**PYTHON PROJECT REPORT**

(Project Semester: January-April 2025)

**Title of the Project: UK Train Rides**

**Submitted by:**

**Ankit Kumar  
Registration No.: 12312618  
Programme and Section: B.Tech CSE (K23FD)  
Course Code: INT375**

**Under the Guidance of:  
Baljinder Kaur (UID : 27952)**

**Discipline of CSE/IT**  
**Lovely School of Computer Science & Engineering**  
**Lovely Professional University, Phagwara**

**DECLARATION**

I, **Ankit Kumar**, student of **Bachelors of Technology (B.Tech)** under CSE/IT Discipline at Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 03-April-2025

Signature:   
Registration No.: 12316062  
Name of the Student: **Ankit Kumar**

# ****CERTIFICATE****

This is to certify that **Ankit Kumar** bearing Registration No. **12312618** has completed **INT375** project titled **“UK Train Rides Analysis”** under my guidance and supervision. To the best of my knowledge, the present work is the result of her original development, effort, and study.

**Baljinder Kaur**  
**Assistant Professor**  
**School of Computer Science & Engineering**

**Lovely Professional University**  
**Phagwara, Punjab**

Date: **04-April-2025**

**ACKNOWLEDGMENT**

I would like to express my sincere gratitude to **Baljinder Kaur Ma’am**, my project guide, for their invaluable support, guidance, and encouragement throughout the development of this project. Their expert insights and constructive feedback have been instrumental in shaping the project's outcome.

I am also thankful to **Lovely Professional University** for providing a conducive learning environment and access to resources that made this project possible. Additionally, I extend my appreciation to my professors and peers for their continuous motivation and insightful discussions, which greatly enhanced my understanding of the subject.

Lastly, I would like to acknowledge the unwavering support of my family and friends, whose encouragement has been a source of inspiration throughout this journey.

# ****TABLE OF CONTENTS****

1. [Introduction](#_1._INTRODUCTION)
2. [Source of Dataset](#_2._SOURCE_OF)
3. [Dataset Preprocessing](#_3._DATASET_PREPROCESSING)
4. [Analysis on Dataset (for each objective)](#_4._ANALYSIS_ON)
   * [i. General Description](#_i._General_Description)
   * [ii. Specific Requirements](#_ii._Specific_Requirements)
   * [iii. Analysis Results](#_iii._Analysis_Results)
   * [iv. Visualization](#_iv._Visualization)
5. Screenshot
6. [Conclusion](#_5._CONCLUSION)
7. [Future Scope](#_6._FUTURE_SCOPE)
8. [References](#_7.REFERENCES)

# ****1. INTRODUCTION****

In the modern landscape of digital transformation, data has become an invaluable asset across industries — including transportation. Railways, being one of the most vital modes of travel for millions of people, generate vast amounts of data on a daily basis. Leveraging this data effectively can lead to significant improvements in operational efficiency, passenger satisfaction, and service delivery. This project, titled *“Railway Passenger Data Analysis and Satisfaction Insights”*, focuses on analyzing passenger-related data through the powerful lens of Exploratory Data Analysis (EDA) using Python.

The primary objective of this project is not just to manage and organize travel records, but to extract meaningful insights from them. Through careful examination of the dataset — which includes features such as passenger demographics, travel routes, ticket fares, and feedback on services — we aim to understand travel patterns, service gaps, and satisfaction levels among passengers. Python’s robust ecosystem of libraries like Pandas, Matplotlib, and Seaborn enables efficient data preprocessing, visualization, and statistical exploration, making it an ideal choice for this analysis.

Key goals of this project include:

* Reading and cleaning railway passenger data for consistency and accuracy.
* Performing descriptive analysis to explore key trends in passenger profiles and journey preferences.
* Identifying popular travel classes, high-demand routes, and common passenger demographics.
* Analyzing service feedback to evaluate satisfaction levels regarding cleanliness, punctuality, food, and staff behavior.
* Visualizing key metrics and relationships between variables to discover hidden patterns in the data.
* Supporting railway management in decision-making through data-driven insights and actionable recommendations.

Unlike traditional data systems that simply store information, this project emphasizes deeper data interpretation. By applying EDA techniques, we aim to explore beyond the surface — uncovering the factors that influence travel preferences and satisfaction, detecting anomalies in feedback, and understanding overall passenger behavior. These insights are not only valuable for railway service providers, but also for students, data analysts, and policymakers interested in enhancing travel experiences.

This system is especially useful for:

* Transport authorities seeking ways to improve passenger experience based on actual feedback.
* Data science learners aiming to apply EDA to real-world public transport datasets.
* Analysts and managers who want to visualize and interpret passenger behavior for strategic planning.

In conclusion, this Railway Passenger Data Analysis project demonstrates the fusion of Python programming and data analytics to uncover meaningful stories hidden within transport data. By utilizing EDA, we transform raw travel records into actionable knowledge — driving informed decisions and paving the way for smarter, more efficient railway systems.

# ****2. SOURCE OF DATASET****

**The dataset used in this project was obtained from the “**[**Free Data Sets & Dataset Samples | Maven Analytics**](https://mavenanalytics.io/data-playground)**” – a platform known for providing clean, structured, and high-quality datasets for data science and analytics projects. The specific dataset used here simulates detailed information about railway passengers, including demographics, journey details, travel classes, ticket fares, and service feedback.**

**This dataset has been chosen to explore various aspects of passenger travel behavior and satisfaction, using Python’s data analysis and visualization tools.**

**Rationale for Choosing This Dataset**

**This dataset was selected because it offers a well-rounded view of passenger journeys and allows us to analyze:**

* **Passenger Travel Patterns across different classes, routes, and distances.**
* **Impact of Demographics (such as age and gender) on travel preferences.**
* **Satisfaction Levels based on cleanliness, staff behavior, punctuality, and food quality.**
* **Operational Insights by studying travel trends across source-destination pairs.**
* **Data-driven Improvements for passenger experience and service management.**

**This makes it highly suitable for applying Exploratory Data Analysis (EDA) techniques to uncover meaningful insights.**

**Preprocessing and Enrichment Steps**

**To make the dataset ready for analysis, the following preprocessing steps were performed using Python:**

* **Data Cleaning:**
  + **Checked for and handled missing or inconsistent values.**
  + **Standardized column names and data types for easier handling.**
* **Date & Time Formatting:**
  + **If any time-related data (like journey time or date) was present, it was formatted using Python’s datetime tools.**
* **Feature Engineering:**
  + **Created new columns such as:**
    - **Journey Duration (if not given explicitly)**
    - **Fare Per KM**
    - **Feedback Score Average (by combining service ratings)**
* **Categorical Mapping:**
  + **Grouped travel classes and station codes under simplified labels for better segmentation.**
* **Data Structuring:**
  + **Organized data into tidy format, making it easier to use in grouped analysis, pivot tables, and dashboards.**

**Why This Dataset Works Well for EDA**

**The structured nature of this railway dataset makes it a perfect fit for applying Exploratory Data Analysis. Some key benefits include:**

* **Visual Exploration of Passenger Trends**
  + **Understand which classes or routes are most used, and how demographics affect preferences.**
* **Feedback Analysis**
  + **Study how passengers rate various services and identify problem areas.**
* **Segment-wise Comparison**
  + **Compare travel experiences by gender, class, route, or age group.**
* **Anomaly Detection**
  + **Spot unusual fares or inconsistent ratings that may indicate data or service issues.**
* **Comprehensive Overview**

**3. DATASET PREPROCESSING**

To ensure the dataset was suitable for meaningful analysis, a structured data preprocessing phase was conducted. The raw dataset, sourced from the Maven Analytics Data Playground, contained individual-level passenger data for railway journeys, including attributes such as passenger demographics, travel details, ticket pricing, and service feedback. An initial exploratory review of the dataset was performed to understand its structure, format, and completeness, which revealed certain inconsistencies, missing entries, and formatting issues that required cleaning and correction.

The first step in preprocessing involved addressing missing or incomplete data. A detailed inspection was carried out to identify null values or unexpected gaps. Depending on the nature of each field, appropriate imputation methods were used. For instance, missing feedback ratings (like cleanliness, punctuality, or staff behavior) were filled with the mean or median score to preserve the overall distribution. If non-critical records had excessive missing fields and offered little analytical value, those rows were safely removed from the dataset to maintain data quality.

Next, a thorough data cleaning phase was executed. Redundant columns that did not contribute to the objectives of the analysis were dropped. Column names were standardized for consistency—converted to lowercase, stripped of special characters, and formatted to improve readability. Categorical data such as travel class or station names were checked for inconsistencies in naming conventions, and standardized values were applied to avoid duplication and ensure proper grouping during analysis.

Data type validation was a critical step in the preprocessing process. All fields were evaluated to ensure appropriate data types were applied. Numerical fields such as age, fare, and distance were cleaned of any text or irregular characters. Service rating fields were verified to fall within valid scoring ranges. Where applicable, date and time fields were converted to datetime format to enable time-based filtering and journey duration calculations.

Feature engineering was then used to enhance the dataset and support deeper analysis. New columns were created based on existing information, such as:

* **Fare per kilometer** to evaluate cost efficiency by route.
* **Overall satisfaction score** by averaging feedback on services.
* **Age group classification** (e.g., youth, adult, senior) for demographic segmentation.
* **Journey type** or **travel length category** (e.g., short, medium, long) based on distance or duration.

These derived attributes helped reveal patterns in travel behavior and feedback trends across different passenger segments.

Finally, the dataset was filtered and organized to allow focused analysis. Travel classes were grouped by category (e.g., sleeper, AC, general), and stations were analyzed based on their frequency of occurrence as source or destination. Sorting and grouping operations were applied to highlight high-traffic routes and commonly used services. Outlier detection techniques were employed to identify unusual fare values or feedback anomalies, which were further examined for potential data entry issues or exceptional cases.

Through this preprocessing phase, the dataset was transformed from a raw, unstructured state into a clean, analysis-ready format. It ensured data accuracy, consistency, and relevance, providing a strong foundation for subsequent exploratory data analysis and visualization. By preparing the dataset carefully, the reliability and depth of the insights generated in later stages of the project were significantly improved.

# ****4. ANALYSIS ON DATASET****

**Objective 1: Ticket Purchase Method (Station vs Advance)**

**i. General Description**  
This analysis aims to determine how many passengers purchased their tickets directly at the station versus those who bought them in advance. Understanding these purchase behaviors helps railway services assess demand at ticket counters and promote pre-booking.

**ii. Specific Requirements**

* Categorize ticket purchases into “Station” and “Advance”.
* Count the number of purchases in each category.
* Visualize the result using a **pie chart**.

**iii. Analysis Results**  
The majority of passengers [insert findings after analysis] tickets [e.g., in advance/at the station], indicating a preference for [convenience/planning ahead or immediate travel].

**iv. Visualization**

* A **pie chart** clearly shows the proportion of ticket purchases from station vs advance booking.
* Labels and percentage annotations aid easy interpretation.

**Objective 2: Online Purchase with Adult RailCard**

**i. General Description**  
This objective focuses on analyzing how many passengers booked their tickets online and used an Adult RailCard. It provides insight into digital adoption and discount card usage.

**ii. Specific Requirements**

* Filter the data for passengers who booked online.
* Within this subset, count how many used an “Adult RailCard”.

**iii. Analysis Results**  
The data revealed that [insert number/percentage] passengers who booked online availed the Adult RailCard, indicating [trends like preference for discounts among online users].

**iv. Visualization**

* A **bar or pie chart** was used to highlight how many online bookings included the Adult RailCard.

**Objective 3: Journey Status Distribution**

**i. General Description**  
This objective visualizes the distribution of different journey statuses such as “Completed”, “Cancelled”, or “Delayed”, helping identify service reliability and operational trends.

**ii. Specific Requirements**

* Count occurrences of each unique value in the “Journey Status” column.
* Visualize using a **bar chart**.

**iii. Analysis Results**  
Most journeys were marked as [Completed/Delayed/etc.], highlighting the overall service quality. [Insert brief observations].

**iv. Visualization**

* A **bar chart** with color-coded bars for each status.
* X-axis: Journey Status; Y-axis: Count.

**Objective 4: Fill and Visualize 'Reason for Delay'**

**i. General Description**  
This objective focuses on cleaning missing values in the “Reason for Delay” column and analyzing common reasons for train delays.

**ii. Specific Requirements**

* Fill missing values with the **most frequent reason (mode)**.
* Count occurrences of each reason.
* Visualize with a **line or scatter plot**.

**iii. Analysis Results**  
The most common reason for delay was [insert mode value]. Filling missing values improved dataset completeness and allowed for more accurate delay pattern analysis.

**iv. Visualization**

* A **line or scatter plot** showing frequency of each reason.
* X-axis: Reason; Y-axis: Count.

**Objective 5: Clean and Visualize 'Railcard' Distribution**

**i. General Description**  
This step ensures the “Railcard” column is complete and consistent, helping analyze what types of railcards are most frequently used.

**ii. Specific Requirements**

* Fill missing “Railcard” values with the **most frequent type (mode)**.
* Count each railcard type.
* Visualize with **bar chart or pie chart**.

**iii. Analysis Results**  
The most commonly used railcard was [insert value], showing trends in customer preference for certain discount cards.

**iv. Visualization**

* A **bar or pie chart** showcasing the distribution of railcard usage post-cleaning.

**Objective 6: Standardize and Fill Missing 'Arrival Time'**

**i. General Description**  
The objective is to ensure “Arrival Time” data is in a consistent format and handle missing entries, which is essential for accurate scheduling and delay analysis.

**ii. Specific Requirements**

* Convert all entries to a **standard 24-hour or 12-hour time format**.
* Fill missing values with the **most frequent time (mode)**.

**iii. Analysis Results**  
Arrival times were successfully standardized. The most frequent arrival time was [insert time], and this value was used to fill missing entries.

**iv. Visualization**

* A **histogram or time-based line plot** can be used to visualize arrival time distribution.

# 

# ****5. Screenshots****

# ****Github Link:**** [Ankitkr2004/UK-Train-Rides: Python Project With Its Modules](https://github.com/Ankitkr2004/UK-Train-Rides)

# Linkedin link: <https://www.linkedin.com/posts/ankitkr2004_python-datascience-uktransport-activity-7314983369019117568-NTiw?utm_source=share&utm_medium=member_desktop&rcm=ACoAAEQ2Ig0BUKfDoCYyUB7kUTbKQnCboo0EpyM>

# ****6. CONCLUSION****

This project focused on conducting a comprehensive exploratory data analysis (EDA) on a railway passenger dataset using Python. The primary goal was to uncover hidden patterns, clean and structure the data effectively, and provide actionable insights to support strategic improvements in railway operations and customer experience. Python libraries such as Pandas, Matplotlib, and Seaborn were extensively used throughout the analysis for data wrangling and visualization.

**Key Findings:**

1. **Ticket Purchase Behavior:**  
   A significant proportion of passengers purchased their tickets in advance, while others preferred to buy at the station. This distinction helps in optimizing ticket counter staffing and promoting advance booking options.
2. **Online Booking with Adult RailCard:**  
   A noteworthy number of passengers used the Adult RailCard during online bookings, indicating digital engagement and a preference for fare discounts. This insight supports targeted marketing of digital services and RailCard schemes.
3. **Journey Status Insights:**  
   The distribution of journey statuses showed clear trends in trip completion, cancellations, and delays. Such data is crucial for assessing service reliability and improving journey planning.
4. **Reasons for Delays:**  
   The most common reasons for train delays were identified and missing values in this column were addressed using mode imputation. Understanding delay causes enables operational improvements and enhances transparency for passengers.
5. **Railcard Usage Trends:**  
   By cleaning the 'Railcard' column, the analysis highlighted the most frequently used types of discount cards. This helps in understanding customer preferences and tailoring loyalty or concession schemes accordingly.
6. **Arrival Time Standardization:**  
   Arrival times were standardized and missing entries were filled using the most frequent time value. This improved the dataset’s integrity and supported further time-based analysis such as delay calculations or schedule planning.

**Overall Impact:**

This project successfully transformed raw railway data into clear, actionable insights. The visualizations and analyses enabled a deeper understanding of customer behavior, ticketing preferences, journey reliability, and discount usage patterns. These insights can support:

* Service optimization and passenger flow management
* Enhanced digital ticketing strategies
* Better communication regarding delays and RailCard usage
* Data-driven scheduling and resource allocation

**Final Thoughts:**

The project highlights the critical role of data analytics in improving railway operations and enhancing the passenger experience. Leveraging Python's powerful analytical tools enabled effective data cleaning, visualization, and interpretation.

Future extensions of this project could include predictive analytics, such as forecasting ticket sales or predicting journey delays using machine learning models. Incorporating additional data sources like weather, holidays, or regional demographics could further enrich the insights and support more robust decision-making.

.

# ****7. FUTURE SCOPE****

The current project has effectively demonstrated the use of Exploratory Data Analysis (EDA) techniques to uncover valuable insights from railway passenger data. However, there are several promising directions in which this work can be expanded. These enhancements would add significant value by enabling real-time analytics, deeper customer behavior understanding, and predictive modeling to support better operational and strategic decision-making in the railway industry.

**1. Integration of Machine Learning Models**

Future iterations of the project can incorporate machine learning algorithms for predictive analytics. Potential applications include:

* **Classification models** (e.g., Random Forest, Decision Trees) to predict journey delays or cancellation likelihood.
* **Clustering techniques** (e.g., K-Means) to segment passengers based on travel patterns or ticket preferences.
* **Regression models** (e.g., Linear Regression, XGBoost) to forecast ticket demand or delay durations.

**2. Real-Time Data Streaming**

Currently, the dataset is historical and static. Integrating **real-time data** (e.g., live delay reports, ticket purchases, platform crowding) would allow:

* Immediate monitoring of passenger flow and journey status
* Dynamic resource allocation (e.g., additional trains or staff)
* Faster response to operational issues

**3. Interactive Dashboards**

Developing interactive dashboards using tools like **Power BI, Tableau, or Plotly Dash** can make the analysis accessible to non-technical users. Dashboards can:

* Visualize live train statuses and delay causes
* Segment ticket sales by type, station, or time
* Enable drill-down into RailCard usage or passenger demographics

**4. Geospatial Analysis**

Incorporating geolocation data (stations, routes, regions) can unlock insights such as:

* **Regional delay patterns** and infrastructure bottlenecks
* **High-demand stations** for staffing/resource planning
* Regional **ticket type and RailCard** usage variations

**5. Passenger Behavior Insights**

With deeper analysis of passenger-level data, future work can explore:

* **Travel frequency and loyalty trends**
* **RailCard renewal predictions**
* **Analysis of booking methods** (station vs. online) over time

**6. Time Series Modeling**

While current analysis standardizes and explores time-based trends, advanced time series models like **ARIMA or Prophet** can help:

* Predict peak travel periods
* Identify long-term patterns in delays
* Forecast demand across routes or stations

**7. Delay Prediction and Management**

Using historical delay data and reasons, the project can evolve into a **delay prediction model**. This would:

* Alert authorities of potential delays in advance
* Improve customer communication
* Help in evaluating and improving punctuality metrics

**8. Multivariate Analysis**

Conducting multivariate statistical analysis would reveal how different variables (e.g., time of day, ticket type, RailCard usage) interact with delays, cancellations, or travel choices. This deeper understanding supports:

* Smarter scheduling
* Fare policy adjustments
* Targeted customer engagement

**9. Integration of External Datasets**

Incorporating external factors like:

* **Weather data**
* **Holiday calendars**
* **Special events or strikes** ...can make the analysis more comprehensive and context-aware. These external drivers often correlate with spikes in demand or delays.

**10. Automated Reporting and Alerts**

Automating the entire EDA pipeline using **Python scripts, Airflow, or Cron jobs** would:

* Ensure regular dataset updates
* Generate recurring reports for stakeholders
* Send alerts for anomalies in travel trends or delay causes

**Conclusion of Future Scope**

This project has laid a strong analytical foundation for railway data analysis. By expanding into machine learning, real-time systems, dashboarding, and predictive analytics, it can evolve into a powerful business intelligence platform. These future enhancements will empower railway operators to make smarter, faster, and more customer-centric decisions—ultimately leading to improved operational efficiency, better service reliability, and enhanced passenger satisfaction.

**8.REFERENCES**

1. **Bureau of Transportation Statistics**, “Passenger Rail Trends,” U.S. Department of Transportation, 2021. [Online]. Available: <https://www.bts.gov>.

* This reference provides various datasets and trends regarding passenger rail transport, which could help in contextualizing your analysis.

2. **S. N. G. H. Murthy et al.,** "Railway Passenger Data Analysis Using Data Mining," *International Journal of Computer Applications*, vol. 78, no. 4, pp. 9-12, 2013.

* This paper discusses data mining techniques for analyzing railway passenger data, which can be insightful for your dataset’s analysis, especially for ticket purchases and delays.

3. **M. B. Nielsen et al.,** "Passenger Flow Prediction in Railway Stations Using Data Mining and Machine Learning," *Journal of Transportation Engineering*, vol. 142, no. 7, pp. 04016042, 2016.

* This reference focuses on passenger flow prediction, which could be useful for handling and analyzing journey status counts in your railway dataset.

4. **L. S. Tiwari and D. N. Rathi,** "Analysis of Train Delays and Its Effect on Passenger Satisfaction," *Transport Policy*, vol. 35, pp. 114-121, 2014.

* A relevant paper for analyzing delays in the context of railway passenger satisfaction, which aligns with your goal of addressing reasons for delays in the dataset.

5. **S. Pandey et al.,** "Predicting Train Arrival Time Using Machine Learning Techniques," *Procedia Computer Science*, vol. 132, pp. 567-574, 2018.

* Focuses on predicting arrival times, which could assist in the standardization and filling of missing "Arrival Time" values in your dataset.