**Airline Transportation Statistics of the United States for December 2020**

**Exploratory Analysis**

Brian Johnston, [bjohnston2@bellarmine.edu](mailto:bjohnston2@bellarmine.edu)

1. **INTRODUCTION**

This data set describes airline travels during the month of December 2020, including origins and destinations, minutes delayed, and reasons for delays. The site where the original dataset can be generated is <https://www.transtats.bts.gov/DL_SelectFields.asp?gnoyr_VQ=FGJ&QO_fu146_anzr=b0-gvzr>

Select the following boxes according to the Field Names: Year, Month, DayofMonth, DayofWeek, FlightDate, Flight\_Number\_Reporting\_Airline, Origin, OriginCityName, OriginState, Dest, DestCityName, DestState, DepTime, DepDelay, DepDelayMinutes, DepartureDelayGroups, WheelsOff, WheelsOn, ArrTime, ArrDelay, ArrDelayMinutes, ArrivalDelayGroups, Cancelled, Diverted, ActualElapsedTime, AirTime, Distance, DistanceGroup, CarrierDelay, WeatherDelay, NASDelay, SecurityDelay, LateAircraftDelay.  
  
I chose this data set because I was curious of the affect of air travel that the coronavirus had, almost a year after the pandemic began.

1. **DATA SET DESCRIPTION**The original data set contained 38 columns and 371,357 rows of data. This means that just during the month of December, there were over 300 thousand flights within the United States. After cleaning the data to only contain flights with complete data that were completed as planned, meaning no cancelled or diverted flights remained in the dataset, we were left with 366,940 flights left to examine. The data is split into three portions, the first being basic information on the flight, such as when and where it began. This is all categorical, or discrete variables. Then from DEP\_TIME to DIST\_GROUP, we have a combination of continuous variables and discrete versions of those continuous variables. The XXX\_GROUP columns are discrete, and their purpose is better grouping for later examination of the data. Finally, we have a grouping of delay causes, in minutes.

**Table 1:**

|  |  |  |
| --- | --- | --- |
| *Variable Name* | *Data Type* | *Missing Data (%)* |
| Day of Week | String, Ordinal | 0% |
| Day of Flight | Datetime64, Ordinal | 0% |
| Origin Airport Code | String, Nominal | 0% |
| Origin City Name | String, Nominal | 0% |
| Origin State Abbreviation | String, Nominal | 0% |
| Destination Airport Code | String, Nominal | 0% |
| Destination City Name | String, Nominal | 0% |
| Destination State Abbreviation | String, Nominal | 0% |
| Departure Time | String (formatted), Interval | 0% |
| Departure Delay | Float, Ordinal | 0% |
| Departure Delay Group | Integer, Ratio | 0% |
| Wheels Off | String (formatted), Interval | 0% |
| Wheels On | String (formatted), Interval | 0% |
| Arrival Time | String (formatted), Interval | 0% |
| Arrival Delay | Float, Ordinal | 0% |
| Arrival Delay Group | Integer, Ratio | 0% |
| Actual Elapsed Time | Float, Ordinal | 0% |
| Air Time | Float, Ordinal | 0% |
| Distance | Float, Ordinal | 0% |
| Distance Group | Integer, Ratio | 0% |
| Carrier Delay | Float, Ordinal | 0% |
| Weather Delay | Float, Ordinal | 0% |
| National Air System (NAS) Delay | Float, Ordinal | 0% |
| Security Delay | Float, Ordinal | 0% |
| Late Aircraft Delay | Float, Ordinal | 0% |

1. **DATA SET SUMMARY STATISTICS**This section will describe each individual variable based on their type, whether they are discrete or continuous. For continuous variables, statistical summary values will be placed in Table 2, and for discrete variables, proportions will be given for each category present in tables following.  
    **Table 2:  
   Summary Statistics for Airline Transportation Statistics of the United States for December 2020**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Variable Name* | *Count* | *Mean* | *Standard Deviation* | *Min* | *25th* | *50th* | *75th* | *Max* |
| Departure Delay | 366940 | 2.9627 | 39.5301 | -54 | -7 | -4 | 0 | 3072 |
| Arrival Delay | 366940 | -2.3364 | 41.7739 | -90 | -17 | -9 | 2 | 3069 |
| Actual Elapsed Time | 366940 | 135.7934 | 68.8934 | 16 | 85 | 121 | 167 | 679 |
| Air Time | 366940 | 112.9247 | 67.9695 | 8 | 63 | 97 | 144 | 658 |
| Distance | 366940 | 809.4888 | 567.8985 | 31 | 391 | 674 | 1050 | 5095 |
| Carrier Delay | 366940 | 3.1125 | 27.1242 | 0 | 0 | 0 | 0 | 2522 |
| Weather Delay | 366940 | 0.5568 | 15.4080 | 0 | 0 | 0 | 0 | 1706 |
| NAS Delay | 366940 | 1.5907 | 11.9999 | 0 | 0 | 0 | 0 | 1343 |
| Security Delay | 366940 | 0.0190 | 1.0013 | 0 | 0 | 0 | 0 | 258 |
| Late Aircraft Delay | 366940 | 1.6993 | 17.0400 | 0 | 0 | 0 | 0 | 2321 |

**Table 3:**

**Proportions for Day of Week**

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| Monday | 51195 | 13.9519 |
| Tuesday | 57195 | 15.5870 |
| Wednesday | 56601 | 15.4251 |
| Thursday | 56050 | 15.2750 |
| Friday | 46231 | 12.5991 |
| Saturday | 47424 | 12.9242 |
| Sunday | 52244 | 14.2377 |

**Table 4:  
Proportions for Day of Flight**

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| 1 | 11858 | 3.2316 |
| 2 | 9930 | 2.7062 |
| 3 | 11873 | 3.2357 |
| 4 | 11920 | 3.2485 |
| 5 | 9962 | 2.7194 |
| 6 | 12069 | 3.2891 |
| 7 | 11875 | 3.2362 |
| 8 | 9273 | 2.5271 |
| 9 | 9753 | 2.6579 |
| 10 | 11858 | 3.2316 |
| 11 | 11913 | 3.2446 |
| 12 | 9898 | 2.6975 |
| 13 | 12001 | 3.2706 |
| 14 | 11850 | 3.2294 |
| 15 | 9301 | 2.5347 |
| 16 | 9511 | 2.5920 |
| 17 | 12586 | 3.4300 |
| 18 | 13703 | 3.7344 |
| 19 | 13844 | 3.7728 |
| 20 | 13835 | 3.7704 |
| 21 | 13699 | 3.7333 |
| 22 | 13541 | 3.6902 |
| 23 | 13896 | 3.7870 |
| 24 | 9862 | 2.6876 |
| 25 | 8695 | 2.3696 |
| 26 | 13720 | 3.7390 |
| 27 | 14339 | 3.9077 |
| 28 | 13771 | 3.7529 |
| 29 | 13222 | 3.6033 |
| 30 | 13511 | 3.6821 |
| 31 | 9889 | 2.6950 |

**Table 5:**

**Proportions for Departure Delay Groups**

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| -2 | 4163 | 1.1345179 |
| -1 | 266584 | 72.65057 |
| 0 | 59449 | 16.201286 |
| 1 | 15093 | 4.1132065 |
| 2 | 7014 | 1.9114842 |
| 3 | 3861 | 1.0522156 |
| 4 | 2435 | 0.6635962 |
| 5 | 1757 | 0.4788249 |
| 6 | 1244 | 0.33902 |
| 7 | 1009 | 0.2749768 |
| 8 | 705 | 0.1921295 |
| 9 | 617 | 0.1681474 |
| 10 | 476 | 0.1297215 |
| 11 | 352 | 0.0959285 |
| 12 | 2181 | 0.5943751 |

**Table 6:**

**Proportions for Arrival Delay Groups**

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| -2 | 105327 | 28.70415 |
| -1 | 153638 | 41.87006 |
| 0 | 64443 | 17.56227 |
| 1 | 20398 | 5.558947 |
| 2 | 8271 | 2.254047 |
| 3 | 4183 | 1.139968 |
| 4 | 2492 | 0.67913 |
| 5 | 1682 | 0.458386 |
| 6 | 1233 | 0.336022 |
| 7 | 951 | 0.25917 |
| 8 | 771 | 0.210116 |
| 9 | 601 | 0.163787 |
| 10 | 459 | 0.125089 |
| 11 | 363 | 0.098926 |
| 12 | 2128 | 0.579931 |

**Table 7:**

**Proportions for Distance Groups**

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| 1 | 45620 | 12.43255 |
| 2 | 81641 | 22.249142 |
| 3 | 73279 | 19.970295 |
| 4 | 62673 | 17.079904 |
| 5 | 44426 | 12.107156 |
| 6 | 17301 | 4.7149398 |
| 7 | 16471 | 4.4887448 |
| 8 | 7550 | 2.0575571 |
| 9 | 5070 | 1.3816973 |
| 10 | 7162 | 1.9518177 |
| 11 | 5747 | 1.5661961 |

The rest of the tables for the discrete variables are saved in the github repository. None of the remaining tables would fit on a single page, as the shortest of them still contains 53 rows. As such I decided it would be more efficient to put those files separately into the repository. This includes the correlation matrix in numerical form, but not the heatmap.

**Figure 1:**

**Heatmap representing the Correlation Matrix of all continuous variables.A picture containing text

Description automatically generated**

1. **DATA SET GRAPHICAL EXPLORATION  
     
   A. Distributions**

**B.**

1. **SUMMARY OF FINDINGS**