Table of Contents

Exploratory Data Analysis (EDA) Project Outline: Analyzing Flight Data

Project Description:

In this project, students will work collaboratively to conduct an exploratory data analysis (EDA) on flight data using Python programming language in a jupyter notebook. The goal is to determine how Covid 19 impacted the domestic airline market.

Topics to Investigate:

Airfare

- How did Covid 19 impact 10 largest US airports in terms of passenger volume?

- How long did it take for passenger volume to normalize after Covid 19 at the 10 largest US airports?

- Did average airfare change at the 10 largest US airports during Covid?

- If it changed, when did it change as compared to when passenger volume changed?

- If it changed when did average airfare normalize?

- Is there a correlation between Covid19 and customer satisfaction?

- Covid19 is defined by the drop off in passenger volume.

- Is there a correlation between Covid19 and the number of flights?

Project Guidelines

Data Cleanup and Analysis

Now that you've picked your data, it's time to tackle development and analysis. This is where

the fun starts! The analysis process can be broken into two broad phases: (1) exploration and

cleanup, and (2) analysis.

As you've learned, you'll need to explore, clean, and reformat your data before you can begin answering your research questions. We recommend keeping track of these exploration and cleanup steps in a dedicated Jupyter notebook to stay organized and make it easier to present your work later.

After you've cleaned your data and are ready to start crunching numbers, you should track your work in a Jupyter notebook dedicated specifically to analysis. We recommend focusing your analysis on multiple techniques, such as aggregation, correlation, comparison, summary statistics, sentiment analysis, and time-series analysis. Don't forget to include plots during both the exploration and analysis phases. Creating plots along the way can reveal insights and interesting trends in the data that you might not notice if you wait until you're preparing for your presentation.
Presentation requirements will be further explained in the next module.

Presentation Guidelines

Each group will prepare a formal, 10-minute presentation that covers the following points.

- An executive summary or overview of the project and project goals:

  o Explain how the project relates to the industry you selected.

IMPORTANT

Whenever you use a dataset or create a new dataset based on other sources (such as existing datasets or information scraped from websites), make sure to use the following guidelines:

1. Check for copyright protections, and make sure that the way you plan to use this dataset is within the bounds of fair use.

2. Document how you intend to use this dataset now and in the future. Find any licenses or terms of use associated with the dataset and review them to confirm that your intended use is in compliance.

3. Investigate how the dataset was collected. Identify any indicators that the data was obtained from a source
   that the compilers were not authorized to access.

An overview of the data collection, cleanup, and exploration processes:

- Describe the source of your data and why you chose it for your

  project.

- Describe the collection, exploration, and cleanup process.

The approach that your group took to achieve the project goals:

- Include any relevant code or demonstrations of the application or analysis.

- Discuss any unanticipated insights or problems that arose and how you resolved them.

The results/conclusions of the application or analysis:

- Include relevant images or examples to support your work.
- If the project goal was not achieved, discuss the issues and how you attempted to resolve them.

  Next steps:

  o Briefly discuss potential next steps for the project.

It's crucial that you find time to rehearse before presentation day.
On the day of your presentation, each member of your group is required to submit the URL of your GitHub repository
for grading.

Project Team:

- Team Name: Group 6
- Team Members:
  1. Christoph Guenther
  2. Ava Lee
  3. Jed Murphy
  4. Silas Phillips

Project Timeline:

Duration: April 9th - April 22nd (3 weeks)

Agile Project Management:
- The project will be managed using Agile methodologies within GITHUB, with regular sprints and checkpoints to track progress, address challenges.
- In Class Stand-up Meetings: Each class, the team will have a brief stand-up meeting to discuss progress, obstacles, and plans for the evening
- We will communicate via slack huddles and zoom calls.

Project Steps:

1. Week 1: Planning and Setup (April 9th)

- Kick-off meeting (April 9th) to discuss project objectives, roles, and responsibilities.
- Set up GitHub repository for collaborative development.- https://github.com/silas-ph/Group_6_EDA
- Elevator pitch
- Assign roles to team members based on their expertise and interests.
- Research data sources. Make sure to use data sources available under either Creative Commons or Open Data Commons.

2. Week 2: Data Collection and Preprocessing (April 9th- April 14th)

- Complete research on data sources
- Refine project scope
- Extract relevant datasets and store them in the project repo.
- Clean and preprocess the data to handle missing values, outliers, and duplicates
- Collaborate to understand data requirements and formats.

- Write Python scripts/functions for data loading and preprocessing tasks.

3. Week 3: Exploratory Data Analysis (April 16th- April 21st)

Analysis:
- Explore the cleaned dataset to identify patterns, trends, and correlations.
- Perform descriptive statistics and generate summary insights.
- Visualization
  - Create visualizations (e.g., histograms, scatter plots, heatmaps) to visualize key aspects of the flight data.
- Use libraries like Matplotlib, Seaborn, and Plotly for visualization tasks.

4. Week 4: Documentation and Presentation (April 22nd)

Team
- Ensure proper version control and collaboration using Git and GitHub.
- Manage documentation including README.md, code comments.
- Coordinate final documentation and presentation preparations.
- Compile findings, insights, and visualizations into a comprehensive report.
- Presentation (April 22nd):
- Present the project findings, analysis process, and insights to the class.

Deliverables:

- GitHub repository with project code, data, documentation, and README.md.
- Final project report documenting the analysis process, findings, and insights.
- Presentation slides for presenting the project to the class.

Elevator Pitch

Title: Analyzing Flight Data

Value Proposition: We leverage the power of big data analytics to extract valuable insights from flight data, empowering consumers to optimize and enhance the passenger experience.

Key Areas of Focus:

- **Predictive Delay Mitigation:** Identify factors contributing to flight delays, enabling consumers to be proactive before purchasing the ticket.
- **Dynamic Fare Optimization:** By analyzing market trends and booking patterns, we help consumers develop data-backed pricing strategies to maximize the ticket buying experience.
- **Route Network Optimization:** We utilize flight data to assess network efficiency and identify opportunities for the consumer to make route adjustments to optimize the travel experience.

Impact: The goal of the analysis is to deliver these benefits to the passenger(s).

- Experience fewer delays
- Benefit from potentially lower fares
- Enjoy a more seamless travel journey

Exploratory Data Analysis (EDA) Project Guide: Analyzing Flight Data

Data Sources:-

1. Bureau of Transportation Statistics (BTS): Website Link
   - Provides extensive data on US flights, including on-time performance, delays, cancellations, and more.
2. Aviation Edge: Website Link
   - Offers a suite of APIs providing access to various aviation-related data, including flight schedules, tracking, airport, and aircraft data.
3. Aviationstack: Website Link
   - Provides a powerful flight data API offering real-time and historical flight information, including tracking, status, routes, and airline details.
4. FAA: Website Link
   - Offers a comprehensive dataset of airline routes, airports, and schedules freely available for download, useful for researchers and developers.
5. Transportation.gov: Website Link
   - Offers multiple data sets.

Exploratory Data Analysis Processes and Methods:

Data Cleaning and Preprocessing:

- Check for missing values, outliers, and duplicates.
- Handle missing data through imputation or deletion.
- Normalize or scale numerical features if necessary.
- Convert categorical variables into dummy variables if needed.
- [Data Cleaning with Python and Pandas: Detecting Missing Values](#)
- [Data Preprocessing in Python: How to Import CSV, Impute Missing Values](#)

Descriptive Statistics:

- Calculate summary statistics for numerical features.
- Explore the distribution of numerical features using histograms, box plots, or violin plots.
- Examine the frequency distribution of categorical features using bar plots or count plots.
- [Descriptive Statistics in Python with Pandas](#)

Exploring Relationships:

- Investigate pairwise relationships between numerical features using scatter plots or pair plots.
- Analyze correlations between features using correlation matrices or heatmaps.
- Explore relationships between categorical and numerical features using box plots or violin plots.
- [Exploratory Data Analysis (EDA) Techniques for Kaggle](#)

Time Series Analysis:

- Plot time series data to identify trends, seasonality, and outliers.
- Decompose time series data into trend, seasonal, and residual components.
- Conduct autocorrelation analysis to detect patterns within time series data.
- [Time Series Analysis in Python with Pandas](#)

Advanced Visualization Techniques:

- Utilize interactive visualizations for exploring complex datasets.

- Create geographical visualizations using maps to analyze spatial data.
- Use advanced plotting libraries like Plotly or Seaborn for more customized visualizations.
- [Advanced Data Visualization with Python](#)