

Carrera:Data Analytics

Modulo 5

Proyecto Integrador

Lineas Aereas Rusas

Cohorte: DAFT 06

Fecha de entrega: Octubre 2024

**Autores:**

* *Bosch Clara- Analista de Datos*
* *Osorio Miguel - Analista de Datos*
* *Soto Yanina - Analista de Datos*
* *Ortega Lopez Samuel - Data Engineer*

**Institución:**

LA Línea Aérea Rusa de vuelos comerciales al estar en un ambiente cada vez más competitivo, necesita ordenar su base de datos para analizar e identificar áreas de mejora que incrementen sus ingresos operativos e insights de valor para posibles futuros acontecimientos importantes que puedan identificar con anterioridad para así poder tener conocimiento y dar con una posible solución.

**Introducción:**

La problemática a analizar de la empresa es la disminución en la venta de tickets y una baja en la tasa de ocupación de vuelos

Se debe identificar rutas y aviones más rentables, comportamiento de compra de clientes y Prever capacidad ociosa para mejorar los ingresos

**Desarrollo del proyecto:**

Comenzamos con la organización:

* Realizamos un email y una cuenta de github grupal para poder subir todo a un repositorio en común, creamos las carpetas a utilizar con sus archivos pertinentes y se comenzo a elaborar el readme , de esta manera comenzamos armando el repositorio
* Se comenzo a realizar una visualizacion interactiva del proyecto para su presentacion ante los clientes
* Utilizamos la herramienta Jira para el seguimiento y la gestión de las tareas.

Luego ,preparamos los datos:

* Se descargo el dataset travel.sqlite
* Para comenzar a analizarlo se utilizo la libreria de python sqlite3 para leer el archivo
* Se importaron todas las demas librerias necesarias para comenzar con el análisis
* Comprobamos los datos obtenidos y verificamos la existencia de datos nulos,duplicados y outliers
* Creamos la base de datos en SQL
* Creamos el servidor en la nube
* Realizamos las tablas en SQL
* Creamos el modelo relacional
* Elaboramos los procedimientos para la carga futura de datos (Automatización de la actualización de datos)
* Validamos el acceso a la base de datos desde los diferentes dispositivos del grupo y la subimos a la nube de AWS
* Realizamos las visualizaciónes de la información mediante graficos en jupiter notebook
* Realizamos el dashboard en POWER BI

**Conclusiónes :**

Comenzando con el analisis observamos que las tablas contienen informacion de un solo mes completo : de Julio 2017, tambien de fines de Junio 2017 y de principios de Agosto de 2017

En primer lugar se pudo observar que ninguna tabla presenta nulos que haya que rellenar ,ni filas duplicadas,aunque en la tabla 'bookings' encontramos un %2 de outliers,pero consideramos que son eventos puntuales de la operatoria normal de la empresa, pero de ser necesario, se identificaran mas adelante en el proceso ETL agregando una nueva columna para realizar su filtrado del dashbord en POWER BI

Las tablas boarding\_passes,tickets\_flights y seats no tienen identificador unico, se les creara los ID a cada uno para poder relacionar las tablas

EDA - Clientes:

Frecuencia : No se puede analizar frecuencia de viajes por cliente porque no tenemos base con datos de pasajero; aquí el passenger\_id es un identificador único. Por cada ticket es un pasajero.

Cambiamos, entonces a análisis del comportamiento de compra.

Comportamiento de Compra

Estacionalidad :tenemos una base muy acotada en fechas, solo contiene información del mes de Julio 2017 completa.

Se analizaron las ventas por día de la semana, pero no se encontró ningún patrón.

Compras individuales o grupales:La gran mayoria compra 1/2 tickets por booking, no se compran en grupos. Son compras individuales.

Anticipación de compra: los clientes hacen su compra con una anticipación promedio de 20 dias.

Tasa de conversión booking a boarding : se registran un 35% de pasajeros que compraron y no abordaron.

Conclusiones EDA Aeropuertos:

Podemos concluir que los aeropuertos DME, SVO y LED son los aeropuertos donde hay mayor presencia de clientes de la aerolínea y estos le generan la mayor cantidad de ingresos.

Se le puede recomendar a la aerolínea que en estos aeropuertos mejore sus salas de espera y priorice la atención tanto para Economy como para Business, como una estrategia para mejorar la tasa de satisfacción de los clientes y generar algún tipo de fidelización de clientes.

Por otro lado, para los aeropuertos que presentan menor cantidad de clientes se pueden generar estrategias de marketing como la instalación de vallas publicitarias, entrega de flyers, generar descuentos para impulsar la marca de la aerolínea con el fin de incrementar la captación de clientes.

Conclusiones EDA Vuelos:

Inicialmente contamos con 33121 en donde la duración de la mayoría de vuelos de la aerolínea no superan las 4 horas. Y la mayor concentración de vuelos está entre una (1) y 2:45:00 min.

El vuelo más corto dura 25:00 min y el más largo dura 8:50:00 min. La duración promedio de los vuelos es de 2:08:00 min. Cabe resaltar que la aerolínea también cuenta con vuelos entre las seis (6) y ocho (8) horas pero son muy pocos los vuelos con esta duración de tiempo, considerándolos de esta manera como datos atípicos.

observamos que la mayoría de vuelos presenta retrasos en la salida de los vuelos, sin embargo la mayoría de estos retrasos son de 5 a 10 minutos. Podemos interpretar que esto se debe a problemas con el control aéreo y tráfico aéreo normales en aeropuertos de gran envergadura.

También podemos observar que se presentan algunos retrasos de más de 150 minutos, no es una cantidad relevante pero vale la pena prestar atención a estos retrasos ya que pueden generar costos adicionales para la aerolínea, debido a que los pasajeros pueden solicitar compensaciones adicionales.

Conclusiones EDA tickets:

Por Valor : el valor promedio de un ticket fue de 20.000, el 75% está por debajo de los 23000.

En Business el promedio es 51,143.42, en Comfort es 32,740.55y en Economy 15,959.81.

Se analizaron los outliers por mes y por día del mes, sin encontrarse nada para destacar, los outliers se encuentran distribuidos de forma bastante uniforme a lo largo del mes.

Por clase de asiento : el 70% de los ingresos viene de los tickets de Economy, que tiene el 88% de ventas de tickets de la compañía. Sin embargo, la tasa de ocupación en Economy es baja.

Combinaciones: la gran mayoría se venden con combinaciones de tickets, entre 2 y 4 vuelos por ticket.

Conclusiones EDA- Aeronaves:

La aerolínea cuenta con 9 tipos de aeronaves, con diferentes capacidades. La más pequeña tiene disponibilidad de 12 pasajeros (Cessna 208 Caravan, ocupada para recorridos cortos) y la más grande tiene disponibilidad de 402 pasajeros (Boeing 777 - 300, ocupada para recorridos largos).

Podemos observar que las aeronaves más utilizadas son la CN1 (Cessna 208 Caravan), CR2 (Bombardier CRJ-200) y SU9 (Sukhoi SuperJet-100). Como característica general de estas aeronaves es que son las tres con menor rango de recorrido de toda la flota (desde 1200 hasta 3000 millas), es decir, se utilizan para recorridos cortos.

observamos que la tasa de ocupación promedio de las aeronaves está alrededor del 50%. Lo anterior quiere decir que se está desperdiciando la mitad de la capacidad de la aeronave en su utilización. Se le recomienda a la aerolínea generar estrategias de mercadeo para vender más tiquetes de vuelo, o utilizar aeronaves con menor capacidad en sus recorridos que generen menos costos operativos y así optimizar la utilización de las aeronaves.

Las aeronaves que más generan ingresos a la aerolínea, son SU9 (Sukhoi SuperJet-100), 763 (Boeing 767-300), 773 (Boeing 777-300), 319 (Airbus A319-300). Estos últimos tres modelos se caracterizan por ser aeronaves con capacidad entre 116 y 402 pasajeros, y realizar recorridos más largos (entre 2 y 6 horas de vuelo).

Cabe resaltar que los modelos CR2 y CN1 a pesar de ser las aeronaves más utilizadas por la aerolínea no generan tantos ingresos para la compañía.

Observamos que la tarifa promedio del tiquete de las aeronaves de acuerdo con su condición (Business, Economy o Comfort). Se puede resaltar que las tres aeronaves mencionadas en la gráfica anterior ( 319, 763, 773) que generan la mayor cantidad de ingresos para la aerolínea, también tienen la tarifa promedio más alta por tiquete con respecto a las demás aeronaves.

**Tecnologias utilizadas:**

Utilizamos las siguientes herramientas ya que nos permiten una sencilla integracion y escalabilidad

* Limpieza de datos = Pandas y Numpy
* Modelado = SQL SERVER
* Visualizacion = Power BI y Plotlib