Analysis of Airline Delay and Cancellation Data, 2009 – 2018

Paul Smith

Western Governors University

**Table of Contents**

[Project Overview 4](#_Toc128352127)

[A. Project Highlights 4](#_Toc128352128)

[1. Research Question 4](#_Toc128352129)

[2. Project Scope 4](#_Toc128352130)

[3. Solution Overview 4](#_Toc128352131)

[a. Tools 4](#_Toc128352132)

[b. Methodologies 4](#_Toc128352133)

[Project Plan 4](#_Toc128352134)

[B. Project Execution 4](#_Toc128352135)

[1. Project Plan 4](#_Toc128352136)

[2. Project Planning Methodology 5](#_Toc128352137)

[3. Project Timeline and milestones 5](#_Toc128352138)

[Methodology 6](#_Toc128352139)

[C. Data Collection Process 6](#_Toc128352140)

[1. Actual data selection vs. planned collection process 6](#_Toc128352141)

[2. Obstacles to data collection 6](#_Toc128352142)

[3. Unplanned data governance handling 6](#_Toc128352143)

[C1. Advantages and Limitations of Dataset 6](#_Toc128352144)

[D. Data Extraction and Preparation Processes 7](#_Toc128352145)

[E. Data Analysis Process 8](#_Toc128352146)

[E1. Data Analysis Methods 8](#_Toc128352147)

[E2. Advantages and Limitations of Tools/Techniques 8](#_Toc128352148)

[E3. Application of Analytical Methods 8](#_Toc128352149)

[Results 10](#_Toc128352150)

[F. Project Success 10](#_Toc128352151)

[F1. Statistical Significance 10](#_Toc128352152)

[F2. Practical Significance 10](#_Toc128352153)

[F3. Overall Success 11](#_Toc128352154)

[G. Key Takeaways 11](#_Toc128352155)

[G1. Summary of Conclusions 11](#_Toc128352156)

[G2. Effective Storytelling 11](#_Toc128352157)

[G3. Findings-based Recommendations 12](#_Toc128352158)

[H. Panopto Presentation 12](#_Toc128352159)

[Appendices 12](#_Toc128352160)

[I. Evidence of Completion 12](#_Toc128352161)

[Sources 13](#_Toc128352162)

# Project Overview

# Project Highlights

### Research Question

The research question that this project aimed to answer is whether late aircraft delay is the overall cause of flight delays. Flight delays cost $32.9 billion in 2007 (Ball et al., 2010), so clearly these delays are a serious and widespread problem. Reducing this cost would be a win for both airlines and passengers.

### Project Scope

This project’s scope was to create a Jupyter notebook that allowed the project data to be loaded, cleansed, and analyzed. The analysis focused on flight delays and not on cancelled flights or diverted flights. The notebook revealed the causes of flight delays and how these causes ranked in terms of the number of flights delayed.

### Solution Overview

### Tools

A Jupyter notebook was used because it allowed both textual results and graphical plots to be viewed together. Cells within the notebook held text or Python programming code. The data supplied to the notebook were a collection of CSV files. The Python code within the notebook loaded the data, cleansed the data in order to provide a good quality dataset, and then was used to analyze the dataset. The results were displayed as both text and graphical plots in order to maximize understanding.

### Methodologies

This project used four different methodologies – project. data cleansing, analytical, and statistical. These methodologies played vital roles in this project. The corresponding sections below provide further detail regarding their role.

# Project Plan

# Project Execution

### Project Plan

The project plan was executed without change. All goals, objectives, and deliverables listed below were completed exactly as described in task 2.

The goal of this project was to find the overall cause of flight delays. To do this a Jupyter Notebook was used to perform data analysis on the flight delay and cancellation data from 2009 – 2018.

The objectives for this goal were:

* Concatenate the data into a single dataset, so that the data analysis can be performed on a single dataset.
  + The deliverable is to return a single dataset containing all the years of data.
* Cleanse the dataset, so that missing or unknown data does not compromise the results.
  + The deliverable is to return a single dataset free from unknown or missing data.
* Analyze the dataset for the cause of flight delays.
  + The deliverable is to list the cause of flight delays.

### Project Planning Methodology

The Waterfall project methodology was used by this project. The phases of this project are Requirements, Design, Implementation, Verification, and Maintenance. This methodology was chosen because each phase must be completed before the next is attempted. The phases were:

**Requirements:** All customer requirements are gathered before any other phase is begun. In this phase, the project scope is determined, the user expectations are decided, and the resources needed to complete the project are finalized.

**Design:** The tasks needing to be completed, in order to achieve the project objectives, are determined in this phase. Some of these tasks include determining what data cleansing will be necessary, the steps needed to analyze the dataset, and the visualizations required for the results.

**Implementation:** The tasks needed to achieve the objectives and test the Jupyter Notebook to ensure it is producing the desired results are completed in this phase.

**Verification:** In this stage, I will complete a standalone file for this project, so that it can be implemented by anyone else who has access to suitable hardware and software, i.e., a Jupyter development environment is installed.

**Maintenance:** This stage will not apply to this project, as it will not be in production in any companies. However, it could be uploaded to Kaggle, and in that case bug fixes and modifications could be requested.

This project’s execution did not change its methodology from its start to its end.

### Project Timeline and milestones

The actual project timeline and milestones followed the same pattern as was initially proposed. The milestones were completed without change. The timeline was changed. Those changes are shown below the table.

Present a table showing for each milestone its projected start and end dates, and its projected duration:

|  |  |  |  |
| --- | --- | --- | --- |
| **Milestone** | **Projected Start Date** | **Projected End Date** | **Duration (hours)** |
| Establish requirements for analytics process | 03/01/2024 | 03/03/2024 | 24 |
| Download dataset | 03/04/2024 | 03/04/2024 | 2 |
| Code notebook – loading data | 03/04/2024 | 03/04/2024 | 6 |
| Code notebook – cleansing data | 03/05/2024 | 03/08/2024 | 32 |
| Code notebook – data analysis | 03/09/2024 | 03/14/2024 | 48 |
| Test notebook | 03/15/2024 | 03/18/2024 | 32 |
| Create html file showing all notebook code and results | 03/19/2024 | 03/19/2024 | 1 |

The timeline of the project changed in the following minor ways:

* Downloading the dataset took one hour instead of the projected two hours.
* Loading the data was completed faster than expected, it took two hours instead of the projected six hours.

# Methodology

# Data Collection Process

### Actual data selection vs. planned collection process

Sourcing data from Kaggle was recommended. This website offers thousands of different datasets for download. The chosen data collection from Kaggle had been modified from data provided by the Bureau of Transportation Statistics, a US Government department. This department was founded in 1966 and its mandate is to collect and disseminate transportation statistics. The Kaggle data was downloaded as mentioned in task 2 with no changes to the process.

### Obstacles to data collection

There was one minor obstacle when performing the data collection process. A Kaggle account was required to download the data. An account was created, and the download was performed successfully.

### Unplanned data governance handling

Due to the nature of the data, there were no data governance issues that needed to be considered.

# C1. Advantages and Limitations of Dataset

The advantages of this dataset are that it was created so that flights could be more easily analyzed since it follows a standard format for every year released, and that it contained the fields that were necessary for this analysis, such as the type of delay for a flight and the late aircraft delay in minutes. Furthermore, the size of the data provides a great many data points to analyze, as each year of data contains millions of flights. The quality of the data was good, as a small number of rows were discarded due to missing data.

The limitations of this dataset are that it contains a great deal of data that has to be removed because it covers so many options, such as flight delays, flight cancellations, and flight diversions, as well as adhering to a standard format which means that some analyses require modifications to the data such as additional columns when breaking out the flight date into year, month or day to be able to utilize it. Furthermore, the size of the data sometimes constrained the operations that could be performed in a reasonable time, such as a pair plot taking over fifteen minutes to generate on what is regarded as a very fast PC. Adding a regression line to that pair plot took the run time to many hours, and eventually it had to be discarded as its generation time was not practical in the time allowed.

# Data Extraction and Preparation Processes

A zip file was downloaded from Kaggle. Once uncompressed, ten CSV files were available for use. Each file contained the data for one year of flights. The years covered were from 2009 to 2018. This was a reasonable delivery system as the uncompressed files were large. For each of the following tasks, Python code was executed within a Jupyter Notebook cell.

Each of the CSV files were loaded into a Pandas data frame. Once all ten files were loaded into their own data frame, the number of rows and columns were displayed for each data frame. A sample of five rows from each data frame was displayed to show the kind of data that was available. All ten data frames were concatenated into a single data frame. This became the main dataset. The number of columns and rows in this dataset was displayed, as well as a sample of five rows. The structure of the dataset was displayed. These tasks were appropriate for this data, because a single dataset was required for analysis, and so had to be built from the ten data files. Showing the results (e.g., data samples) from various tasks allowed for a more in-depth understanding of the data, as well as proving that these tasks had worked successfully.

The pattern of NaNs in the dataset was determined. This allowed for a data cleansing strategy to be formulated. A function that displayed the percentage of NaNs in each column of the dataset was written. This function was executed to show the percentage of NaNs in each column. Several observations regarding the quality of the data were made. Columns with a low percentage (less than two percent) of NaNs corresponded to missing data in their respective rows. These rows were removed from the dataset. A column with no name was composed totally of NaNs. This column was removed from the dataset. Several columns that held delay data had a high percentage of NaNs and the percentage was the same for all these columns. This corresponded to flights that were not delayed (although they could be cancelled or diverted). To make the analysis easier, NaNs in these columns would be replaced with zeroes, since a NaN really meant zero minutes of delay. Columns that contained data regarding cancelled or diverted flights were removed from the dataset since this analysis was concerned delayed flights. Once these changes were made the previously described function was used to show that percentage of NaNs in each remaining column. No NaNs remained in the dataset. These cleansing tasks were necessary for the dataset, so that missing or unknown data did not cause problems with the data analysis.

All IATA codes used within the dataset were replaced with the corresponding airline’s business name. This was done to provide greater readability when viewing the data. Non-delayed flights were removed from the dataset, as the analysis was concerned with flight delays. The flight date column was converted to a DateTime data type to allow for future analysis to work correctly. Finally, the structure of the remaining dataset was displayed. This showed that previous tasks had worked correctly, as well as showing that the size of the dataset was reduced. This allowed for faster execution of tasks, due to the reduced dataset size.

# Data Analysis Process

# E1. Data Analysis Methods

This project used a descriptive data analysis method. This method was most suitable for this project because the goal is to gain additional information from historical data.

Several data analysis techniques were used. Pearson correlation coefficients for the dataset were calculated and then displayed on a heatmap. This allowed columns that showed multicollinearity to be determined. These columns were removed from the dataset, and the Pearson correlation coefficients were generated again and displayed in another heatmap. This showed those columns that were correlated – late aircraft delay, carrier delay, NAS delay, and weather delay were all correlated with the arrival delay. A confidence test (Student’s t-test) was used to determine the statistical significance of the correlated columns. Although the columns that did have a correlation were determined, the relative rank of the number of delays for each column could not be determined from the correlation. This was determined by plotting the number of delays versus the flight date for each of the significant columns. From this plot, each kind of delay was able to be ranked relative to each other. This showed that the late aircraft delay caused the greatest number of delays. Based on this result, the null hypothesis was disproved.

# E2. Advantages and Limitations of Tools/Techniques

A Jupyter Notebook was the tool used to apply the analytic techniques used in this project. This tool uses Python code to perform the analysis and manipulation of data. Many different packages (e.g., pandas) are available for data analysis. This environment is robust and flexible. It allows the code output to be mixed with text so that visually pleasing reports can be produced. A limitation of this tool is that the underlying programming language, Python, can be slower to perform certain operations compared to other languages. Using larger datasets can reduce the available options for displaying certain plots, as it can take a long time for them to be processed and so render them unsuitable for use.

Pearson correlation is a technique that not only determines the degree to which two variables are correlated, but also the presence or absence of correlation. It also determines whether the correlation between two variables is positive or negative. Limitations of this technique are that it is comparatively difficult to calculate, and it is affected by the values of extreme items. It is also based on several assumptions that may not always be valid. Student’s t-test is the most powerful test (i.e., the one most likely to reject the null hypothesis if the null hypothesis is false) if both samples are normal distributed with the same variance, but possible different mean. It also allows you to get a confidence interval for the difference in means. They are also easy to interpret, and easy to calculate. Limitations of this technique include sensitivity to sample sizes, being less robust to violations of the equal variance and normality assumptions when sample sizes are unequal.

# E3. Application of Analytical Methods

A function to perform a Pearson correlation was defined. This function also displayed a heatmap of the correlation coefficients. The function was executed with the cleansed dataset. Several assumptions about the data are made when a Pearson correlation is applied:

* Both variables are on an interval or ratio level of measurement (data can be categorized, ranked, evenly spaced, and has a natural zero).
* Data from both variables follow normal distributions.
* Your data has no outliers.
* Your data is from a random or representative sample.
* You expect a linear relationship between the two variables.

These assumptions were verified as follow:

* All the variables that were used for correlation purposes were able to be categorized, ranked, evenly spaced, and all had a natural zero. The data cleansing process replaced NaNs with zeroes in those cases where the missing data context required a zero.
* Outliers were reduced or minimized by the removal of extraneous data during the data cleansing process.
* In the case of the dataset used, it was deliberately large, so that the sample mean will approach the population mean, and the sampling distribution of the mean is normally distributed, even if the original variable is not normally distribution for each variable. This means that we can say that the data from both variables that are correlated follow normal distributions.
* The dataset is a representative sample, based on the quality of the data.
* A linear relationship is expected between the variables that show correlation.

Several correlations stand out in the heatmap. They have a correlation coefficient of greater than 0.8. These are variables that will naturally have a high correlation. For example, the flight distance and time in the air will always be correlated since the time in the air increases with the distance flown. This kind of linear predictability is called multicollinearity. Several other variables, have a smaller correlation, but are still multicollinear. All these variables were removed from the dataset. The structure of the dataset was displayed to show that the corresponding columns of the dataset had been removed. The Pearson correlation function is again executed and the heatmap examined. The remaining correlated variables are most likely the causes of flight delays. To ensure that is the case, these variables are tested for statistical significance using the stats.pearsonr Python library function. Python code is used to display the correlation coefficient and the p-value for each of the correlated variables (carrier delay, weather delay, NAS delay, and late aircraft delay). The stats.pearsonr function uses the Student’s t-test. Several assumptions about the data are made when a t-test is applied:

* The scale of measurement applied to the data follows a continuous or ordinal scale.
* The data is collected from a representative, randomly selected portion of the total population.
* The data when plotted results in a normal distribution.
* A reasonably large sample size is used.
* Homogenuous, or equal variance, exists when the standard deviations of samples are approximately equal.

The assumptions were verified as follows:

* The data followed an ordinal scale, once various data were removed for various reasons (such as cancelled flights, diverted flights, etc.).
* The data were effectively a representative, randomly selected portion of the total population.
* As detailed earlier in this section the data can be considered to follow a normal distribution due to the dataset’s large size.
* The large size also implies that equal variance exists.

The p-values for each variable were less than 0.05 so the correlation between arrival delay and the various delay types were statistically significant. Although some correlation coefficients were larger than others, this does not imply the relative rank of the number of delays caused by each variable. A plot was displayed to visualize the number of delays for each delay variable versus the flight date. This plot enabled the relative rank to be determined. In decreasing rank – late aircraft delay, carrier delay and NAS delay – were found. The weather delay was responsible for a much smaller number of delays and so was not included in the conclusion, particularly since reducing weather delays would be very difficult to do. Also, it is likely that weather affects certain airports at certain times of the year due to their climates.

# Results

# Project Success

# F1. Statistical Significance

To determine if the project met the goal for statistical significance for any flight delay causes that were found, the four types of delay (carrier delay, weather delay, NAS delay, and late aircraft delay) that were shown to have a correlation with the arrival delay had their significance tested. The statistical significance of these delay types needed to be determined. To prove significance two hypotheses were formulated for each delay type – the first stated this result was significant and the second was a null hypothesis that stated the result was not significant. Python code was written to calculate the Pearson correlation coefficient and the p-value for each delay type. The Python library function stats.pearsonr was used to perform the calculation – it calculates both the Pearson correlation coefficient, r, and also the p-value for r. It was noted that the value of r generated by the function matched the value of r that was displayed on the second heatmap. This was checked to ensure the accuracy of the calculations. A significance level(α) of 0.05 was chosen. Each p-value for each delay type was less than 0.05, which means that the null hypothesis was disproven. This showed that each delay type was statistically significant.

# F2. Practical Significance

The practical significance of the results means that it is now known which delay types to target for reduction strategies. Knowing that late aircraft delay is the biggest cause of flight delays means that reducing this delay will benefit airlines and passengers. Airlines could allow aircraft to speed up, as well as flying at a different altitude to get benefit from the wind, all to increase groundspeed which would eliminate or reduce arrival delays. Knowing that carrier delays are the second biggest cause of flight delays means that reducing that delay would help to reduce flight delays. Airlines could improve their operational efficiency to reduce carrier delays, e.g., cleaning aircraft could be optimized for speed rather than cost.

While the previous causes of flight delays are under the control of the airlines, the third biggest one is not – NAS delay. Some of these kinds of delays could be reduced. Air traffic control systems could be modified or upgraded so that they can handle more aircraft and respond more quickly. Airport operations could be improved to reduce NAS delays, e.g., more runways could be built to deal with an increase in air traffic.

The weather delay was responsible for a much smaller number of delays and so was not considered in the conclusion, particularly since reducing weather delays would be very difficult to do. Also, it is likely that weather delays affect certain airports at certain times of the year due to their climates and so has a more localized effect. Further investigation on this subject is needed.

# F3. Overall Success

This project was a success. All the criteria laid out in task two were met. Several causes of flight delays were identified, and these were shown to be statistically significant. The biggest cause of flight delays was identified as late aircraft delay which confirmed the initial premise of the project. Now that the causes have been determined it is possible to take steps to reduce these delays. In the future this project can be expanded, and more in-depth analysis can be performed to answer questions such as whether different airports have different causes of flight delays, or whether more flight delays occur on certain days of the week or months of the year.

# Key Takeaways

# G1. Summary of Conclusions

This project set out to determine whether the overall cause of flight delays is late aircraft delay. Certain objectives were required to be accomplished for the project goal to be a success. The following chart summarizes whether these objectives were accomplished:

|  |  |
| --- | --- |
| **Objective** | **Success (Yes or No)** |
| Concatenate the data into a single dataset, so that data analysis can be performed on a single dataset. | Yes |
| Cleanse the dataset so that missing, or unknown does not compromise the results. | Yes |
| Analyze the dataset for the cause of flight delays. | Yes |

All the objectives were achieved. There were a several criteria used to determine success of the project. The following chart summarizes whether these criteria were successful:

|  |  |
| --- | --- |
| **Criterion** | **Success (Yes or No)** |
| Are causes of flight delays identified? | Yes |
| Are causes statistically significant? | Yes |
| Are the number of delays for each cause ranked? | Yes |

All the criteria were successful. So, the project was a success.

The biggest cause of flight delays is determined to be late aircraft delay, followed by carrier delay. NAS delay also contributes to flight delays, but to a lesser extent. From these results we can conclude that the null hypothesis is disproved.

# G2. Effective Storytelling

The tools and visual representations chosen were the best methods for compelling storytelling. Using a heatmap to display the Pearson correlation coefficients allowed for a lot of information to be displayed in a small amount of space, while enabling the multicollinear variables to be easily identified in the first heapmap, as well as the correlated variables in the second heapmap.

Using a Jupyter Notebook enabled easy embedding of relevant graphical representations, while giving flexibility in running Python code or displaying informative text.

The graphical plot of the number of delays for each delay type versus flight date allowed a very large dataset to be used, while displaying the plot in a relatively small area. The rank of each delay type was determined easily as this plot was easy to understand.

# G3. Findings-based Recommendations

The research question asks whether the overall cause of flight delays is late aircraft delay. This is shown to be correct, as a ranked list of flight delays is determined, and late aircraft delay is the delay that is ranked first. Carrier delay is shown to be ranked second, and NAS delay third.

The implications of this analysis are that the two biggest causes of flight delays are within the control of the airlines.

The biggest cause, late aircraft delay, is when an aircraft is late to an airport due to its late arrival at a previous airport. The ripple effect of an earlier delay at previous airports is referred to as delay propagation.

The second biggest cause, carrier delay, is caused by one or more carrier tasks that do not complete on time. For example, aircraft cleaning, fueling, maintenance, and cargo loading.

Two recommended courses of action based on these findings are:

* Airlines should allow aircraft to increase speed or change altitude (to benefit from wind) in order to eliminate or reduce the late aircraft delay.
* Airlines should do their best to reduce carrier delay, e.g., calculate the optimal number of cleaning staff needed to clean an aircraft quickly.

# Panopto Presentation

My Panopto presentation can be found at the following link:

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=f2117795-6013-4de5-a790-afb5009aa3a4>

# Appendices

# Evidence of Completion

The following files will serve as evidence of completion for this project:

1. Jupyter Notebook.
2. CSV data files that are compressed in the 7-Zip format.
3. HTML file generated from Jupyter Notebook showing Python code results and markdown text.

# Sources

Ball, M. and Barnhart, C. and Dresner, M. and Hansen, M. and Neels, K. and Odoni, A. and Peterson, E. and Sherry, L. and Trani, A. and Zou, B. (2010, October 16). *Total delay impact study: a comprehensive assessment of the costs and impacts of flight delay in the United States.* Institute of Transportation Studies, University of California, Berkeley. https://worldcat.org/title/671248487