Data Appendix File

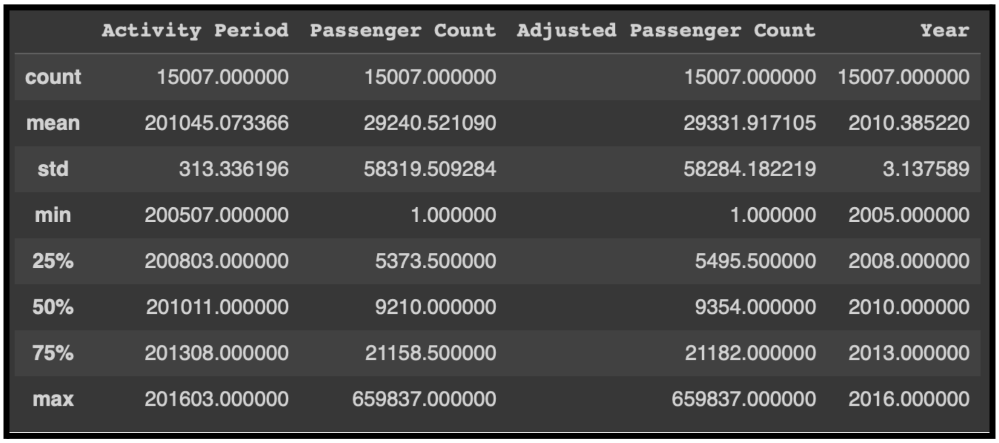
Section 1: Air_Traffic_Passenger_Statistics.csv

Unit of Observation: Each row in this dataset gives detailed information on monthly airline passengers, prior to cleaning and analysis.

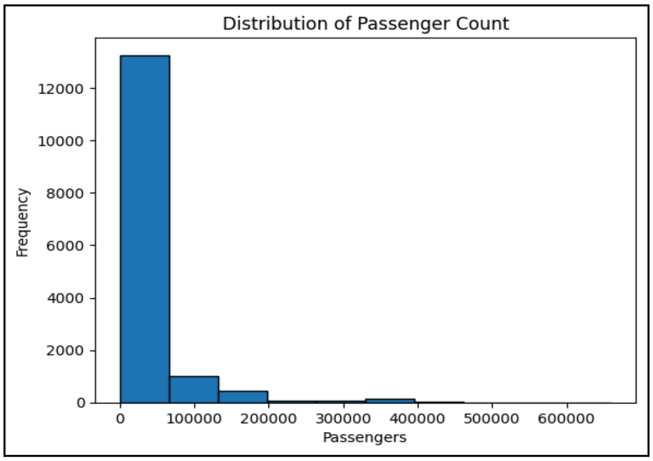
Variables:

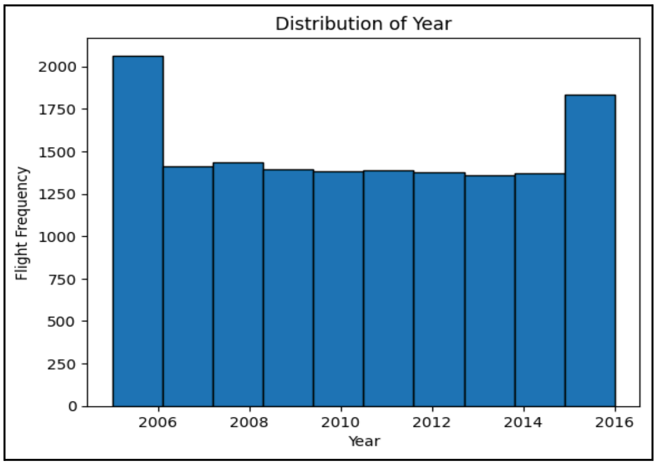
- Activity_period : This numerical variable is the year and month when the flight occurred, contained as an integer. There are no missing observations out of 15,007 entries.
- Operating_airline : This categorical variable details the airline name the flight was completed by, contained as a string. There are no missing observations out of 15,007 entries.
- Operating_airline_IATA_code : This categorical variable tells the International Air Transport Association (IATA) code for the operating airline, contained as a string. There are 54 missing observations out of 15,007 entries.
- Published_airline : This categorical variable gives the airline name that issued the ticket and book revenue for passenger activity, contained as a string. There are no missing entries out of 15,007 entries.
- Operating_airline_iata_code : This categorical variable tells the IATA code for the operating airline, contained as a string. There are 54 missing observations out of 15,007 entries.
- Geo_summary : This categorical variable tells whether the flights to / from the San Francisco airport were international or domestic, contained as a string. There are no missing entries out of 15,007 entries.
- Geo_region : This categorical variable gives a detailed breakdown of the geo_summary variable, telling the region the flight arriving from or departing to, contained as a string. There are no missing entries out of 15,007 entries.
- Activity_type_code : This categorical variable tells the physical action the passenger was taking in relation to the flight, contained as a string. There are no missing entries out of 15,007 entries.
- Price_category_code : This categorical variable gives the categorization of whether the published airline was a low cost and not low cost carrier, contained as a string. There are no missing entries out of 15,007 entries.
- Terminal : This categorical variable gives the name of the airport terminal, contained as a string. There are no missing entries out of 15,007 entries.

- Boarding_area : This categorical variable tells the letter representing the flight boarding area, contained as a string. There are no missing entries out of 15,007 entries.
- Passenger_count : This numerical variable gives the total number of passengers that were on the flight, contained as an integer. There are no missing entries out of 15,007 entries.
- Adjested_activity_type_code : This categorical variable gives the adjusted activity codes, now detailing if the flight was round trip or not, contained as a string. There are no missing entries out of 15,007 entries.
- Adjusted_passenger_count : This numerical variable gives the adjusted passenger counts, doubling the values for flights that were marked as being round trip, and is contained as an integer. There are no missing entries out of 15,007 entries.
- Year : This numerical variable tells the year the flight occurred, contained as a year. There are no missing entries out of 15,007 entries.
- Month : This categorical variable tells the month that the flight occurred, contained as a string. There are no missing entries out of 15,007 entries.



	Activity Period	Passenger Count	Adjusted Passenger Count	Year
count	15007.000000	15007.000000	15007.000000	15007.000000
mean	201045.073366	29240.521090	29331.917105	2010.385220
std	313.336196	58319.509284	58284.182219	3.137589
min	200507.000000	1.000000	1.000000	2005.000000
25%	200803.000000	5373.500000	5495.500000	2008.000000
50%	201011.000000	9210.000000	9354.000000	2010.000000
75%	201308.000000	21158.500000	21182.000000	2013.000000
max	201603.000000	659837.000000	659837.000000	2016.000000



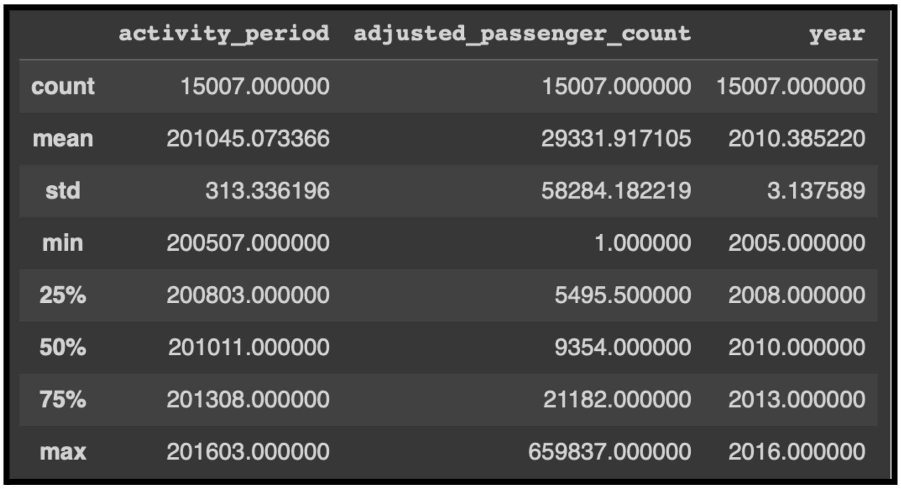


Section 2: Cleaned_Air_Traffic_Data.csv

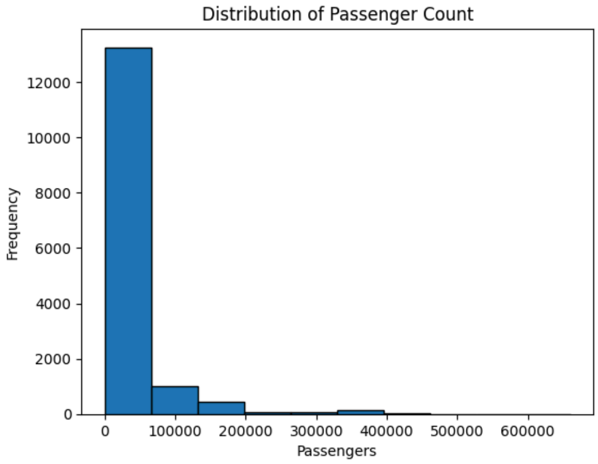
Unit of Observation: Each row in this dataset contains cleaned, detailed information on monthly airline passengers prior to analysis.

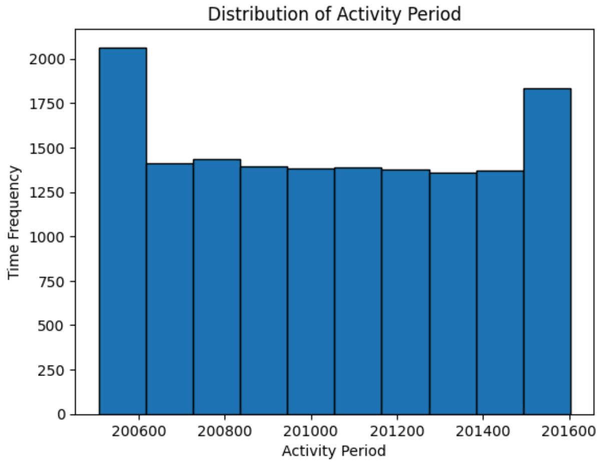
Variables:

- Activity_period: This numerical variable is the year and month when the flight occurred, contained as an integer. There are no missing observations out of 15,007 entries.
- Operating_airline: This categorical variable details the airline name the flight was completed by, contained as a string. There are no missing observations out of 15,007 entries.
- Operating_airline_iata_code: This categorical variable tells the International Air Transport Association (IATA) code for the operating airline, contained as a string. There are 54 missing observations out of 15,007 entries.
- Geo_summary: This categorical variable tells whether the flights to / from the San Francisco airport were international or domestic, contained as a string. There are no missing entries out of 15,007 entries.
- Geo_region: This categorical variable gives a detailed breakdown of the geo_summary variable, telling the region the flight arriving from or departing to, contained as a string. There are no missing entries out of 15,007 entries.
- Adjusted_passenger_count: This categorical variable gives the adjusted activity codes, now detailing if the flight was round trip or not, contained as a string. There are no missing entries out of 15,007 entries.
- Year: This numerical variable tells the year the flight occurred, contained as a year. There are no missing entries out of 15,007 entries.
- Month: This categorical variable tells the month that the flight occurred, contained as a string. There are no missing entries out of 15,007 entries.



	activity_period	adjusted_passenger_count	year
count	15007.000000	15007.000000	15007.000000
mean	201045.073366	29331.917105	2010.385220
std	313.336196	58284.182219	3.137589
min	200507.000000	1.000000	2005.000000
25%	200803.000000	5495.500000	2008.000000
50%	201011.000000	9354.000000	2010.000000
75%	201308.000000	21182.000000	2013.000000
max	201603.000000	659837.000000	2016.000000





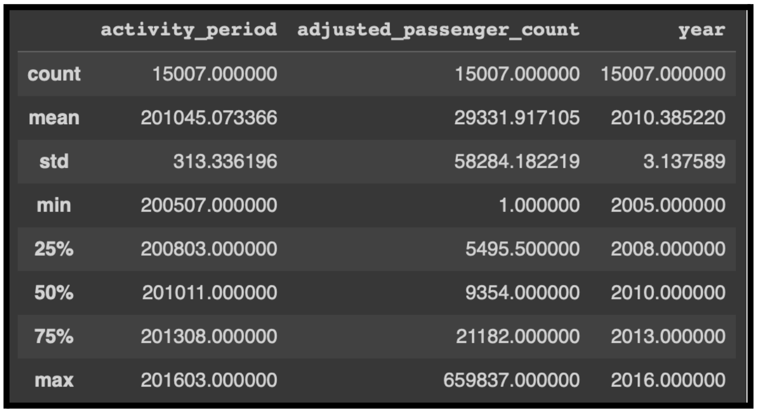
Section 3: Complete_Air_Traffic_Data.csv

Unit of Observation: Each row in this dataset contains cleaned, detailed information on monthly airline passengers for analysis.

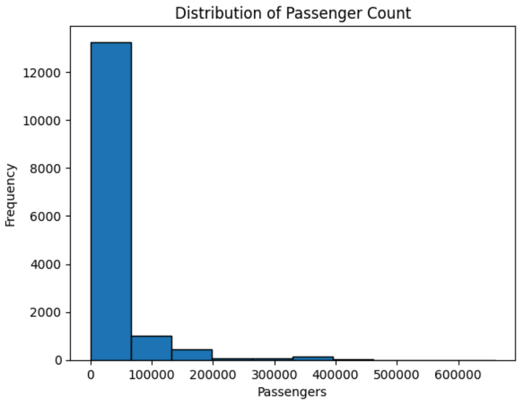
Variables:

- Activity_period: This numerical variable is the year, month, and day when the flight occurred, contained as a string (2005-07-01 for example). There are no missing observations out of 15,007 entries.
- Operating_airline: This categorical variable details the airline name the flight was completed by, contained as a string. There are no missing observations out of 15,007 entries.
- Operating_airline_iata_code: This categorical variable tells the International Air Transport Association (IATA) code for the operating airline, contained as a string. There are 54 missing observations out of 15,007 entries.
- Geo_summary: This categorical variable tells whether the flights to / from the San Francisco airport were international or domestic, contained as a string. There are no missing entries out of 15,007 entries.
- Geo_region: This categorical variable gives a detailed breakdown of the geo_summary variable, telling the region the flight arriving from or departing to, contained as a string. There are no missing entries out of 15,007 entries.
- Adjusted_passenger_count: This categorical variable gives the adjusted activity codes, now detailing if the flight was round trip or not, contained as a string. There are no missing entries out of 15,007 entries.
- Year: This numerical variable tells the year the flight occurred, contained as a year. There are no missing entries out of 15,007 entries.

- Month: This categorical variable tells the month that the flight occurred, contained as a string. There are no missing entries out of 15,007 entries.



	activity_period	adjusted_passenger_count	year
count	15007.000000	15007.000000	15007.000000
mean	201045.073366	29331.917105	2010.385220
std	313.336196	58284.182219	3.137589
min	200507.000000	1.000000	2005.000000
25%	200803.000000	5495.500000	2008.000000
50%	201011.000000	9354.000000	2010.000000
75%	201308.000000	21182.000000	2013.000000
max	201603.000000	659837.000000	2016.000000

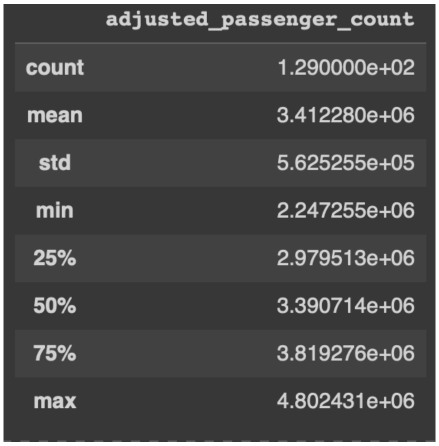


Section 4: monthly_passenger_data.csv

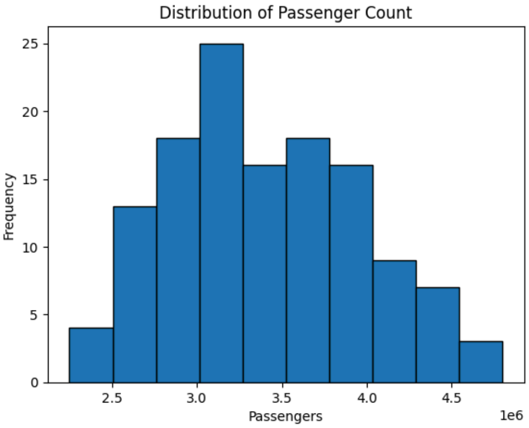
Unit of Observation: Each row in this dataset contains extracted data from that dataset that will contribute to the model prediction and analysis.

Variables:

- Activity_period: This numerical variable is the year, month, and day when the flight occurred, contained as a string (2005-07-01 for example). There are no missing observations out of 15,007 entries.
- Adjusted_passenger_count: This categorical variable gives the adjusted activity codes, now detailing if the flight was round trip or not, contained as a string. There are no missing entries out of 15,007 entries.



	adjusted_passenger_count
count	1.290000e+02
mean	3.412280e+06
std	5.625255e+05
min	2.247255e+06
25%	2.979513e+06
50%	3.390714e+06
75%	3.819276e+06
max	4.802431e+06

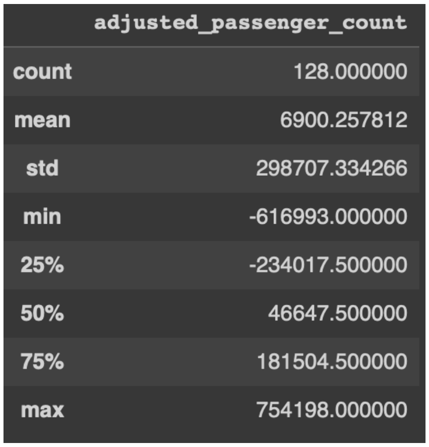


Section 5: monthly_passenger_data_diff.csv

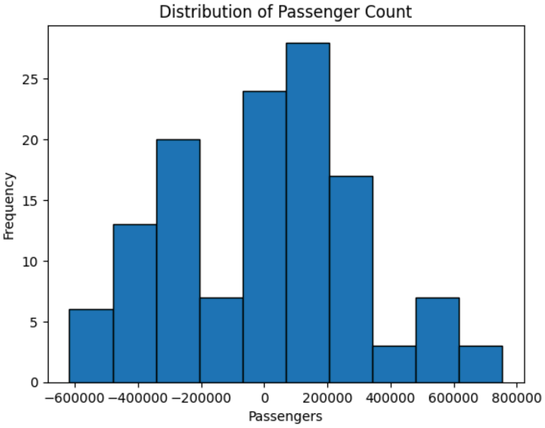
Unit of Observation: Each row in this dataset contains extracted, differenced information from the dataset for analysis.

Variables:

- Activity_period : This numerical variable is the year, month, and day when the flight occurred, contained as a string (2005-07-01 for example). There are no missing observations out of 15,007 entries.
- Adjusted_passenger_count : This numerical variable gives the output after first order differencing was applied to remove trends that previously failed the Augmented Dickey-Fuller (ADF) test, contained as an integer. There are no missing observations out of 15,007 entries.



	adjusted_passenger_count
count	128.000000
mean	6900.257812
std	298707.334266
min	-616993.000000
25%	-234017.500000
50%	46647.500000
75%	181504.500000
max	754198.000000


Distribution of Passenger Count
Frequency
Passengers