A Project Report on

**Air Fly Insights: Data Visualization and Analysis of Airline Operations**

For the Partial Fulfillment of

Infosys Spring board 6.0 Internship

Submitted by

**Valiveti K T Rama Tulasi Lakshmi**

*Reg No: 22NE1A12C4*

**Executive Summary: Air-Fly Insights - Flight Data Analysis Project**

The **Air-Fly Insights** project is a comprehensive data visualization and analysis initiative focused on examining large-scale airline operational data to uncover meaningful patterns in flight operations, delays, and cancellations. This project represents a systematic approach to airline industry analytics using modern data science techniques and visualization methodologies.

**Project Scope and Objectives**

The primary objective is to analyse extensive airline flight data to identify operational trends, delay patterns, and cancellation reasons through advanced data visualization techniques. The project aims to deliver actionable insights for stakeholders including airline operators, airport management, and aviation analysts to optimize operational efficiency and improve passenger experiences.

**Dataset and Scale**

The analysis utilizes the Kaggle Airlines Flights Data, containing over 484,000 flight records with 29 key operational variables including departure/arrival times, delays, cancellations, routes, and carrier information. The dataset encompasses comprehensive flight operations data from January through June 2019, providing substantial depth for trend analysis and operational insights.

**Technical Implementation and Methodology**

The project follows a structured 8-week implementation approach divided into four major milestones:

**Data Foundation and Preprocessing**: Comprehensive data cleaning involving memory optimization techniques that reduced dataset memory usage from 334.41 MB to 40.17 MB through strategic data type optimization and categorical encoding. Missing values were systematically handled using domain-specific logic, with delay columns filled with zeros (indicating no recorded delays) and categorical fields appropriately managed.

**Feature Engineering**: Creation of derived variables including temporal features (month, day of week, hour), route combinations, and enhanced datetime processing to enable comprehensive temporal analysis.

The preprocessing pipeline includes sophisticated datetime handling for flight schedules and overnight flight adjustments.

**Analysis Framework**: Multi-dimensional analytical approach covering univariate and bivariate analysis, delay cause analysis, seasonal trends, cancellation patterns, and route-level performance metrics.

**Key Technical Achievements**

* **Memory Optimization**: Achieved 88% reduction in memory usage through intelligent data type casting and categorical encoding
* **Data Quality Enhancement**: Implemented robust missing value handling strategies specific to aviation domain requirements
* **Temporal Data Processing**: Developed sophisticated datetime conversion functions to handle HHMM format time data and overnight flights
* **Scalable Analysis Framework**: Created reusable preprocessing functions and sampling strategies for large-scale data analysis

**Expected Deliverables and Impact**

The project delivers a comprehensive suite of visualizations including bar charts, time series analysis, heat maps, route maps, and performance comparisons. Key analytical outputs include identification of delay-prone carriers, seasonal operation patterns, route congestion analysis, and cancellation trend insights.

**Technology Stack and Tools**

The implementation leverages a robust data science ecosystem including pandas and numpy for data handling, matplotlib, seaborn, and plotly for visualization, with optional dashboard capabilities through PowerBI. All analysis is documented through Jupyter notebooks with comprehensive GitHub repository organization.

**Strategic Value Proposition**

This project provides aviation industry stakeholders with data-driven insights to optimize flight operations, improve on-time performance, reduce operational costs, and enhance passenger satisfaction through better understanding of delay patterns and operational bottlenecks. The analytical framework can be extended to support real-time operational decision-making and long-term strategic planning initiatives.

**Module 1 Progress**

**Data Acquisition and Understanding**

The **Week 1: Data Acquisition and Understanding** module has been successfully completed with significant accomplishments across all planned objectives. This foundational phase established a robust data infrastructure for subsequent analysis phases of the Air-Fly Insights project.

**Module Objectives Completion**

**1.Project Initialization and Dataset Setup**

**Dataset Loading and Schema Analysis**: Successfully loaded the comprehensive flight dataset containing 484,551 flight records spanning January through June 2019 with 29 core operational variables. The dataset encompasses essential aviation metrics including departure/arrival times, delays, cancellations, carrier information, routes, and operational performance indicators.

**Data Structure Assessment**: Conducted thorough exploration of the dataset schema, identifying data types, null value patterns, and structural characteristics. The analysis revealed critical insights about data completeness, with minimal missing values in airport name fields (1,177 in OrgAirport and 1,479 in DestAirport) and complete data coverage across all operational metrics.

**2.Memory Optimization Achievements**

**Advanced Memory Management**: Implemented sophisticated memory optimization techniques achieving **88% reduction** in memory usage from 334.41 MB to 40.17 MB. This optimization involved strategic data type casting, converting integer columns from int64 to appropriate smaller formats (int8, int16, int32) and transforming categorical string fields to efficient category data types.

**Performance Enhancement**: The memory optimization framework included custom functions for automated dtype mapping and memory reporting, establishing a scalable foundation for handling large-scale aviation datasets. These optimizations enable efficient processing of the full dataset without memory constraints.

**3.Data Quality and Preprocessing**

**Comprehensive Missing Value Analysis**: Systematically addressed missing values using domain-specific logic appropriate for aviation data. Delay-related columns were filled with zeros (indicating no recorded delays), categorical airport fields were handled with "Unknown" placeholders, and cancellation codes were appropriately managed based on flight status.

**Data Cleaning Validation**: Successfully eliminated duplicate records (reducing from 484,551 to 484,549 entries) and implemented robust data validation procedures. All preprocessing steps were documented with clear rationale and verification checks.

**4.Advanced Datetime Processing**

**HHMM Format Conversion**: Developed sophisticated datetime processing functions to handle aviation time formats, converting HHMM integer representations to proper datetime objects. This included handling edge cases such as missing values, proper zero-padding, and time zone considerations.

**Temporal Feature Engineering**: Created comprehensive temporal features including departure hour, minute, month, day of week, and route combinations. Successfully handled overnight flights with automatic date adjustment logic for cross-midnight arrivals.

**Full Datetime Integration**: Established complete datetime columns for departure, arrival, and scheduled arrival times, enabling temporal analysis capabilities for subsequent analytical phases.

**5.Feature Engineering and Enhancement**

**Derived Variables Creation**: Generated essential analytical features including route strings (Origin-Destination pairs), temporal components (month, hour, day of week), and enhanced datetime processing. These features provide the foundation for trend analysis and operational insights.

**Data Export and Documentation**: Successfully saved preprocessed datasets in multiple formats with comprehensive documentation of transformation logic. Created feature dictionaries and preprocessing summaries for project continuity and reproducibility.

**Technical Deliverables Achieved**

* **Cleaned Dataset**: Production-ready dataset with optimized memory usage and complete data validation
* **Preprocessing Logic Documentation**: Comprehensive documentation of all data transformation procedures
* **Feature Dictionary**: Complete mapping of original and derived variables with descriptions
* **Memory-Optimized Functions**: Reusable code framework for efficient large-scale data processing
* **Temporal Processing Pipeline**: Advanced datetime handling system for aviation time formats