10)
```

```
##      CLASS    AGE   SEX  SURVIVED
## 2192  crew Adulto Mujer Sobrevive
## 2193  crew Adulto Mujer Sobrevive
## 2194  crew Adulto Mujer Sobrevive
## 2195  crew Adulto Mujer Sobrevive
## 2196  crew Adulto Mujer Sobrevive
## 2197  crew Adulto Mujer Sobrevive
## 2198  crew Adulto Mujer Sobrevive
## 2199  crew Adulto Mujer     Muere
## 2200  crew Adulto Mujer     Muere
## 2201  crew Adulto Mujer     Muere
```

5. Representación de los resultados a partir de tablas y gráficas.

Los gráficos y el resto de la resolución se encuentran mas detallados en el html adjunto de la práctica.

6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

A partir del árbol de decisión que generamos, se pueden extraer las siguientes reglas de decisión:

SEX = "Hombre" → Muere. Validez: 80,2%

CLASS = "3a" → Muere. Validez: 75.1%

CLASS "1ª", "2ª" o "Crew" y SEX = "Mujer" → Sobrevive. Validez: 90,5%

Por tanto, podemos concluir con el conocimiento extraído que al momento de evacuar si fueron primero los niños y las mujeres a excepción de la 3ra clase, lastimosamente sea la edad o el sexo que sea, mientras estuviera en 3ra clase tenía mayor probabilidad de morir frente al resto del barco.

Efectivamente los resultados realizar su cometido de responder la pregunta inicial mediante procesos estadisticos.

y cruzado con el análisis visual se resume en "las mujeres y los niños primero a excepción de que fueras de 3ª clase".

7. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.

Recursos

Los siguientes recursos son de utilidad para la realización de la práctica:

- Calvo M., Subirats L., Pérez D. (2019). Introducción a la limpieza y análisis de los datos. Editorial UOC.
- Megan Squire (2015). *Clean Data*. Packt Publishing Ltd.
- Jiawei Han, Micheine Kamber, Jian Pei (2012). *Data mining: concepts and techniques*. Morgan Kaufmann.
- Jason W. Osborne (2010). *Data Cleaning Basics: Best Practices in Dealing with Extreme Scores.* Newborn and Infant Nursing Reviews; 10 (1): pp. 1527-3369.
- Peter Dalgaard (2008). *Introductory statistics with R*. Springer Science & Business Media.
- Wes McKinney (2012). *Python for Data Analysis*. O'Reilley Media, Inc.
- Tutorial de Github https://guides.github.com/activities/hello-world.

Criterios de valoración

Todos los apartados son obligatorios. La ponderación de los ejercicios es la siguiente:

- Los apartados 1, 2 y 6 valen 0,5 puntos.
- Los apartados 3, 5 y 7 valen 2 puntos.
- El apartado 4 vale 2,5 puntos.

Se valorará la idoneidad de las respuestas, que deberán ser claras y completas. Las diferentes etapas deberán justificarse y acompañarse del código correspondiente. También se valorará la síntesis y claridad, a través del uso de comentarios, del código resultante, así como la calidad de los datos finales analizados.

Formato y fecha de entrega

Durante la semana del 25 de mayo el grupo podrá entregar al profesor una entrega parcial opcional. Esta entrega parcial es muy recomendable para recibir asesoramiento sobre la práctica y verificar que la dirección tomada es la correcta. Se entregarán comentarios a los estudiantes que hayan efectuado la entrega parcial pero no contará para la nota de la práctica. En la entrega

parcial los estudiantes deberán entregar por correo electrónico (mcalvogonza@uoc.edu o xvivancos@uo.edu) el enlace al repositorio Github con el que hayan avanzado.

En referente a la entrega final, hay que entregar un único fichero que contenga el enlace Github donde haya:

1. Una Wiki con los nombres de los componentes del grupo y una descripción de los ficheros.
2. Un documento PDF con las respuestas a las preguntas y los nombres de los componentes del grupo. Además, al final del documento, deberá aparecer la siguiente tabla de contribuciones al trabajo, la cual debe firmar cada integrante del grupo con sus iniciales. Las iniciales representan la confirmación de que el integrante ha participado en dicho apartado. Todos los integrantes deben participar en cada apartado, por lo que, idealmente, los apartados deberían estar firmados por todos los integrantes.

Contribuciones	Firma
Investigación previa	Integrante 1, Integrante 2, ...
Redacción de las respuestas	Integrante 1, Integrante 2, ...
Desarrollo código	Integrante 1, Integrante 2, ...

3. Una carpeta con el código generado para analizar los datos.
4. El fichero CSV con los datos originales.
5. El fichero CSV con los datos finales analizados.

Este documento de entrega final de la Práctica 2 se debe entregar en el espacio de Entrega y Registro de AC del aula antes de las **23:59** del día **9 de junio**. No se aceptarán entregas fuera de plazo.