**ETL CODE WITH COMMENTS**

**import** pandas **as** pd  
  
*#We are reading the csv file of 2017's data in a dataframe df\_2017  
#Here we are removing the unnecesary columns from the csv file, by explicitly supplying the list of columns to be used/included in the Dataframe.*df\_2017 = pd.read\_csv(**"/Users/nadeemalam/Downloads/airlines\_data/2017.csv"**, usecols=[**'FL\_DATE'**,  
**'OP\_CARRIER'**,  
**'OP\_CARRIER\_FL\_NUM'**,  
**'ORIGIN'**,  
**'DEST'**,  
**'CRS\_DEP\_TIME'**,  
**'DEP\_TIME'**,  
**'DEP\_DELAY'**,  
**'CRS\_ARR\_TIME'**,  
**'ARR\_TIME'**,  
**'ARR\_DELAY'**,  
**'CANCELLED'**,  
**'CANCELLATION\_CODE'**,  
**'DIVERTED'**,  
**'DISTANCE'**,  
**'CARRIER\_DELAY'**,  
**'WEATHER\_DELAY'**,  
**'NAS\_DELAY'**,  
**'SECURITY\_DELAY'**,  
**'LATE\_AIRCRAFT\_DELAY'**])  
  
*#We are reading the csv file of 2018's data in a dataframe df\_2018  
#Here we are removing the unnecesary columns from the csv file, by explicitly supplying the list of columns to be used/included in the Dataframe*df\_2018 = pd.read\_csv(**"/Users/nadeemalam/Downloads/airlines\_data/2018.csv"**, usecols=[**'FL\_DATE'**,  
**'OP\_CARRIER'**,  
**'OP\_CARRIER\_FL\_NUM'**,  
**'ORIGIN'**,  
**'DEST'**,  
**'CRS\_DEP\_TIME'**,  
**'DEP\_TIME'**,  
**'DEP\_DELAY'**,  
**'CRS\_ARR\_TIME'**,  
**'ARR\_TIME'**,  
**'ARR\_DELAY'**,  
**'CANCELLED'**,  
**'CANCELLATION\_CODE'**,  
**'DIVERTED'**,  
**'DISTANCE'**,  
**'CARRIER\_DELAY'**,  
**'WEATHER\_DELAY'**,  
**'NAS\_DELAY'**,  
**'SECURITY\_DELAY'**,  
**'LATE\_AIRCRAFT\_DELAY'**])  
  
*#here we have merged the data of 2017 and 2018 into a single dataframe df*df = pd.concat([df\_2017, df\_2018])  
  
size = len(df)  
  
*#Based on the length we have divided the merged dataframe into 5 equal sized dataframes  
#This file split approach is necessary because the original file is too huge to be loaded in the table directly.*df1 = df.iloc[:int(size/5)]  
df2 = df.iloc[int(size/5):int(2\*size/5)]  
df3 = df.iloc[int(2\*size/5):int(3\*size/5)]  
df4 = df.iloc[int(3\*size/5):int(4\*size/5)]  
df5 = df.iloc[int(4\*size/5):]  
  
*#the 5th dataframe was further divided into 2 equal sized dataframe to ensure data was loaded smoothly into the table*size2 = len(df5)  
df5\_1 = df5.iloc[:int(size2/2)]  
df5\_2 = df5.iloc[int(size2/2):]  
  
*#These divided dataframes will then be exported to separate csv's, so that these small csv's can then be imported in our table.*df1.to\_csv(**r"/Users/nadeemalam/Downloads/airlines\_data/part1\_vf.csv"**, header=True,index=False)  
df2.to\_csv(**r"/Users/nadeemalam/Downloads/airlines\_data/part2\_vf.csv"**, header=True,index=False)  
df3.to\_csv(**r"/Users/nadeemalam/Downloads/airlines\_data/part3\_vf.csv"**, header=True,index=False)  
df4.to\_csv(**r"/Users/nadeemalam/Downloads/airlines\_data/part4\_vf.csv"**, header=True,index=False)  
df5\_1.to\_csv(**r"/Users/nadeemalam/Downloads/airlines\_data/part5\_1.csv"**, header=True,index=False)  
df5\_2.to\_csv(**r"/Users/nadeemalam/Downloads/airlines\_data/part5\_2.csv"**, header=True,index=False)

Source of Dataset : <https://www.kaggle.com/yuanyuwendymu/airline-delay-and-cancellation-data-2009-2018>