**Team Project Report: Plane Crash Analysis**

| *Saphyra*  *Dulcio* | *Noah*  *Growney* | *Shrihari*  *Gunasekaran* | *Madhan*  *Saravanan* | *Janny*  *Valasquez* |
| --- | --- | --- | --- | --- |

**I. Report**

**Project Summary: Analysis of Plane Crash**

**Purpose:** The Analysis of Plane Crash project aimed to fill a critical void in aviation safety analysis by leveraging advanced data analytics techniques to understand and predict aviation crashes. This involved solving complex problems related to identifying trends and causal factors in historical plane crash data.

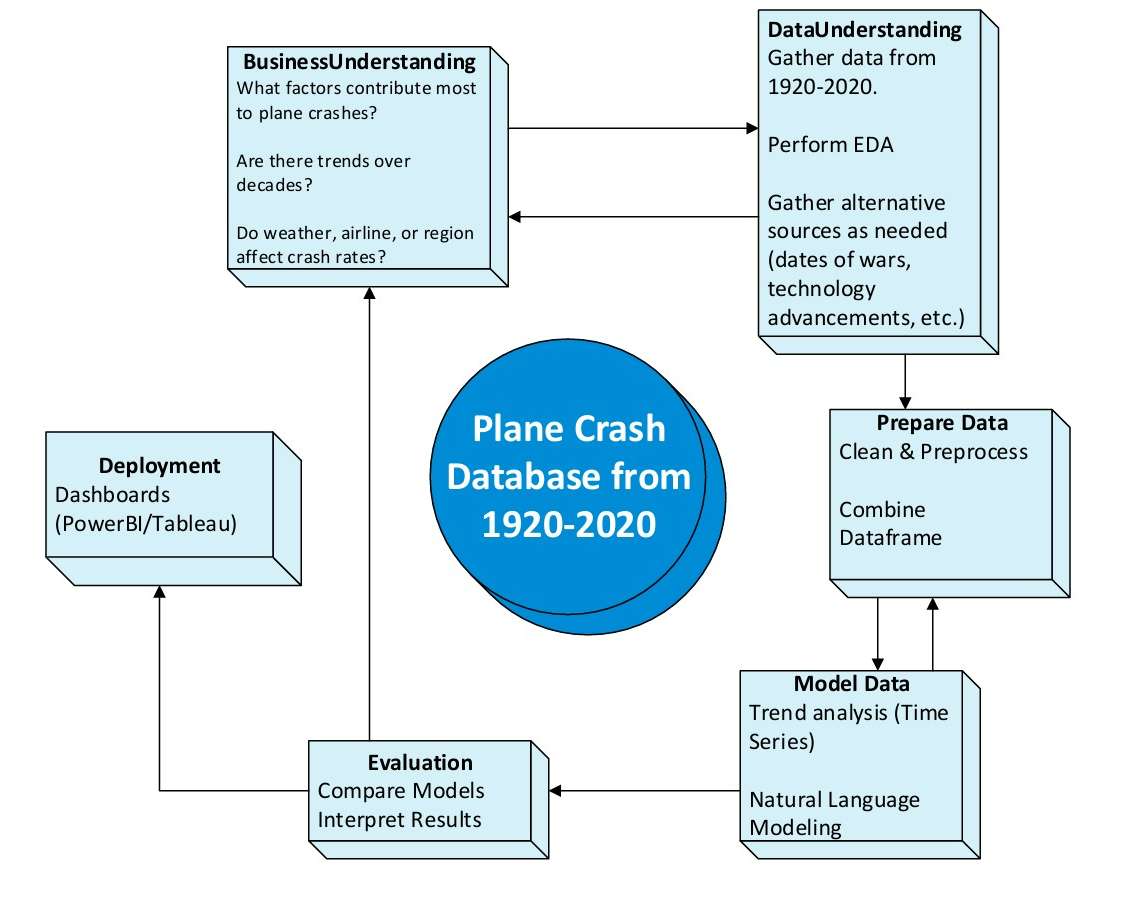
**Stakeholders and Business Understanding:** In the “problem identification and business understanding” phase of our project, our team was able to identify several key stakeholders who could potentially benefit from our analysis:

* **Airlines:** Airlines can utilize insights from the project to identify potential risk factors and trends, enabling them to implement more effective safety protocols and training programs. This proactive approach can lead to a reduction in accidents and improved passenger safety.
* **Military:** Military aviation units can enhance their operational safety measures by applying the findings from the project, ensuring better preparedness and risk mitigation during missions.
* **Regulatory Bodies:** Organizations such as the Federal Aviation Administration (FAA) and International Civil Aviation Organization (ICAO) can use the data to refine safety regulations and standards, promoting a safer aviation environment globally.
* **Aircraft Manufacturers:** Companies like Boeing and Airbus can leverage the analysis to improve aircraft design and manufacturing processes, addressing identified vulnerabilities and enhancing overall safety.

One of the key contributions of this project is its impact on aviation safety research. By providing a comprehensive analysis of historical crash data, the project offers valuable insights that help in identifying patterns and trends in aviation accidents, predicting potential future crashes, understanding the root causes of crashes, and developing more robust safety measures. These insights can be used to inform future safety measures and policies, allowing stakeholders to proactively address risks and ultimately reduce the incidence of aviation accidents.

By addressing these complex problems, the project is able to contribute in creating a safer aviation industry, benefiting all stakeholders involved and contributing to the overall goal of reducing aviation-related fatalities and incidents.

**II. High-Level Architecture**

****

This high-level architecture above represents our group’s approach to analyzing a comprehensive plane crash database covering the years 1920 to 2020. We begin with Business Understanding, where we defined the key questions that guide our analysis. Our goal was to identify the primary factors contributing to plane crashes, observe trends across different decades, and explore how elements like weather, airline, or region influence crash rates. This stage set the foundation for our entire project by ensuring that we remain focused on relevant and insightful objectives.

Next, we moved into the Data Understanding/Exploration phase, where we gathered and explored the historical data. We performed exploratory data analysis (EDA) and supplement our dataset with external sources such as wartime events or technological advancements to provide context. Once we understood the structure and scope of the data, we transitioned into theData Preparation stage. Here, we cleaned, preprocessed, and merged all data sources into a unified dataframe, preparing it for modeling.

In the Data Modeling phase, we applied techniques such as time series analysis to uncover long-term trends, and we incorporated natural language processing to analyze narrative crash reports. We then evaluated our models in the Evaluation phase, comparing different approaches and interpreting the results to determine the most meaningful insights.

Finally, we focused on Deployment, where we use tools like Power BI or Tableau to create dashboards that communicate our findings. These visualizations allow stakeholders to explore the data interactively and help inform decisions, completing our group’s analytical pipeline.

**III. Project Goals**

**Summary:** The Plane Crash Analysis Project explores documented aviation crashes from 1920-2020. By cleaning the data, exploring the data, utilizing Natural Language Processing (NLP), and predictive modeling, the project aims to uncover trends, root causes, and provide insights into how aviation has progressed throughout the years.

**Project Requirements**

This section of the report details the various project requirements:

**Functional Requirements**

These requirements detail what the system should do:

* + Data Cleaning and Aggregation. The project must clean and aggregate the data.
  + Crash Classifications: The project must classify the Plane Crashes into different categories.
  + Trend Analysis: The project must provide meaningful insights based on historical data.

**Performance Requirements**

These requirements detail how well the system should perform:

* + Predictive Models: The project must provide significant performance with predictive modeling.
  + Dashboard: The project must provide an interactive dashboard.

**Design Constraints**

These requirements detail how the project was built:

* + Technical: The project’s code must be written in Python or R, using various libraries.
  + Visual: The project team must visualize the data in Python, R, or Power BI.

**Quality Requirements**

These requirements detail the expectations around clarity:

* + Output: The project must have readable, well-organized outputs.

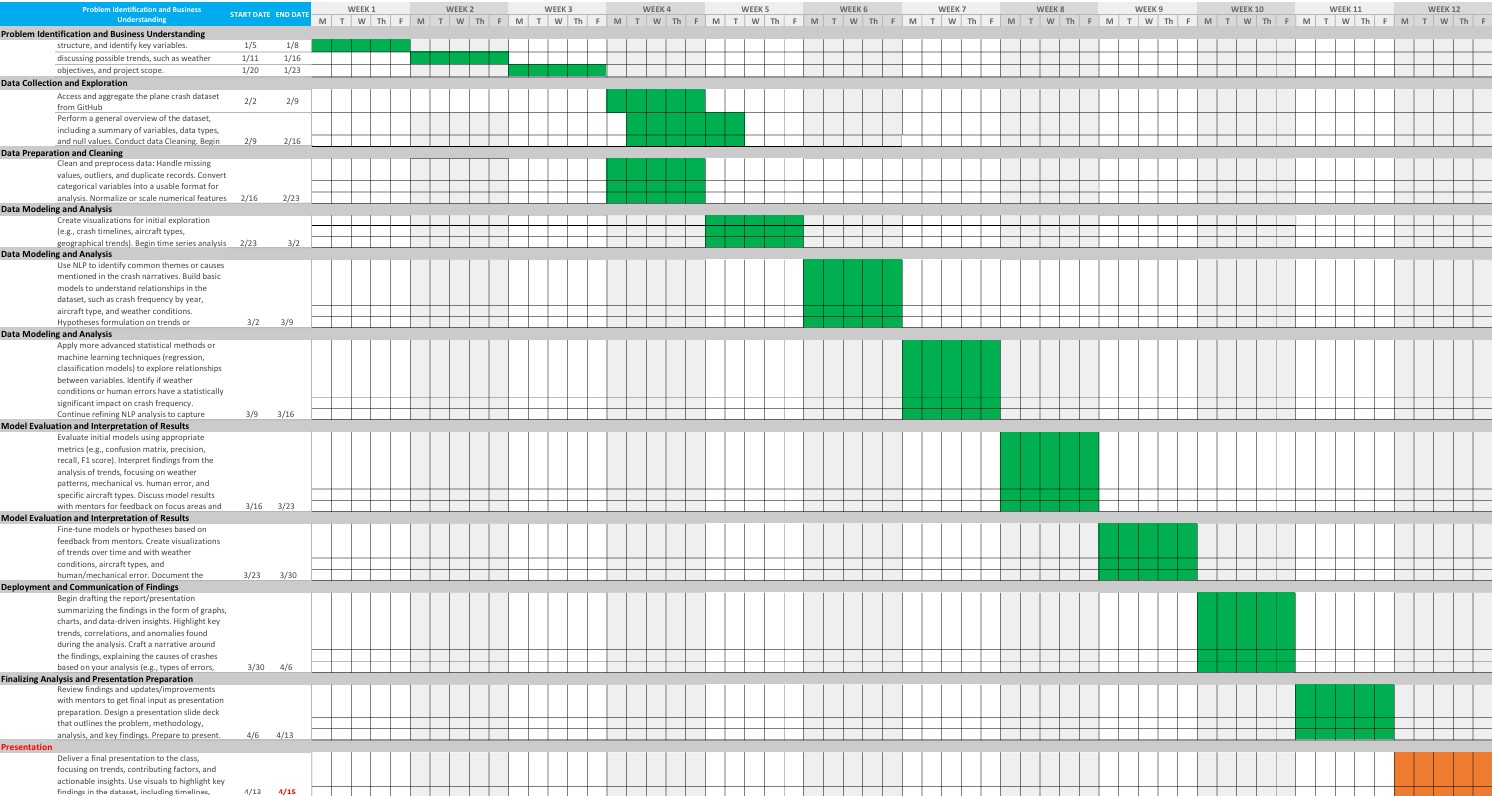
**Other Requirements**

This section details miscellaneous requirements:

* + Schedule: The project team must follow the schedule outlined in the Project Plan.
  + Final Project Report: The project team must provide a Final Project Report.
  + Presentation: The project team must present the findings.

**IV. Project Plan**

The project plan follows the High-Level Architecture of the project to ensure consistency and progress.



The Plane Crash Analysis Project began with problem identification and business understanding. The team worked to understand the project’s scope, which allowed them to focus on specific patterns, such as weather involvement, human error, mechanical error, and crash frequency. Weekly team meetings were established for consistent progress planning throughout the semester.

At the beginning of the data collection and exploration phase, the team reviewed the data and assessed the data structure. The dataset consisted of 100 separate files that required aggregation into a single format. After successfully aggregating the data, the team conducted an initial review of the dataset to explore key variables, identify data types, and highlight missing or inconsistent entries. Early exploratory data analysis (EDA) was used to uncover potential trends, such as crashes throughout time or variations by aircraft model.

The data preparation and cleaning stage focused on transforming the raw dataset. This included handling missing values. Feature engineering also took place to ensure the data was ready for analysis.

When the team moved into the data modeling and analysis portion, visual explorations such as crash timelines and word maps were created. These visual tools helped uncover broader patterns in the data. Natural Language Processing (NLP) techniques were applied to crash summaries to extract themes and keywords related to crash causes, such as "engine” "pilot" ,and "weather." This led the team to do further feature engineering to potentially gain insights into plane crash summary. This included creating new features, such as “manufacturer” and “Root Cause”. The team then sought to uncover potential relationships between the root causes of the crashes and other features of the data, such as total fatalities and aircraft manufacturer.

To determine the relationships between aircraft summaries and other features of the data, the team implemented multiple predictive modeling techniques in a variety of ways, including expanding summary keywords, and other feature manipulation methods. The team also explored several modeling techniques, such as logistic regression, decision trees, and random forests. These models were then tested based on accuracy and refined.

Finally, the team moved to the deployment and communication of findings phase of the project. Here is when the team created an interactive dashboard, using power BI. This dashboard included geographical graphs of the plane crashes, as well as summaries and insights into the data. The team also created a presentation and a two-minute video to share the project with peers.

**V. Risk Analysis**

| # | Risk | Prob | Impact | Problem | Possible Solutions | Why It Matters |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | Poor or Incomplete Data | High | High | 6.6% missing data and 5% duplicates. Inconsistent aircraft names (e.g., “Douglas DC-3” vs. “DC-3”) and unstructured crash summaries reduce analysis reliability. | Cleaned using Python (Pandas, Regex); could improve with NLP and fuzzy matching. Suggesting automated checks to catch errors early. | Bad data leads to wrong conclusions—similar to PMI warnings. Like construction errors from missing specs. Accurate data is essential. |
| 2 | Overusing Keywords Related to Older Aircraft | Medium | High | Older aircraft like Douglas and Curtiss dominate crash data due to historical use, making them seem more dangerous than modern aircraft. | Adjusting crash rates based on flight hours or active fleet size. Comparing aircraft across relevant time periods. | Using outdated data without context leads to misleading insights. Like judging modern software from 1990s bug reports—context matters. |
| 3 | AdaptableAI Models | Medium | High | Current model has around 71% accuracy and has issues with rare models like Ilyushin. Our code relies on TF-IDF keyword matching and other NLP methods | Upgrading to BERT or LLMs. Balancing training data. Using SHAP for model explainability. | Moderate models = some mistaken insights. AI misses deeper meanings. Better models improve reliability and decision-making. |
| 4 | Unreliable Forecasting | Medium | Medium | ARIMA models overfit to historical crashes and ignore emerging risks like drones, cyberattacks, and climate change. | Combining ARIMA with scenario planning to consider future risk scenarios such as increasing drone traffic. | Poor forecasting = planning risk. Like bad project plans, it causes delays and cost overruns. Scenario planning helps anticipate changes. |
| 5 | Language and Cultural Bias | Medium | Medium | Crash reports are mostly in English, causing underrepresentation of non-English incidents. Cultural differences affect reporting accuracy. | Incorporating multilingual reports. Using better translation tools. Analyzing for cultural bias. | Language gaps = misinterpretation. PMI links this to safety issues in global projects. Multilingual data = more inclusive analysis. |

**VI. MVP -Minimum Viable Product**

As part of our Plane Crash analysis project, we developed a Minimum Viable Product (MVP) that uses data science and Natural Language Processing (NLP) to better understand and predict airplane crash factors. Our goal was to show how machine learning, text analysis, and statistical modeling can be helpful in improving aviation safety and help decision-makers take action based on data. These features were created to help airlines, regulators, and safety analysts quickly spot risks and get useful insights from large volumes of crash data.

**What the MVP provides...**

The MVP focused on four main tasks:

* **Time Series Forecasting**: We used models like ARIMA and STL decomposition to analyze crash trends over time. This helped us understand patterns and seasonal factors in accident rates.
* **Crash Report Classification**: Using simple NLP models like TF-IDF and logistic regression, we tried to predict the aircraft manufacturer based on crash descriptions.
* **Topic Extraction**: We used keyword extraction and Named Entity Recognition (NER) to find common causes in crash reports, like “engine failure” or “pilot error.”
* **Interactive Dashboard Creation**: Consolidating inferences and data predictions into better visualization using Power BI, which provides stakeholders a clear distinct picture, understanding of risks and key factors related to Plane crash Analysis .

| **Hypothesis** | **Result** | **Notes** |
| --- | --- | --- |
| Crash rates have changed over time due to safety improvements | Satisfied | The time series models showed a drop in crash numbers over the years. |
| NLP can classify crash reports by manufacturer | Almost Satisfied | It worked, but the accuracy was low. We’d need better models like BERT in future versions. |
| NLP can extract causes from crash text | Satisfied | The NLP tools successfully found key causes and patterns in the reports. |
| The tool has real-world use for airlines or regulators | Satisfied | Based on our business model canvas, the idea fits well into the aviation safety ecosystem. |
| Crash prediction tools can help in prevention | Satisfied | Forecasting models can help build early warning systems to avoid future accidents. |

**VII. Final contemplation**

**Address any issues that you had in developing your MVP. Did you complete all  
of what you expected to accomplish?**

Developing the Minimum Viable Product (MVP) for my CAP6942 project was both a challenging and rewarding experience. One of the primary issues We encountered was balancing the scope of the project with the time and resources available. One of the key challenges we faced was coordinating consistent meeting times in an online environment. This inconsistency in communication sometimes led to confusion around deadlines and task delegation. Establishing a fixed schedule and clearer accountability from the beginning would have likely improved our overall productivity and team cohesion

**What could have made your project more meaningful and your results more  
accurate?**

In terms of project outcomes, we successfully modeled key aspects of the manufacturer's side of the aviation data. However, our model may not fully represent the complete picture. For instance, while we could infer that more planes in the air might lead to more accidents, the current dataset lacked the depth and scale needed to validate this hypothesis. Access to broader, more detailed datasets—including air traffic volume, airline policies, and flight durations would have helped improve the accuracy and reliability of our results.

**What would be appropriate next steps for your project?**

To enhance the project moving forward, the first priority would be to cross-reference our dataset with verified real-world sources to resolve inconsistencies and potential biases. For instance, while U.S. crash data often included detailed state-level information, international entries typically listed only city and country, creating a disparity that may affect the accuracy of analysis. Additionally, exploring alternative modeling techniques and machine learning approaches could help improve prediction accuracy and adaptiveness of the model. A longitudinal Natural Language Processing (NLP) analysis on aviation incident reporting over the years could also provide meaningful trends and context. Further steps would include optimizing the current model for performance, expanding the simulation to accommodate larger and more complex datasets, and enhancing the visualization to make insights more intuitive. Developing an interactive interface would not only improve usability but also support deeper exploration of the data by end-users.

**What did you learn from your CAP6942 experience?**

Through this project and the broader CAP6942 experience, we gained valuable hands-on skills in model design, algorithm development, and data analysis, along with a deeper appreciation for the iterative nature of research and experimentation. We also learned the importance of clear documentation and communication when building computational models. Overall, this experience has strengthened our problem-solving abilities and has better prepared us for tackling real-world challenges in data science and analytics.