Fundamentos matemáticos del análisis de datos.

Curso 2021-22.

Universidad Pontificia Comillas. Master en Big Data.

Examen del martes 5 de Octubre de 2021.

Instrucciones

- Crea una carpeta para los ficheros del examen llamada:

  `examen_tuapellido_tunombre`

- Guarda el fichero `2021-Examen-FMAD.Rmd` disponible en Moodle en esa carpeta. Es el fichero que usarás para introducir tus respuestas. Los bloques de código de este fichero deben contener el código completo que produce tus respuestas. Se corregirá sobre la salida en pdf o HTML, pero en caso de duda lo que prevalece siempre es el código de este fichero Rmarkdown. Si tienes problemas para descargarlo desde Moodle avisa a tu profesor.

- La carpeta debe contener una subcarpeta llamada `data` en la que guardaremos el fichero de datos. La tabla de datos para el examen es distinta para aada uno de vosotros. Las variables son siempre las mismas, pero para descargar la tabla que te corresponde debes:

  - asegurarte de que el directorio de trabajo es la carpeta que contiene este fichero Rmd.
  - introducir en el siguiente bloque tu código de alumno y ejecutarlo desde aquí:

```r
# Comprobación del directorio de trabajo
getwd()

codigo = ""
serverURL = "https://hungry-stonebraker-03db5b.netlify.app/"
if(codigo == "") stop("INTRODUCE TU CÓDIGO EN EL FICHERO RMD.")

dataFile =  paste0(codigo, ".csv")

dataURL = paste0(serverURL, dataFile)
varURL = paste0(serverURL, "variables.txt")

localFile = paste0("./data/", dataFile)
localVar = "./data/variables.txt"
if(!dir.exists("data"))
  stop("CUIDADO: No existe la carpeta data en el directorio de trabajo.")
if(!file.exists(localFile)){
  download.file(dataURL, destfile = localFile)
  download.file(varURL, destfile = localVar)
} else {
```

```
  warning(paste0("Cuidado: el fichero de datos ",
                 dataFile,
                 " ya existe en tu carpeta de datos"))
}
```

Si ese enlace no funciona o no recuerdas tu código avisa a tu profesor.

- Esta tabla contiene información sobre los empleados de una empresa. El significado de las variables se explica en el documento *variables.txt* que se ha descargado a tu carpeta data.

Entrega.

- Cuando termines de responder a las preguntas del examen, comprime la carpeta de tu proyecto en formato `zip`, de manera que el archivo comprimido tenga el nombre de esa carpeta y extensión `.zip`. Enviáme por correo ese fichero comprimido a la dirección `fsansegundo@comillas.edu`.
- Se valorará adicionalmente el uso de funciones del tidyverse (dplyr, ggplot) para las representaciones gráficas y las operaciones con tablas.

Preguntas del examen

Preliminares

- Lee el fichero de datos y crea un `data.frame` o `tibble` de R llamado `empleados`, que usaremos para el resto del examen.

Apartado 1

- C1: ¿Cuántas observaciones hay en la tabla? ¿Cuántas variables? ¿Hay datos ausentes? Si la respuesta a la última pregunta es afirmativa, localiza esos datos ausentes indicando en qué filas y columnas se encuentran. Después, antes de seguir adelante, elimina de la tabla las filas que contienen esos datos.
- **Respuesta:**
- C2: Vamos a fijarnos en las tres últimas columnas de la tabla, que son:
  `RateType, Rate, Education_Field`
  En particular verás que las columnas `Rate` y `RateType` incumplen una de las condiciones que definen a los conjuntos de datos limpios (*tidy*), porque la columna `Rate` mezcla los valores de más de una variable. La variable correspondiente se indica en la columna `Rate_type`. Por otro lado, `Education_Field` incumple otra de esas condiciones, porque cada fila de esa tabla combina los valores de dos variable. En este apartado debes limpiar el conjunto de datos utilizando las herramientas del `tidyverse`. Visualiza el resultado usando `select` para que la salida del código muestre exclusivamente las variables que se han modificado en este apartado (¡cuidado, no modifiques los datos al hacer esto!)
- **Respuesta:**
- C3: Una vez limpios los datos: ¿de qué tipo son las variables del conjunto de datos? Usa la función `mutate_if` de *dplyr* (o alternativamente usa *across*) para asegurarte de que todas las variables de esta tabla que se han leído como `character` se convierten en factores. Mira la ayuda de `mutate_if` para hacer esto si lo necesitas. Después haz una lista de variables cuantitativas y otra de variables cualitativas (factores). ¿Cuáles, dentro de estas últimas, son factores ordenados (no es necesario incluir el orden en el proceso de transformación de estos factores)?
  Notas:

  - si no consigues usar correctamente `mutate_if` asegúrate, en los siguientes apartados, de convertir en factores al menos las variables necesarias para el trabajo de ese apartado.

- **Respuesta:**

Apartado 2

- C4: Representa gráficamente la distribución de la variable edad `Age` mediante un histograma con la curva de densidad de la variable superpuesta. Recuerda que en este y otros apartados se valorará positivamente el uso de ggplot. Representa también un boxplot de la variable edad. Opcionalmente puedes usar un violinpot o añadir los puntos de la muestra (asegúrate en ese caso de que no impiden ver el boxplot).

- **Respuesta:**

- C5: Estudia gráficamente (por ejemplo con boxplots) la relación entre la variable `Age` y la variable `Attrition`. ¿Influye el género en esa relación? Usa los recursos gráficos de ggplot para discutir la respuesta, lo que se busca es un juicio inicial basado en una exploración gráfica.

- **Respuesta:**

- C6: Haz la tabla de frecuencias absolutas del factor `JobSatisfaction`. Después haz una representación gráfica adecuada de esa tabla. Opcionales:

  - juega con los argumentos `fill` y `position` de `ggplot` para incorporar al gráfico la información de `Attrition` mediante colores.
  - ten en cuenta que `JobSatisfaction` es un factor ordenado y trata de incorporar esa ordenación a la representación gráfica. Indicación: usa la función `factor` y sus argumentos `levels` y `ordered` para esto.

- **Respuesta:**

Apartado 3

- C7: Calcula la mediana del salario mensual para cada departamento y cada nivel dentro de ese departamento. Ordena la respuesta de mayor a menor y asegúrate de que en la tabla de salida se muestran las columnas relevantes.

- **Respuesta:**

- C8: Si elegimos al azar un empleado cuyo nivel de satisfacción en el trabajo `JobSatisfaction` es `VeryHigh`, calcula la probabilidad de que sea soltero. Si elegimos un empleado de la empresa al azar ¿cuál es la probabilidad de que sea una mujer que lleva más de cinco años en la empresa?

- **Respuesta:**

- C9: Si elegimos 12 empleados de esta compañía al azar y con remplezamiento, ¿cuál es la probabilidad de que 4 de ellos hayan trabajado en 3 o más compañías (`NumCompaniesWorked`)?

- **Respuesta:**

- C10: Haz una tabla de contingencia (dos por dos) de `Attrition` frente a `Overtime`, la variable que nos dice si un empleado hace o no horas extras. Supongamos que queremos usar `Overtime` como una especie de *prueba diagnóstica* de los valores de `Attrition`, equiparando enfermo/sano con `Attrition` Yes/No y test positivo/negativo con `Overtime` Yes/No. ¿Cuál es la tasa de falsos positivos de este test? ¿Cuál es la precisión del test? Opcional: ¿cuáles son su sensibilidad y especificidad?

- **Respuesta:**

Apartado 4

- C11: Asumiendo la normalidad de los datos, calcula un intervalo de confianza al 95% para la edad medio de los empleados del departamento más numeroso de la empresa. Opcional: ¿crees que está justificada la hipótesis de normalidad de esos datos?
- **Respuesta:**
- C12: Toma una muestra aleatoria y con remplazamiento de 20 empleados del departamento de ventas y haz un contraste (al 95% de significación) de la *hipótesis nula:* la distancia media a su domicilio es de 9 km. Asegúrate de mantener el comando `set.seed` como primera línea de tu respuesta en este apartado para garantizar la reproducibilidad.
- **Respuesta:**

```
set.seed(2021)
```

- C13: Considerando los empleados del departamento *Research_Development* vamos a hacer un modelo de regresión lineal para las dos variables:
  x : YearsAtCompany y : MonthlyIncome
  Construye el modelo usando la función `lm` y además dibuja el diagrama de dispersión de esas variables junto con la recta de regresión obtenida. ¿Qué porcentaje de la variabilidad en el salario mensual se explica con el modelo? ¿Cuánto se incrementa el salario mensual de un empleado de ese departamento por cada año de antigüedad en la empresa?
- **Respuesta:**