

BIG DATA

PROYECTO FINAL

Samuel Porcel Rodríguez



SAMUEL PORCEL RODRÍGUEZ

RESPONSABLE DEL DESARROLLO DEL TRABAJO FINAL DE CURSO

Fecha: 13/11/2024

ÍNDICE

| | |
|---|----|
| 1. DESARROLLO..... | 2 |
| 1.1. Categorizar según el tipo de dato y su estructura..... | 2 |
| 1.2. Cassandra..... | 3 |
| 1.3. Dataframe de PySpark..... | 4 |
| 1.4. Análisis descriptivo, matriz de correlación y logaritmo..... | 5 |
| 1.4.1. ANÁLISIS DESCRIPTIVO..... | 5 |
| 1.4.2. ANÁLISIS DE CORRELACIÓN..... | 11 |
| 1.4.3. ALGORITMO: ÁRBOL DE DECISIÓN..... | 13 |
| 2. CONCLUSIONES..... | 15 |



PROYECTO FINAL: BIG DATA

1. DESARROLLO

INFORME CIENTÍFICO

TAREAS:

- 1.1. Categorizar según el tipo de dato y su estructura:

| Nombre del campo | Tipo de dato |
|---------------------------------|--------------------|
| 1. Activity Period | Numérico discreto |
| 2. Operating Airline | Categorico nominal |
| 3. Operating Airline IATA Code | Categorico nominal |
| 4. Published Airline | Categorico nominal |
| 5. Published Airline IATA Code | Categorico nominal |
| 6. GEO Summary | Categorico nominal |
| 7. GEO Region | Categorico nominal |
| 8. Activity Type Code | Categorico nominal |
| 9. Price Category Code | Categorico nominal |
| 10. Terminal | Categorico ordinal |
| 11. Boarding Area | Categorico ordinal |
| 12. Passenger Count | Numérico discreto |
| 13. Adjusted Activity Type Code | Categorico nominal |
| 14. Adjusted Passenger Count | Numérico discreto |
| 15. Year | Numérico discreto |
| 16. Month | Categorico ordinal |

Categorizando los datos de los campos del CSV "Air_Traffic_Passenger_Statistics" encontramos tres tipos de datos:

- Numérico discreto: 4 campos
- Categorico nominal: 9 campos
- Categorico ordinal: 3 campos

1.2. Cassandra:

Se ha insertado un conjunto de registro para mostrar peticiones. A continuación, mostraré el conjunto de registro y dichas peticiones:

| Operating Airline | Year | Activity Period | Activity Type Code | Adjusted Activity Type Code | Adjusted Passenger Count | Boarding Area | GEO Region | GEO Summary | Month |
|-----------------------------|-----------------|---------------------|--------------------|-----------------------------|--------------------------|---------------|---------------|---------------|-----------|
| Operating Airline IATA Code | Passenger Count | Price Category Code | Published Airline | Published Airline IATA Code | Terminal | | | | |
| Air China | 2015 | 201512 | Enplaned | Enplaned | 9341 | G | Asia | International | December |
| CA | 9341 | | Other | Air China | CA | International | | | |
| Air China | 2016 | 201601 | Enplaned | Enplaned | 7125 | G | Asia | International | January |
| CA | 7125 | | Other | Air China | CA | International | | | |
| LAN Peru | 2014 | 201403 | Enplaned | Enplaned | 2886 | A | South America | International | March |
| LP | 2886 | | Other | LAN Peru | LP | International | | | |
| Air Berlin | 2010 | 201010 | Deplaned | Deplaned | 1455 | G | Europe | International | October |
| AB | 1455 | | Other | Air Berlin | AB | International | | | |
| Air Berlin | 2012 | 201210 | Enplaned | Enplaned | 1487 | A | Europe | International | October |
| AB | 1487 | | Other | Air Berlin | AB | International | | | |
| Miami Air International | 2015 | 200509 | Deplaned | Deplaned | 166 | A | US | Domestic | September |
| GL | 166 | | Other | Miami Air International | GL | International | | | |
| Air France | 2016 | 201603 | Enplaned | Enplaned | 10574 | A | Europe | International | March |
| AF | 10574 | | Other | Air France | AF | International | | | |
| Asiana Airlines | 2005 | 200507 | Deplaned | Deplaned | 5041 | A | Asia | International | July |
| OZ | 5041 | | Other | Asiana Airlines | OZ | International | | | |

(8 rows)
cq1sh>

a) Recuperar todos los registros de la aerolínea “Air China”:

```
cqlsh> SELECT * FROM statistics.traffic WHERE "Operating Airline" = 'Air China' ALLOW FILTERING;
```

| Operating Airline | Year | Activity Period | Activity Type Code | Adjusted Activity Type Code | Adjusted Passenger Count | Boarding Area | GEO Region | GEO Summary | Month | Operating Airline IATA Code | Passenger Count | Price Category Code | Published Airline | Published Airline IATA Code | Terminal |
|-------------------|------|-----------------|--------------------|-----------------------------|--------------------------|---------------|------------|---------------|----------|-----------------------------|-----------------|---------------------|-------------------|-----------------------------|----------|
| Air China | 2015 | 201512 | Enplaned | Enplaned | 9341 | G | Asia | International | December | CA | 9341 | | Other | Air China | CA |
| Air China | 2016 | 201601 | Enplaned | Enplaned | 7125 | G | Asia | International | January | CA | 7125 | | Other | Air China | CA |

(2 rows)
cqlsh>

b) Recuperar todos los vuelos de la compañía “Air Berlín” embarcados por la puerta “G”:

```
cqlsh> SELECT * FROM statistics.traffic
... WHERE "Operating Airline" = 'Air Berlin'
... AND "Boarding Area" = 'G' ALLOW FILTERING;
```

| Operating Airline | Year | Activity Period | Activity Type Code | Adjusted Activity Type Code | Adjusted Passenger Count | Boarding Area | GEO Region | GEO Summary | Month | Operating Airline IATA Code | Passenger Count | Price Category Code | Published Airline | Published Airline IATA Code | Terminal |
|-------------------|------|-----------------|--------------------|-----------------------------|--------------------------|---------------|------------|---------------|---------|-----------------------------|-----------------|---------------------|-------------------|-----------------------------|----------|
| Air Berlin | 2010 | 201010 | Deplaned | Deplaned | 1455 | G | Europe | International | October | AB | 1455 | | Other | Air Berlin | AB |

(1 rows)
cqlsh>

1.3. Dataframe de Pyspark:

Cargamos el conjunto de datos “Air_Traffic_Passenger_Statistics” en un dataframe de PySpark. Y a su vez responderemos los siguientes puntos:

- ¿Cuántas compañías diferentes aparecen en el fichero?

```
DESARROLLO, ejercicio 3:
Cargar el conjunto de datos en un dataframe

[13] data_path = '/content/drive/MyDrive/TokioSchool/data_curso/'
df = spark.read.options(header=True, inferSchema=True).csv(data_path + 'Air_Traffic_Passenger_Statistics.csv')

¿Cuántas compañías diferentes aparecen en el fichero?

[14] df.select('Operating Airline').distinct().count()

77
```

En el conjunto de datos nos encontramos con 77 compañías diferentes.

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

```
¿Cuántos pasajeros tienen de media los vuelos de cada compañía?

[15] df_compañia = df.groupBy('Operating Airline').agg(F.mean('Passenger Count').alias('mean_passenger_count'))
df_compañia.show()

+-----+-----+
| Operating Airline | mean_passenger_count |
+-----+-----+
| Icelandair       | 2799.7                |
| Ameriflight      | 5.0                   |
| Cathay Pacific   | 17121.325581395347    |
| Aeromexico       | 5463.822222222222     |
| Etihad Airways   | 6476.088235294118     |
| Philippine Airlines | 10248.635658914729    |
| United Airlines  | 48915.46758232126     |
| Turkish Airlines | 8162.416666666667     |
| Swiss International | 6061.640287769784    |
| Independence Air | 6391.3                |
| Miami Air Interna... | 107.375              |
| Air France       | 11589.077519379845    |
| Japan Airlines   | 6470.332046332046    |
| Midwest Airlines | 3883.0                |
| Atlas Air, Inc    | 34.0                  |
| JetBlue Airways  | 35261.13963963964     |
| China Eastern    | 5498.402777777777     |
| Mexicana Airlines | 7993.806451612903    |
| Air Canada       | 18251.560109289618   |
+-----+-----+

top5_compañias = df_compañia.orderBy(F.desc('mean_passenger_count')).limit(5)
top5_compañias.show()

low5_compañias = df_compañia.orderBy(F.asc('mean_passenger_count')).limit(5)
low5_compañias.show()

+-----+-----+
| Operating Airline | mean_passenger_count |
+-----+-----+
| American Airlines | 127164.38970588235    |
| Southwest Airlines | 81188.15857605178    |
| Virgin America    | 74405.35359116022    |
| United Airlines   | 72732.05829596413    |
| Delta Air Lines   | 68498.49740932643    |
+-----+-----+

+-----+-----+
| Operating Airline | mean_passenger_count |
+-----+-----+
| Evergreen Interna... | 2.0                   |
| Ameriflight         | 5.0                   |
| Boeing Company      | 18.0                  |
| Atlas Air, Inc       | 34.0                  |
| Xtra Airways        | 73.0                  |
+-----+-----+
```

Analizando los resultados obtenidos en la imagen podemos ver como las cinco compañías con más pasajeros por vuelo de media son: American Airlines, Southwest Airlines, Virgin America, United Airlines, Delta Air Lines.

Mientras que las cinco compañías con menos pasajeros por vuelo de media son: Evergreen International, Ameriflight, Boeing Company, Atlas Air, Inc, Xtra Airways.

- Eliminaremos los registros duplicados por el campo 'GEO Región', manteniendo únicamente aquel con mayor número de pasajeros.

```
Eliminaremos los registros duplicados por el campo "GEO Región", manteniendo únicamente aquel con mayor número de pasajeros.
```

```
[17] max_passengers = df.groupby("GEO Región").agg(F.max("Passenger Count").alias("max_passenger_count"))
max_passengers.show()
```

| GEO Región | max_passenger_count |
|---------------------|---------------------|
| Europe | 48136 |
| Central America | 8970 |
| US | 659837 |
| South America | 3685 |
| Mexico | 29286 |
| Middle East | 14769 |
| Canada | 39798 |
| Australia / Oceania | 12973 |
| Asia | 86398 |

Las regiones donde han registrado mayor número de pasajeros son:

- US, Asia y Europe.

En cambio, las regiones con menor número de pasajeros fueron:

- South America, Central America y Australia/Oceania.

La causa de estos resultados puede deberse al poder adquisitivo de los habitantes en dichas regiones. Ya que donde menos pasajeros hay es en el sur y centro de América donde la situación económica se encuentra en peores condiciones que las regiones de US y Europa.

1.4. Análisis descriptivo, matriz de correlación y logaritmo:

1.4.1. ANÁLISIS DESCRIPTIVO:

El análisis descriptivo será realizado sobre los campos pasajeros, región, categoría y mes. Ya que considero que el resto de los campos no seleccionados para este análisis carecen de sentido y no ayudarían a identificar patrones en el modelo. Campos como 'Activity Period' o 'Terminal' son de carácter identificativo y organizativo en las distintas compañías.

Calcularemos la media y la desviación estándar de los campos pasajeros según la región, categoría y mes:

- MEDIA: A continuación, podéis ver el código con PySpark para calcular la media de pasajeros según la región, categoría y mes.

```
• MEDIA DE CADA ELEMENTO

Media del número de pasajeros según la región, categoría de precio y el mes

[15] df_region = df.groupBy('GEO Region').agg (F.mean('Passenger Count').alias('mean_passenger_count')).show()

df_price = df.groupBy('Price Category Code').agg (F.mean('Passenger Count').alias('mean_passenger_count')).show()

df_month = df.groupBy('Month').agg (F.mean('Passenger Count').alias('mean_passenger_count')).show()
```

| GEO Region | mean_passenger_count |
|---------------------|----------------------|
| Europe | 12755.652465294399 |
| Central America | 4946.715328467153 |
| US | 58330.34345351044 |
| South America | 2786.0111111111111 |
| Mexico | 7173.62062780269 |
| Middle East | 8658.61214953271 |
| Canada | 9777.9682651622 |
| Australia / Oceania | 6417.016282225238 |
| Asia | 13435.00458295142 |

| Price Category Code | mean_passenger_count |
|---------------------|----------------------|
| Other | 27787.546114464734 |
| Low Fare | 39144.21041666667 |

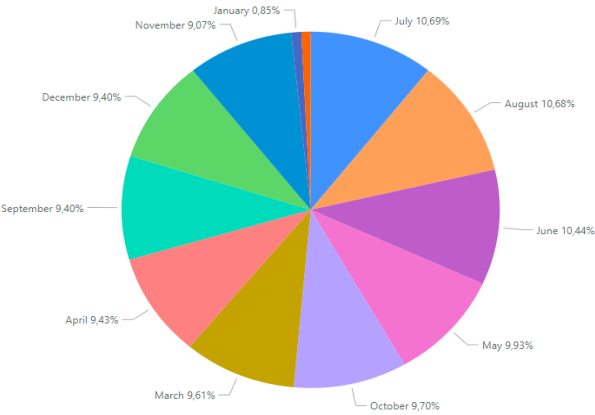
| Month | mean_passenger_count |
|-----------|----------------------|
| July | 32657.033768227167 |
| November | 27702.751385589865 |
| February | 24258.40796812749 |
| January | 26094.37066246057 |
| March | 29359.807661612133 |
| October | 29645.857915057913 |
| May | 30340.417235494882 |
| August | 32636.380916030535 |
| April | 28813.577893820715 |
| June | 31886.650887573964 |
| December | 28723.198570293884 |
| September | 28725.348496530456 |

A continuación, para entender con más facilidad estos datos y comprender la información, permitiendo visualizar patrones de forma intuitiva podéis analizar los siguientes gráficos circulares:



Category
Month

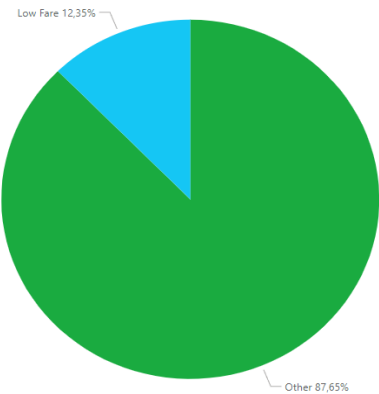
Average passengers per Month



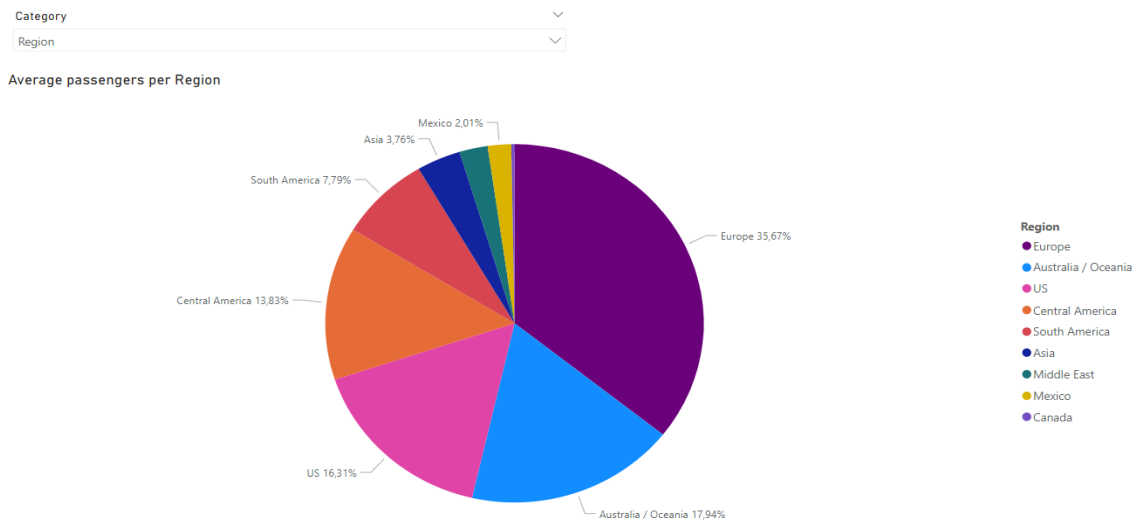
- Month
- July
 - August
 - June
 - May
 - October
 - March
 - April
 - September
 - December
 - November
 - January
 - February

Category
Price

Average passengers per Price



- Price
- Other
 - Low Fare



Si analizamos estos tres gráficos, donde están representados la media según 3 variables, podemos sacar las siguientes conclusiones:

- **Mes:** junio julio y agosto, es decir, los meses de verano, son los meses donde más pasajeros viajan de media. Siendo el pico más alto concretamente en el mes de agosto. Estos meses coinciden generalmente con la estación del año donde trabajadores y estudiantes disfrutan las vacaciones, además de ser un clima caluroso. Esto incentiva a la gente a viajar más.
 - **Región:** Europa es la región donde más pasajeros viajan de media, seguida de Australia/Oceanía y US. Esto puede darse por el alto nivel económico y de adquisición de los habitantes de dichos continentes. Además, las grandes dimensiones territoriales de US generan que los habitantes de esta región dependan de las compañías de aerolíneas para desplazarse de estado a estado.
Por otro lado, Europa es una región que, a excepción del sur de Europa, tiene un clima muy frío y con muchas precipitaciones. Por esta razón, en los meses de verano aprovechan para viajar a zonas más cálidas y con costa.
 - **Precio:** la categoría de precio 'other' de media es mayor que 'low fare', este es un dato que puede impactar ya que lo más esperado hubiera sido que 'low fare' fuera mayor. Pero podemos intuir que los pasajeros prefieren viajar en mejores condiciones independientemente del precio. Además, como analizamos en el anterior punto, las regiones con más pasajeros se caracterizaban por el alto nivel adquisitivo, dato que justifica perfectamente que 'other' sea la categoría de precio más escogida por los pasajeros.
- DESVIACIÓN ESTÁNDAR: A continuación, podéis ver el código con PySpark para calcular la desviación estándar de pasajeros según la región, categoría y mes.

```
• DESVIACIÓN TÍPICA DE CADA ELEMENTO

Desviación estandar del número de pasajeros según la región, categoría de precio y el mes

[16] df_region = df.groupby('GEO Region').agg (F.std('Passenger Count').alias('std_passenger_count')).show()
df_price = df.groupby('Price Category Code').agg (F.std('Passenger Count').alias('std_passenger_count')).show()
df_month = df.groupby('Month').agg (F.std('Passenger Count').alias('std_passenger_count')).show()

+-----+-----+
| GEO Region | std_passenger_count |
+-----+-----+
| Europe | 8634.076411562175 |
| Central America | 1220.8403125914656 |
| US | 84951.31664013123 |
| South America | 396.7586506195526 |
| Mexico | 5336.223001980255 |
| Middle East | 2732.7195183986 |
| Canada | 7833.110588404248 |
| Australia / Oceania | 2799.0406500183883 |
| Asia | 16188.148775860833 |
+-----+-----+

+-----+-----+
| Price Category Code | std_passenger_count |
+-----+-----+
| Other | 59368.099858611706 |
| Low Fare | 49486.13871114107 |
+-----+-----+

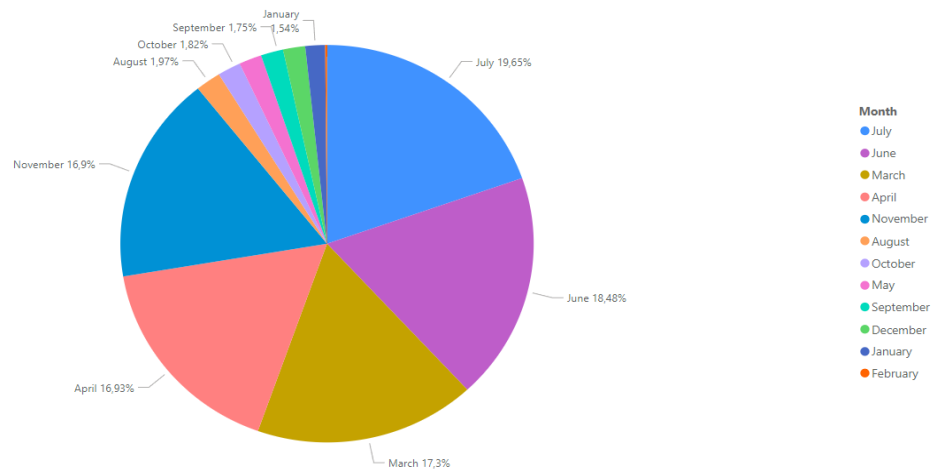
+-----+-----+
| Month | std_passenger_count |
+-----+-----+
| July | 65309.479556884944 |
| November | 56175.549821247325 |
| February | 47569.0069987234 |
| January | 51294.88007998184 |
| March | 57519.653306710045 |
| October | 60490.12293144308 |
| May | 59000.17436267259 |
| August | 65568.50655076625 |
| April | 56293.933065065205 |
| June | 61424.496549347394 |
| December | 57566.02013447425 |
| September | 58236.29488536921 |
+-----+-----+
```

A continuación, para entender con más facilidad estos datos y comprender la información, permitiendo visualizar patrones de forma intuitiva podéis analizar los siguientes gráficos circulares:



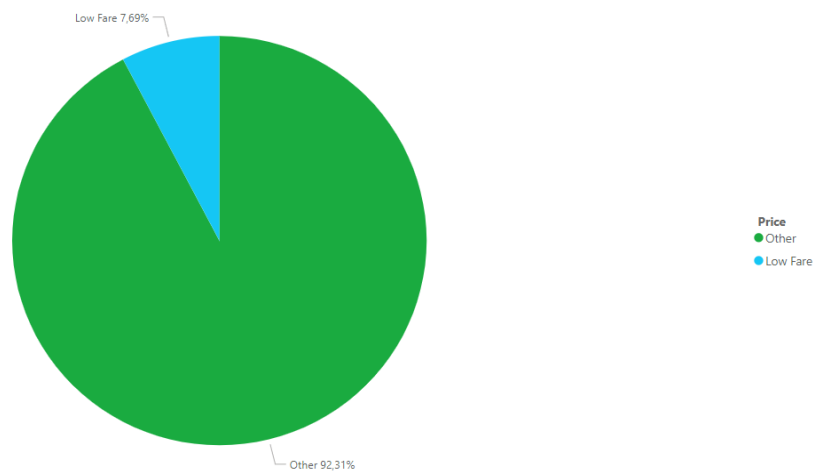
Category

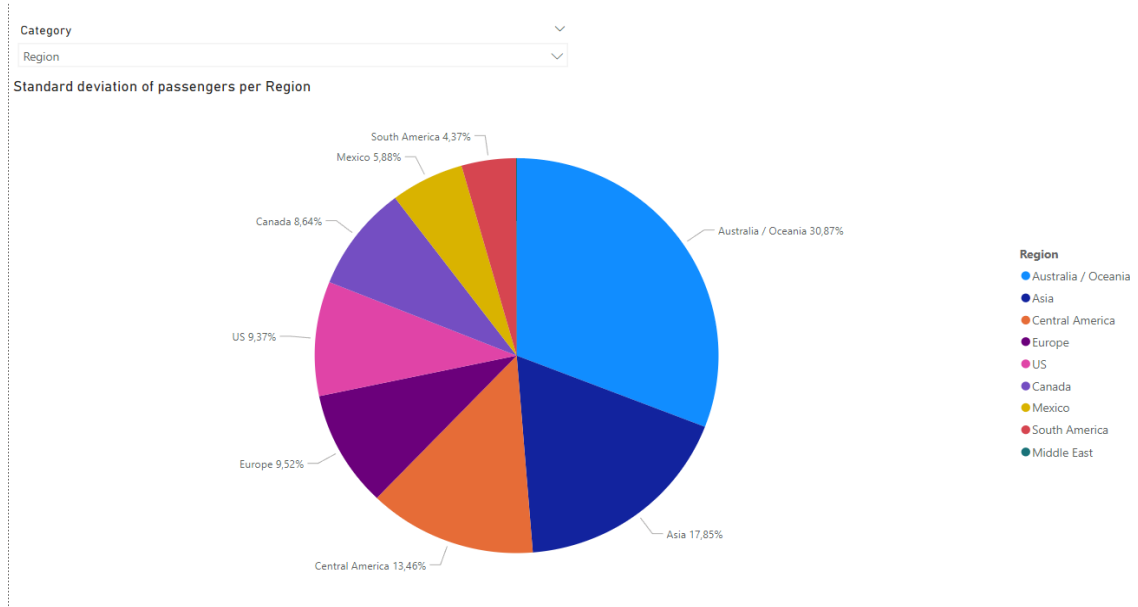
Standard deviation of passengers per Month



Category

Standard deviation of passengers per Price





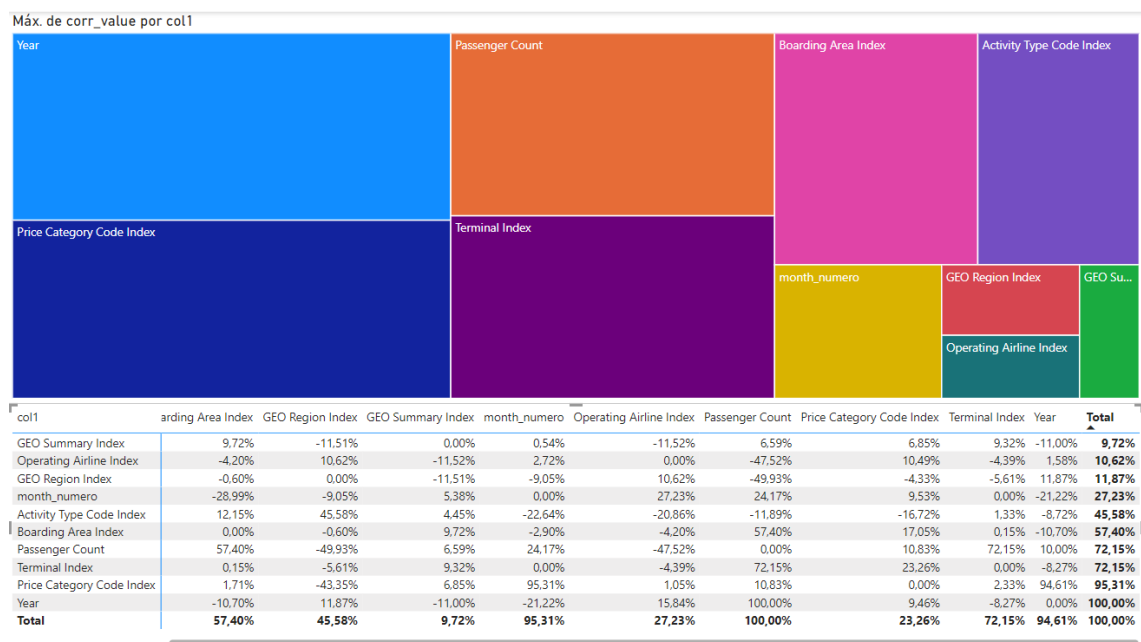
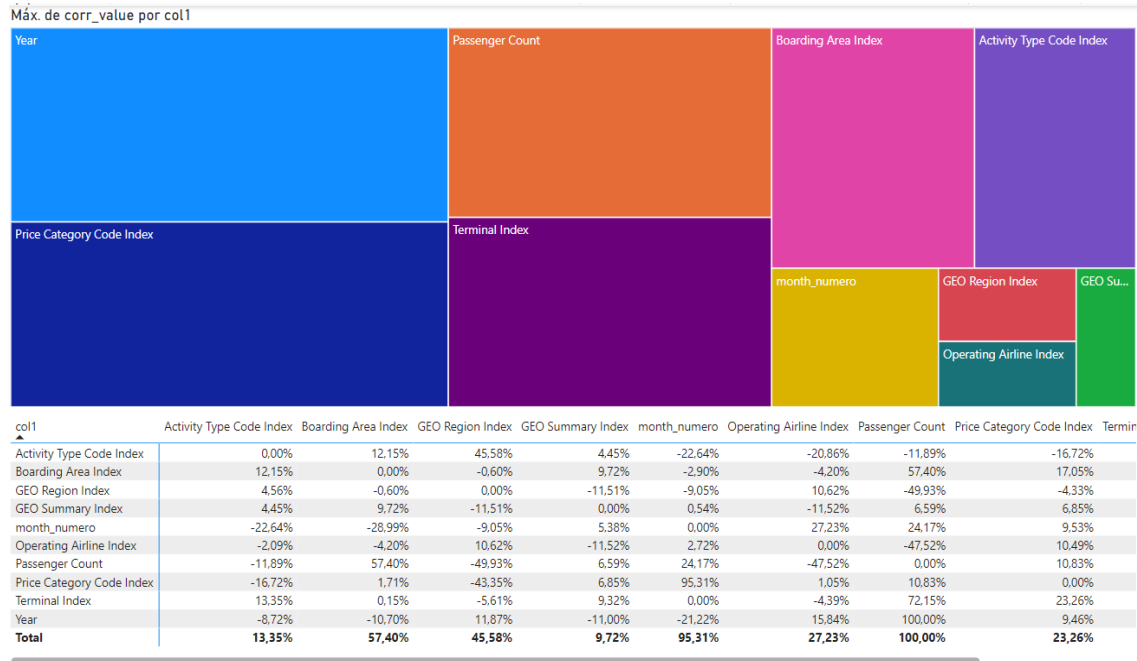
Si analizamos estos tres gráficos, donde están representados la desviación estándar según 3 variables, podemos sacar las siguientes conclusiones:

- **Mes:** Julio, junio y marzo son los meses que presentan mayor desviación estándar, es decir, una mayor variabilidad o dispersión de los datos. En cambio, enero y febrero tienen una desviación estándar baja, esto significa que los datos son más uniformes.
- **Región:** Australia/Oceanía es la región con mayor desviación estándar, es decir, hay grandes diferencias en el número de pasajeros. Mientras que 'South America' es la región en la que el número de pasajeros está muy cerca de la media.
- **Precio:** la categoría de precio con mayor variabilidad es 'other', algo normal ya que este ocupa aproximadamente un 87% de los pasajeros. Por esta razón, muestra una gran dispersión de los datos.

1.4.2. ANÁLISIS DE CORRELACIÓN:

Calcularemos una matriz de correlación que represente la manera en la que están relacionadas los diferentes campos del conjunto de datos.

A continuación, para entender con más facilidad estos datos y comprender la información, permitiendo visualizar patrones de forma intuitiva podéis analizar el siguiente mapa de árbol o treemap:



Si analizamos este treemap, donde está representada la matriz de correlación, podemos sacar las siguientes conclusiones:

- 1.5. Correlación positiva:** el campo con un nivel de coeficiente de correlación mayor es 'Passenger Count'. Este campo presenta fuerte correlación con 'Year', 'Terminal' y 'Boarding Area', respectivamente. Esto significa que estos campos tienen a aumentar o disminuir juntos.

Seguidamente, otro campo que presenta una gran correlación es 'Price Category Code'. Este tiene una gran relación con los campos 'Month' y 'Year', respectivamente. Aquí podemos identificar un patrón entre estas variables.

1.6. Correlación negativa: el campo con menor nivel de coeficiente de correlación es 'Passenger Count'. Este tiene muy poca relación con 'GEO Region' y 'Operating Airline', por tanto, cuando un campo aumenta los otros tienden a disminuir.

Otro campo con correlación negativa es 'Month'. Tiene poca relación con 'Activity Type Code' y 'Year'. La poca relación entre 'Year' y 'Month' se debe a una razón natural, ya que los meses llevan un ciclo fijo, donde siempre serán los mismo independientemente del paso del tiempo de los años.

1.4.3. ALGORITMO: ÁRBOL DE DECISIÓN

Entre los distintos algoritmos dados a lo largo de esta formación, he elegido el modelo de regresión basado en un árbol de decisión.

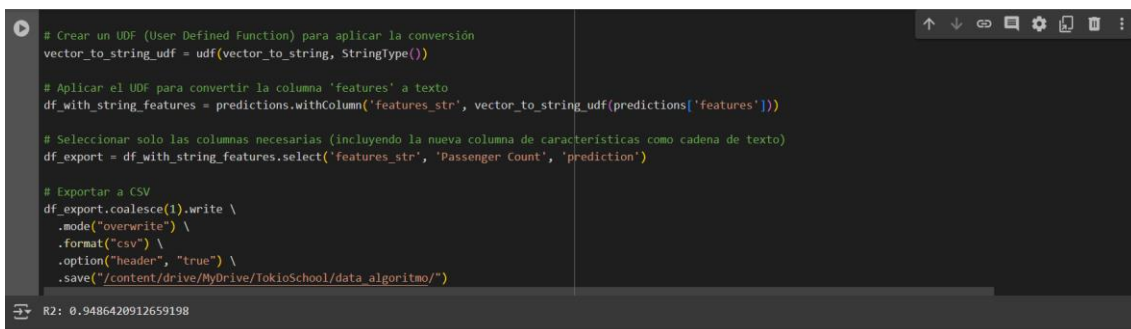
Mi decisión se ha visto afectada por la complejidad que presentaba el modelo de datos y el árbol de decisión:

- Simplicidad e interpretabilidad
- Capacidad para manejar datos mixtos
- Flexibilidad y poca preparación de datos
- Facilidad para identificar características importantes

Concretamente he elegido el GBRegressor (abreviatura de Gradient Boosting Tree Regressor) es un modelo de aprendizaje supervisado utilizado para problemas de regresión. Este modelo se basa en la técnica de Gradient Boosting, que combina múltiples árboles de decisión (árboles débiles) en secuencia para construir un modelo fuerte que mejore de forma iterativa.

De esta manera he conseguido que el algoritmo presente un coeficiente de correlación muy elevado y así la predicción pueda ser más exacta.

A continuación, podéis ver el coeficiente de correlación del algoritmo (R^2) que comentaba, el cual tiene un valor de 0,9486:



```
# Crear un UDF (User Defined Function) para aplicar la conversión
vector_to_string_udf = udf(vector_to_string, StringType())

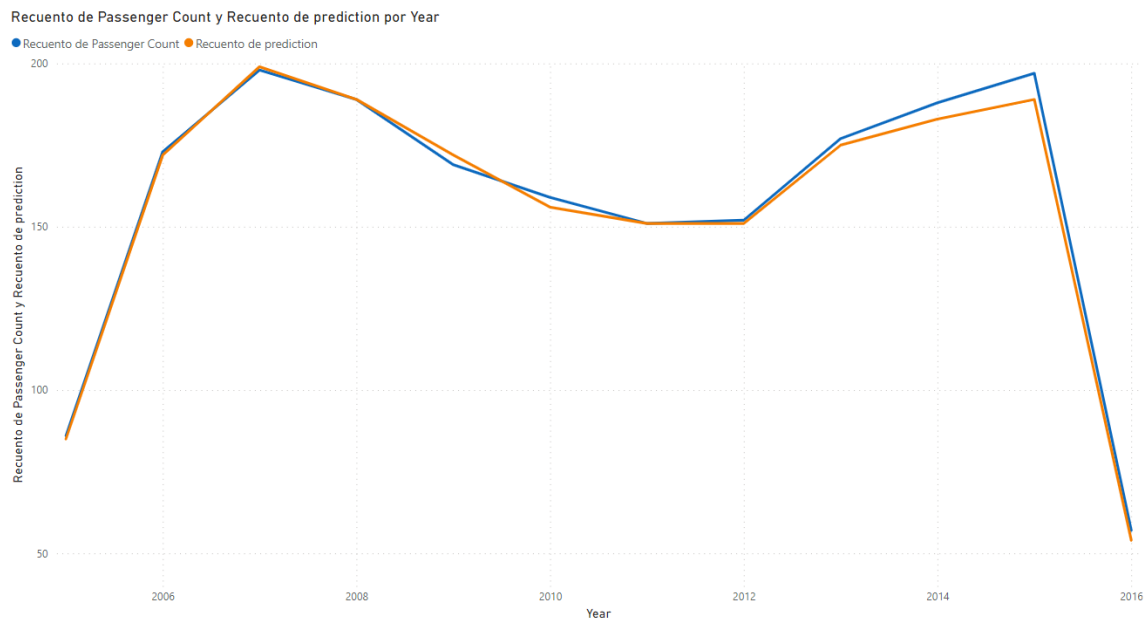
# Aplicar el UDF para convertir la columna 'features' a texto
df_with_string_features = predictions.withColumn("features_str", vector_to_string_udf(predictions["features"]))

# Seleccionar solo las columnas necesarias (incluyendo la nueva columna de características como cadena de texto)
df_export = df_with_string_features.select("features_str", "Passenger Count", "prediction")

# Exportar a CSV
df_export.coalesce(1).write \
    .mode("overwrite") \
    .format("csv") \
    .option("header", "true") \
    .save("/content/drive/MyDrive/TokioSchool/data_algoritmo/")
```

R2: 0.9486420912659198

Para entender con más facilidad estos datos y comprender la información, permitiendo visualizar patrones de forma intuitiva podéis analizar el siguiente gráfico de líneas:



La línea azul representa los datos reales del conjunto de datos y la línea naranja representa la predicción que realiza el algoritmo, estas están muy cercanas entre sí. Esto indica que el modelo ha logrado una predicción precisa, con un error mínimo entre lo pronosticado y los valores observados. La consistencia en su alineación con los puntos reales refleja la efectividad del algoritmo, lo que sugiere que ha capturado correctamente las tendencias y patrones del conjunto de datos.



2. CONCLUSIONES

Este proyecto ha tenido como objetivo analizar un conjunto de datos sobre el tráfico en el distintos aeropuertos y compañías. El análisis ha sido posible gracias al desarrollo de los siguientes aspectos: análisis descriptivo, análisis de correlación y un algoritmo para predecir el número de pasajeros en el futuro.

A través del análisis realizado, se ha logrado demostrar lo siguiente:

- En los meses de verano aumentan los pasajeros de media que viajan en las regiones de Europa y US. Debido principalmente a la buena situación económica de estas regiones y en el caso de US de las grandes dimensiones territoriales.

Además, los pasajeros generalmente prefieren viajar con tarifas caras ante la preferencia de disfrutar de buenas condiciones.

- En las variables donde hay más porcentaje de participación como es los meses de verano y las tarifas más caras podemos ver mayor variabilidad en los resultados.

En cambio, en meses como enero y febrero encontramos mayor uniformidad en los datos, ya que son meses donde viajan menos pasajeros de media.

- Los campos del conjunto de datos con una correlación positivo han sido 'Passenger Count' junto con 'Year', 'Terminal' y 'Boarding Area'.

Mientras que los campos con una correlación negativa han sido 'Month' con 'Year', ya que los meses llevan un ciclo fijo, donde siempre serán los mismo independientemente del paso del tiempo de los años. Y 'Passenger Count' con 'GEO Region' y 'Operating Airline', presentando una relación indirectamente proporcional.

- El modelo de regresión basado en un árbol de decisión GBRegressor indica que el modelo ha logrado realizar una predicción exacta, mostrando una mínima diferencia entre los valores estimados y los valores reales observados. La constante alineación con los datos reales demuestra la efectividad del algoritmo, indicando que ha logrado captar correctamente las tendencias y patrones presentes en el conjunto de datos.

Los hallazgos de este proyecto son significativos debido a la posibilidad de predecir patrones y comportamientos del mercado aeronáutico, pudiéndose utilizar por sectores que tengan influencia de forma indirecta como son: hostelería y turismo; comercio y logística; industria tecnológica y de innovación; energético y de combustible y finalmente el sector financiero.