| **Project Title** | **Project Title: U.S. Airline Performance & Delay Analysis ✈️📊** |
| --- | --- |
| **Skills take away From This Projec** | Welcome to your capstone project! This project challenges you to analyze a large dataset of U.S. domestic flights to uncover patterns in delays and cancellations, evaluate airline and airport performance, and provide data-driven insights.   * Advanced SQL Querying (joins, aggregations, filtering complex data) * Business Intelligence with Power BI/Tableau (dashboard creation, data storytelling) * Key Performance Indicator (KPI) Development & Analysis * Data Cleaning, Transformation, and Integration * Exploratory Data Analysis (EDA) * Performance Benchmarking * Analytical Report Writing & Presentation |
| **Domain** | Aviation Analytics / Transportation Management |

# **Problem Statement:**

Flight delays and cancellations are significant issues in the U.S. aviation industry, impacting passengers, airlines, and the economy. Your task is to perform an in-depth analysis of historical flight data to identify key drivers of these disruptions, assess performance, and propose actionable insights.

# **Business Use Cases:**

The project's solution can be implemented in various scenarios, including:

* Analyzing the primary causes and patterns of flight delays and cancellations for strategic operational adjustments.
* Benchmarking the on-time performance, delay severity, and cancellation rates of different airlines for competitive analysis and improvement initiatives.
* Evaluating the operational performance of various U.S. airports to identify bottlenecks and areas for infrastructure or process improvement.
* Investigating how factors like time of day, day of week, month, and specific routes affect flight operations to optimize scheduling and resource allocation.
* Providing data-driven recommendations to stakeholders (airlines, airports, regulatory bodies) to enhance passenger experience and operational efficiency.

# **Approach:**

**Phase 1: Project Setup & Data Ingestion**

* Set up your SQL database environment.
* Create appropriate table structures for flights, airlines, and airports data. Consider data types carefully.
* Import the data from the CSV files into your SQL database.
* Perform initial verification (e.g., row counts, sample data review) to ensure data is loaded correctly.

**Phase 2: Data Cleaning, Preparation & Integration (SQL)**

* **Time/Date Handling:** Convert text-based time and date fields (e.g., SCHEDULED\_DEPARTURE) into proper datetime formats. This may involve creating new columns. Think about how to combine year, month, day, and HHMM time strings.
* **Missing Values:** Investigate and handle missing values (NULLs) in key columns (e.g., delay columns, cancellation reason). Decide on an appropriate strategy (e.g., replace with 0 where logical, leave as NULL, or flag).
* **Data Enrichment:**
  + Create a more descriptive CANCELLATION\_REASON\_DESC column based on the CANCELLATION\_REASON code.
  + Consider creating a FLIGHT\_DATE column for easier time-based analysis.
* **Integration:** Create a unified analytical dataset (e.g., a comprehensive SQL View or a new table) by joining flights data with airlines and airports (for both origin and destination) to include descriptive names and locations.

**Phase 3: Exploratory Data Analysis (EDA) & KPI Definition (SQL)**

* Using SQL queries on your integrated dataset, explore:
  + Overall flight volumes, cancellations (total, by reason), and diversions.
  + Basic statistics for departure and arrival delays (average, median, min, max).
  + The distribution of different types of delays (airline, weather, NAS, etc.).
* **Define Key Performance Indicators (KPIs)** relevant to the project objectives. Examples include:
  + On-Time Performance (OTP) Rate (e.g., flights arriving within 15 minutes of schedule).
  + Average Arrival/Departure Delay (in minutes).
  + Cancellation Rate (%).
  + Percentage contribution of each delay type.
* Perform initial aggregations to understand how these KPIs vary by:
  + Airline
  + Origin/Destination Airport
  + Month, Day of Week, Time of Day (requires robust time parsing)

**Phase 4: Dashboard Development (Power BI / Tableau)**

* Connect your BI tool to your SQL database (or the integrated view/table).
* **Data Modeling in BI Tool:**
  + Define necessary relationships if you load multiple tables.
  + Create calculated measures/fields required for your KPIs (e.g., Total Flights, OTP Rate, Avg Delay, Cancellation Rate). *You will need to translate your KPI definitions into the BI tool's formula language (DAX for Power BI, or Tableau's functions).*
* **Design and build your interactive dashboard. Consider including pages/views for:**
  + **Overview:** Key KPIs, overall delay/cancellation reasons.
  + **Airline Performance:** Comparative analysis of airlines against KPIs.
  + **Airport Performance:** Comparative analysis of airports; consider a map visualization.
  + **Temporal Trends:** How KPIs change by month, day of week, or hour.
* Ensure your dashboard includes filters/slicers for interactive exploration (e.g., by airline, airport, date).

**Phase 5: Insight Generation & Recommendation Formulation**

* Analyze your dashboards and SQL query results deeply. Ask "why?" and "so what?".
* Identify significant trends, patterns, anomalies, and correlations.
* Based on your findings, formulate 2-3 actionable and data-driven recommendations for potential stakeholders (e.g., airlines, airport authorities). Consider what changes could realistically lead to improvements.

**Phase 6: Reporting & Presentation**

* **Final Report:** Structure your report logically (Introduction, Methodology, Key Findings & Analysis, Dashboard Overview, Recommendations, Conclusion). Use visuals from your dashboard to support your findings.
* **Presentation:** Prepare a concise and engaging presentation summarizing your project, highlighting key insights, and showcasing your dashboard.

# **Results:**

* Analyze the primary causes and patterns of flight delays and cancellations.
* Benchmark the on-time performance, delay severity, and cancellation rates of different airlines.
* Evaluate the operational performance of various U.S. airports.
* Investigate how factors like time of day, day of week, month, and route affect flight operations.
* Translate your findings into meaningful recommendations for stakeholders.

# **Project Evaluation metrics:**

* **SQL Proficiency:** Correctness, efficiency, and clarity of your SQL scripts.
* **Data Handling:** Accuracy and completeness of your data cleaning and preparation.
* **Dashboard Effectiveness:** Clarity, usability, insightfulness, and interactivity of your Power BI/Tableau dashboards.
* **Analytical Depth:** Actionability and depth of the insights derived from your analysis.
* **Report & Presentation Quality:** Professionalism, clarity, and comprehensiveness of your final deliverables.

# **Technical Tags:**

* **Primary:** SQL, Power BI (or Tableau), Microsoft Excel (for supplementary tasks).
* **Minimal/Optional:** Python (e.g., for initial bulk data handling if preferred, but not a core requirement for analysis).

# **Data Set:**

* **Source:** Google Drive: [2015 Flight Delays and Cancellations](https://drive.google.com/file/d/1_Mt-OR_IxoIy7HVkvD4bW0-fm4MRY6og/view?usp=drive_link)
* **Format:** CSV (Comma Separated Values)

# **Data Set Explanation:**

* **Content:** This dataset contains detailed records of domestic flights in the USA for 2015, including scheduled/actual times, delays, cancellations, airlines, and airports.
* **Key Files:**
  + flights.csv: Main flight transaction data.
  + airlines.csv: Lookup table for airline names.
  + airports.csv: Lookup table for airport details (including location).
* **Important Columns to Explore (in flights.csv):** YEAR, MONTH, DAY, DAY\_OF\_WEEK, AIRLINE, FLIGHT\_NUMBER, ORIGIN\_AIRPORT, DESTINATION\_AIRPORT, SCHEDULED\_DEPARTURE, DEPARTURE\_TIME, DEPARTURE\_DELAY, SCHEDULED\_ARRIVAL, ARRIVAL\_TIME, ARRIVAL\_DELAY, CANCELLED, CANCELLATION\_REASON, AIR\_SYSTEM\_DELAY, SECURITY\_DELAY, AIRLINE\_DELAY, LATE\_AIRCRAFT\_DELAY, WEATHER\_DELAY, DISTANCE.

**Key preprocessing steps required include:**

* **Handling Time and Date:** Converting string representations of dates and times (e.g., "0855" for HHMM) into proper datetime formats to enable temporal analysis. This often involves combining year, month, day, and time components.
* **Managing Missing Values:** Addressing NULLs in critical columns such as delay durations (which might be NULL if not delayed or if cancelled) and cancellation reasons.
* **Data Type Conversion:** Ensuring all columns are of appropriate data types for analysis.
* **Data Enrichment:** Creating new descriptive columns, for example, mapping CANCELLATION\_REASON codes (A, B, C, D) to their textual descriptions (Airline/Carrier, Weather, etc.).
* **Data Integration:** Joining the flights.csv data with airlines.csv and airports.csv to incorporate descriptive names for airlines and airports (origin and destination), as well as airport geographical information (city, state, latitude, longitude) into a unified analytical dataset or view.

# **Project Deliverables:**

1. **SQL Scripts Package:** Well-commented SQL scripts used for data cleaning, transformation, integration, and all analytical queries performed to derive insights.
2. **Interactive Dashboard (Power BI/Tableau):** The BI tool file (e.g., .pbix for Power BI, .twbx for Tableau) containing a multi-page dashboard that visually presents the analysis, key findings, and allows for interactive exploration.
3. **Final Project Report (Word/PDF Document):** A comprehensive document detailing the project's introduction (problem statement, objectives), methodology (data sources, tools, analytical approach), key findings and analysis (supported by charts/tables), a walkthrough of the dashboard, actionable recommendations, and a conclusion with any limitations.
4. **Presentation (PowerPoint/Google Slides):** A concise slide deck summarizing the project's objectives, methodology, most impactful findings, dashboard highlights, and key recommendations, suitable for presenting to stakeholders.
5. Recorded Video

# **Project Guidelines:**

* Write clean, well-commented, and reasonably efficient SQL code.
* Clearly document all data cleaning and transformation decisions and steps.
* Design dashboards that are intuitive, informative, visually appealing, and effectively tell a story with the data.
* Focus on deriving actionable recommendations that are clearly supported by your data analysis.
* Ensure data consistency and validation at various stages of processing and analysis.
* Use version control (e.g., Git) for managing SQL scripts and documentation if you are comfortable with it, though not mandatory.
* Manage your time effectively across the different project phases to meet deadlines.

# **Timeline:**

The project is expected to be completed following a phased approach.

* **Phase 1: Project Setup & Data Ingestion** (e.g., 1-2 Days)
* **Phase 2: Data Cleaning, Preparation & Integration** (e.g., 1-2 Days)
* **Phase 3: Exploratory Data Analysis (EDA) & KPI Definition** (e.g., 2 Days)
* **Phase 4: Dashboard Development & Refinement** (e.g., 2 Days)
* **Phase 5: Insight Generation & Recommendation Formulation** (e.g., 1 Day)
* **Phase 6: Final Report Writing & Presentation Preparation** (e.g., 1 Day)

Deadline → 7 days

# Phase 5: Insight Generation

1. Alaska Airlines leads in total flights  
 Alaska Airlines operated the highest number of flights, showcasing its dominant presence in U.S. domestic airspace.

2. Seattle-Tacoma (SEA) is the busiest origin airport  
 SEA airport had the most departures, confirming it as a major hub in the western United States.

3. Weather & Late Aircraft are top delay causes  
 Delay analysis shows that late arriving aircraft and weather delays were the most frequent, indicating scope for schedule optimization and better forecasting.

# Phase 6: Recommendations

1. Optimize Aircraft Rotation  
 Reduce cascading delays by allocating buffer time for aircraft turnaround at major airports.

2. Weather-Responsive Scheduling  
 Improve delay prediction using historical weather data, especially for regions like SEA and DEN.

3. Focus on Evening Flights  
 Delays peak in late evening slots — reschedule or add spare capacity to improve on-time performance.