Universidad de Buenos Aires
Facultad De Ingeniería
Año 2023 - 1er Cuatrimestre

Organización de Datos (75.06 /95.58)
Trabajo Practico 1: Reservas de Hotel

CHECKPOINT 1: Análisis Exploratorio y Preprocesamiento de Datos
FECHA: 14/04/23

INTEGRANTES:

Gonzalez,Tomás	#108193
<togonzalez@fi.uba.ar>	
Moreno del Ruvo, Valentina	#107948
<vmdr18@gmail.com>	
Pol, Juan Manuel	#108448
<jpol@fi.uba.ar>	

1. Introducción

El objetivo de este trabajo es analizar un set de datos relacionados a reservas de hoteles para finalmente estimar si una futura reserva se va a rechazar o no. Para este primer checkpoint se buscaba cumplir 4 requisitos fundamentales:

- Exploración inicial
- Visualización de los datos
- Manejo de datos faltantes
- Manejo de valores atípicos

2. Exploración Inicial

Para el segmento de exploración inicial lo que buscábamos era saber con qué exactamente estábamos tratando. Es decir, queríamos saber la cantidad de filas y columnas que teníamos, que tipo de variable era cada una, que representaba cada una, su distribución y valores más significativos.
Luego de tener un vistazo general decidimos revisar para los valores numéricos lo correspondiente a cuantiles, mínimos, máximos, media y varianza. Y para las variables cualitativas la cantidad de valores únicos, el más frecuente y cual es esa frecuencia.

En ambos casos usamos el *.describe*().
Finalmente, buscamos la matriz de correlación entre todas las variables numericas y las graficamos en el heatmap de la figura 2. Los valores obtenidos fueron excepcionalmente bajos por lo que no pudimos llegar a una conclusión significativa.

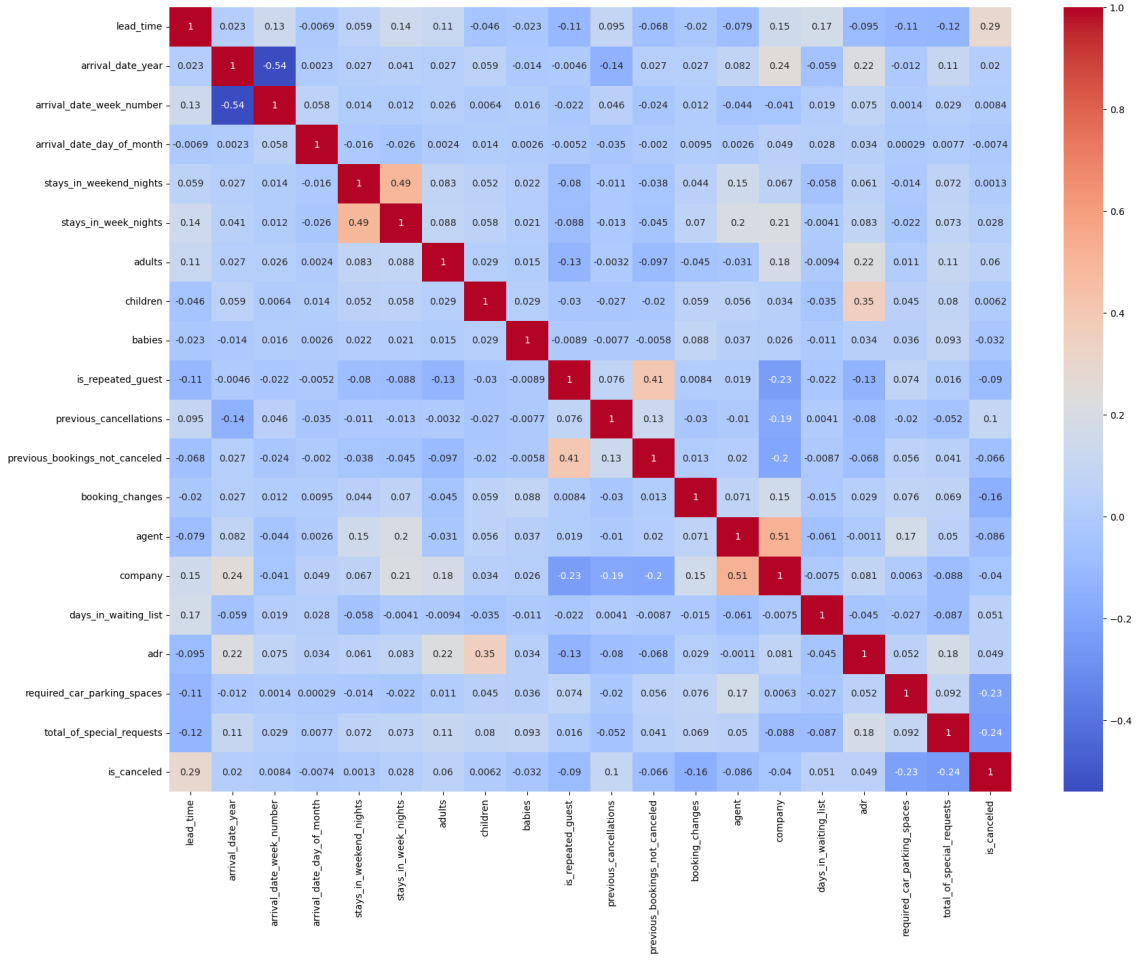

Figura 1: Heatmap con la correlación entre variables numéricas

3. Visualización de los datos

Para la visualización utilizamos las librerías de Seaborn y Matplotlib, realizamos tres gráficos iniciales que consideramos de suma importancia, y son por un lado el gráfico *scatter* de todas las variables numéricas frente a nuestra variable a estimar *is_canceled*. Y por otro lado, para las variables cualitativas decidimos hacer un *barplot* representando para cada valor la proporción de *is_canceled* y también para

cada una hicimos otro para su distribución en cantidad. Luego decidimos tratar ciertas variables en especifico para buscar si había algún tipo de relación entre la variable *target* y las mismas. Utilizamos las variables *arrival date month*, *market segment country*.

4. Limpieza de datos

En lo que respecta al manejo de valores faltantes, decidimos hacer uso del paper provisto con el set de datos para asegurarnos que aquellos datos faltantes no tengan un significado por si mismos. Efectivamente este era el caso para las variables *country*, *agent* y *company*, donde el valor faltante representaba que no era aplicable ninguna categoría.
Par cada uno de estos decidimos crear una propia para significar la falta llamándose

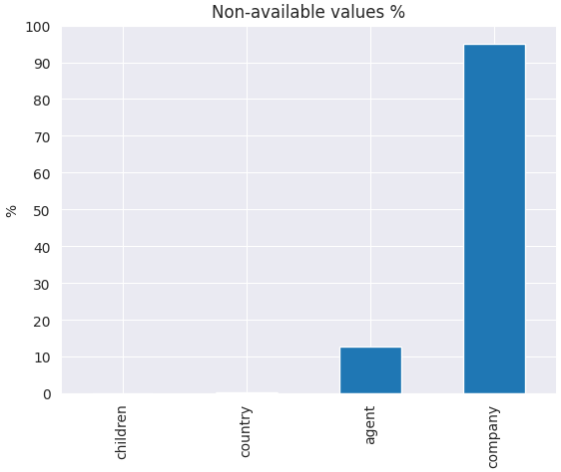

Figura 2: Barplot con el porcentaje de valores no disponibles por variable

"NC"(No country) o "NA"(No agent). Solamente 4 variables tenían valores nulos, faltando solo *children* decidimos remplazarlos por 0, dado que es el valor más frecuente.

5. Manejo de valores atípicos

El manejo de valores atípicos o *outliers* lo realizamos en dos segmentos, por un lado de forma univariada y por el otro lado multivariada. Para el análisis univariado utilizamos el z-score modificado, luego graficamos para cada variable la distribución de los score. Decidimos considerar a cualquier valor mayor en modulo a 3 como un

outlier y analizamos específicamente casos donde el valor fue excepcionalmente alto. Por ejemplo el caso de la reserva con 9 *babies*.
Para el análisis multivariado primero elegimos ciertas variables cuya correlación fue relativamente considerable, aun así de un orden bajo, y las graficamos enfrentadas para intentar hallar gráficamente algún valor extraño. Sin embargo, para los 3 casos realizados las variables no presentaron ninguna anomalía.
Luego, razonando en términos del contexto del dataset, comparamos valores de la variable *adults* con otras 3 que nosotros consideramos significativas y que podrían presentar outliers. Notamos que había un número importante de registros con 0 adultos, 0 niños y 0 bebes, y también registros que contaban con niños y/o bebes pero ningún adulto, que en el contexto del dataset resulta extraño.