Analysis of Airline Delay and Cancellation Data, 2009 – 2018

Paul Smith

Western Governors University

**Table of Contents**

[Project Overview 3](#_Toc127972428)

[A. Project Highlights 3](#_Toc127972429)

[1. Research Question 3](#_Toc127972430)

[2. Project Scope 3](#_Toc127972431)

[3. Solution Overview 3](#_Toc127972432)

[a. Tools 3](#_Toc127972433)

[b. Methodologies 3](#_Toc127972434)

[Project Plan 3](#_Toc127972435)

[B. Project Execution 3](#_Toc127972436)

[Methodology 3](#_Toc127972437)

[C. Data Collection Process 3](#_Toc127972438)

[C1. Advantages and Limitations of Data Set 3](#_Toc127972439)

[D. Data Extraction and Preparation Processes 3](#_Toc127972440)

[E. Data Analysis Process 4](#_Toc127972441)

[E1. Data Analysis Methods 4](#_Toc127972442)

[E2. Advantages and Limitations of Tools/Techniques 4](#_Toc127972443)

[E3. Application of Analytical Methods 4](#_Toc127972444)

[Results 4](#_Toc127972445)

[F. Project Success 4](#_Toc127972446)

[F1. Statistical Significance 4](#_Toc127972447)

[F2. Practical Significance 4](#_Toc127972448)

[F3. Overall Success 4](#_Toc127972449)

[G. Key Takeaways 4](#_Toc127972450)

[G1. Summary of Conclusions 4](#_Toc127972451)

[G2. Effective Storytelling 4](#_Toc127972452)

[G3. Findings-based Recommendations 4](#_Toc127972453)

[H. Panopto Presentation 5](#_Toc127972454)

[Appendices 5](#_Toc127972455)

[I. Evidence of Completion 5](#_Toc127972456)

[Sources 5](#_Toc127972457)

# 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

Discuss these elements; offer examples.

* Actual data selection vs. planned collection process
* Obstacles to data collection
* Unplanned data governance handling

# C1. Advantages and Limitations of Data Set

Include examples of both advantages and limitations.

# Data Extraction and Preparation Processes

Explain this process in terms of the tools and techniques used. Discuss why these processes were appropriate for your data.

# Data Analysis Process

# E1. Data Analysis Methods

Discuss the methods used for data analysis, including why these are appropriate for this project.

# E2. Advantages and Limitations of Tools/Techniques

Discuss the advantages and limitations of the tools and techniques used for data analysis.

# E3. Application of Analytical Methods

The submission includes a thorough step-by-step explanation of how the analytical methods were applied to the data and how *all* assumptions or requirements were verified.

# Results

# Project Success

# F1. Statistical Significance

A thorough evaluation of the statistical significance of the analysis is provided, and the evaluation uses accurate calculations.

# F2. Practical Significance

What do the results mean in practical terms? Offer examples.

# F3. Overall Success

Offer your view of the overall success and effectiveness of the project. Explain why you believe this.

# Key Takeaways

# G1. Summary of Conclusions

Present your conclusions.

# G2. Effective Storytelling

Include logical reasons why the chosen tools and graphical representations for visually communicating the findings support effective storytelling.

# G3. Findings-based Recommendations

Recommend 2 logical courses of action based on the analysis and findings. Directly address the research question or organizational need of the project.

# Panopto Presentation

Provide a link to your Panopto presentation. Include the following in your summary:

• a summary of your research question or organizational need

• a demonstration of the functionality of any code you used for your data analytics solution

• an outline of the findings and implications of your analysis

The summary should be appropriate for a data-analytics audience.

# Appendices

# Evidence of Completion

Submit at least 3 pieces of evidence related to the project.

# 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