Predictive Analysis of US Flight Delays

Brandon Boyle

bboyle@bellarmine.edu

12 Feb 2026

Before getting into the data itself, I would like to quickly go over the background. For this project I wanted to look at a data set that delt with aviation in some form. After graduation I'll be heading off to flight school in order to become a commercial pilot for someone like Delta, American, United, etc. So when looking, I came across a data set from the Bureau of Transportation Statistics, which ultimately stems from the US Department of Transportation. The data looks at flight delays in the US and is customizable with many variables. They have data dating back to 2018 and as recent as December of 2025 with every domestic flight flown by a regional, major, or legacy airline marked and recorded. With 50 plus variables that range from the date of the flight to the plane's tail number, the dataset was very customizable.

Originally, I was going to do all flights flown between 2022 and 2025, but I ended up narrowing it down to only flights in December of 2024. Since this dataset keeps track of every single flight flown, 2022 to 2025 contained over 26 million rows of data. And for the programs and computer that I was using, this was just too much. So, I narrowed it down to a single month and ended up with 631,944 rows of data. Still more than enough data. I decided to choose ten variables to look at. When looking though the fifty plus to choose from, many variables were of no use to me as they were written in aviation code and didn't protein to flights being delayed. I want to look at what airports and airlines have the most delays and the variables I choose reflect that. There were variables that looked at weather and other factors similar to that, but the amount of missing data would have resulted in getting rid of a majority of my data points. Below is a description of the variables I chose.

Variable	Description	Data Type	Range	% of missing data
Date	Data of flight	Interval/Date	12/1/2024 to 12/31/24	0%
Carrier	Airlines carrier code	Nominal/Categorical	n/a	0%
Airline.Name	Full name	Nominal/Categorical	n/a	0%
Flight_Num	Flight number	Nominal/Integer	n/a	0%
Origin	Airport plane took off from	Nominal/Categorical	n/a	0%
Dest	Airport plane landed	Nominal/Categorical	n/a	0%
Dep_Time	Scheduled departure time	Ratio/Numeric	2 to 2359	0%
Actual_Dep	Actual departure time	Ratio/Numeric	1 to 2400	0.68%

Delay	Minutes delayed	Ratio/Numeric	-59 to 3274	0.68%
Cancelled	If the flight was canceled	Nominal/Binary	n/a	0%

These ten variables will allow me to look at both airports and airlines and the number of times and percentage of times their flights are delayed. As well as days of the week and times of the day that seems to be the most prominent for delays. I can also look at flight routes and how many flights are delayed by x number of minutes. x being a number I can change, for example I can look at number of flights delayed by 10 minutes, 30 minutes, 60 minutes, etc. When looking at the missing data, you can see that for Actual_Dep and Delay, 0.68% of the data is missing. This is due to the fact that when a flight is canceled, there is no departure time.

One of the main things Im interested in this dataset is flight delays. So I decided to run a summary and below is what I found.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 -59.00   -6.00   -2.00   13.62   10.00 3274.00    4309
```

The minimum is -59 which means that there was a flight that took off 59 minutes before it was scheduled to. This seems unusual if you have ever flown before so its something to note and look into, especially with a first quartile value of -6. The average time delayed is 13.62 minutes, but the median value is -2, meaning my data is right skewed. But when looking at the max, 3274 seems very high. 3274 minutes equates to around 54 hours. This is a data point to look into as it could be causing a big disruption to the mean if it's an outlier. Then finally the 4309 missing data points is due to flights being canceled. These are important datapoints so I won't be dropping the n/a values.

Next I looked at the correlation between my variables. I wanted to make sure that there was a correlation between scheduled and actually departure time. I also wanted to see if time of day had anything to do with if the flight will get delayed or not.

```
            Dep_Time Actual_Dep     Delay
Dep_Time   1.0000000  0.9510380 0.0690056
Actual_Dep 0.9510380  1.0000000 0.1101514
Delay      0.0690056  0.1101514 1.0000000
```

As we can see, there is a strong correlation between scheduled and actual departure time. This is good as it indicates that planes typically take off close to their scheduled time. Next, we can see that delay has a weak correlation with

both scheduled and actually departure time. This mean that the time of day is not generally an indicator on if a flight will be delayed or not.

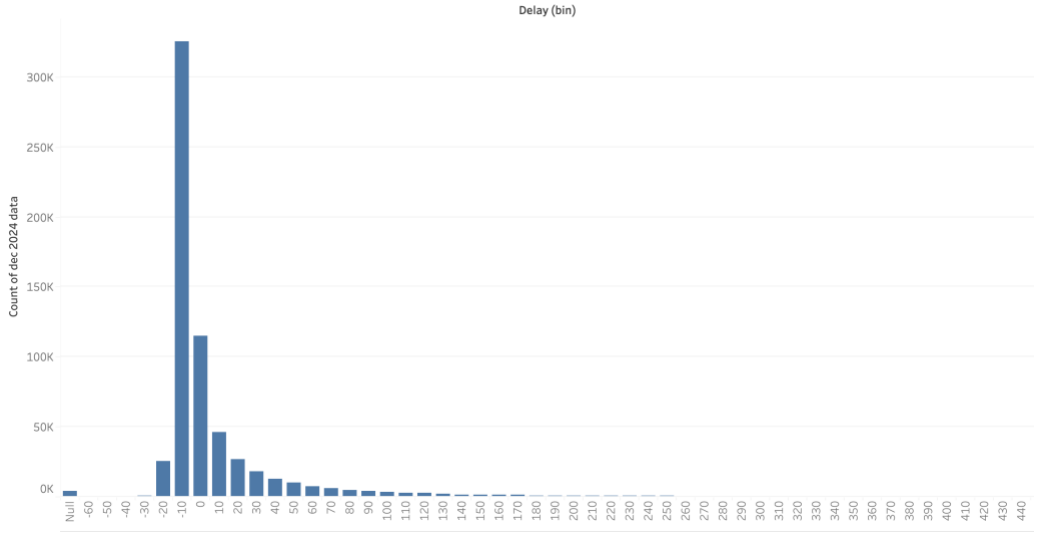
Finally, the last little thing I wanted to look at were the airlines and how many times they appear in my dataset. This is purely out of interest and doesn't have much use right now until I get into more analysis but its still helpful to see what airlines Ill be dealing with.

Air Wisconsin Airlines Corp	Alaska Airlines
3692	20027
Allegiant Air	American Airlines
11084	77737
Comair	CommuteAir LLC dba CommuteAir
20903	6730
Delta	Endeavor Air
83283	17356
Envoy Air	Frontier Airlines
22510	17294
GoJet Airlines LLC d/b/a United Express	Hawaiian Airlines
5619	6699
Horizon Air	Jet Blue
7352	20771
Mesa Airlines	Midwest Airlines
6801	27288
Piedmont Airlines	SkyWest Airlines
11237	66593
Southwest	Spirit Airlines
116276	18662
United Airlines	
64030	

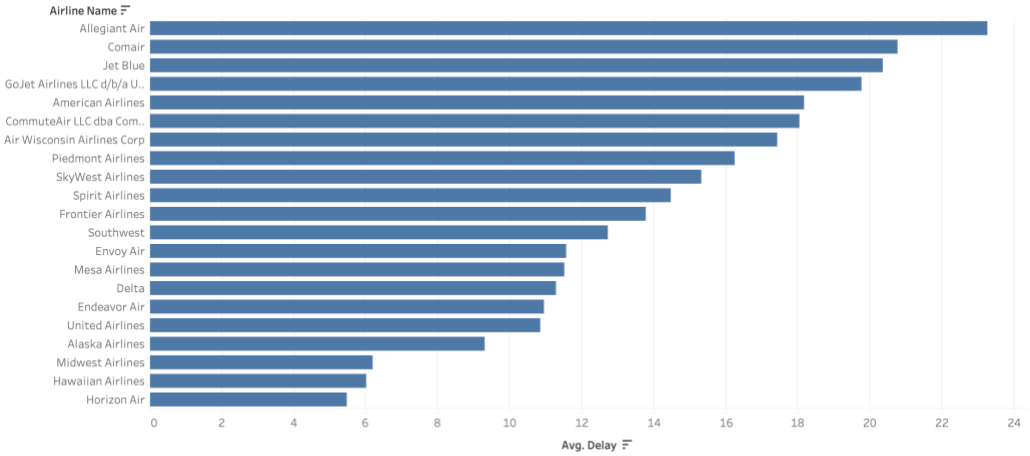
From the table, we can see that Southwest appears the most at 116,276 times, with Delta and American Airlines following behind. This goes with what we know as being some of the biggest airlines, so seeing the data back this up is good.

Below I have added in 5 visualizations of my data. Each is labeled appropriately.

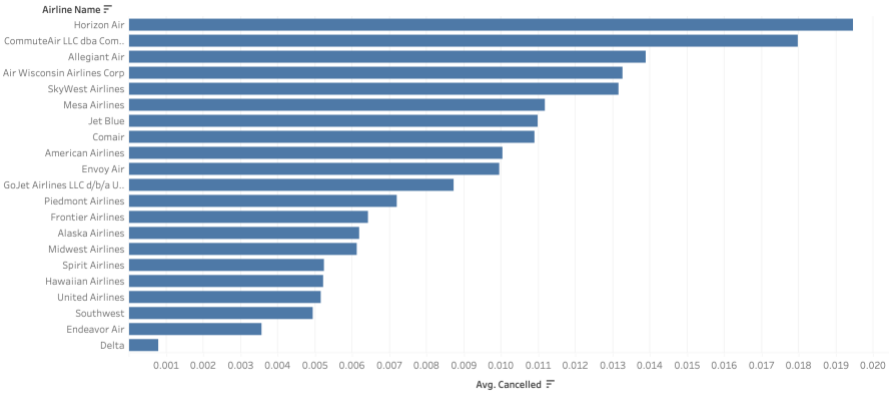
<Histogram of Delays>



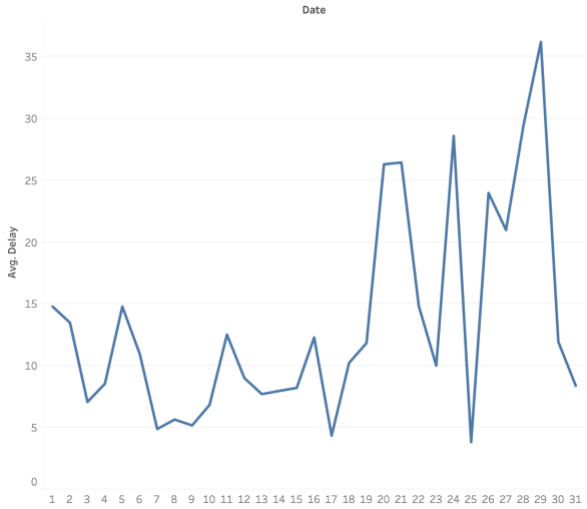
<Average Delay by Airline>



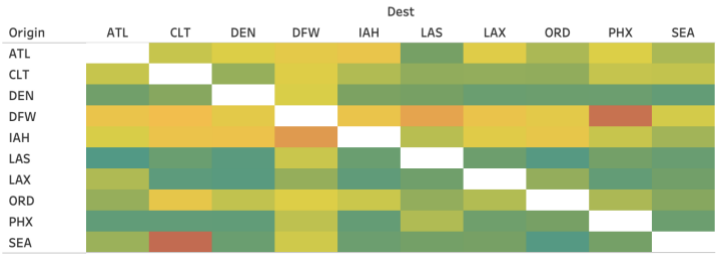
<Flights cancled by airline>



<Average delay based on date>



<Average Delay from the top 10 airports>



When looking at some potential issues in my dataset, the first one that pops up to me is how heavily right skewed my data is. With many points right around that 0 range, this makes the larger values of 200+ have a very big effect on model. Especially with a max value around 3500 minutes. There are a couple solutions to this however. I can look at the number of values that are outliers and from there I can get rid of them, or I can create categories that classify short delay, long delay, and extreme delay.

The next issue I could potentially run into is the distribution of airline carries. While you have your major airlines that show up 60k plus time, there are also some carriers that show up right around a thousand times. This could potentially lead to some unreliable predictions. To fix this, I may need to group the smaller carriers together into an “Other” category.

Finally, one of the last problems I could potentially run into is the use of negative numbers when looking at delayed minutes. With flight that are early showing up as negative numbers, this could potentially lead to issues with models not being able to interpret positive vs negative. To fix this, I may need to class all delays with a negative value as “Early”.