Data Descriptions and Sources

Airline_Delay_Cause.csv
This file contains information on the type of delay for each airline at airports that they serve by quarter. Each type of delay gets its own column with the count of that type of delay that happened for the corresponding year and quarter which are located on the same row.
Source: Bureau of Transportation Statistics (https://www.transtats.bts.gov/ot_delay/OT_DelayCause1.asp?20=E)

Basic_cancelled.csv
This file contains every flight route that was discontinued after a certain quarter, as well as statistics on it. It contains the origin airport, destination airport, distance, average passengers on the route per day, and more. One of the most important columns is the amt_quarters_back column which explains how many (including the quarter that it got dropped after) quarters in a row the route operated. In order to ensure that one is not tracking seasonal flights or other short-term flight routes of the sort, only look at flights where the amt_quarters_back amount is at or above something like 4, as there are 4 quarters in a year. A row with amt_quarters_back = 4, year = 2020, and quarter = 3 wold be a route that had been flying for a year when it was cut at the end of Q3 2020.
Source: custom data file created using data from Consumer_Airfare_Report__Table_6_-_Contiguous_State_City-Pair_Markets_That_Average_At_Least_10_Passengers_Per_Day.csv

Bureau of Transportation Statistics Airfare
This folder is a collection of files that we downloaded from the Department of Transportation website. It includes the average fare for major airports around the United States by quarter, as well as annual statistics.
Source: Bureau of Transportation Statistics (https://www.bts.gov/air-fares)

CAGDP2__ALL_AREAS_2001_2021.csv
This file contains information about GDP growth by county in the United States. It includes the FIPS number which is an identifier, as well as the GDP per county by year.
Source: Bureau of Economic Analysis (https://apps.bea.gov/regional/downloadzip.cfm)

CAGDP9__definition.xml
This file contains all of the information about the different metrics for GDP growth that can be found in the Bureau of Economic Analysis files.
Source: Bureau of Economic Analysis (https://apps.bea.gov/regional/downloadzip.cfm)

Consumer_Airfare_Report__Table_6_-_Contiguous_State_City-Pair_Markets_That_Average_At_Least_10_Passengers_Per_Day.csv
This dataset contains information on all flight routes that have existed in the United States since the 1990's where at least 10 people take the route per day on average. A description of most of the columns can be found in **farereportmetadata1.xlsx**. I added some columns on the end which are self-explanatory.
Source: Department of Transportation (https://data.transportation.gov/Aviation/Consumer-Airfare-Report-Table-6-Contiguous-State-C/yj5y-b2ir)

Crucial_data_and_files
This folder contains files which help map different labels to one another. For instance, some of the files map FIPS labels (used by miscellaneous government branches and agencies) to market_id labels (used by D.O.T. for airline markets).
Source: Various sources / created

Farereportmetadata1.xlsx
This file contains descriptions of each of the columns that exist in the **Consumer_Airfare_Report__Table_6_-_Contiguous_State_City-Pair_Markets_That_Average_At_Least_10_Passengers_Per_Day.csv** dataset which came with the original file. There are some additional columns that I added to the dataset, but they are self-explanatory.
Source: Department of Transportation (https://data.transportation.gov/Aviation/Consumer-Airfare-Report-Table-6-Contiguous-State-C/yj5y-b2ir)

final_natural_disasters_states.csv
This file contains information about natural disasters that have occurred in the United States that have caused at least $1 Billion in damage. Figures are given in millions of dollars with regard to cost in this file. It has been modified from its original format to include extra data about the time frame of each event as well as information about how frequently each event was, according to our program, mentioned in news outlets which we took headlines from.
Source: National Centers for Environmental Information (https://www.ncei.noaa.gov/access/billions/events/US/1980-2023?disasters[]=all-disasters)

National Centers for Environmental Information Daily
This folder is not contained in Google Drive. It contains the daily time series information from several weather stations inside of and around the United States which track information about the weather including precipitation, wind, humidity, and a host of other metrics.
Source: National Centers for Environmental Information (https://www.ncei.noaa.gov/data/normals-daily/)

News Links
This folder contains all of the links that we got from different news sources. Most of the files include the date that the article was published or last modified online as well. One might notice that most of the links that relate to news articles have part or all of the title in the link, so it is easy to parse this to get the title.

This was all collected using a custom-built scraper that accesses news websites' robots.txt file. This means that we effectively do the same thing that internet search engines do, but to a lesser extent that is less taxing on the server as we are not looking for information about the page itself. Every robots.txt file has sitemaps in it which link to every part of the website (including articles). We were primarily looking for articles which didn't include the words "image" or "video" in the url.
Source: independently collected and created

NOAA Disasters
This folder contains information about natural disasters that have occurred in the United States since the 1990's that have caused at least $1 Billion in damage.
Source: National Centers for Environmental Information (https://www.ncei.noaa.gov/access/billions/events/US/1980-2023?disasters[]=all-disasters)

non_dropped_gdp_mapping.csv
This file contains the inflation adjusted GDP per county year over year. It has been altered to fill invalid values and provide projections of expected growth in cases where there are gaps in information.
Source: Bureau of Economic Analysis (https://apps.bea.gov/regional/downloadzip.cfm)

new_gdp_mapping_final.csv
This file contains the inflation adjusted GDP per county year over year. It has been altered to fill invalid values and provide projections of expected growth in cases where there are gaps in information.
Source: Bureau of Economic Analysis (https://apps.bea.gov/regional/downloadzip.cfm)

T_FORM298C_A1.csv
This file contains revenue and capacity statistics for airlines dating back to 1974.
Source: (https://www.transtats.bts.gov/Fields.asp?gnoyr_VQ=FKI)

Us_county_lat_lng.csv

This file contains mappings for the fips ids to their latitude and longitude. We use this to assign each county to their respective market_id as it relates to the D.O.T. labels
Source: Github (https://gist.github.com/russellsamora/12be4f9f574e92413ea3f92ce1bc58e6)

weather_state
The 1991-2020 U.S. Climate Normals are conventional 30-year normals of many weather and climate variables. Normals are organized into hourly, daily, monthly, seasonal and annual normals. This document describes the elements of the daily normals. These observations are compiled from many surface weather station records, predominantly from National Weather Service (NWS) and Federal Aviation Administration stations at airports, the NWS Cooperative Observer Network, and other sources. Divided into 10 big files based on main locations.
Sources: (https://www.ncei.noaa.gov/data/normals-daily/1991-2020/)