

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

**Conclusión :**

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

**Recomendaciones: (en proceso)**

**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