

CURSO: CC50 – ADMINISTRACION DE LA INFORMACION

CLASE: TA1 #3 (TRABAJO ACADEMICO GRUPAL)

TEMA: ANÁLISIS EXPLORATORIO DE UN CONJUNTO DE DATOS EN R/RSTUDIO

PROFESOR/A: Ing. PATRICIA REYES SILVA

Objetivo

Realizar un análisis exploratorio de un conjunto de datos, creando visualizaciones, **preparando los datos** y obteniendo inferencias básicas utilizando R/RStudio como herramienta de software.

Competencias

Serán evaluadas las competencias, según rubrica adjunta EA-RUBCC50-2101-CC51.

Conjunto de Datos

El conjunto de datos motivo de análisis se denomina: **Hotel booking demand**. Su versión original se obtuvo de Kaggle, sin embargo, para esta evaluación, este conjunto de datos ha sido modificado incorporando ruido en los datos, básicamente: datos faltantes (NA) y datos atípicos (outliers). El conjunto de datos se puede descargar desde el siguiente enlace: shorturl.at/dgtCK

En este conjunto de datos se recopilan datos de un hotel urbano y otro de tipo resort. Incluye información de cuándo se realizó la reserva, la duración de la estadía, la cantidad de espacios de estacionamiento disponibles, cantidad de huéspedes adultos, niños y / o bebés, entre otros datos.

El conjunto de datos original proviene del documento: [Hotel booking demand datasets](#)

Documento Entregable

El grupo de estudiantes entregará un único documento de tipo PDF, desarrollando los siguientes temas en este orden propuesto:

1. CASO DE ANALISIS

Explicación sobre el origen de los datos (procedencia de los datos, autor/autores, fecha, país, etc.)

Casos de uso aplicables (describir, por ejemplo: ¿Para quién sería importante el análisis de estos datos?, ¿Quien o quienes se benefician?)

El análisis exploratorio de los datos y las visualizaciones generadas debieran dar respuesta a las siguientes preguntas u otras que analizaron, y que debieran considerarse en las conclusiones (**responder mínimo a cinco de ellas, de preferencia todas**):

- a. ¿Cuántas reservas se realizan por tipo de hotel? o ¿Qué tipo de hotel prefiere la gente?
- b. ¿Está aumentando la demanda con el tiempo?
- c. ¿Cuándo se producen las temporadas de reservas: alta, media y baja?
- d. ¿Cuándo es menor la demanda de reservas?
- e. ¿Cuántas reservas incluyen niños y/o bebés?
- f. ¿Es importante contar con espacios de estacionamiento?
- g. ¿En qué meses del año se producen más cancelaciones de reservas?

Las respuestas deben estar acompañadas de una visualización.

2. CONJUNTO DE DATOS (DATA SET)

- ❖ Descripción de la estructura de los datos

3. ANÁLISIS EXPLORATORIO DE DATOS

Descripción de instrucciones ejecutadas en R/RStudio y resultados obtenidos para:

- ❖ CARGAR DATOS

- La carga del dataset deberá considerar los parámetros `header = TRUE`, `stringsAsFactors = FALSE`

- ❖ INSPECCIONAR DATOS

- Los alumnos deberán explorar los datos del dataset, verificando, por ejemplo, estructura, tipo, valores de los datos, nombre de columnas, etc.

- ❖ PRE-PROCESAR DATOS

- Identificación de datos faltantes (NA).
- Explicación y aplicación de la técnica utilizada para eliminar o completar los datos faltantes.
- Identificación de datos atípicos (Outliers).
- Explicación y aplicación de la(s) técnica(s) utilizada(s) para transformar los datos atípicos.

- ❖ VISUALIZAR DATOS (mínimo cinco visualizaciones)

- Los alumnos decidirán que variables del dataset seleccionarán del conjunto de datos para demostrar las correlaciones existentes, visualizarlas e inferir sus conclusiones.

4. CONCLUSIONES PRELIMINARES

Las conclusiones resultan de las respuestas a las preguntas iniciales del punto 1. Caso de Análisis.

5. Archivar y publicar

- Se deberá contemplar un repositorio en **Github.com** llamado ea-2021-1-cc51 conteniendo dos carpetas:
 - **data:** deberá contener el dataset original y el dataset final resultante (limpio o preparado para análisis).
 - **code:** deberá contener los scripts en R utilizados para el proceso de carga, inspección, pre-procesado y visualización del dataset.
- El archivo .Readme, dentro de Github, deberá contemplar:
 - Objetivo del trabajo
 - Nombre de los alumnos participantes
 - Breve descripción del dataset (se puede adjuntar el archivo PDF)
 - Conclusiones
 - Licencia

Guiarse de estos ejemplos de publicaciones de trabajos en Github:

<https://github.com/fernandoabcampos/titanic-data-cleaning-and-validation>

<https://github.com/navarroyepes/TCVDPRAC2>

- (Opcional) El mismo contenido publicado en Github reproducirlo en la sección Wiki perteneciente al grupo en el Aula Virtual.
- En el documento entregable, **se deberá incluir el enlace a la cuenta de Github.com** desde donde se accede a la publicación de la evaluación.

Consideraciones adicionales

- Se evaluará el orden dentro de la organización del documento, así como la correcta redacción y gramática.
- Se valorarán las respuestas a preguntas que no hayan sido propuestas en la presente evaluación.

Fecha límite de entrega: 02/10/2021