**EDA-EXPLORACION US\_AIRLINES AGATADATA – TEST**

**Primer Acercamiento de los datos**

Se encuentras 3 archivos Data\_transaccional1, Data\_transaccional2, Data\_Transaccional3 y 1 archivo de definición de Origin, de los cuales se les realiza el primer acercamiento de los datos que me proporcionan de los vuelos de US Airlines.

Se realiza el primer acercamiento que es conocer los campos y sus tipos de datos y del negocio a cual hace parte (Periodo de tiempo, los vuelos, retrasos, cancelaciones y plan de vuelos)

Anexo Excel



**Segundo Acercamiento a los datos**

Se realiza un profilling de los datos que me va a permitir conocer la calidad de los datos proporcionados

Para este acercamiento se utilizó las librerías de panda de Python



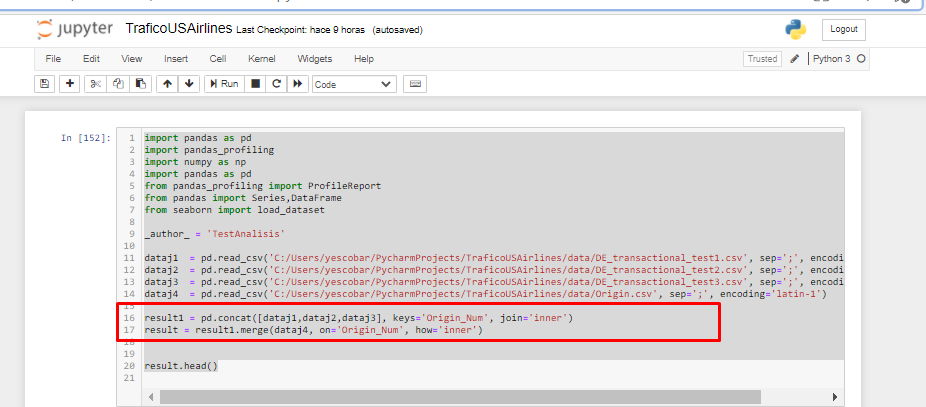
Anexo HTML

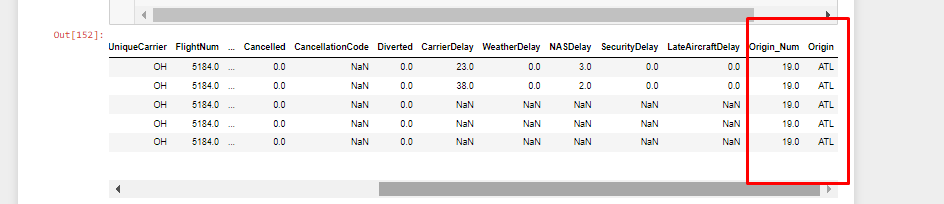




**Análisis de los Datos**

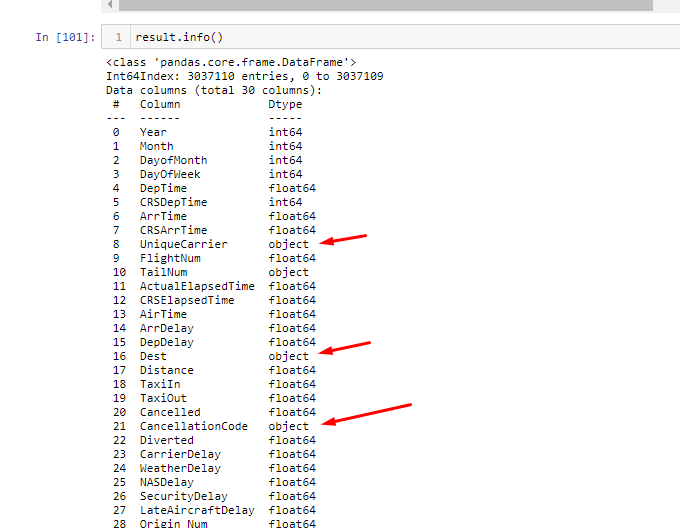
**Relacionar los datos**





**Limpieza de los datos**

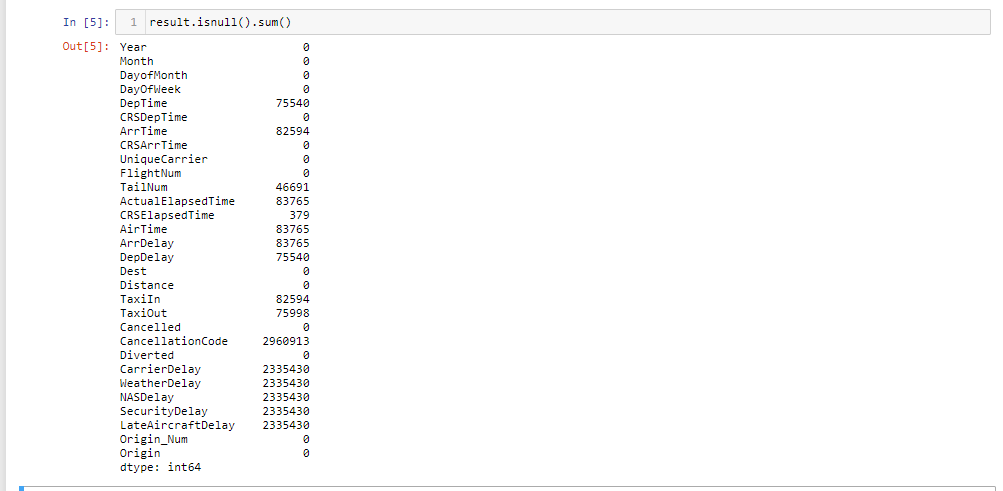
Transformación

Se consulta los tipos de datos de c/una de las columnas y se identifica que hay algunos campos de tipo Object 

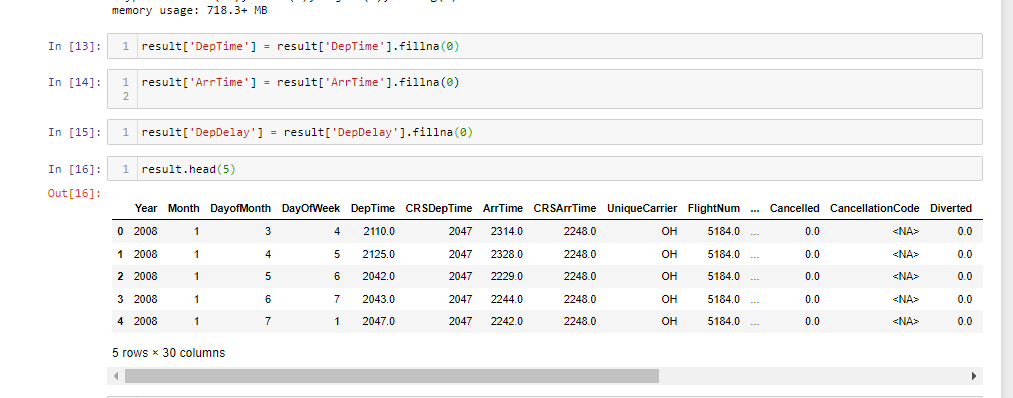
Lo cual se realiza el cambio del tipo de dato para esos campos en específico



Obtenemos la cantidad de registros nan

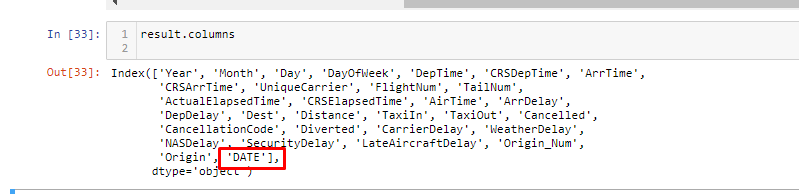


Realizo las transformaciones a los campos con los cuales voy a llevar para el análisis

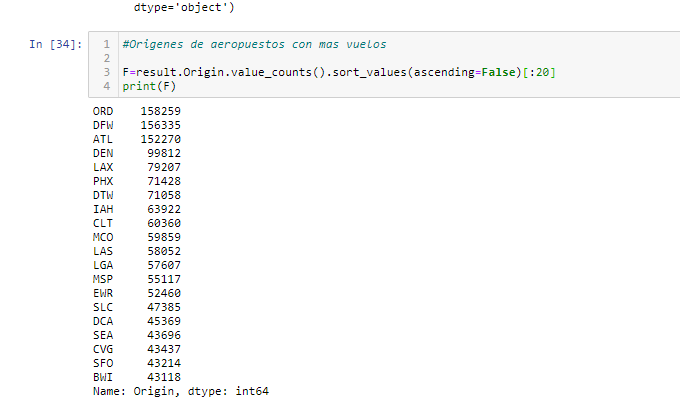


Formatear los campos hh:mm y agregar un nuevo campo con la fecha formateada

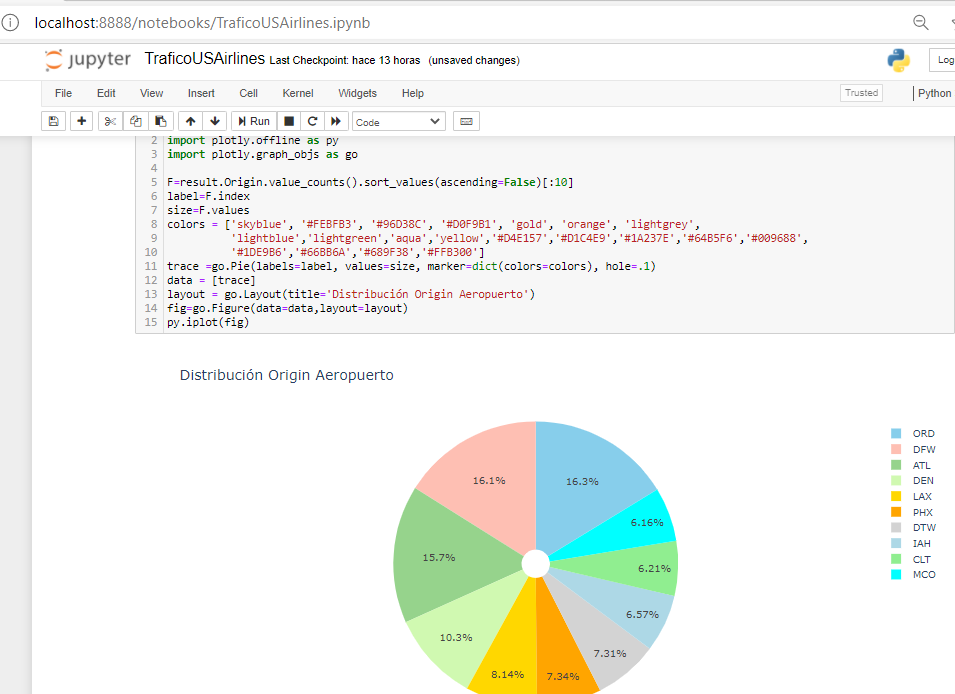




Explotación de los datos



**#Distribución de aeropuertos vs los vuelos**



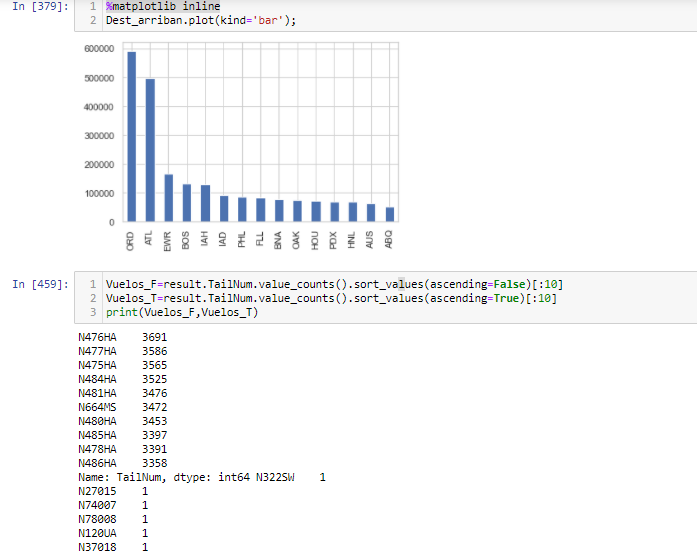
Se pude ver que de la variable Origin de aeropuerto ORD tiene los mayores porcentajes de vuelos.

**#Destino con más vuelos arribar**



Para la Variable Dest se realiza un TOP 15 de cual es el destino de donde mas llegan los vuelos y se evidencia que aeropuerto ORD tiene el mayor conteo

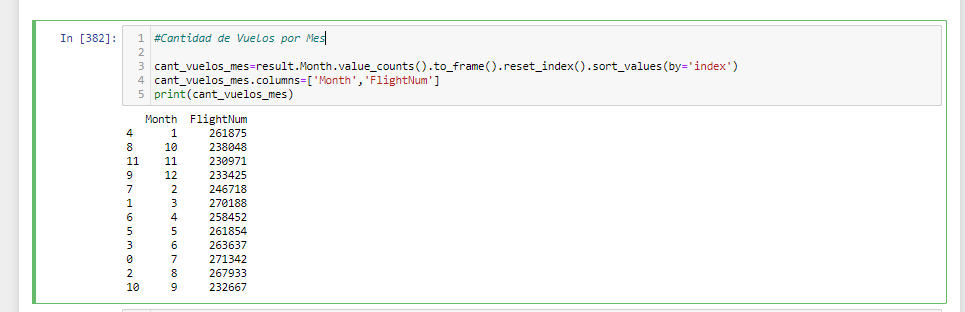
se representa mediante una grafica de barras para que se pueda apreciar



De acuerdo a lo anterior podemos representar por medio de una comparación de líneas los aviones que más tienen vuelos



#CantidadVuelosxMes

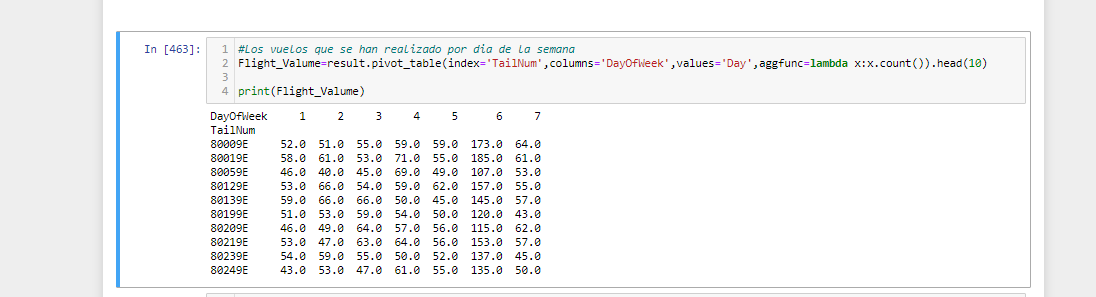


Podemos observar el Mes 7 se realiza la mayor cantidad de vuelos

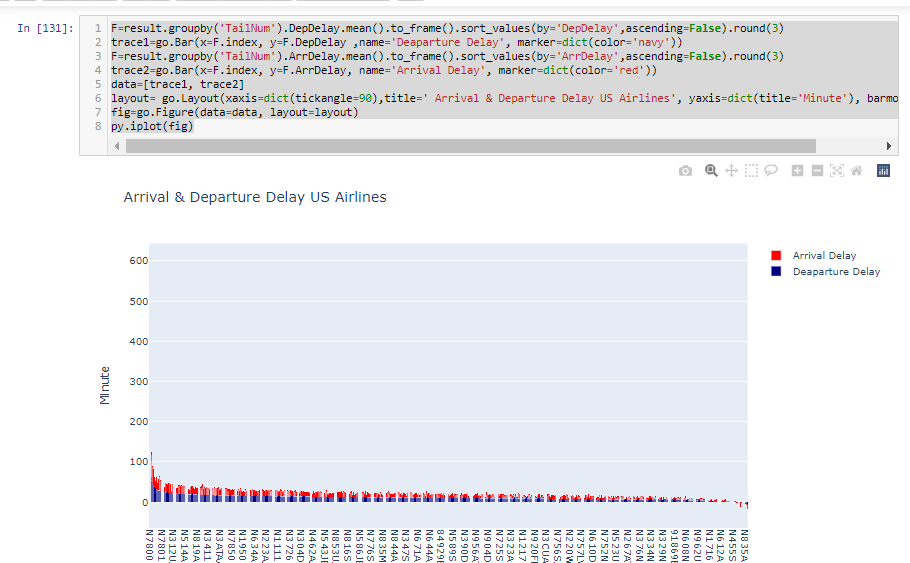
Se puede apreciar mejor en este grafico de barras

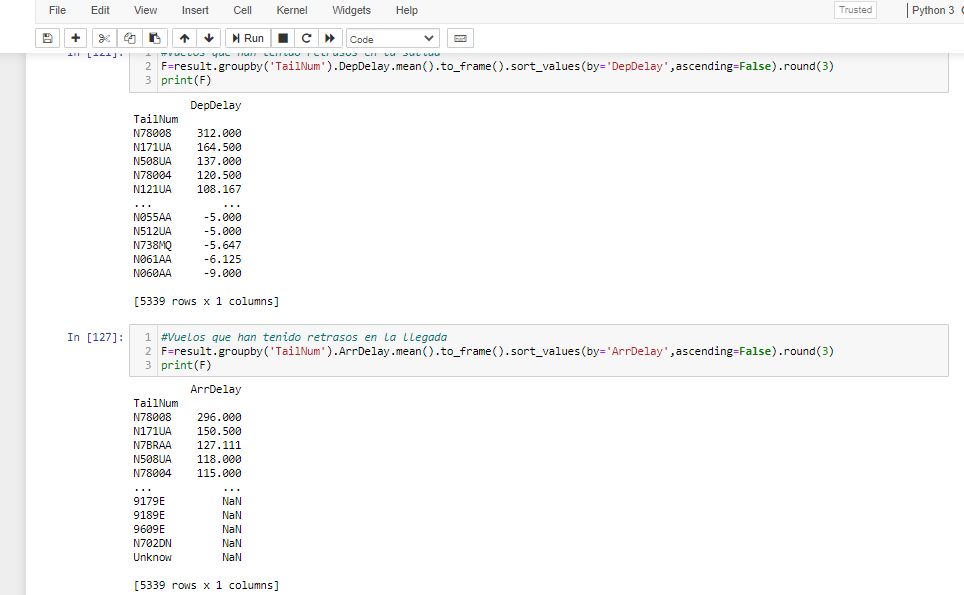


Con la variables TailNum y DayofWeek muestra de manera de pivote un TOP 10 , los vuelos realizados en los días de la semana y el día 7 de la semana



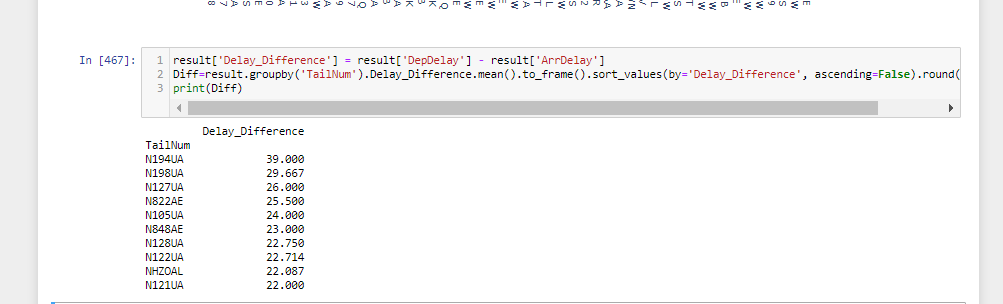
Mediante un gráfico de barras apiladas se evidencia el vuelo que más tiene retraso por medio de las variables ArrDealy y DepDelay que es el vuelo N78008



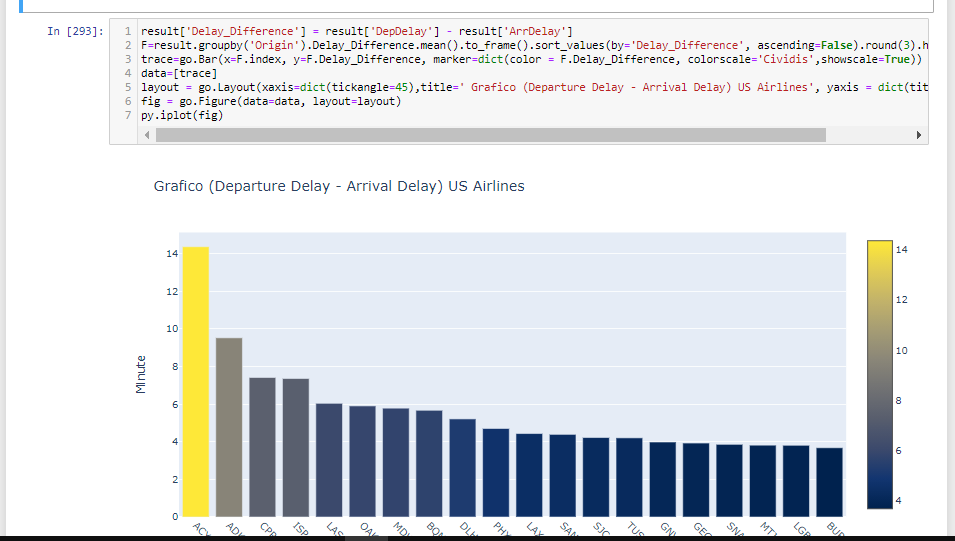


#Diferencia en retrasos

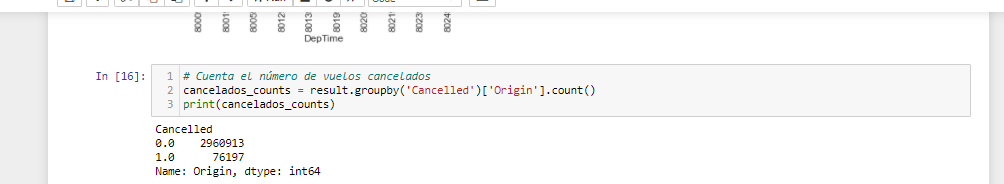
Por medio de la diferencia que se obtiene con las variables DepDelay, ArrDelay se sabe que el vuelo que marca la diferencia en retraso es el vuelo N194UA



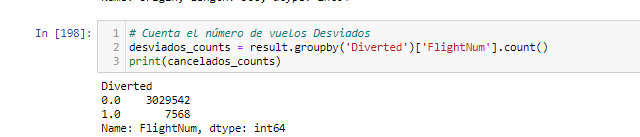
Se puede apreciar en el este grafico de barras apiladas el aeropuerto origen que marca esa diferencia que es el ACY



Muy pocos los vuelos son cancelados

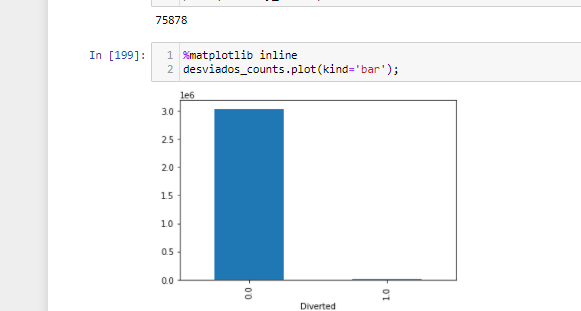


Y muy pocos los vuelos son desviados



Distribucion de los vuelos cancelados % y Deviados en comparativo

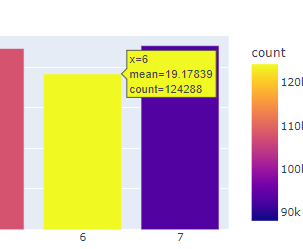




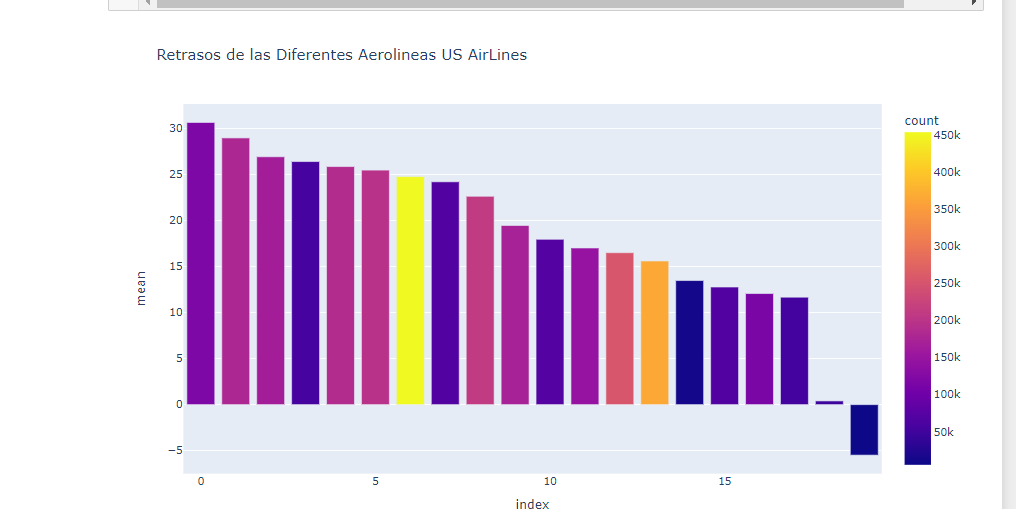
Por medio de la variable LateAircraftDelay podemos saber el dia de la semana el cual se retraso a causa de la llegada tardía de la aeronave



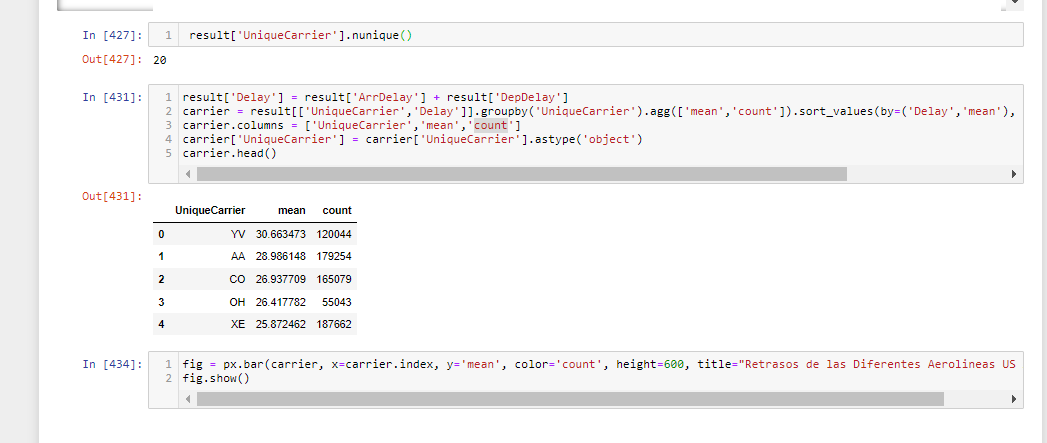
De acuerdo al mapa de calor es el dia 6 de la semana que se presentó retraso por esa causa.



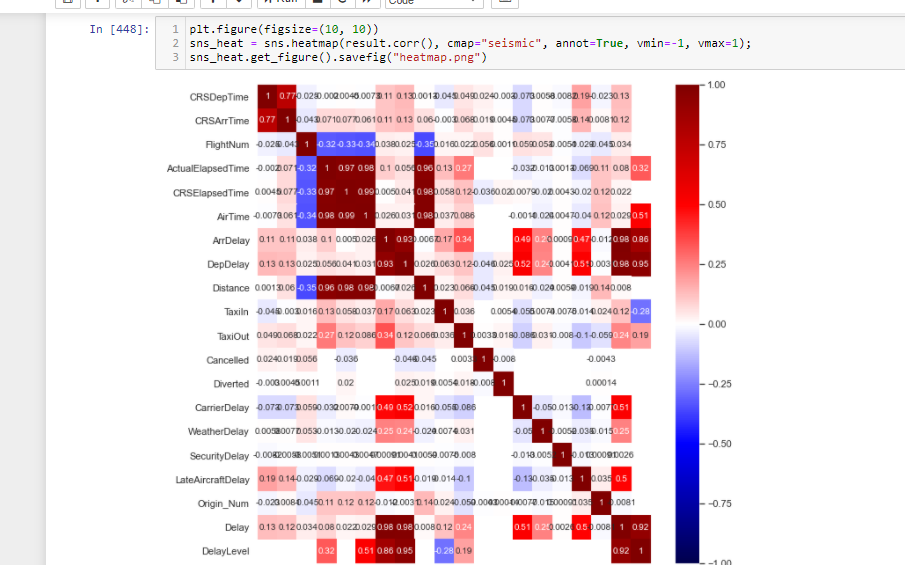
De acuerdo al insumo entregado se cuentan 20 erolineas distintas por medio de la varible UniqueCarrier se puede identificar los diferentes retrasos que tienen las aerolíneas







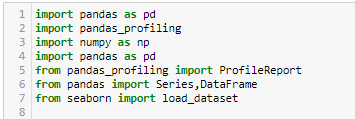
Finalmente podemos ver que de acuerdo a las variables analizadas cuales tienen alta correlación



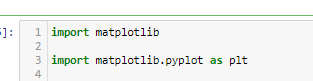
**Tecnologia Utilizada**

Jupiter NoteBook(Anaconda3)

Librerias utilizadas







Se realizó la Instalación de librerías

