Capstone Project 1: Data Wrangling Summary

The analysis for this project will use two different datasets. One dataset contains the On-Time performance data for flights, and the other dataset contains weather data. Data wrangling was performed separately for each dataset. The cleaned data was combined by merging flight information with weather conditions based on location and time. The merged data was written to a csv file to be loaded for analysis.

The data wrangling code can be found here

Flight Data

The On-Time performance data for flights was collected by downloading .csv files from the Bureau of Transportation Statistics website. Each file contains data for one month during the years 2016 and 2017. The data in the files were cleaned, transformed, and combined using python.

The pre-processing involved the following steps:

1. The input csv files were read into a list of pandas DataFrames with the following settings:
    - Several fields in the csv files contain 4 digit numbers that represent time values, using the format HHMM. The values in these fields were read and stored as strings, in order to preserve the format and to store each missing value as NaN.
    - The csv files indicate whether each flight was diverted using a field populated by zeroes and ones. There is a similar field for cancellations. The values in these fields were read and stored as booleans rather than integers.
2. Each DataFrame was filtered individually. The full unfiltered dataset contains millions of data points, and the filters reduce the number of data points in a way that highlights the most valuable information for a streamlined analysis. This project will focus on data for the top 10 busiest U.S. airports (See **Appendix A**) and the top 8 largest U.S. airlines (See **Appendix B**). This analysis will focus on departure delays from the origin airport. DataFrame rows were kept if they met the following conditions:
    - The airline is among the largest.
    - The origin airport is among the busiest.
    - The departure delay is not null.
    - The flight date is not null.
    - The flight was not cancelled.
3. The data was transformed to create three new columns with departure information, listed below. These columns were then filtered to remove all missing values from them.
    - Scheduled departure time, represented as a Timestamp object.
    - Actual departure time, represented as a Timestamp object.
    - Wheels off time, represented as a Timestamp object.

4. All of the dates and times associated with departures are displayed in local time for the origin airport. These values are kept at local time, rather than converting to UTC, in order to preserve information about the time of day. A new column was created that contains the timezone of the origin airport as a string for each observation.
5. The DataFrames were concatenated into a single DataFrame. Unnecessary columns were dropped, and the index was reset to a default RangeIndex.

Missing values were excluded in the fields related to departure dates, times, and durations. Missing values in other fields were not removed, in order to preserve the departure data.

Weather Data

The original plan was to obtain the weather data from the NOAA website. Unfortunately, at the time of writing this, there is an ongoing U.S. government shutdown and the NOAA website is not providing data files. It is unclear when the current shutdown will end, and so weather data has been obtained from an alternate source in order to proceed with the project.

Weather text files (.txt) with comma separated values are downloaded from the following website: https://mesonet.agron.iastate.edu/request/download.phtml
One file was downloaded for each station, corresponding to the top 10 busiest airports. Each file contains METAR data for all of 2016 through 2017, presented in local time for the station. The data in the files were cleaned, transformed, and combined using python.

The weather data was pre-processed using the following procedure:

1. The input .txt files were read into a list of pandas DataFrames with the following settings:
    - The observation timestamp values were read and stored as Timestamp objects.
    - The text files indicate invalid data with 'M'. All instances of 'M' in the input files were converted to the appropriate missing values ('NaN', 'NaT', …).
2. Each DataFrame was filtered individually. DataFrame rows were kept if they met the following conditions:
    - The station is located at one of busiest airports.
    - The observation timestamp is not null.
3. A single observation from DEN for May 11, 2017 at 14:53 local time was removed. See the Outliers section for a discussion.
4. The wind directions with values of 360 degrees were converted to 0 degrees.
5. All of the observation timestamps are displayed in local time for the station. These values are kept at local time, rather than converting to UTC, in order to match readily with the flight times. A new column was created that contains the timezone of the station as a string for each observation.
6. The DataFrames were concatenated into a single DataFrame. Unnecessary columns were dropped, and the index was reset to a default RangeIndex.

Missing observation timestamps were removed from the DataFrame. All other missing values were kept, in order to preserve the number of observations.

Merged Data

The pre-processed flight and weather DataFrames were merged into a single DataFrame. Each scheduled departure time for a particular origin airport was matched with the nearest observation time for the corresponding station, as long as the observation time was either an exact match or was at most 1 hour earlier than the scheduled departure time. The rows without a successfully matched observation time were excluded.

The final merged DataFrame was written to a .csv file so that the pre-processing can be performed separately from the analysis.

Outliers

Descriptive statistics for the merged DataFrame were inspected using the method describe(). In the initial round of data cleaning, these statistics revealed a suspicious data point. Maximum wind speeds of 70 knots occurred at DEN on May 11, 2017 at 14:53 local time. The weather data at Denver International Airport was examined around that time, and the wind speed value of 70 knots seemed suspicious. The wind speed abruptly jumps to 70 knots for a single hourly report, with smaller wind speeds before and after. There are no special notes in the raw METAR suggesting severe weather conditions, such as a hurricane. This observation was removed from the weather dataset for the final cleaned version.

After removing the weather data point, the minimum and maximum values as well as the 25th, 50th, and 75th percentiles for all columns looked reasonable. The largest distances and flight times were consistent with flights from New York to Hawaii. The largest taxi times were on the order of a few hours. The most suspicious remaining values were a maximum departure delay of roughly 36 hours, and a similar maximum arrival delay. These delays occurred during the same flight, and they are associated with a reported Carrier Delay of 35.7 hours.

Box plots were examined for the departure delay duration at each origin airport. There were many outliers, based on the deviation of 1.5 times the interquartile range beyond the 25th or 75th percentile, but none of the inspected outlier values looked unreasonable. All of the outliers, except for the one that was removed for having suspiciously high wind speed, are retained for the full analysis. Some data points may be excluded from certain parts of the analysis. For example, data points with high security delays will be excluded from the portion of analysis that examines the relationship between weather variables and delay duration, but they will be included in the portion of analysis that examines the relationship between each airport and delay duration. The treatment of the outliers in the analysis will be explained in later documentation for the project.

Appendix A - Busiest U.S. Airports

The top 10 busiest U.S. airports by total passenger traffic in 2016 are displayed in the table below. The list in the table is based on data compiled by Airports Council International - North America (ACI-NA). The data can be found in the ACI-NA 2016 North American Airport Traffic Summary (Passenger).

Rank	Airport name	IATA Code	Total Passengers
1	Hartsfield–Jackson Atlanta International Airport	ATL	104,171,935
2	Los Angeles International Airport	LAX	80,921,527
3	Chicago O'Hare International Airport	ORD	77,960,588
4	Dallas/Fort Worth International Airport	DFW	65,670,697
5	John F. Kennedy International Airport	JFK	59,105,513
6	Denver International Airport	DEN	58,266,515
7	San Francisco International Airport	SFO	53,099,282
8	McCarran International Airport	LAS	47,496,614
9	Seattle–Tacoma International Airport	SEA	45,736,700
10	Miami International Airport	MIA	44,584,603

Table 1: Top 10 busiest U.S. airports by total passenger traffic in 2016

Appendix B - Largest U.S. Airlines

The top 8 largest U.S. airlines based on passengers carried in 2017 are listed below. The list is based on data from the Bureau of Transportation Statistics (BTS) tables T-100 Domestic Market (U.S. Carriers) and T-100 Market (US Carriers Only). Note that SkyWest Airlines and Republic Airline are excluded from the list because they are regional airlines that operate service for other airlines.

Airline	Enplaned Passengers (Domestic)	Enplaned Passengers (Domestic & International)
Southwest Airlines Co.	153,859,080	157,727,005
Delta Air Lines Inc.	120,928,953	145,436,827
American Airlines Inc.	116,528,317	144,919,764
United Air Lines Inc.	80,554,287	107,161,566
JetBlue Airways	32,395,833	40,013,934
Alaska Airlines Inc.	24,089,158	26,110,618
Spirit Air Lines	21,971,273	23,812,748
Frontier Airlines Inc.	15,970,347	16,799,968

Table 2: Top 8 largest U.S. airlines based on passengers carried in 2017