Analysis of Airline Arrivals in the US (%)

MSC-325
5/4/18

Madeline Warndorf, Najd Alayid, Rahel Hailemariam

Overview

The general idea of this project is to analyze the percentage of on-time airline arrivals in the United States. We were interested in this project because some of us travel a lot by plane and we wanted to know if it has gotten better or not. We were also interested in finding a data set that is not perfect and is missing data from a really credible source. We chose the time frame of 2012 through 2016. This gives us five years of data to analyze as well as provided some holes for us to use our skills to fill in the missing data. We wanted to select this data set because it allowed us to use three different software as well as a variety of techniques learned in this class.

Project Goal

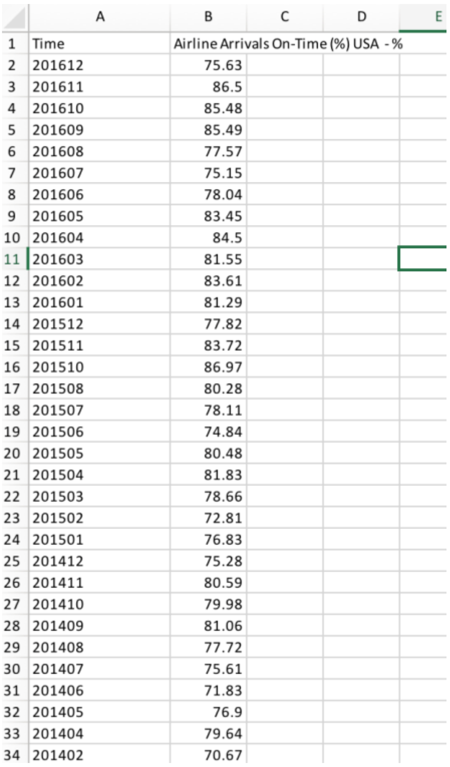
In this section, we stated the goal associated with the project, and also came up with some questions that our analysis was eventually supposed to answer. The goal of this project was to learn how to handle missing data as well as answer these questions:

1. Which month has most delays?
2. Which month is on time the most?
3. Has it gotten better over the 5 years?

Describe the Data

We found our data from the Marymount University's Articles and Databases Statistics category. It was found on the database called Data-Planet. The data itself is provided from the Bureau of Transportation Statistics (BTS) which was established as a Statistical agency in 1992.

The original raw data[1] was downloaded from Data-Planet's website as a .CSV file. The original file only contained two columns which were Time (yyyymm)[2] and Airline Arrivals On-Time (%) USA - %. The original data was missing entries for September 2015 and March 2014. The figure below shows the original data that was downloaded from Data-Planet. We wanted to have a full data set (January 2012-December 2016) so we had to fill in the missing data as well as fix the headings of the data to better analyze it in Tableau.



	A	B	C	D	E
1	Time	Airline Arrivals On-Time (%) USA - %			
2	201612	75.63			
3	201611	86.5			
4	201610	85.48			
5	201609	85.49			
6	201608	77.57			
7	201607	75.15			
8	201606	78.04			
9	201605	83.45			
10	201604	84.5			
11	201603	81.55			
12	201602	83.61			
13	201601	81.29			
14	201512	77.82			
15	201511	83.72			
16	201510	86.97			
17	201508	80.28			
18	201507	78.11			
19	201506	74.84			
20	201505	80.48			
21	201504	81.83			
22	201503	78.66			
23	201502	72.81			
24	201501	76.83			
25	201412	75.28			
26	201411	80.59			
27	201410	79.98			
28	201409	81.06			
29	201408	77.72			
30	201407	75.61			
31	201406	71.83			
32	201405	76.9			
33	201404	79.64			
34	201402	70.67			

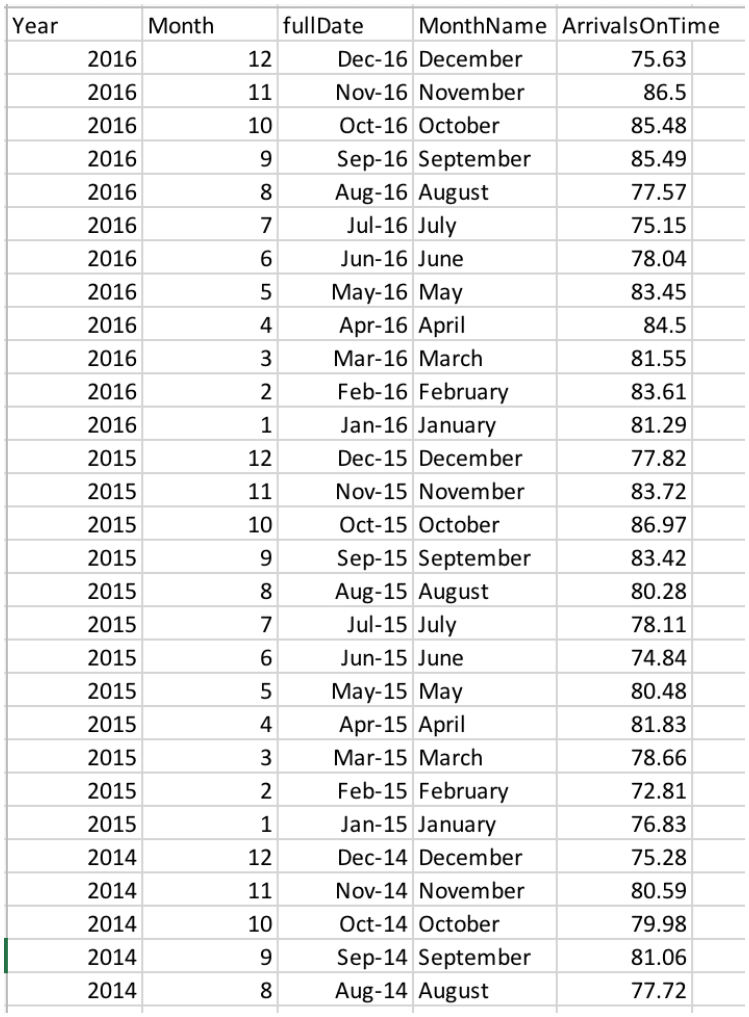
Data Sample

The data sample was created from the two columns by breaking down the Time column and filling in the missing data. Time was broken down into adding Year, Month, FullDate (mm/yyyy) and MonthName. We also renamed the column Airline Arrivals On-Time (%) USA

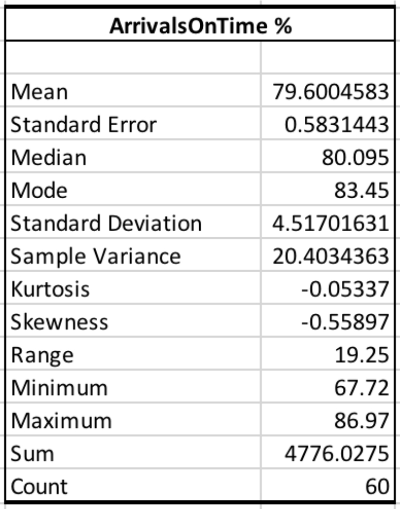
[1] Bureau of Transportation Statistics (2018-02-22). Airline Performance - Arrivals: Airline Arrivals On-Time (%), 01/2012 - 12/2016. Data-Planet™ Statistical Datasets by Conquest Systems, Inc. [Data-file]. Dataset-ID: 007-001-001
https://doi.org/10.6068/DP162636486E716

[2] Example: 201805 would be May 2018

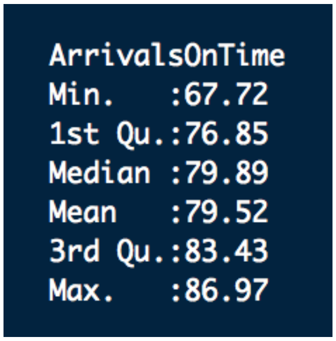
- % to just ArrivalsOnTime. To fill in the missing two months (March 2014 and September 2015), we used Excel to find the mean for all of March and all of September across the four other years and used that mean to fill in the missing months. The data sample that we used is shown below along with the Descriptive Summary from both R and Excel.



Year	Month	fullDate	MonthName	ArrivalsOnTime
2016	12	Dec-16	December	75.63
2016	11	Nov-16	November	86.5
2016	10	Oct-16	October	85.48
2016	9	Sep-16	September	85.49
2016	8	Aug-16	August	77.57
2016	7	Jul-16	July	75.15
2016	6	Jun-16	June	78.04
2016	5	May-16	May	83.45
2016	4	Apr-16	April	84.5
2016	3	Mar-16	March	81.55
2016	2	Feb-16	February	83.61
2016	1	Jan-16	January	81.29
2015	12	Dec-15	December	77.82
2015	11	Nov-15	November	83.72
2015	10	Oct-15	October	86.97
2015	9	Sep-15	September	83.42
2015	8	Aug-15	August	80.28
2015	7	Jul-15	July	78.11
2015	6	Jun-15	June	74.84
2015	5	May-15	May	80.48
2015	4	Apr-15	April	81.83
2015	3	Mar-15	March	78.66
2015	2	Feb-15	February	72.81
2015	1	Jan-15	January	76.83
2014	12	Dec-14	December	75.28
2014	11	Nov-14	November	80.59
2014	10	Oct-14	October	79.98
2014	9	Sep-14	September	81.06
2014	8	Aug-14	August	77.72



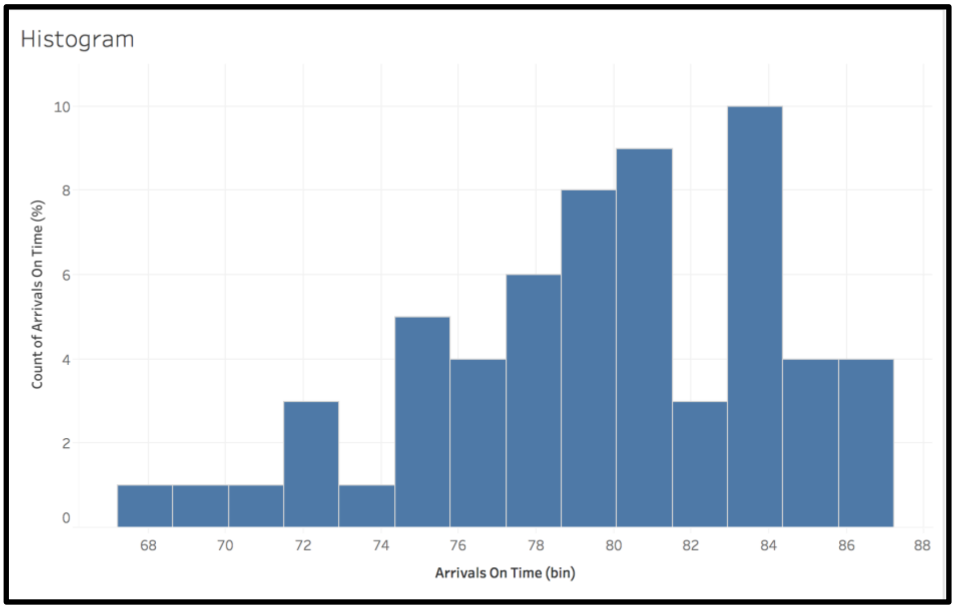
ArrivalsOnTime %	
Mean	79.6004583
Standard Error	0.5831443
Median	80.095
Mode	83.45
Standard Deviation	4.51701631
Sample Variance	20.4034363
Kurtosis	-0.05337
Skewness	-0.55897
Range	19.25
Minimum	67.72
Maximum	86.97
Sum	4776.0275
Count	60



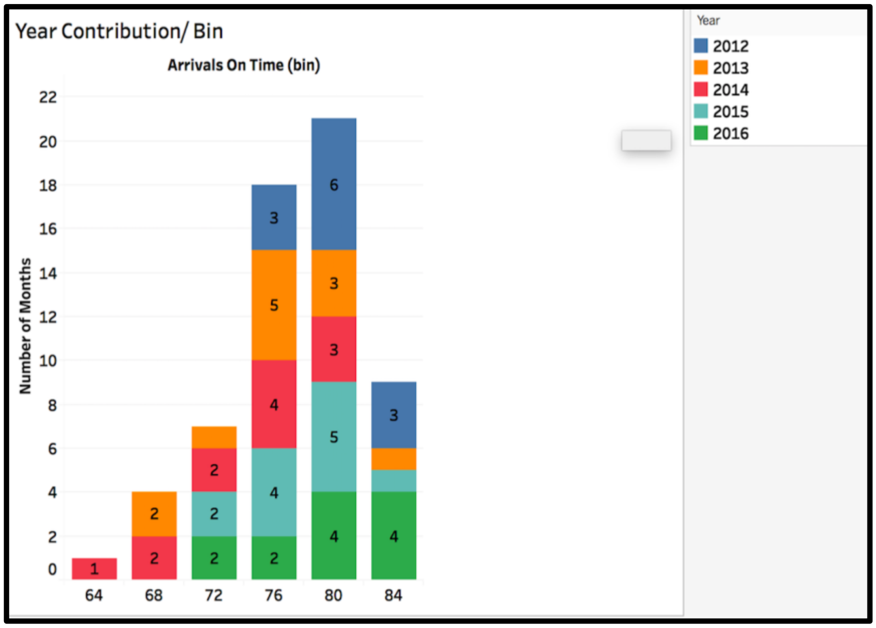
```
ArrivalsOnTime
Min.   :67.72
1st Qu.:76.85
Median :79.89
Mean   :79.52
3rd Qu.:83.43
Max.   :86.97
```

Methodology

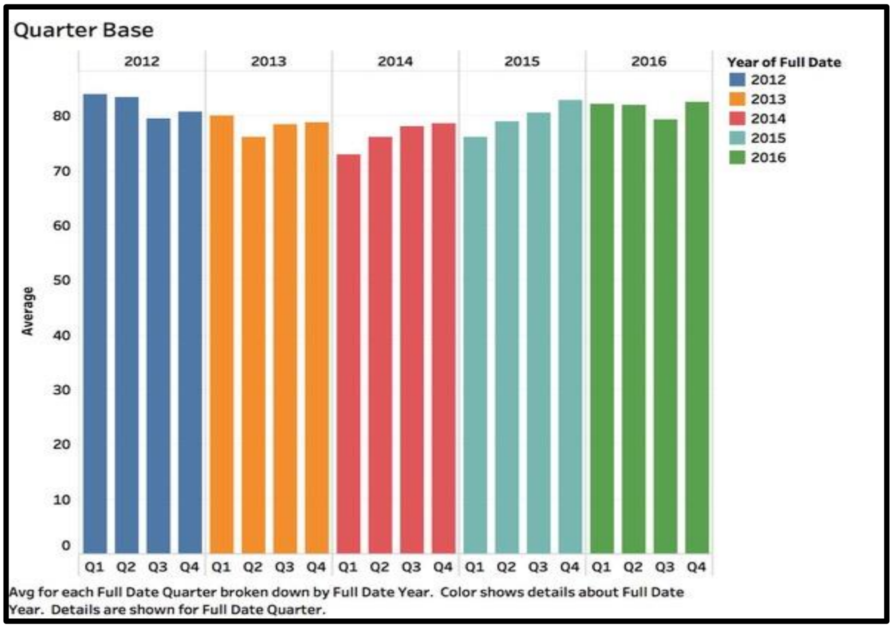
After the data sample was created in R, it was saved as a .CSV file and edited in Excel. From there it was uploaded into Tableau Public. In Tableau, the graphics were made to show the histograms, pie chart, box-plots, trend lines, and forecasting line. The histogram below shows the frequency of the percentage of arrivals on time. It shows that a majority of the months had the percentage of arrivals falling in the 84% bin. This is a right skewed distribution.



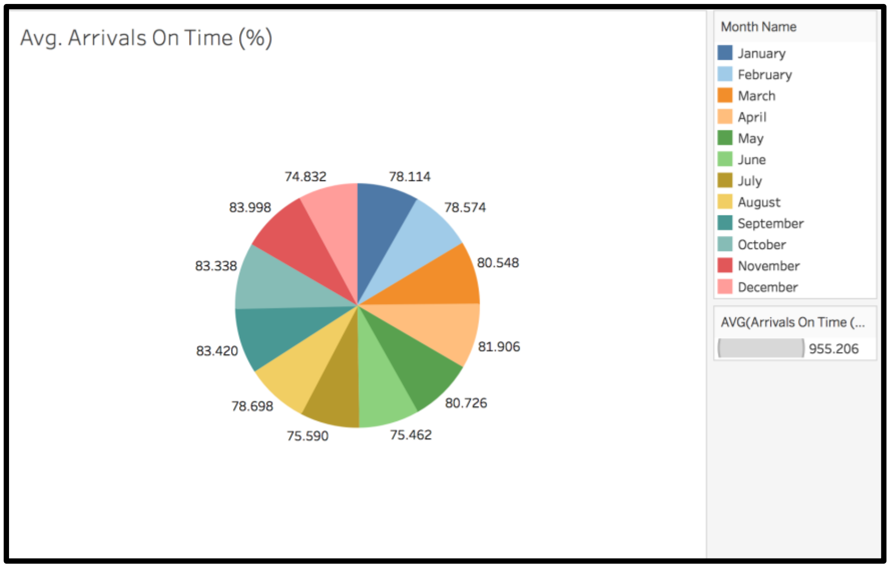
The histogram below shows the arrivals on time (%) based on the number of months per year. It shows that in 2014 a majority of the flights were in the bin ranges 68%-80%. The numbers inside the blocks on the graph show the number of months that specific year. Meaning, in 2013 there were 5 months that had an average arrival on-time of 76%.



The histogram below shows the same histogram but broken down by quarters and years.



The pie chart below shows the average on-time arrival percentage by month for all five years. The reason that we made a pie chart to look at the breakdown of the averages by month was to see how close the months were. The range of the averages is 74.832% to 83.998%. The distribution as seen above in the descriptive summary is very close.



Software Used

Presentation software used: Google Slides and downloaded as PowerPoint

Data compiled into, manipulated, and visualized: Excel, R, and Tableau

R: https://github.com/madelinew/MSC325Project

Tableau Public:

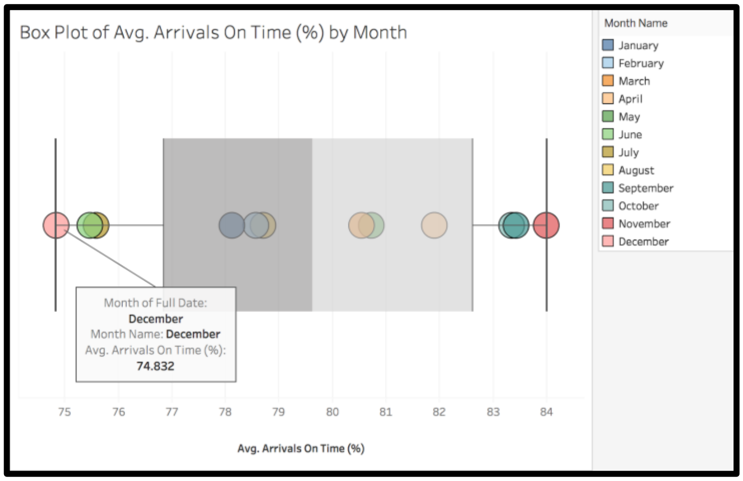
https://public.tableau.com/views/RealMSC325/Dashboard1?:embed=y&:display_count=yes&publish=yes

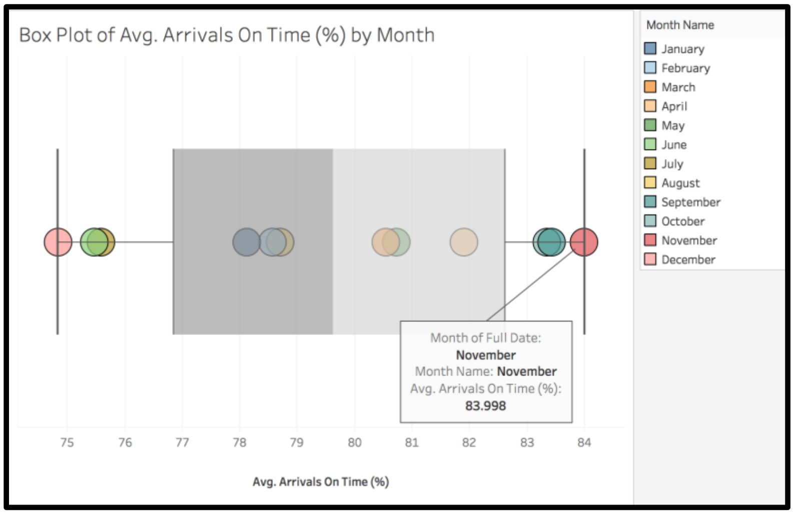
File format from the data: CSV files

Analysis and Result

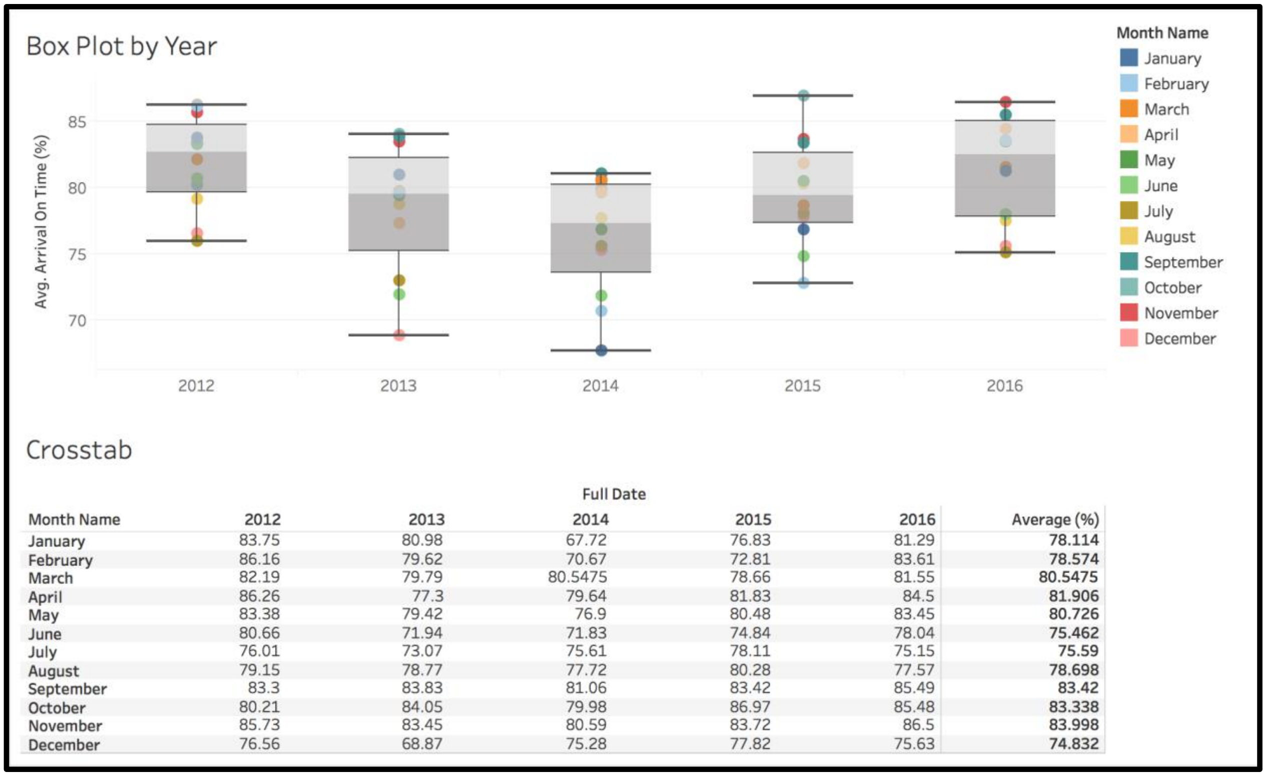
Month with the Lowest and Highest Arrival %

The two figures above show the lowest and highest arrival percentage on average across the five years. The first figure shows the box plot with the month on average that had the lowest percentage of arrivals on time. Not to our surprise it was December. It has the average percentage of 74.832%. The second figure shows the month on average that had the highest percentage of arrivals on time. However, surprisingly it was November. We were expecting it not to be a month that had traditionally higher flight reputation. But, November had an average percentage of 83.998%.





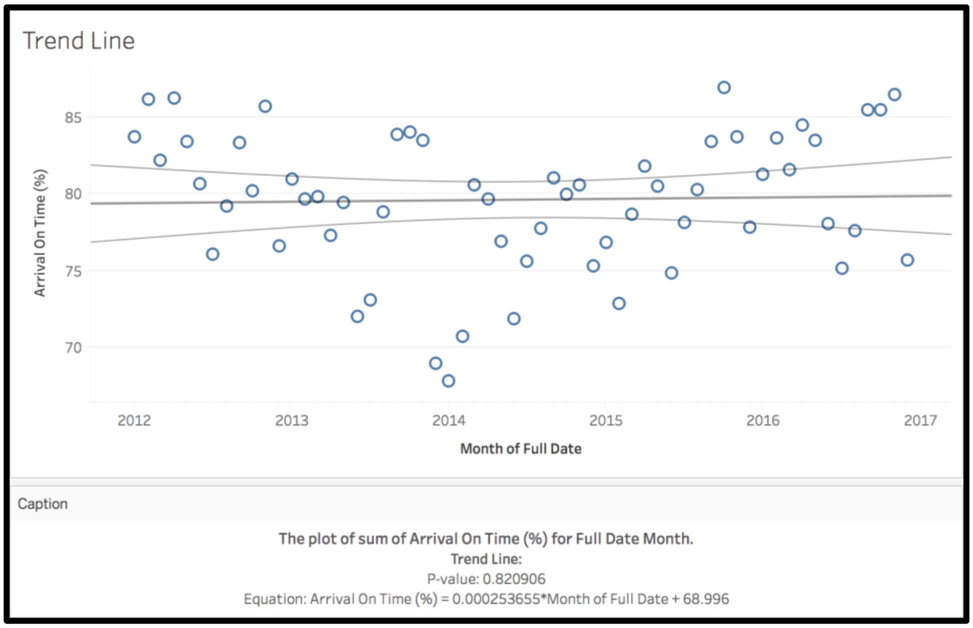
To analyze why the two months had averages this way, we did box plots broken down by years. We also looked at the crosstab of the data with the average of the months at the end of the chart. In 2012, April has the highest percentage of on time arrival when compare to other months. February has the second highest percentage which shows it’s in the median range. In 2013, October has the highest percentage of on time arrival and September has the second highest percentage. In 2014, September has the highest percentage of on time arrival when compare to all other months in all airlines. Year 2016 shows the most outliers in percentage of on time arrival. November was the month with the height percentage of on time arrival. The image below shows the box plot by year along with the crosstab.



Crosstab

	Full Date					
Month Name	2012	2013	2014	2015	2016	Average (%)
January	83.75	80.98	67.72	76.83	81.29	**78.114**
February	86.16	79.62	70.67	72.81	83.61	**78.574**
March	82.19	79.79	80.5475	78.66	81.55	**80.5475**
April	86.26	77.3	79.64	81.83	84.5	**81.906**
May	83.38	79.42	76.9	80.48	83.45	**80.726**
June	80.66	71.94	71.83	74.84	78.04	**75.462**
July	76.01	73.07	75.61	78.11	75.15	**75.59**
August	79.15	78.77	77.72	80.28	77.57	**78.698**
September	83.3	83.83	81.06	83.42	85.49	**83.42**
October	80.21	84.05	79.98	86.97	85.48	**83.338**
November	85.73	83.45	80.59	83.72	86.5	**83.998**
December	76.56	68.87	75.28	77.82	75.63	**74.832**

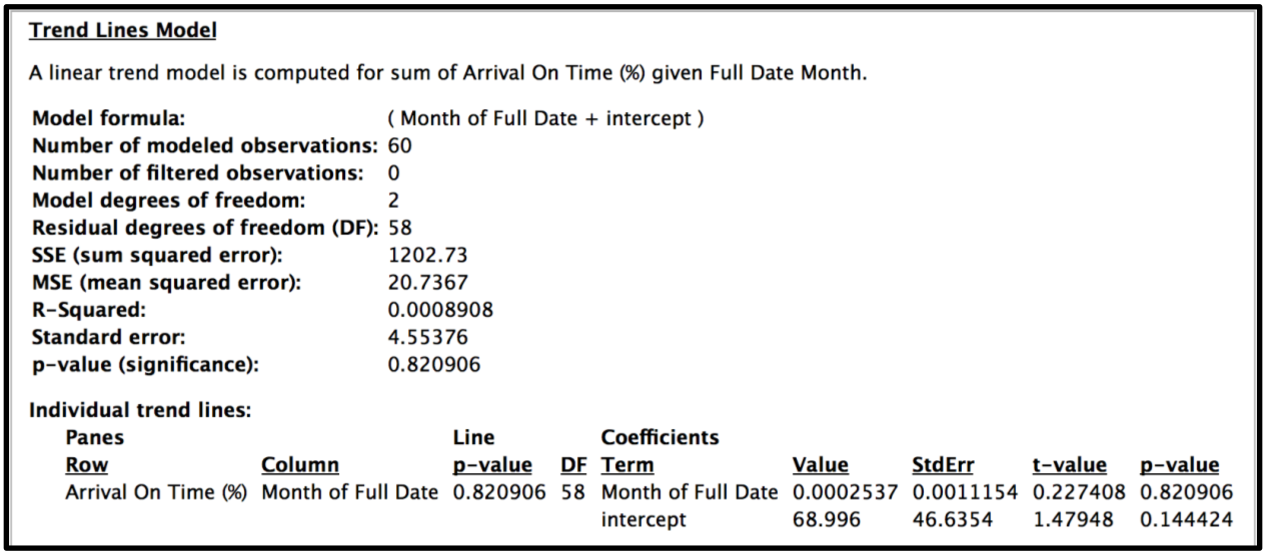
Trend Line



Caption

The plot of sum of Arrival On Time (%) for Full Date Month.
Trend Line:
P-value: 0.820906
Equation: Arrival On Time (%) = 0.000253655*Month of Full Date + 68.996

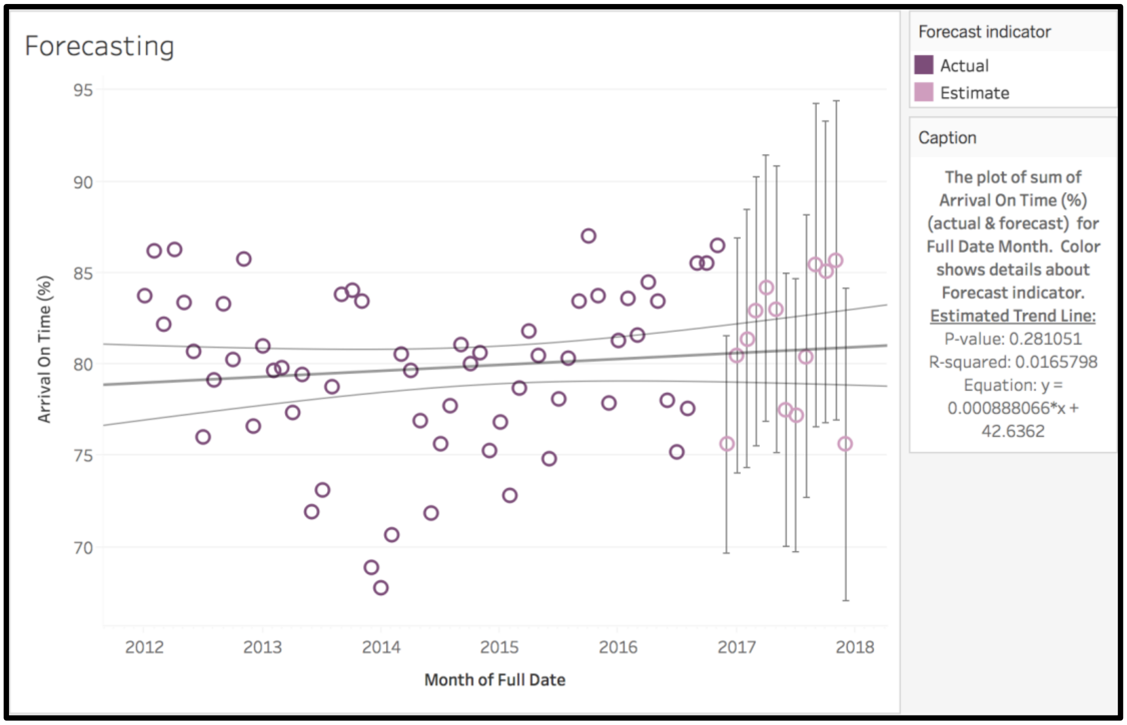
We used a trend line to determine if the percentage of arrivals on time have gotten better over the past five years. The image above has the trend line along the p-value and the regression line equation. It shows that the positive slope means that it is getting better, however, the large p-value means that the increase is not statistically significant. The image below the is the trend line model.



```
Trend Lines Model

A linear trend model is computed for sum of Arrival On Time (%) given Full Date Month.

Model formula:                     ( Month of Full Date + intercept )
Number of modeled observations:    60
Number of filtered observations:   0
Model degrees of freedom:          2
Residual degrees of freedom (DF):  58
SSE (sum squared error):           1202.73
MSE (mean squared error):          20.7367
R-Squared:                         0.0008908
Standard error:                    4.55376
p-value (significance):            0.820906

Individual trend lines:
    Panes                                     Line              Coefficients
    Row                 Column                p-value    DF   Term                  Value       StdErr      t-value     p-value
    Arrival On Time (%) Month of Full Date    0.820906   58   Month of Full Date    0.0002537   0.0011154   0.227408    0.820906
                                                              intercept             68.996      46.6354     1.47948     0.144424
```

Forecasting Model

Below is the forecasting model to see how Tableau forecasts the trend of the data for the year 2017. The graph shows that it predicts that the slope of the estimated trend line does increase a little bit compared to the actual trend line. The p-value also gets better compared to the actual p-value.



The plot of sum of Arrival On Time (%) (actual & forecast) for Full Date Month. Color shows details about Forecast indicator.
Estimated Trend Line:
P-value: 0.281051
R-squared: 0.0165798
Equation: y = 0.000888066*x + 42.6362

Conclusion

During our analysis of our data sample we found that the month that on average had the lowest percentage of arrivals on time was December. The month that on average had the highest percentage of arrivals on time was November. We also wanted to know if the percentage of airline arrivals in the US have gotten better in the past 5 years. Even though the p-value is not statistically significant, the slope was a positive value. In general, our group had problems finding time in each other’s schedules to be able to meet. Our project brought other questions to the surface like why is November the month with the highest on average arrival time (%). Or would November not be delayed by weather compared to December. Finally, what would the trend line look like if the data was broken down by days instead of months.