CC216-Fundamentos de Data Science

Trabajo Parcial

2024-01

I. OBJETIVO

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

II. 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 AQUI

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

III. DOCUMENTO ENTREGABLE

El grupo de estudiantes entregará un único documento de acuerdo a la nomenclatura de archivos, 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 mínimamente a las siguientes preguntas u otras que analicen, y que debieran considerarse en las conclusiones:

i. ¿Cuántas reservas se realizan por tipo de hotel? o ¿Qué tipo de hotel prefiere la gente?
ii. ¿Está aumentando la demanda con el tiempo?
iii. ¿Cuándo se producen las temporadas de reservas: alta, media y baja?
iv. ¿Cuándo es menor la demanda de reservas?
v. ¿Cuántas reservas incluyen niños y/o bebes?
vi. ¿Es importante contar con espacios de estacionamiento?
vii. ¿En qué meses del año se producen más cancelaciones de reservas?

Las respuestas deben estar acompañadas de una visualización sustentada en los datos.

2. CONJUNTO DE DATOS (DATA SET)

- Descripción de la estructura de los datos (tabla conteniendo la estructura y descripción de cada uno 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

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

IV. CONCLUSIONES PRELIMINARES

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

V. ARCHIVAR Y PUBLICAR

- Se deberá contemplar un repositorio en Github.com llamado:
  CC216- TP-2024-1 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, preprocesado 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.

Nomenclatura de Archivos:

upc-pre-2401-nrogrupo-tp1 (.docx o .pdf)
upc-pre-2401-nrogrupo-tp1.R
upc-pre-2401-nrogrupo-tp1(pptx u otros tipos de presentación)

Sólo se debe entregar tres archivos por grupo.

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.
- Durante la semana 7 se realizará la exposición grupal.
- Cada grupo contará con 10-15 minutos para que cada integrante detalle cómo se obtuvieron las visualizaciones y conclusiones (preparar una presentación de no más de diez diapositivas).
- El orden de la exposición será mediante sorteo.
- El alumno que no se presente, perderá la calificación oral.
- La calificación podrá ser diferenciada por alumno si su participación / respuestas durante la exposición es insuficiente.