Informe y presentación

Teresa Álvarez de Portugal

May 2023

1 Introduction

Aqui transmitiré los resultados del los análisis realizados. Voy a explicar paso a paso cada uno de los apartados con las colclusiones correspondientes de las tareas que he realizado. Incluiré principalmente los resultados de las tareas realizadas sobre los datos obtenidos a través de la ejecución del código contenido en el documento técnico.

1.1 Ejercicio 1

Esta tarea consiste en conocer con qué tipo de datos contamos. Para ello, he tenido que categorizarlos según su estructura y presentarlos de la siguiente manera:

NOMBRE DEL CAMPO 1	TIPO DE DATO
Activity Period	Integer
Operating Airline 4	Object
Operating Airline IATA Code	Object
Published Airline	Object
Published Airline IATA Code	Object
GEO Summary	Object
GEO Region	Object
Activity Type Code	Object
Price Category Code	Object
Terminal	Object
Boarding Area	Object
Passenger Count	Integer
Adjusted Activity Type Code	Object
Adjusted Passenger count	Integer
Year	Integer
Month	Object

1.2 Ejercicio 2

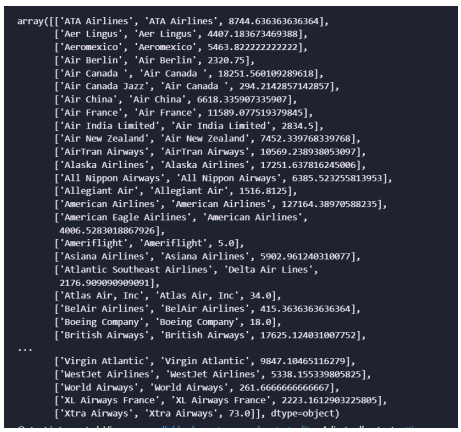
Este ejercicio consiste en desarrollar una serie de puntos que incluyen preguntas que son las que contestaré aqui.

-¿Cuántas compañias diferentes aparecen en el fichero?

Aparecen un total de 77 compañias. Para poder contestar a esta pregunta combiné dos columnas que contenían los nombres de las compañias y saqué los valores unicos.

-¿Cuántos pasajeros tiene de media los vuelos de cada compañía?

Como en el caso anterior, habia dos columnas que contenian número de pasajeros, pero en este caso las comparé y elejí una de ellas. Utilicé la lista en la que habia guardado las compañias



anteriormente, filtré el dataframe para las aerolíneas presentes en esa lista y calculé el número medio de pasajeros por compañia. El resultado es el siguiente:

Como se aprecia son varios datos debido a que son 77 compañias, por lo que en el notebook se puede comprobar la respuesta mejor.

-Eliminar los registros duplicados por el campo GEO Region, manteniendo únicamente aquel con mayor número de pasajeros

Para realizar esta tarea agrupé los datos de la columna pedida. Despues calculé el valor máximo de las columnas de pasajeros para cada grupo de GEO Region. Y reestablecí el índice porque despues me pide que lo guarde en un dataset nuevo. Los resultados se pueden observar bien en la siguiente imagen.

	GEO Region	Passenger Count	Adjusted Passenger Count
0	US	659837	659837
1	Canada	39798	39798
2	Asia	86398	86398
3	Europe	48136	48136
4	Australia / Oceania	12973	12973
5	Mexico	29206	29206
6	Central America	8970	8970
7	Middle East	14769	14769
8	South America	3685	3685

1.3 Ejercicio 3

En este ejercicio se nos pide realizar un análisis descriptivo de los datos usando Dask y calculando la media y la desviación estándar de cada elemento del conjunto de datos. Para esto primero observé que solo habia cuatro columnas con valores númericos, por lo que entendí que tenía que codificar el dataframe para poder trabajar comodamente. Elegí entre la técnica de codificación de Label Encoding y One-Hot Encoding, porque ambas son usadas comunmente en el procesamiento de datos antes de realizar el análisis. Para este caso que tenia que calcular tanto media como desviación estándar usé Label Encoding, el cual asigna valores numéricos únicos a cada categoría, lo que haría mas fácil el cálculo.

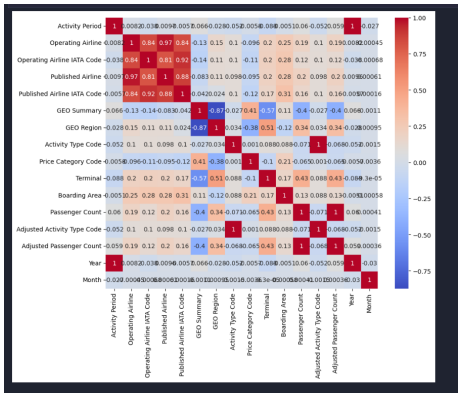
-Mis conclusiones fueron las siguientes:

A parte de calcular la media y la desviación estandar, he calculado la moda porque creo que es mas relevante en el caso de analizar algunas columnas.

La media representa el valor promedio de los datos de una variable. En este caso sería efectiva sobre las columnas como la de 'Passenger count', 'ADjusted passenger count', 'Price code category', 'GEO summary'. En las demas son datos en los que necesariamente los valores no tienen por que tener un orden, con lo que sería mas efectivo usar la moda para ver que valor es el que mas se repite.

En cuanto a la desviación, es una medidia de dispersión que indica cuánto se alejan los datos de la media. Los que tienen desviación estandar alta quiere decir que tiene los valores mas dispersos y hay mayor variabilidad. Donde es baja significa que los valores estan mas cercanos a la media. Para interpretar todo esto tambien hay que tener en cuenta la cantidad de observaciones. En las columnas donde hay pasajeros por ejemplo, hay una desviación muy alta porque los valores van desde 2 hasta un número muy alto. En cambio en columnas commo el precio o la actividad hay una desviacion baja porque hay pocos valores por lo que estarán mas cerca de la media.

Después nos pide tambien realizar un análisis de la correlación cuyo resultado debía ser una mariz de correlación de datos que represetne de qué manera están relacionadas las diferentes variables. Como ya tenía el dataframe codificado continue programando la matriz de correlación que me daba como resultado lo siguiente:



-De donde saqué las siguientes conclusiones:

Las celdas en las que hay correlación alta, en este caso de 0,8 para arriba, son las que tienen una correlación positiva muy fuerte. En este caso se puede observar que las 4 columnas que hablan sobre las aereolíneas y sus códigos IATA estan altamente correlacionadas ya que hablan de los mismos datos. Lo mismo ocurre con las columnas 'Activity Type Code' con 'Adjusted Activity Type Code' y 'Passenger Count' con 'Adjusted Passenger count'. De la misma manera ocurre con la correlación entre 'Activity Period' y 'Year' ya que la primera de las mencionadas contiene tanto el año como el mes pero unido y por esa razon no reconoce la correlación con la columna del mes, aunque se puede apreciar cuando se compara.

Donde hay correlación negativa significa que las variables tiendes a moverse en direcciones opuestas. Entre las columnas GEO Summary y Geo region es donde está mas cerca de 1 por lo que son las mas correlacionadas negativamente. La siguiente con estas características serian GEO Summary con Terminal pero no es tan fuerte.

En el resto se puede observer correlación nula o cerca de ella lo que significa que no hay una relación lineal clara entre las variables. Esto no significa que no haya una relación entre las variables, ya que podría haber relaciones no lineales.

Por último, este ejercicio nos pedía seleccionar uno de los algoritmos vistos durante el curso y aplicarlo para despues explicar los resultados obtenidos.

Aunque no he podido finalizar la aplicación del algoritmo, mi decisión fué que la mejor opción era usar Apache Spark como framework de procesamiento paralelo y distribuido. En mi opinión es una buena opción graciazs a la capacidad que tiene para procesar grandes volúmenes de datos de manera escalable. Esto significa que puede servir bien con el dataset de grandes dimensiones con el que estabamos trabajando. También lo seleccioné porque proporciona una amplia gama de bibliotecas útiles para aplicar algoritmos y análisis específicos al dataset y porque se integra con python a travéz de la biblioteca PySpark.