**PDA-2**

Name: Ahmed Khawar  
 Azka Falak Sher

Reg No# L1F20BSSE0558

L1F20BSSE0624

Subject: Big Data Analytics

Submitted to: Ma’am Mubashera Anwar

**Checking for duplicate columns**

df = df.loc[:, ~df.columns.duplicated()]

**Check for duplicate observations**

df = df.drop\_duplicates()

**Checking for NaN-values**

nan\_columns = df.columns[df.isna().any()].tolist()

df.dropna(inplace=True)

**Renaming columns with names that match SQL syntax/ raw data;**

df.columns = df.columns.str.replace('Route', 'PathWay').str.lower()

**Conversion of values to lowercase letters**

df = df.apply(lambda Location: Location.astype(str).str.lower())

**Removal of extra spaces**

df = df.apply(lambda Location: Location.str.strip() if Location.dtype == "object" else Location)

**Conversion of binary values to 1 and 0**

binary\_columns = ['aboard'] # Specify columns with binary values

df[binary\_columns] = df[binary\_columns].replace({r'^-?\d+$': lambda x: int(x) if pd.notna(x) else x}, regex=True)

**Transfer of cleaned data to a new table to preserve the integrity of the original dataset.** cleaned\_df = df.copy()

**Data Transformation: Describe any transformations applied to the data.**

I have used .info () function on our initial dataset as well as the new dataset prepared there is the clear different in the values of these two datasets



Entries were cut down to 944 from 5268 while all the column names were changed into lowercase memory usage was also cut down to 103.2 KB from 535.2 KB indicating a clear change in the size of the data