**Indian Railways Data Analysis Report**

**Prepared by:** Tamal Majumdar  
**Date:** 05 September 2025

**1. Introduction**

This report presents a comprehensive exploratory data analysis (EDA) on the Indian Railways dataset, which includes information about stations, trains, and schedules. The goal is to understand the structure, distribution, and key insights of the data, and to visualize geographical and operational patterns.

**Datasets Used:**

* stations.json – Station information including code, name, zone, state, and coordinates
* trains.json – Train information including type, route, distance, and duration
* schedules.json – Train schedules including station halts, arrival and departure times

**Tools & Libraries:**

* Python: pandas, numpy, matplotlib, seaborn, folium

**2. Data Loading & Cleaning**

**2.1 Stations Dataset**

* Loaded from stations.json
* Cleaned columns: station\_code, station\_name, state, zone, address, longitude, latitude
* Split coordinates into separate longitude and latitude columns
* Handled missing coordinates by replacing invalid values with [None, None]

**Quick Overview:**

| **Column** | **Missing Values** |
| --- | --- |
| station\_code | 0 |
| station\_name | 0 |
| state | 4532 |
| zone | 4532 |
| address | 4532 |
| longitude | 293 |
| latitude | 293 |

**Observation:**  
Many stations have missing zone/state information and a few missing coordinates. Cleaning was necessary for geospatial visualizations.

**2.2 Trains Dataset**

* Loaded from trains.json
* Cleaned columns: train\_number, train\_name, train\_type, zone, from\_station\_code, to\_station\_code, distance\_km, duration\_hr, arrival\_time, departure\_time

**Missing Values:**

| **Column** | **Missing Values** |
| --- | --- |
| distance\_km | 15 |
| duration\_hr | 15 |

**Observation:**  
Overall, the train dataset is clean. Minor missing values exist in distance and duration.

**2.3 Schedules Dataset**

* Loaded from schedules.json
* Cleaned columns: train\_number, train\_name, station\_code, station\_name, arrival\_time, departure\_time, day
* Replaced invalid entries such as NA, ?, or null with pd.NA

**Missing Values:**

| **Column** | **Missing Values** |
| --- | --- |
| train\_name | 8 |
| station\_name | 2 |
| day | 22561 |

**Observation:**  
Large dataset with 417,080 records. Missing days indicate incomplete scheduling info for some trains.

**3. Stations Analysis**

**3.1 Basic Stats**

* Total Stations: 8,990
* Unique Zones: 18
* Unique States: 30

**3.2 Stations per State**

Top 10 states by number of stations:

| **State** | **Count** |
| --- | --- |
| Uttar Pradesh | 529 |
| Rajasthan | 451 |
| Gujarat | 422 |
| Maharashtra | 378 |
| West Bengal | 345 |
| Madhya Pradesh | 316 |
| Karnataka | 301 |
| Tamil Nadu | 253 |
| Punjab | 212 |
| Bihar | 210 |

**Observation:**  
Uttar Pradesh, Rajasthan, and Gujarat have the highest concentration of railway stations.

**3.3 Geospatial Map**

* Stations plotted on an interactive map using Folium
* Stations represented as blue circle markers

**Observation:**  
Most stations are concentrated in northern and western states. Sparse coverage is observed in northeastern India.

**4. Trains Analysis**

**4.1 Basic Stats**

* Total Trains: 5,208
* Columns: train\_number, train\_name, train\_type, zone, distance\_km, duration\_hr, etc.

**4.2 Train Types Distribution**

| **Train Type** | **Count** |
| --- | --- |
| Pass | 2,459 |
| Exp | 1,288 |
| SF | 719 |
| MEMU | 297 |
| Hyd | 121 |
| Others | 324 |

**Observation:**  
Passenger trains dominate, followed by express (Exp) and superfast (SF) trains.

**4.3 Trains per Zone**

| **Zone** | **Count** |
| --- | --- |
| NR | 628 |
| SR | 606 |
| WR | 470 |
| SCR | 437 |
| NER | 394 |
| ER | 369 |

**Observation:**  
Northern Railway (NR) and Southern Railway (SR) operate the most trains.

**4.4 Journey Distances**

* Average distance: 545 km
* Maximum distance: 4,279 km

**Observation:**  
Trains cover a wide range of distances from short regional trips to long inter-state journeys.

**4.5 Average Speed**

* Calculated as distance\_km / duration\_hr
* Summary:

| **Metric** | **Value** |
| --- | --- |
| Mean | 47 km/h |
| Min | 10.5 km/h |
| Max | 225 km/h |

**Observation:**  
Most trains operate below 60 km/h. High-speed trains (SF) achieve up to 225 km/h.

**5. Schedules Analysis**

**5.1 Busiest Stations (Most Train Halts)**

Top 10 busiest stations:

| **Station Name** | **Train Halts** |
| --- | --- |
| SABARMATI JN | 342 |
| KANPUR CENTRAL | 312 |
| ITARSI JN | 293 |
| GHAZIABAD | 287 |
| SAHIBABAD | 285 |
| HOWRAH JN | 283 |
| VIJAYAWADA JN | 264 |
| KOPAR ROAD | 262 |
| MUGHAL SARAI JN | 259 |
| VADODARA JN | 254 |

**Observation:**  
Busiest stations are mostly major junctions in northern and western India.

**5.2 Train Coverage**

* Number of stations served per train:
  + Median: ~35 stations
  + Maximum: 186 stations

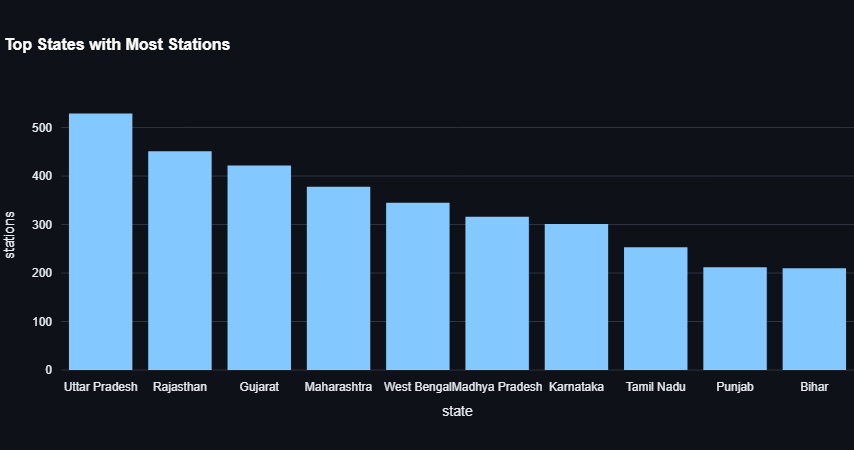
**Observation:**  
Long-distance trains cover a large number of stations, while regional trains have fewer stops.

**6. Key Insights**

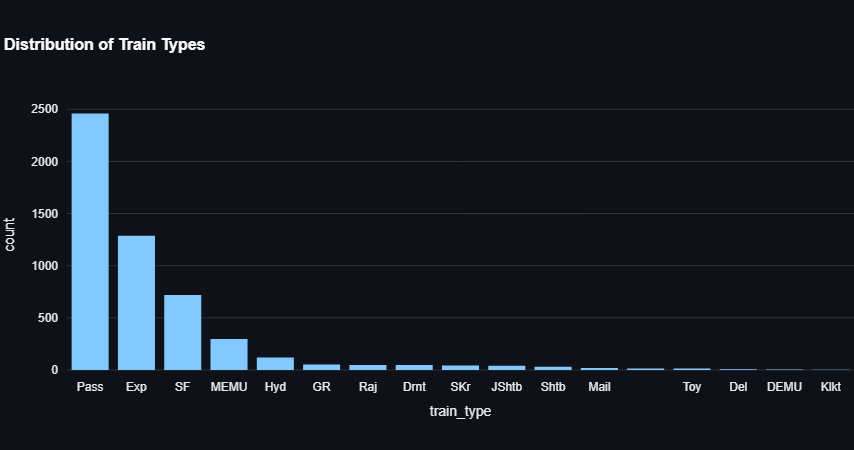
1. Northern and western states have the highest density of railway stations.
2. Passenger trains dominate the Indian Railways network.
3. Major hubs like SABARMATI, KANPUR, and HOWRAH handle the highest traffic.
4. Average train speed is 47 km/h, indicating slow regional trains and few high-speed services.
5. Long-distance trains serve a broad range of stations, providing connectivity across states.

**7. Visualizations**

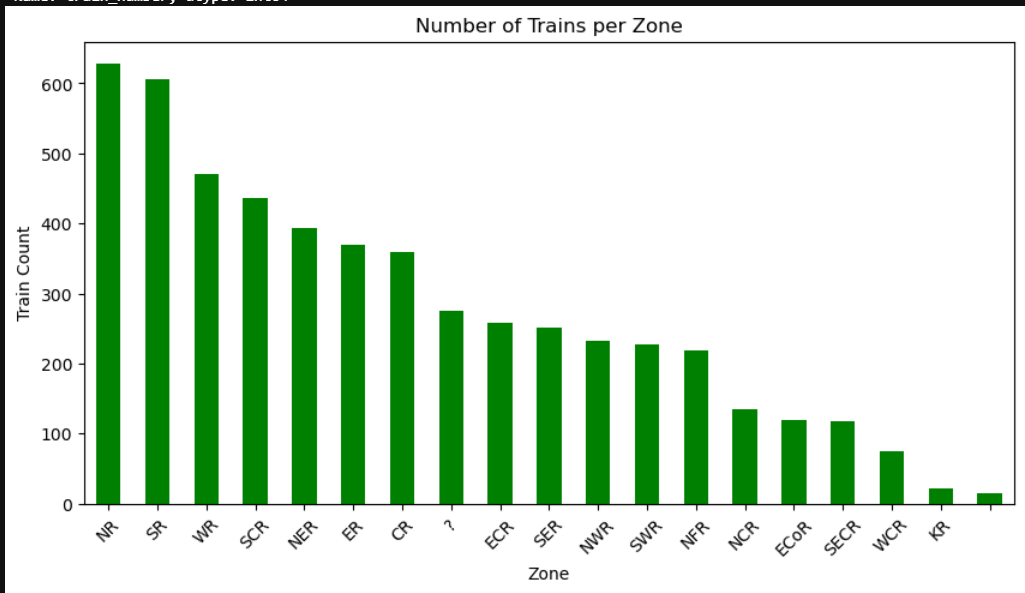
* **Stations per State:** Bar chart



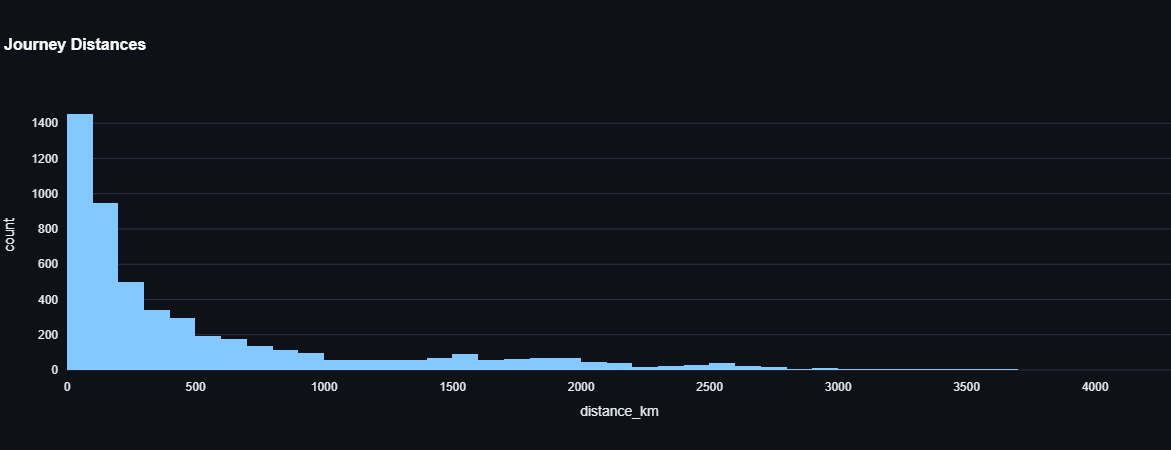
* **Train Types Distribution:** Bar chart



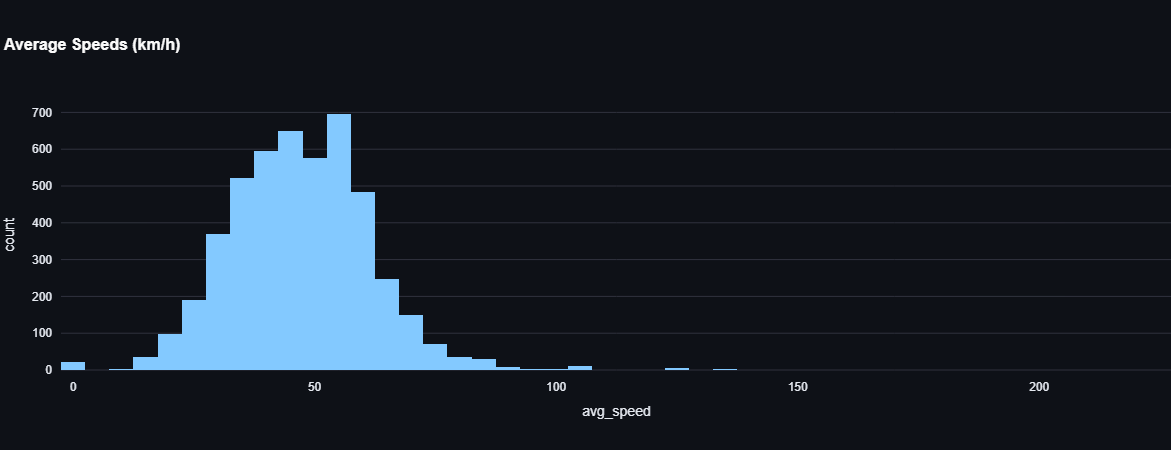
* **Trains per Zone:** Bar chart



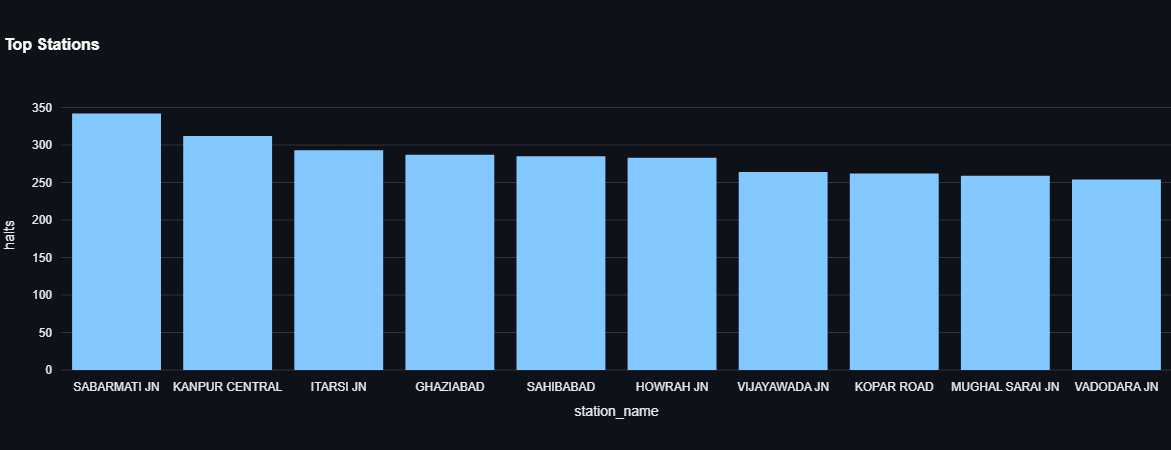
* **Journey Distances:** Histogram



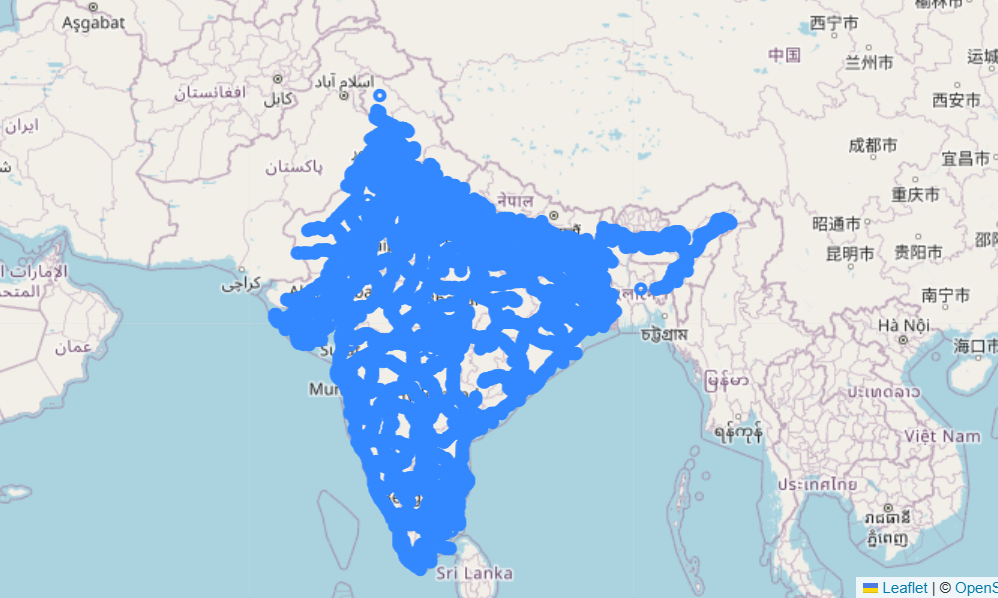
* **Average Speed:** Histogram



* **Busiest Stations:** Bar chart



* **Geospatial Map:** Interactive stations\_map.html



**8. Conclusion**

The exploratory analysis provides a clear view of the Indian Railways network:

* Coverage is dense in major states but sparse in the northeast.
* Passenger and express trains dominate, but high-speed trains are limited.
* Certain stations act as hubs for high traffic, offering opportunities for network optimization.
* The geospatial map enables interactive visualization of station locations.

This analysis lays the foundation for deeper studies such as train traffic optimization, predictive modeling, and geospatial planning.