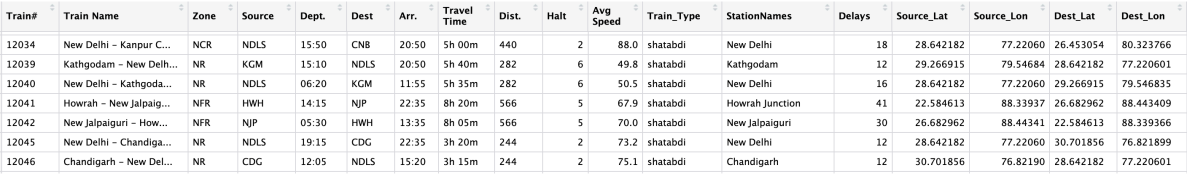
Report on Indian Railway Analysis

Table of Contents

1. Data

Our Data looks like:



Train#	Train Name	Zone	Source	Dept.	Dest	Arr.	Travel Time	Dist.	Halt	Avg Speed	Train_Type	StationNames	Delays	Source_Lat	Source_Lon	Dest_Lat	Dest_Lon
12034	New Delhi – Kanpur C...	NCR	NDLS	15:50	CNB	20:50	5h 00m	440	2	88.0	shatabdi	New Delhi	18	28.642182	77.22060	26.453054	80.323766
12039	Kathgodam – New Delh...	NR	KGM	15:10	NDLS	20:50	5h 40m	282	6	49.8	shatabdi	Kathgodam	12	29.266915	79.54684	28.642182	77.220601
12040	New Delhi – Kathgoda...	NR	NDLS	06:20	KGM	11:55	5h 35m	282	6	50.5	shatabdi	New Delhi	16	28.642182	77.22060	29.266915	79.546835
12041	Howrah – New Jalpaig...	NFR	HWH	14:15	NJP	22:35	8h 20m	566	5	67.9	shatabdi	Howrah Junction	41	22.584613	88.33937	26.682962	88.443409
12042	New Jalpaiguri – How...	NFR	NJP	05:30	HWH	13:35	8h 05m	566	5	70.0	shatabdi	New Jalpaiguri	30	26.682962	88.44341	22.584613	88.339366
12045	New Delhi – Chandiga...	NR	NDLS	19:15	CDG	22:35	3h 20m	244	2	73.2	shatabdi	New Delhi	12	28.642182	77.22060	30.701856	76.821899
12046	Chandigarh – New Del...	NR	CDG	12:05	NDLS	15:20	3h 15m	244	2	75.1	shatabdi	Chandigarh	12	30.701856	76.82190	28.642182	77.220601

Figure 1: Data

Our Data contains 2942 rows and 18 columns and description of each column is given below:

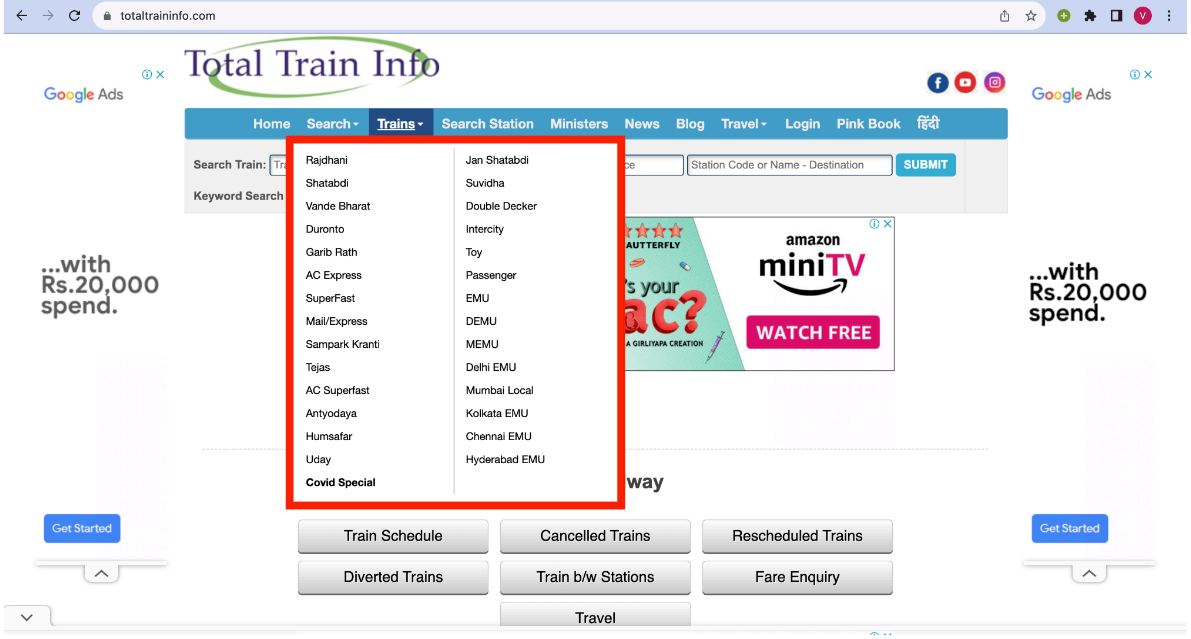
Table 1: Explanation of Dataset

Column Name	Explanation
Train#	Unique identification number assigned to each train.
Train Name	The designated name of the train
Zone	The railway zone associated with the train's route
Source	The originating station of the train journey
Dept.	Departure time from the source station
Dest:	The final destination station of the train
Arr.	Arrival time at the destination station
Travel Time	The total time taken for the journey
Dist.	The distance covered by the train during the journey
Halt	Number of scheduled halts or stops
Avg Speed	The average speed of the train
Train_Type	Indicates the type or category of the train
StationNames	Contain Full name of departure station
Delays	Avg Delay of train in past one year
Source_Lat	The latitude coordinate of the source station
Source_Lon	The longitude coordinate of the source station
Dest_Lat	The latitude coordinate of the destination station
Dest_Lon	The longitude coordinate of the destination station

2. Obtaining the Data

2.1 Initial Data Extraction

We started with the website Total Train Info and extracted links for different types of trains.

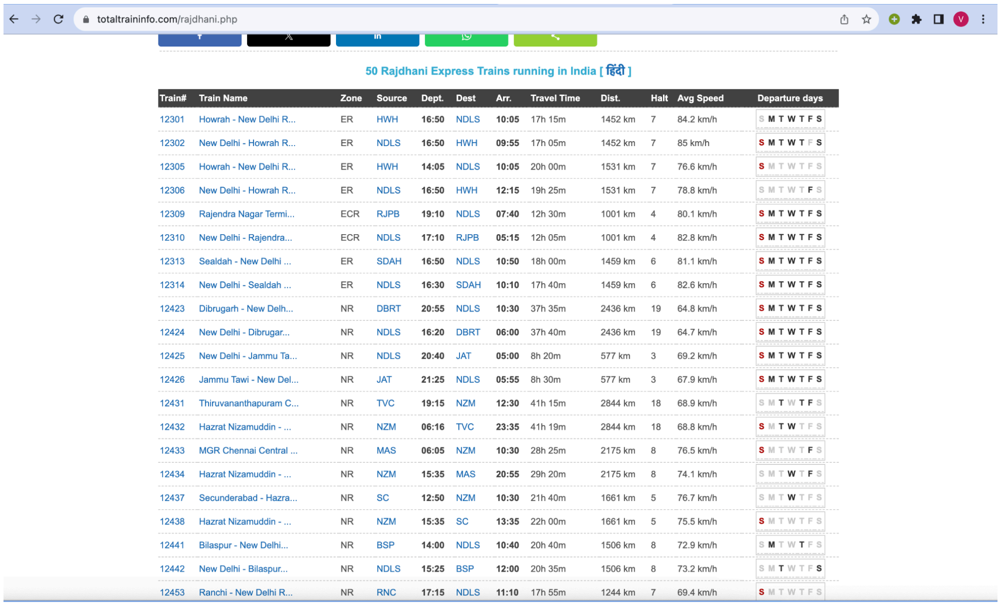


2.2 Train Name Extraction

Next, we extracted the names of different types of trains from the obtained links. This information will be used in subsequent steps.

2.3 Webpage Scraping

We moved to the webpage of each train and extracted the data table for each type of train, storing the results in a list. The following screenshot illustrates the process:



totaltraininfo.com/rajdhani.php

50 Rajdhani Express Trains running in India [हिंदी]

Train#	Train Name	Zone	Source	Dept.	Dest	Arr.	Travel Time	Dist.	Halt	Avg Speed	Departure days
12301	Howrah - New Delhi R...	ER	HWH	16:50	NDLS	10:05	17h 15m	1452 km	7	84.2 km/h	S **M T W T F S**
12302	New Delhi - Howrah R...	ER	NDLS	16:50	HWH	09:55	17h 05m	1452 km	7	85 km/h	**S M T W T** F **S**
12305	Howrah - New Delhi R...	ER	HWH	14:05	NDLS	10:05	20h 00m	1531 km	7	76.6 km/h	**S** M T W T F S
12306	New Delhi - Howrah R...	ER	NDLS	16:50	HWH	12:15	19h 25m	1531 km	7	78.8 km/h	S M T W T **F** S
12309	Rajendra Nagar Termi...	ECR	RJPB	19:10	NDLS	07:40	12h 30m	1001 km	4	80.1 km/h	**S M T W T F S**
12310	New Delhi - Rajendra...	ECR	NDLS	17:10	RJPB	05:15	12h 05m	1001 km	4	82.8 km/h	**S M T W T F S**
12313	Sealdah - New Delhi ...	ER	SDAH	16:50	NDLS	10:50	18h 00m	1459 km	6	81.1 km/h	**S M T W T F S**
12314	New Delhi - Sealdah ...	ER	NDLS	16:30	SDAH	10:10	17h 40m	1459 km	6	82.6 km/h	**S M T W T F S**
12423	Dibrugarh - New Delh...	NR	DBRT	20:55	NDLS	10:30	37h 35m	2436 km	19	64.8 km/h	**S M T W T F S**
12424	New Delhi - Dibrugar...	NR	NDLS	16:20	DBRT	06:00	37h 40m	2436 km	19	64.7 km/h	**S M T W T F S**
12425	New Delhi - Jammu Ta...	NR	NDLS	20:40	JAT	05:00	8h 20m	577 km	3	69.2 km/h	**S M T W T F S**
12426	Jammu Tawi - New Del...	NR	JAT	21:25	NDLS	05:55	8h 30m	577 km	3	67.9 km/h	**S M T W T F S**
12431	Thiruvananthapuram C...	NR	TVC	19:15	NZM	12:30	41h 15m	2844 km	18	68.9 km/h	S M **T** W **T F** S
12432	Hazrat Nizamuddin - ...	NR	NZM	06:16	TVC	23:35	41h 19m	2844 km	18	68.8 km/h	**S** M **T W** T F S
12433	MGR Chennai Central ...	NR	MAS	06:05	NZM	10:30	28h 25m	2175 km	8	76.5 km/h	**S** M T W T **F** S
12434	Hazrat Nizamuddin - ...	NR	NZM	15:35	MAS	20:55	29h 20m	2175 km	8	74.1 km/h	S M T **W** T **F** S
12437	Secunderabad - Hazra...	NR	SC	12:50	NZM	10:30	21h 40m	1661 km	5	76.7 km/h	S M T **W** T F S
12438	Hazrat Nizamuddin - ...	NR	NZM	15:35	SC	13:35	22h 00m	1661 km	5	75.5 km/h	**S** M T W T F S
12441	Bilaspur - New Delhi...	NR	BSP	14:00	NDLS	10:40	20h 40m	1506 km	8	72.9 km/h	S **M** T W **T** F S
12442	New Delhi - Bilaspur...	NR	NDLS	15:25	BSP	12:00	20h 35m	1506 km	8	73.2 km/h	S M **T** W T F **S**
12453	Ranchi - New Delhi R...	NR	RNC	17:15	NDLS	11:10	17h 55m	1244 km	7	69.4 km/h	**S** M T W T F S

Figure 2: Webpage of specific Train Type

2.4 Data Cleaning and Merging

After obtaining the data tables, the next step involved cleaning the data to ensure its quality and consistency. The cleaning process included:

- We addressed missing values (NA) in the dataset by either imputing them using appropriate statistical methods or removing rows or columns containing these missing values.

- In the process of cleaning, we noticed that certain columns contained additional text, such as "km/h" and "km," in the entries for average speed ('avg_speed') and distance ('Dist'). To ensure uniformity and facilitate numerical analysis, we removed these extraneous text elements from the respective columns.

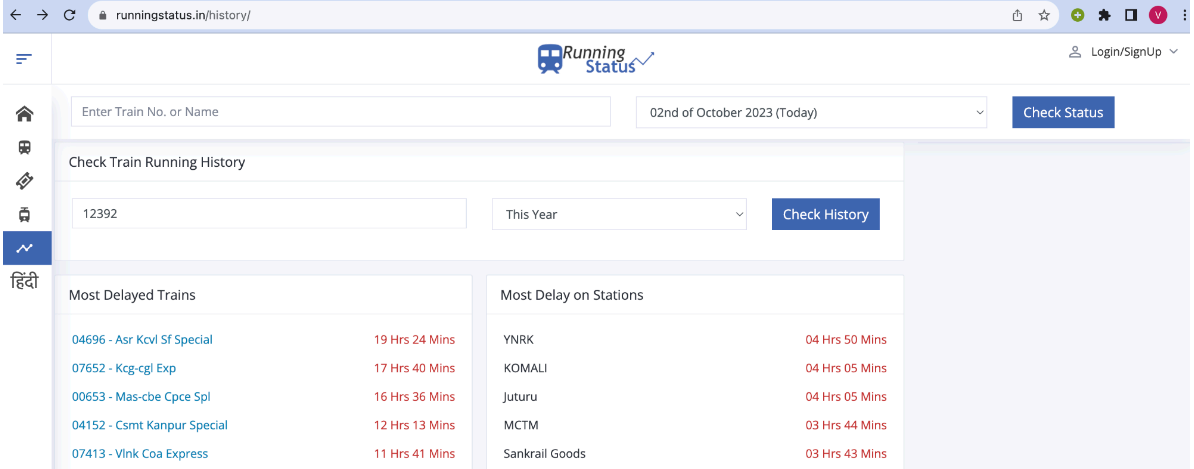
  For example, if the 'avg_speed' column contained entries like "60 km/h," we extracted the numerical value (60) and stored it in the column, discarding the unit information.

- To enhance the interpretability of the dataset and enable categorical analysis, we added a new column, 'train_type', to each data table. This column indicates the type of train associated with each record, allowing for easy segmentation and comparison.

- With the data now cleaned and standardized across all tables, we proceeded to merge them into a single cohesive dataset. This consolidated dataset serves as the foundation for further analysis, providing a comprehensive overview of the various train types, their characteristics, and relevant information for subsequent investigations.

2.5 Scraping Average Delays

Now we wanted to scrape average delays of train in past one year forw which we used this site runningstatus.in .



Enter Train No. or Name

02nd of October 2023 (Today)

Check Train Running History

12392

This Year

Most Delayed Trains	
04696 - Asr Kcvl Sf Special	19 Hrs 24 Mins
07652 - Kcg-cgl Exp	17 Hrs 40 Mins
00653 - Mas-cbe Cpce Spl	16 Hrs 36 Mins
04152 - Csmt Kanpur Special	12 Hrs 13 Mins
07413 - Vlnk Coa Express	11 Hrs 41 Mins

Most Delay on Stations	
YNRK	04 Hrs 50 Mins
KOMALI	04 Hrs 05 Mins
Juturu	04 Hrs 05 Mins
MCTM	03 Hrs 44 Mins
Sankrail Goods	03 Hrs 43 Mins

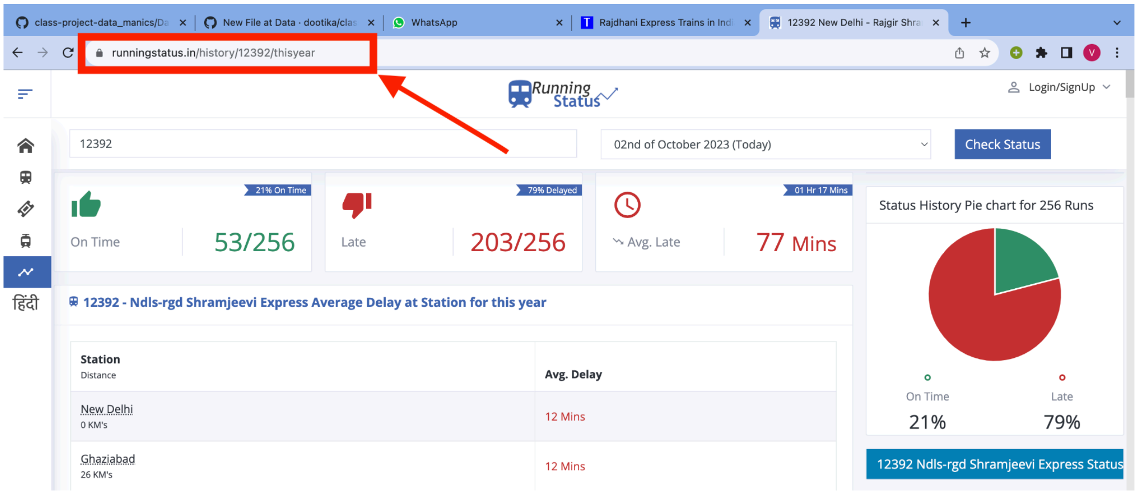
Figure 3: Website used for scraping Average Delays

2.6 Average Delays Extraction

Then we created links of each train and went to that page and then scraped the average delays of the trains.

2.7 Getting Latitudes and Longitudes of Railways Stations

For latitudes and longitudes we used csv file since we were not able to scrape out this data. Because the only site containing it for free was out of date.


class-project-data_manics/D
New File at Data · dootika/cla
WhatsApp
Rajdhani Express Trains in Ind
12392 New Delhi - Rajgir Shra
Running Status
Login/SignUp
हिंदी
Status History Pie chart for 256 Runs
On Time
21%
Late
79%

runningstatus.in/history/12392/thisyear

12392

02nd of October 2023 (Today)

Check Status

21% On Time	79% Delayed	01 Hr 17 Mins
On Time 53/256	Late 203/256	Avg. Late 77 Mins

12392 - Ndls-rgd Shramjeevi Express Average Delay at Station for this year

Station Distance	Avg. Delay
New Delhi 0 KM's	12 Mins
Ghaziabad 26 KM's	12 Mins

12392 Ndls-rgd Shramjeevi Express Status

Figure 4: Webpage of a particular train

3. Biases in Data

3.1 Data was collected from third party sites

The data of trains is scraped out from sites like Total Train Info , runningstatus.in and mapsofindia are 3rd party sites and has no affiliation with the Indian Railways Officially site.

3.2 Fast changing and Uncertain dynamic data

Another inherent bias in the data is the dynamic and fast-changing nature of train information. Given the dynamic conditions of the rail network. For instance, the dynamic nature of train information may manifest in scenarios where a train's scheduled departure time is altered abruptly, or in some cases, a train may be entirely discarded from the schedule.

3.3 Incomplete Data and Sampling Bias

In our data we have only considered the data of source location and final destination of our train. Due to the above reason in actual India has 7,325 (refrence1) stations but our data contains only 562 stations which lead to sampling bias since the origin and final destination stations are usually the large stations.

4. Questions

1. When planning a journey between two different locations in India, faced with multiple train options offering varying travel times and ticket prices, how can one determine the most suitable train with least average delays?

2. Which train types are most widely distributed across various regions of India?

3. Are there specific geographical areas where certain train types are concentrated?

4. Why do certain regions in India, such as Karnataka, Andhra Pradesh, Himachal Pradesh, etc., experience limited passenger train connectivity, and what factors may contribute to this restriction in railway development?

5. Is there any relation between duration of train journey and the delay in that journey. Is it True that long routes train are having more delays?

6. In what aspects new Train services like Vande Bharat are different from the old ones?

7. Which train services are more prominent in a particular zone of India?

8. Which zones of India are more dense comparative to others. Are they dense for a particular train type or as a whole?

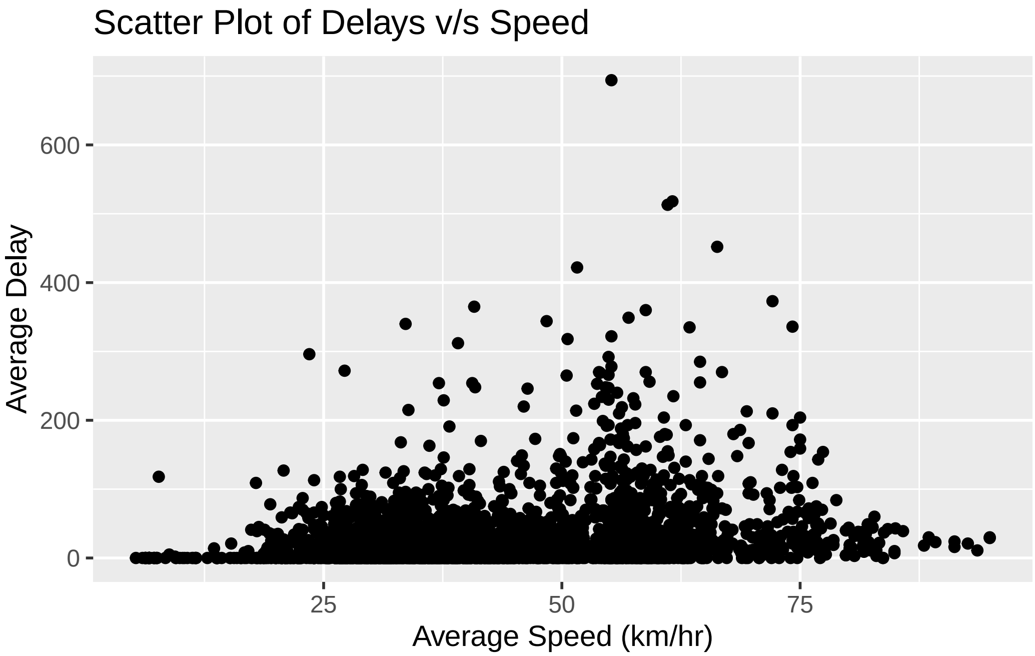
9. In a particular zone which Train are faster than other and do the same trend continues in other zones of India?

5. Visualizations

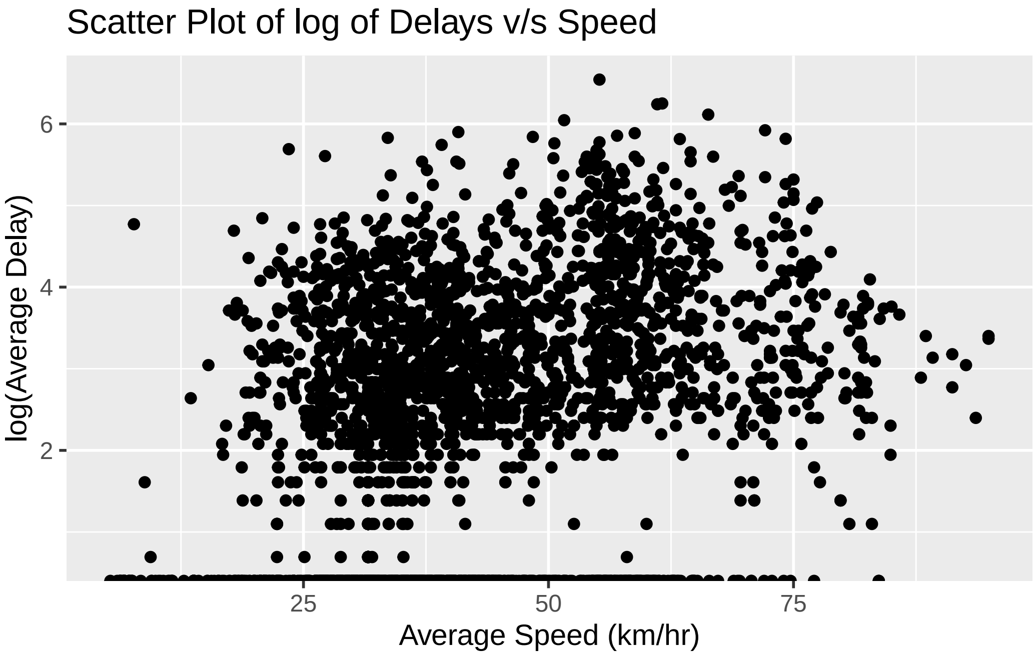
There were multiple visualizations based on cross-sectional rows as well as columns of our data.

5.1 Delays, Average Speed and Duration Group

First, we plotted a scatter plot between the average delays and the average speed of the trains and it came out to be something like this:

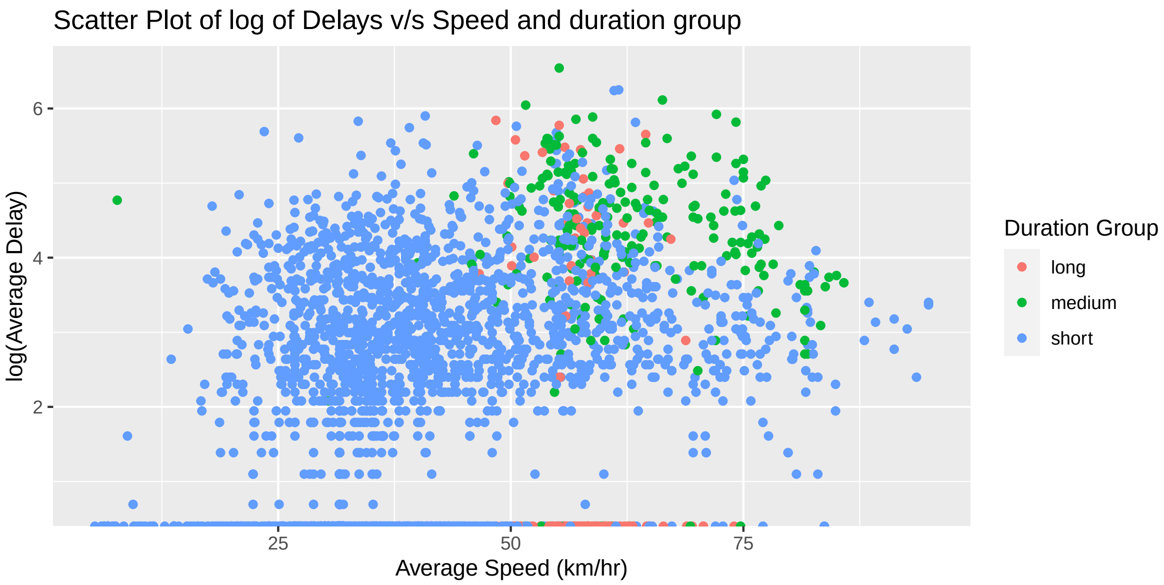


However, since a lot the data points were accumulated close to zero, we decided to plot the logarithm of average delays v/s the average speed.



However, as is clear from the plot, there was no discernible or clear relationship visible between average speed and the delays of the trains. So, we decided to color code the points based on the duration group of the trains.
We divided all the trains into 3 groups: long duration trains, medium duration trains and short duration trains with 30-60 hours, 15-30 hours and 0-15 hours long journey respectively.

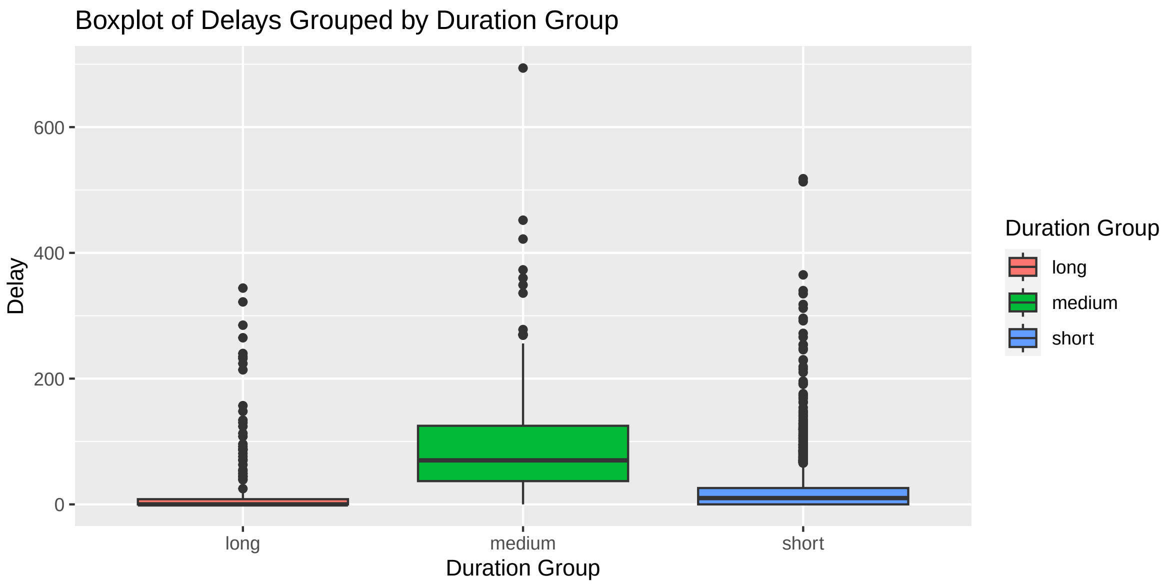
And, the result was this:



Based on this graph, here are our findings:

- Short duration trains: The number of short duration trains is much more than both long and medium duration trains combined. However, the spread average delay is contained but has some outliers.
- Medium duration trains: These trains have longer delays with a larger spread when compared with the other two categories.
- Long duration trains: The average delays are shorter than both short and medium duration trains. However, the number of data points is also lesser than the other two.

However, even after categorizing the data into different groups, we still couldn't identify any relationship between the Average speed and delays of the trains.
Thus, we created a box plot of various duration groups with respect to average delays and our previous findings were confirmed:



The median as well as the spread of the medium trains is the most. The spread as well as the mdeian of the long duration trains is the least. However, that can be attributed to the fact that the number of long duration trains is also the least.
Here is a summary of average delays based on the duration group:

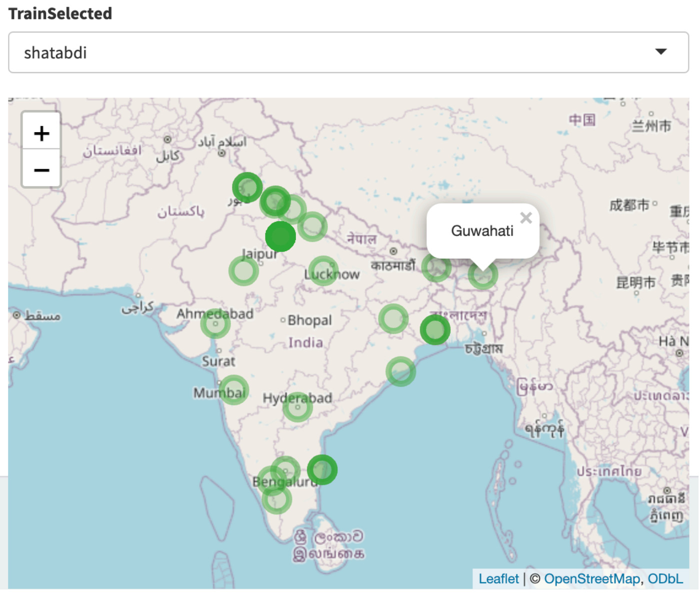
```
  Duration Group Avg_delay Count
1           long  28.83333   162
2         medium  94.62449   245
3          short  21.78760  2535
```

5.2 Geographical Analysis

Next we plotted the source stations of various train types on the map of India as shown below:

We analysed different train types based on the location of their source stations and came across the following two conclusions:

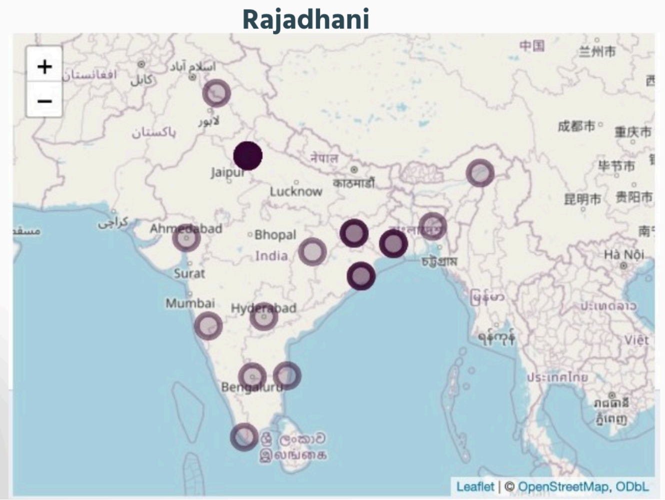
- Rajdhani:
  As is visible from the following map, rajdhani generally moves from state capitals. However, the source stations for this train are not evenly distributed across the nation.



TrainSelected

shatabdi

Figure 5: Map



.

- Passenger: The passenger train serves as the backbone of the Indian Railways as it is affordable and accessible for a large portion of the population. Also, as can be seen from the graph, the source stations for passenger trains are more evenly distributed when compared with other much popular trains like rajdhani or shatabdi.
  However, it was interesting to note that the stations are less in the western as well as the southern region of the nation.

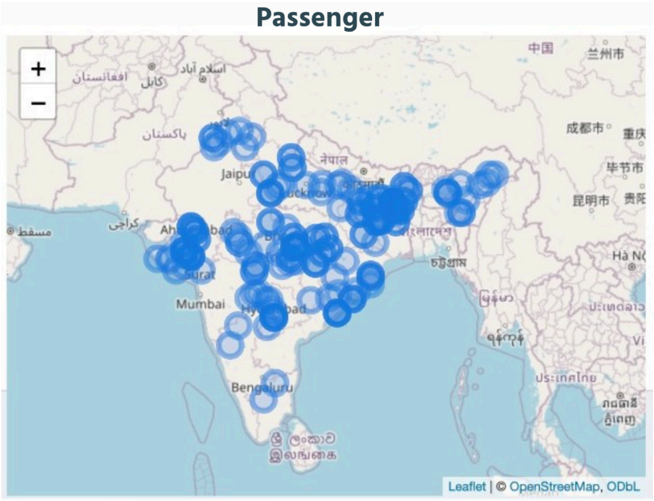
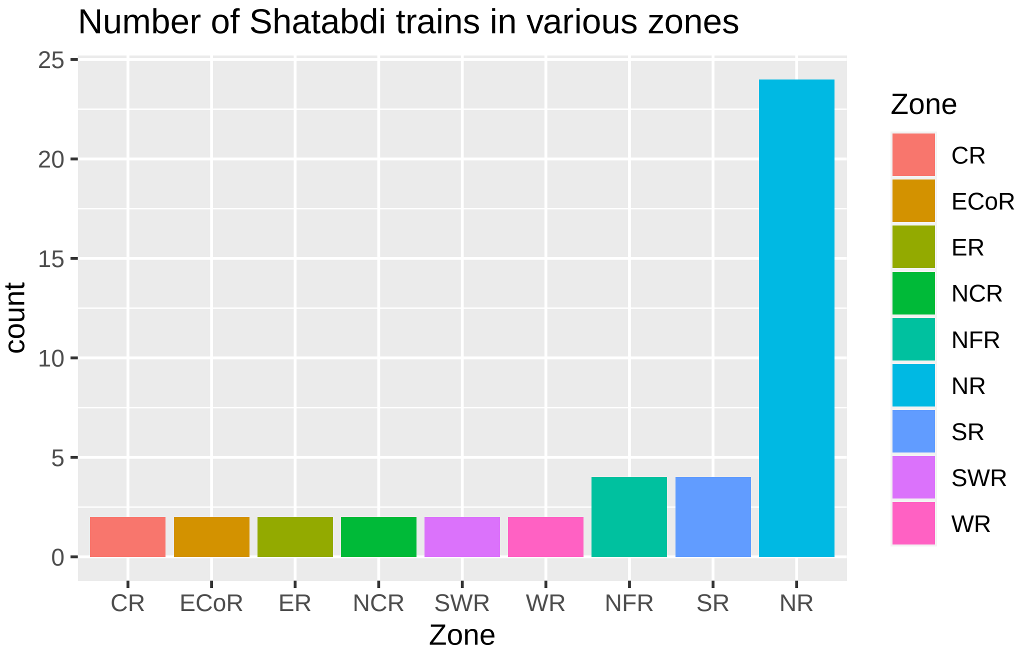


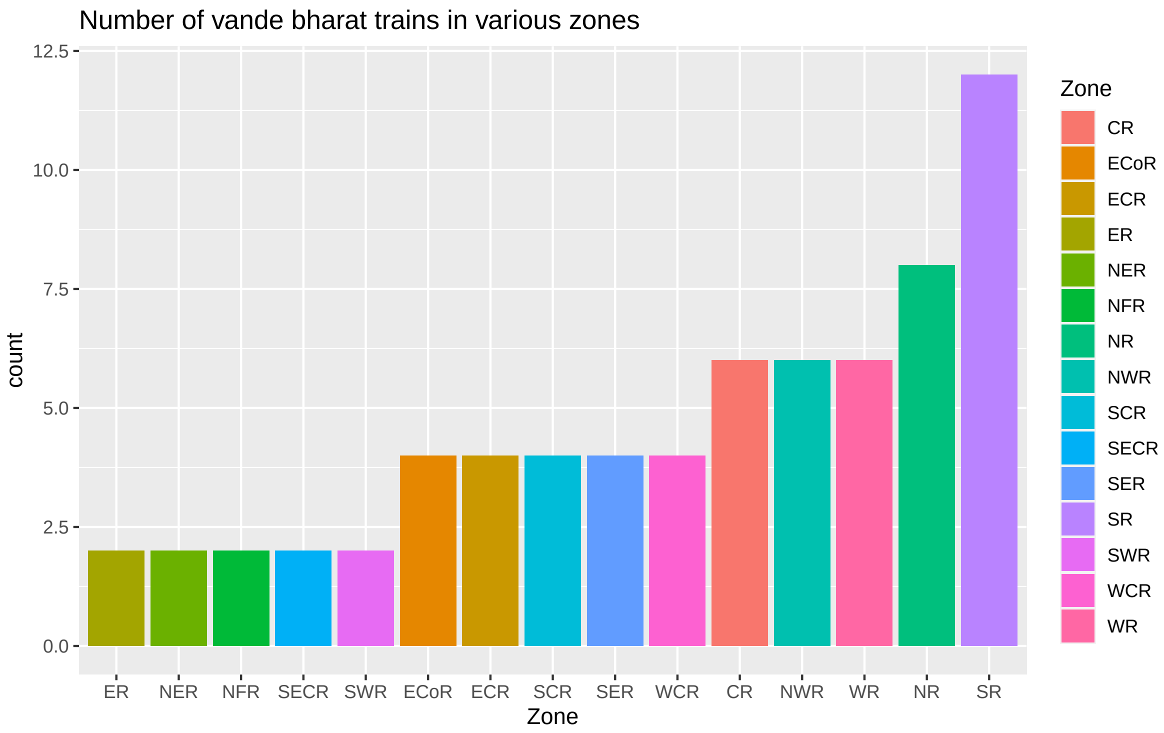
Figure 6: Passenger map

We further plotted a bar graph based on the number of trains in each zone for various train types:

1. Shatabdi



2. Vande Bharat



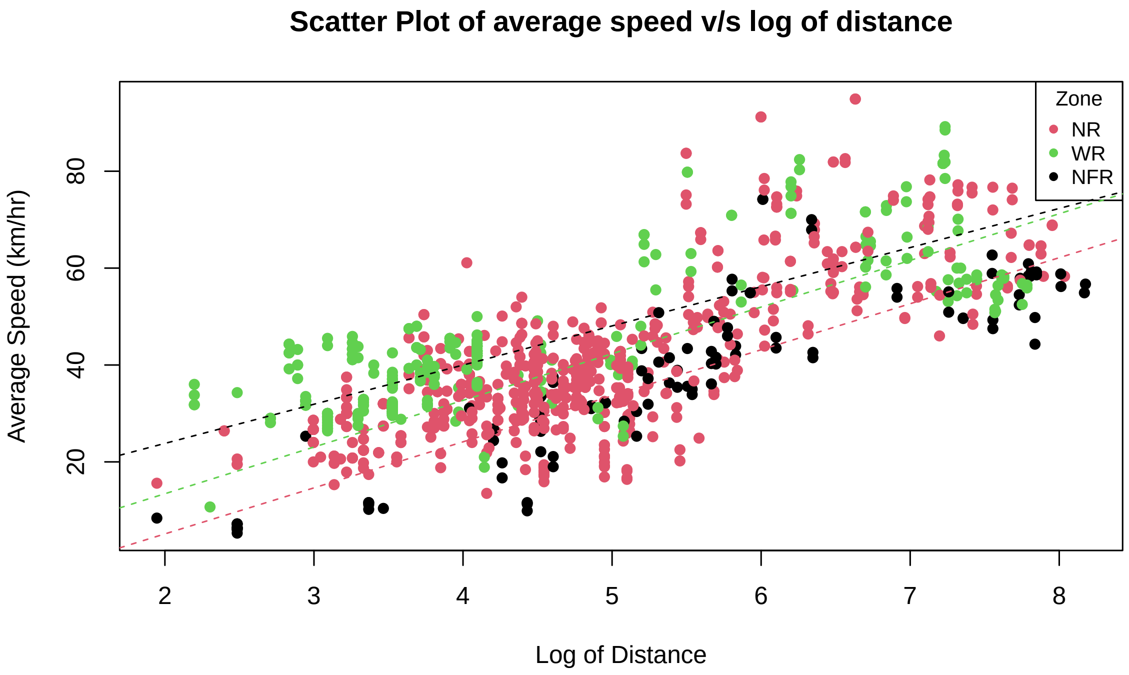
As is clearly visible from the graphs, the northern railways comprises most dense zone of Indian Railways.

Also, Vande Bharat train is comparatively more evenly distributed among different zones than Shatabdi.

5.3 Average Speed v/s Distance of the journey

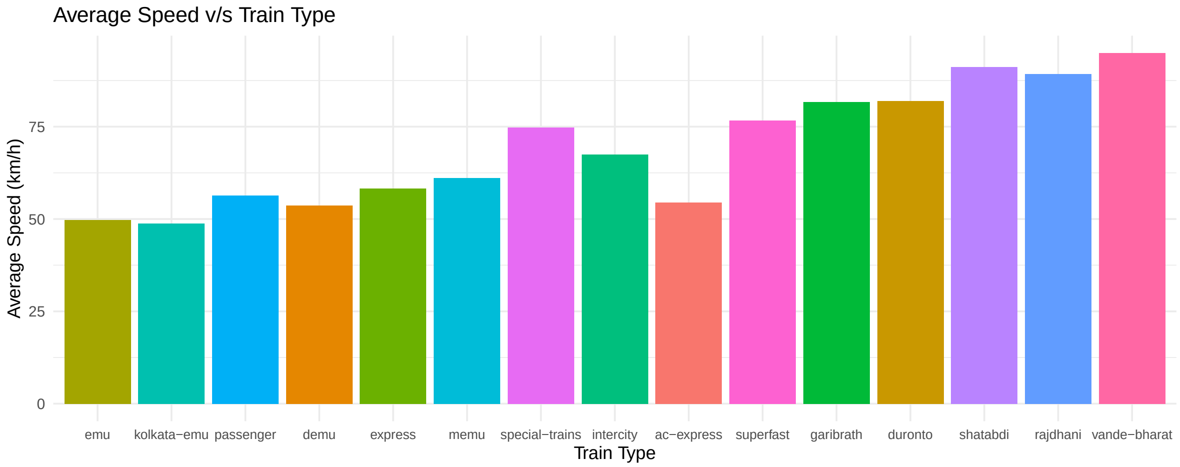
Next, we plotted the average speed of various trains against the distance covered by them on a give journey. Further, we color coded the data based on various zones.

It was clear as per our findings that as the distance covered by a train increases, its average speed also increases. Thus, if a user tries to find suitable trains for a longer journey, the train with greater speed are most likely to occur on the top.



5.3 Average Speed based on train type

Lastly, we plotted a bar graph based on average speed of various trains by categorising them with their type.



The graph indicates Vande Bharat as the fastest at 90 km/h, followed by Rajdhani Express. MEMU, EMU, and DEMU trains are the slowest. Within each type, there's a wide speed range due to factors like distance, stops.

6. Conclusion

6.1 In conclusion, the project has provided valuable insights into the factors contributing to delays in our railway system. Through meticulous data collection, analysis, and interpretation, we have identified patterns, trends, and potential root causes of delays.

6.2 From the visualization of delays and average speed , we can say that there is no clear relationship between them. The correlation between delays and average speed is 0.27.Now after dividing the plot into different areas by highlighting various duration groups(long (30-60 hrs), medium(15-30 hrs), short(0-15 hrs)), then we noticed that the number of short duration train is significantly more than medium and long duration trains. However, the range of delays is still limited. Even, within different subgroups, there is no clear relation between average speed and train delays.From the boxplot of different duration groups with respect to delays, the spread for medium duration trains is the most. The number of outlier is more in medium duration trains.

6.3 Rajdhani express a prominent train services in India as the name suggest it believe to be cover all the states of India but from the route map we see either it cover them unequally or does not provide service to those states.

6.4 On the other hand from the map we see Passenger train services in India is very dense and well distributed which helps rural population of India to travel better. From the route map of passenger train we observed that density of train is less in South and North zone, this is due to hill and plateau terrain and due to the local trains in southerns part of country.

6.5 After analyzing all train types we come with a statement that the new train Vande Bharat is evenly distributed across whole India. There is a positive correlation between distance of the journey and average speed of the train.So, for the long route journey the fastest trains are used more.

6.6 After analyzing the speed of all train types, we conclude that average speed of Rajdhani Express is more than any other train type.

This information is instrumental in formulating targeted strategies to enhance the efficiency and reliability of our train services.We can help to improve the punctuality of trains and to reduce the inconvenience caused by delays for passengers.

7.References

1. Total Train Info

2. Running Status

3. Travel Khana