Take Home Assignment: Data Analysis with Python

March 2025

Dataset

To complete this technical exercise, you will need to access the sample datasets for the Aviation Safety Information System (ASIS) available from the Transportation Safety Board of Canada website. There are five CSV files (Occurrence, Aircraft, Injuries, Events and Phases, Survivability) available to download here:

https://www.tsb.gc.ca/eng/stats/aviation/data-5.html

Instructions

Please create a single Python file named `analysis.py`. This file should exclusively utilize the following standard data science libraries: `pandas`, `matplotlib`, `scikit-learn`, and `numpy`. You may also use additional libraries that are part of a standard Python installation, such as `os` and `datetime`.

Your script must run independently, without the need for any external libraries or an Internet connection. All outputs generated by the script should be saved in a directory named `outputs` within your working directory. The contents of this `outputs` directory will be considered the results of your analyses.

In addition to the script, please provide a markdown file named `response.md`. This file should include text responses summarizing your answers to the selected questions, along with any graphical outputs from your analysis.

Your submission should be organized in a GitHub repository with the following structure (note that you may include other `.py` files if necessary, but `analysis.py` should be the primary executable script):

```
technical_exercise/
 |
 +-- analysis.py
 |
 +-- response.md
 |
 +-- outputs/
 |
 +-- data/
```

During your interview, you will have the opportunity to present your findings and approach in a 5-minute presentation.

Please note that no additional clarifying information will be provided.

Questions

Please provide answers for **any three of the six** following questions**.**

1. Understanding data quality is essential in data science, and data quality encompasses a variety of dimensions. Examine the Occurrence table and evaluate the completeness of the dataset. Use these

findings to provide recommendations about what types of analysis could be conducted with this data to better understand aviation accidents and safety incidents as reported in the ASIS data.

2. How common are aircraft collisions? Are there specific airports or regions that seem to have an unexpected risk associated with aircraft collisions?

3. Describe in detail how you would leverage the Summary field in the Occurrence data to predict future occurrences. You may consider use of additional libraries in your approach. What would the main challenges be, and what technical approaches would you take to overcome these? Demonstrate your approach with relevant data examples.

4. Develop a model that predicts the probability of surviving a safety incident. What are the key factors associated with survivability? What are the strengths and weaknesses of your analysis and what would your next steps be with this model?

5. What ICAO event categories are most common at Canadian airports? Is there any trend or pattern evident in these Canadian events?

6. Create a forecasting model to predict the number of incidents at an airport of your choosing for the 2025 calendar year. How would you validate your model?