WEB SCRAPING PROJECT

COMPARATIVE ANALYSIS OF TRAIN DELAY

Group 7

Avanish Mishra
Shivangi
Shivani Shukla

CONTENT

01 PROBLEM STATEMENT

02 OBJECTIVE

03 APPROACH

04 WHAT WE SCRAPED?

05 FINAL DATASET

06 RECOMMENDATION CHART

07 LIMITATIONS

PROBLEM STATEMENT

- Train delays in India significantly impact passengers, leading to psychological distress due to uncertainty of travel time. It influences efficient decisions on travel plans. It is of utmost concern when one has a connecting journey.

- Travelers face challenges deciding on trains in any particular route when train delays are of significant concern.

- While it's easy to check one train's delay, there's no simple way to compare delays for multiple trains on the same route at once.

OBJECTIVE

- This project aims to provide a platform for comparing train delays by scraping the average delay data of all trains departing from Mumbai from the **NTES** (National Train Enquiry System) website.

- To implement a feature on a small scale where users can enter the name of destination and day of journey and receive recommendation chart with trains ranked on the basis of past 7-day average delay.

APPROACH

Scraped raw data for each train

Cleaned the data and merged it into a single dataset

Built a recommendation chart

DATA COLLECTION

Website: https://enquiry.indianrail.gov.in/mntes/

Libraries: Selenium, BeautifulSoup, Pandas, Time

Steps:

1. Open Google Chrome
2. Open the website
3. Click on ‘More Info’
4. Click on ‘Average Delay’
5. Enter train number in the input box ‘Train No./Name’
6. Click on ‘Go’
7. Scrap the loaded page.

WHAT WE SCRAPED?

Scraped average delay data (past 7-days) for all 234 trains (as of 21st Sep) that originates from Mumbai

National Train Enquiry System

Average Delay (Last 7 Days)

Train No./Name

22221 X Search Go

22221 NZM RAJDHANI

Days of Run: Daily

Type: RAJDHANI

Sr.	Station	Code	Avg. Arr. Delay	Avg. Dep. Delay
1	CHHATRAPATI SHIVAJI MAHARAJ TERMINUS	CSMT		On Time
2	KALYAN JN	KYN	00:08	00:08
3	NASIK ROAD	NK	00:21	00:22
4	JALGAON JN	JL	00:14	00:18
5	BHOPAL	BPL	00:01	00:07
6	VIRANGANA LAKSHMIBAI JHANSI	VGLJ	00:05	00:07
7	GWALIOR	GWL	00:13	00:14
8	AGRA CANTT	AGC	00:15	00:16
9	HAZRAT NIZAMUDDIN JN	NZM	00:13	

RAW DATA

	sr_no	station	station_code	avg_arr_delay	avg_dep_delay
0	Train No./Name	NaN	NaN	NaN	NaN
1	22221 NZM RAJDHANI	NaN	NaN	NaN	NaN
2	Days of Run:Daily	Type:RAJDHANI	NaN	NaN	NaN
3	Sr.	Station	Code	Avg. Arr. Delay	Avg. Dep. Delay
4	1	CHHATRAPATI SHIVAJI MAHARAJ TERMINUS	CSMT	NaN	On Time
5	2	KALYAN JN	KYN	00:08	00:08
6	3	NASIK ROAD	NK	00:21	00:22
7	4	JALGAON JN	JL	00:14	00:18
8	5	BHOPAL	BPL	00:01	00:07
9	6	VIRANGANA LAKSHMIBAI JHANSI	VGLJ	00:05	00:07
10	7	GWALIOR	GWL	00:13	00:14
11	8	AGRA CANTT	AGC	00:15	00:16
12	9	HAZRAT NIZAMUDDIN JN	NZM	00:13	NaN
13	02-Oct 11:22	NaN	NaN	NaN	NaN

CLEANED DATA

	Unnamed: 0	sr_no	station	station_code	avg_arr_delay	avg_dep_delay	train_no	days	type
0	0	1	CHHATRAPATI SHIVAJI MAHARAJ TERMINUS	CSMT	0	00:00	22221	Daily	RAJDHANI
1	1	2	KALYAN JN	KYN	00:08	00:08	22221	Daily	RAJDHANI
2	2	3	NASIK ROAD	NK	00:21	00:22	22221	Daily	RAJDHANI
3	3	4	JALGAON JN	JL	00:14	00:18	22221	Daily	RAJDHANI
4	4	5	BHOPAL	BPL	00:01	00:07	22221	Daily	RAJDHANI
5	5	6	VIRANGANA LAKSHMIBAI JHANSI	VGLJ	00:05	00:07	22221	Daily	RAJDHANI
6	6	7	GWALIOR	GWL	00:13	00:14	22221	Daily	RAJDHANI
7	7	8	AGRA CANTT	AGC	00:15	00:16	22221	Daily	RAJDHANI
8	8	9	HAZRAT NIZAMUDDIN JN	NZM	00:13	0	22221	Daily	RAJDHANI

FINAL DATASET

	station	avg_arr_delay	train_no	days	type
0	MUMBAI CENTRAL	0.0	20901	Mon,Tue,Thu,Fri,Sat,Sun	VANDE BHARAT
1	BORIVALI	13.0	20901	Mon,Tue,Thu,Fri,Sat,Sun	VANDE BHARAT
2	VAPI	2.0	20901	Mon,Tue,Thu,Fri,Sat,Sun	VANDE BHARAT
3	SURAT	0.0	20901	Mon,Tue,Thu,Fri,Sat,Sun	VANDE BHARAT
4	VADODARA JN	4.0	20901	Mon,Tue,Thu,Fri,Sat,Sun	VANDE BHARAT
...	...	...	...	...	...
5009	ARA	23.0	13202	Daily	MAIL EXPRESS
5010	BIHTA	30.0	13202	Daily	MAIL EXPRESS
5011	SADISOPUR	36.0	13202	Daily	MAIL EXPRESS
5012	DANAPUR	42.0	13202	Daily	MAIL EXPRESS
5013	PATNA JN	0.0	13202	Daily	MAIL EXPRESS

5014 rows × 5 columns

RECOMMENDATION SYSTEM

- Input parameters:
  - Destination
  - Day of Journey

- The user input of destination can be city or specific station name (if name of station and city differs)

- The output is a data frame with the recommended trains matched with user input of destination and day of journey, which is then ordered by delay time.

- This recommendation chart can help the user choose a train with minimal delays for the given destination on a particular day.

- A few charts generated using the system are shown in the next slide.

Sample output 1

Destination input - Gorakhpur
Day of journey input - Monday

```
Recommendation chart for Journey to GORAKHPUR on Monday
```

	Train Number	Station	Delay Time (in Minutes)	Type
1	19091	GORAKHPUR JN	0.0	MAIL EXPRESS
2	20103	GORAKHPUR JN	6.0	SUPERFAST
3	19037	GORAKHPUR JN	9.0	MAIL EXPRESS
4	11055	GORAKHPUR JN	15.0	MAIL EXPRESS
5	22538	GORAKHPUR JN	76.0	SUPERFAST
6	12165	GORAKHPUR JN	99.0	SUPERFAST
7	15066	GORAKHPUR JN	213.0	MAIL EXPRESS
8	15017	GORAKHPUR JN	228.0	MAIL EXPRESS

Sample output 2

Destination input - Tatanagar
Day of journey input - Tuesday

```
Recommendation chart for Journey to TATANAGAR on Tuesday
```

	Train Number	Station	Delay Time (in Minutes)	Type
1	18029	TATANAGAR JN	215.0	MAIL EXPRESS
2	12859	TATANAGAR JN	233.0	SUPERFAST
3	12261	TATANAGAR JN	246.0	DURONTO EXPRESS
4	12101	TATANAGAR JN	298.0	SUPERFAST
5	12809	TATANAGAR JN	300.0	SUPERFAST

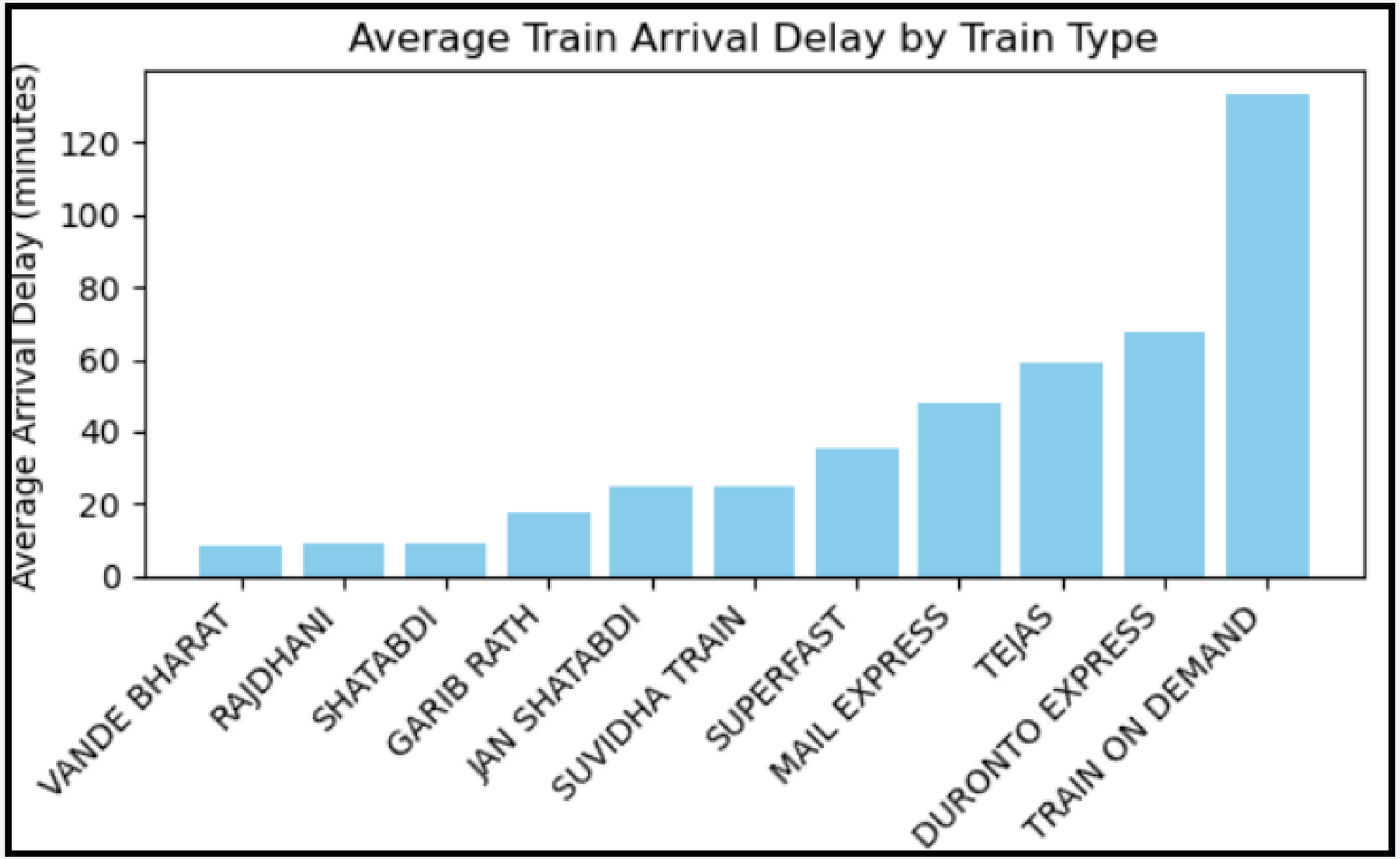
Sample output 3

Destination input - Hazrat Nizamuddin
Day of journey input - Wednesday

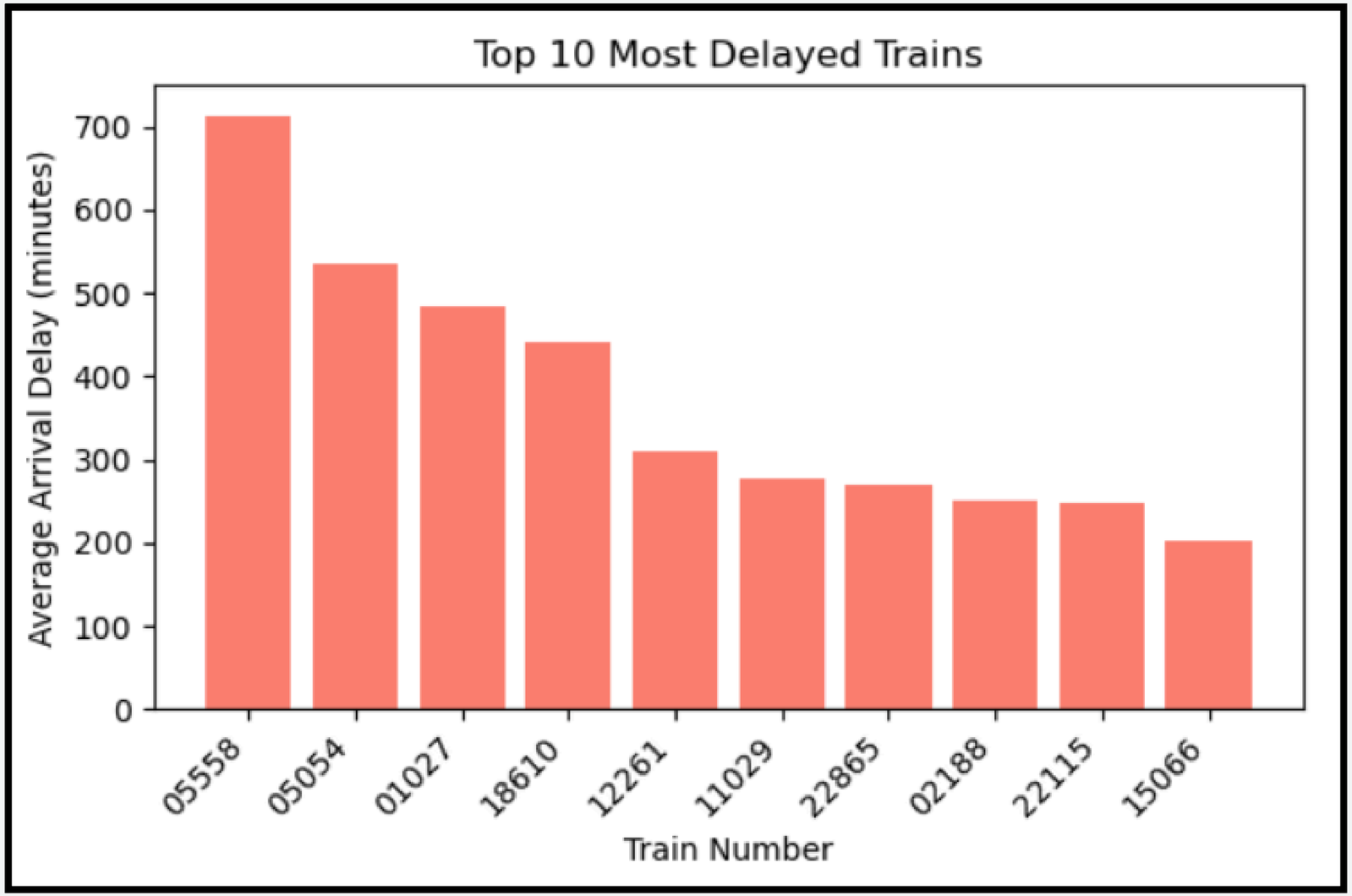
```
Recommendation chart for Journey to HAZRAT NIZAMUDDIN on Wednesday
```

	Train Number	Station	Delay Time (in Minutes)	Type
1	22917	HAZRAT NIZAMUDDIN JN	6.0	SUPERFAST
2	12953	HAZRAT NIZAMUDDIN JN	7.0	RAJDHANI
3	12903	HAZRAT NIZAMUDDIN JN	7.0	SUPERFAST
4	22221	HAZRAT NIZAMUDDIN JN	13.0	RAJDHANI
5	12925	HAZRAT NIZAMUDDIN JN	17.0	SUPERFAST
6	19019	HAZRAT NIZAMUDDIN JN	20.0	MAIL EXPRESS
7	12907	HAZRAT NIZAMUDDIN JN	23.0	SUPERFAST
8	11057	HAZRAT NIZAMUDDIN JN	48.0	MAIL EXPRESS
9	12137	HAZRAT NIZAMUDDIN JN	55.0	SUPERFAST

DELAY COMPARISON BY TRAIN TYPE



DELAY COMPARISON BY TRAIN NUMBER



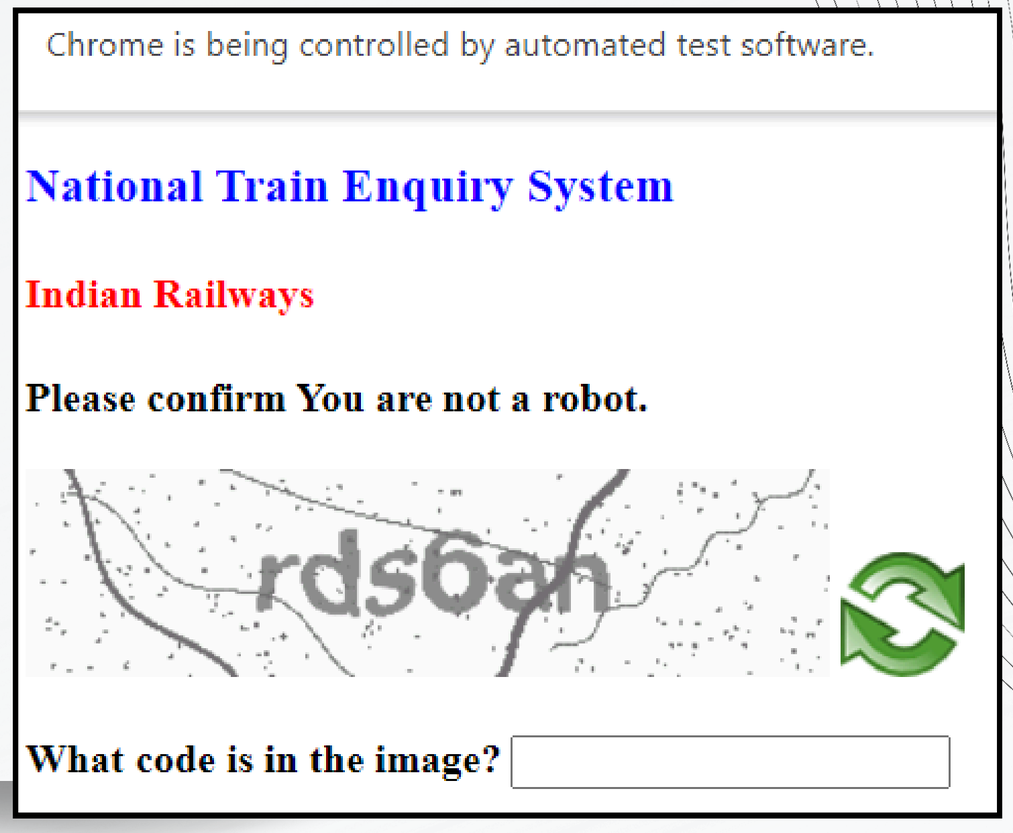
- Train no. 05558, 05054, 01027, 02188 are Trains on demand.
- Train no. 18610 is mail express
- Train no. 12261 is Duronto express.

CHALLENGES AND SCOPE

- During the scraping process, we encountered CAPTCHAs that require human interaction. This step cannot be avoided due to legal and ethical concerns.

- This project includes trains originating from Mumbai only (as of 21 September 2024).

- **Future Scope:** It can be extended to cover other routes and trains.



THANK YOU!