Exploratory Data Analysis on

Passport Services

by

Group 17

Dhyey Patel
ID: 202103053
Course: B.Tech MnC

Mit Desai
ID: 202103013
Course: B.Tech MnC

Aarzoo Khambhoo
ID: 202103026
Course: B.Tech MnC

Course Code: IT 462
Semester: Winter 2024

Under the guidance of

Dr. Gopinath Panda

Dhirubhai Ambani Institute of Information and Communication Technology

May 2, 2024

Acknowledgment

I wanted to take a moment to express my sincere appreciation for the exceptional guidance and support you provided me during my project "Passport Services" Your mentorship has been invaluable and played a pivotal role in ensuring the success of this endeavor.

I consider myself very lucky to have had the opportunity to benefit from your expertise and mentorship. Your insightful advice, coupled with extensive knowledge in the field, has significantly influenced the quality and scope of the project. Your constructive feedback and suggestions not only helped me navigate challenges but deepened my understanding of the subject.

I would also like to show my gratitude to the entire team at DAIICT for fostering a culture of collaboration and innovation. The resources and facilities provided by the institution have been instrumental in facilitating comprehensive research and analysis, thereby enriching the outcome of the project.

Moreover, I am thankful to my peers and colleagues for their unwavering support and camaraderie throughout this journey. Their contributions have undeniably enhanced the development of the "MSP WHEAT" project.

Entering into the "Employment" project has been an immensely fulfilling experience for me. I am confident that the knowledge and skills acquired during this endeavor will serve as a solid foundation for my future pursuits.

Once again, I want to express my deepest gratitude for your invaluable guidance and support. Your mentorship was indispensable, and I am genuinely apprecative of the opportunity to learn from you.

Sincerely,

Dhyey Patel, 202103053
Mit Desai, 202103013
Aarzoo Khambhoo, 202103026

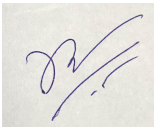
DECLARATION

We, [202103053,202103013,202103026] now declare that the EDA project work presented in this report is our original work and has not been submitted for any other academic degree. All the sources cited in this report have been appropriately referenced.

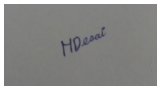
We acknowledge that the data utilized in this project has been sourced from data.gov.in and data.gov.in. We affirm that we have complied with the terms and conditions specified on the website for accessing and using the dataset. We hereby confirm that the dataset employed in this project is accurate and authentic to the best of our knowledge.

We acknowledge that we have received no external help or assistance in conducting this project except for the guidance provided by our mentor, Prof. Gopinath Panda. We declare no conflict of interest in conducting this EDA project.

We have now signed the declaration statement and confirmed the submission of this report on 29 April 2024.



Dhyey Patel
ID: 202103053
Course: B.Tech MnC



Mit Desai
ID: 202103013
Course: B.Tech MnC

Aarzoo Khambhoo
ID: 202103026
Course: B.Tech MnC

CERTIFICATE

This is to certify that Group 17 comprising Dr. Gopinath Panda and Dr. Gopinath Panda has completed an exploratory data analysis (EDA) project on the PROJECT, which was obtained from data.gov.in, data.gov.in, US Reports and Statistics.

The EDA project presented by Group 17 is their original work. It was completed under the guidance of the course instructor, Prof. Gopinath Panda, who provided support and guidance throughout the project. The project is based on a thorough analysis of the PROJECT dataset, and the results presented in the report are based on the data obtained from the dataset.

This certificate is issued to recognize the successful completion of the EDA project on the PROJECT, which demonstrates the analytical skills and knowledge of the students of Group 17 in the field of data analysis.

Signed,
Dr. Gopinath Panda,
IT 462 Course Instructor
Dhirubhai Ambani Institute of Information and Communication Technology
Gandhinagar, Gujarat, INDIA.

May 2, 2024

Contents

List of Figures

Chapter 1. Introduction

1.1 Project idea

This project aims to conduct an Exploratory Data Analysis (EDA) on the number of passports issued in India as well as the United States. The project uses exploratory data analysis techniques to explore and understand the patterns, trends, and factors influencing the number of passports issued across different countries.

1.2 Data Collection

Data Sources

- The data source for this project is a dataset containing the number of passports issued (category-wise) in different parts of India as well as the United States across different countries.

1.3 Dataset Description

Dataset:

- The data source for this project is a dataset containing the number of passports issued in different parts of India as well as the United States across different countries.

Features of Dataset 1:

- **Service Name:** It portrays that the applications received by the RPO are whether Scheme-wise or Gender-wise received.
- **RPO Name:** It shows that the passport data received are from the RPO of which location.
- **Scheme Type:**
  - **Gender-wise number of passport applications received by RPO:** The genders of the passport applications received by RPO which include MALE, FEMALE, and TRANSGENDER in the gender-wise scheme.
  - **Scheme-wise passport applications received by RPO:** This includes time duration of whether the application is based on First cum first serve basis or Tatkal(Immediate) basis.
- **Last week count:**

  - **Count of Applications received per week:** It gives us the numerical value of total applications that RPO received last week.

- **Last Month Count:**
  - **Count of Applications received per month:**It gives us the numerical value of total applications that RPO received last month.
- **Year Till Date:**
  - It indicates the number of passport applications received by the RPO from the past year till the date of registration.
- **Date:**
  - Date on which the applications are received by the RPO.

	ServiceName	RpoName	SchemeType	LastWeekCount	LastMonthCount	YearTillDate	Date
0	Applications Received - Scheme wise	RPO Ahmedabad	Normal	7174	51414	567081	2019-11-03 04:29:09.831353
1	Applications Received - Scheme wise	RPO Ahmedabad	Tatkaal	160	1368	14930	2019-11-03 04:29:09.831353
2	Applications Received - Gender wise	RPO Ahmedabad	FEMALE	3029	21273	233635	2019-11-03 04:29:09.831353
3	Applications Received - Gender wise	RPO Ahmedabad	MALE	4305	31498	348282	2019-11-03 04:29:09.831353
4	Applications Received - Gender wise	RPO Ahmedabad	TRANSGENDER	0	1	15	2019-11-03 04:29:09.831353

Figure 1.1: Dataset 1

Features of Dataset 2:

- **RPO Name:** It portrays the location or name of the RPO by which the applications are received.
- **Dispatch on Same Day:** It gives us the value of how many passports RPO dispatch on the same day the application is received.
- **Dispatch in 1 days:**
  - It gives us the value of how many passports RPO dispatch in 1 day after the application is received.
- **Dispatch in 2-3 days:**
  - It gives us the value of how many passports RPO dispatch in 2-3 days after the application is received.
- **Dispatch in 4-7 days:**
  - It gives us the value of how many passports RPO dispatch in 4-7 days after the application is received.
- **Dispatch in 8-14 days:**

  - It gives us the value of how many passports RPO dispatch in 8-14 days after the application is received.

- **Dispatch in 15-21 days:**
  - It gives us the value of how many passports RPO dispatch in 15-21 days after the application is received.
- **Dispatch in 22-30 days:**
  - It gives us the value of how many passports RPO dispatch in 22-30 days after the application is received.
- **Dispatch in more than 30 days:**
  - It gives us the value of how many passports RPO dispatch in more than 30 days after the application is received.
- **Total:**
  - It gives us the total value of how many passports RPO dispatches.

	RPO Name	Dispatch on Same Day	Dispatch in 1 days	Dispatch in 2 – 3 days	Dispatch in 4 – 7 days	Dispatch in 8 – 14 days	Dispatch in 15 – 21 days	Dispatch in 22 – 30 days	Dispatch in more than 30 days	Total
0	CPV Delhi	107	274	301	664	421	259	108	636	2770
1	RPO Ahmedabad	5996	734	18252	84677	86716	103933	134725	109180	544213
2	RPO Amritsar	88	7625	9488	7248	29579	41864	15055	17802	128749
3	RPO Bangalore	444	26349	106424	88492	65579	68181	89498	173448	618415
4	RPO Bareilly	43	1843	939	2822	8665	15132	55842	83717	169003

Figure 1.2: Dataset 2

Features of Dataset 3:

- **Fiscal Year:** It gives us the fiscal year for which we are analyzing our data.
- **US passports issued:** It gives us the value of how many US passports were issued in a given fiscal year.

	Fiscal Year	US passports issued
0	1974	24,15,003
1	1975	23,34,359
2	1976	28,16,678
3	1977	31,07,122
4	1978	32,34,471

Figure 1.3: Dataset 3

1.4 Packages required

Pandas (pd)

- **Why it's Required:** Pandas is critical to our research because it provides a solid framework for processing structured data, which is required for analysing the passport dataset. Its characteristics ensure that the dataset is ready for comprehensive study by facilitating data preparation chores such as cleaning, conversion, and aggregation.
- **Uses for the Report:** With Pandas, you can effectively import the passport dataset into a DataFrame and get powerful capabilities for indexing, slicing, and filtering data to extract pertinent information for analysis. Pandas facilitates the exploration of significant statistics and trends within the dataset with operations like describe(), mean(), median(), and value counts(). This allows for a comprehensive understanding of the dynamics of food price inflation across national borders.

Matplotlib (plt)

- **Why it's Required:** The success of this project depends on Matplotlib's extensive capabilities for creating various plot types and visualisations, which are essential for examining trends, patterns, and correlations within the passport dataset. Its versatility allows for the creation of static, interactive, and publication-quality visualisations that effectively aid in the comprehension and sharing of research findings.
- **Uses for the Report:** A variety of visualisations, such as line graphs, bar charts, scatter plots, and histograms, may be made with Matplotlib to compare inflation rates between countries, display trends in passport data over time, and look at correlations with other data. Matplotlib can be used to visualise regional variances in passports and uncover potential explanations influencing this data when used with Basemap or Cartopy.

datetime

- **Why it's Required:** Because it offers a variety of classes and methods for working with dates, times, time zones, and durations, the DateTime library is indispensable. Applications that need to precisely modify, compare, or display dates and timings must have this.
- **Uses for the Report:** Because it offers a variety of classes and methods for working with dates, times, time zones, and durations, the DateTime library is indispensable. Applications that need to precisely modify, compare, or display dates and timings must have this.

statsmodels.api

- **Why it's Required and Uses for the report:** For projects including econometric analysis, predictive modelling, time series forecasting, experimental design, hypothesis testing, and other statistical analyses, statsmodels.api is necessary because to its strong statistical modelling capabilities. These studies enable data-driven insights and decision-making.

sklearn.preprocessing

- **Why it's Required and Uses for the report:** In order to ensure data readiness and compatibility with machine learning algorithms, we use sklearn.preprocessing in our project for feature engineering, missing value imputation, feature scaling, and categorical variable encoding. This requires sklearn.preprocessing for its essential data preprocessing capabilities.

Chapter 2. Data Cleaning

The requirement to clean the data arose only in the first dataset as there were a lot of cells in the columns having value allotted zero, Last Week Count, Last Month Count, and Year Till Date. Those zeroes were first converted to **NaN** after which the individual column mean was taken and in each column, the zeroes were then replaced by the respective mean of those columns. That was the case for numerical data. Now, for the categorical data if the problem would have arisen then we would replace all the NaN values of a column with the mode of each respective column.

2.1 Missing data analysis

The missing data can be analyzed also by:

- **Deletion of the data:** This option should only be used when there are very few columns with non-null values filled in, as it may result in bias and information loss if any missing data is present.
- **Model Based filling of data:** Several models based on predictive regression may be employed to forecast the column's missing data values.
- **Assess Impact on Analysis:**To determine the possible consequences of missingness on findings, we can compare results with and without traded values or perform sensitivity studies, which involve determining the data's correctness, precision, specificity, and recall values.

Initially, dataset_2 had 0 NULL values. But there were significant values which were '0'. We assumed that these values were 0 because of data inconsistency and missing data. Hence, we first replaced all the 0 values with NaN so that we can treat it as missing values analyze the missing values and then further impute the values.

```
ServiceName          0
RpoName              0
SchemeType           0
LastWeekCount      148
LastMonthCount     138
YearTillDate       117
Date                 0
dtype: int64
```

Figure 2.1: Null values in dataset 2

We also plotted the heap map for the missing values to visualize it better. We found missing values in the numerical columns- LastWeekCount, LastMonthCount and YearTillDate.

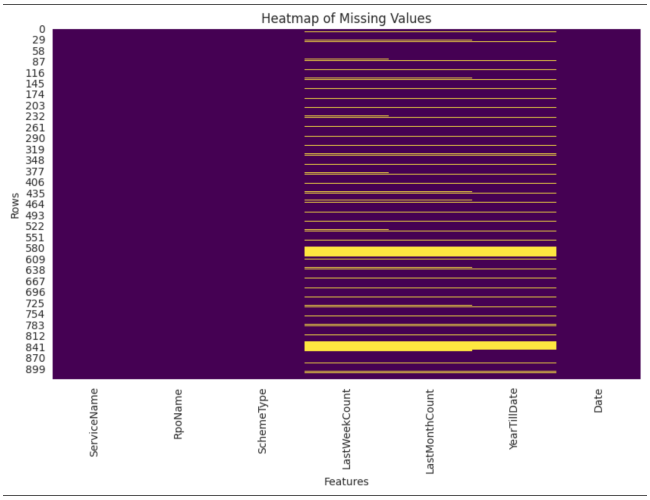


Figure 2.2: Heat map for missing values

2.2 Imputation

imputation is the process of forecasting the missing data values using the column's previously observed data. The information that is currently accessible is used to fill in the missing values.

- **Last Observation Carried Forward (LOCF):** IIn a time series or longitudinal set of data, impute missing values with the value that was most recently seen.
- **Regression Imputation:** Regression models can be used to forecast missing values in a dataset by taking into account other variables. Although it depends on linearity and might be susceptible to outliers, this approach captures the relationships between the variables.

- **K-Nearest Neighbors (KNN) Imputation:** Based on the values of related observations in the dataset, forecast missing values. In order to determine the closest neighbours and impute missing values appropriately, KNN imputation takes into account the distances between observations in the feature space.
- **Multiple Imputation:** Create datasets with multiple imputed values by taking a sample from the missing value distribution. In comparison to single imputation techniques, this approach gives more accurate estimates and takes into consideration imputed value uncertainty.

We have replaced all the NULL values with the mean of that particular feature and below is the heatmap after imputing the NULL values.

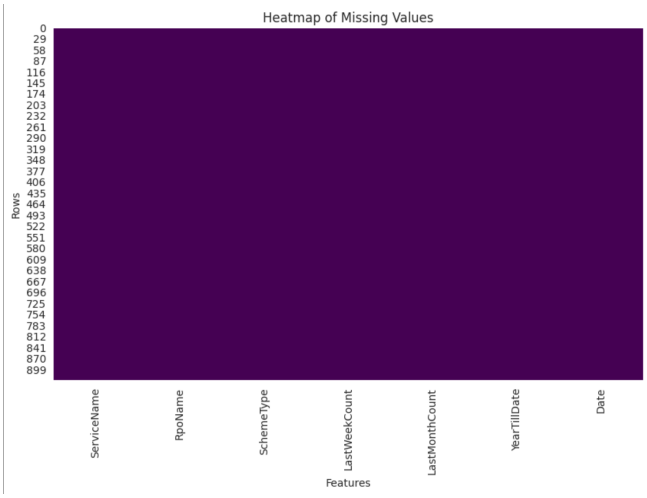


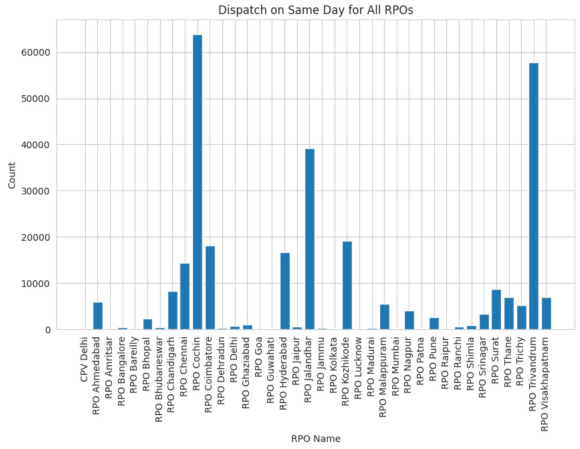
Figure 2.3: Heatmap after imputing null values

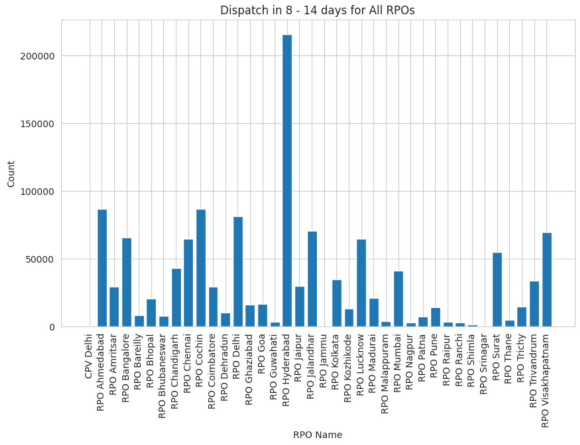
Chapter 3. Visualization

Visualizations help in gaining a deeper understanding of the dataset by revealing its structure and potential patterns. Techniques like scatter plots, histograms, box plots, and pie charts provide a visual summary of the data distribution, central tendency, and spread.

3.1 DataSet-1

3.1.1 Bar Graphs and Histograms





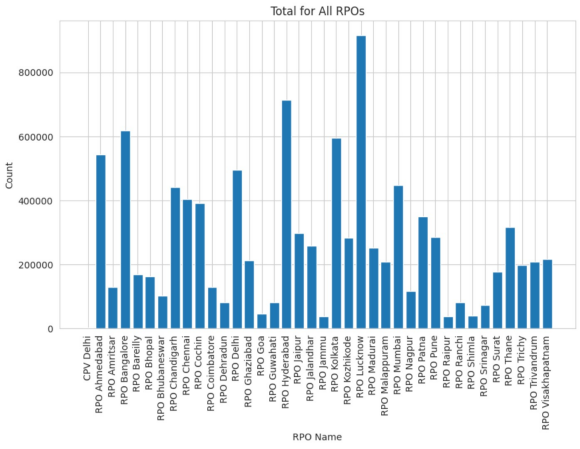
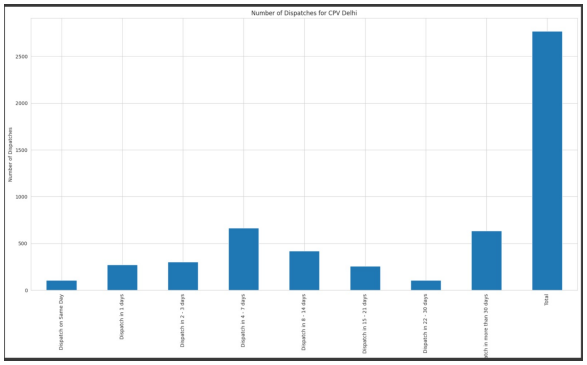
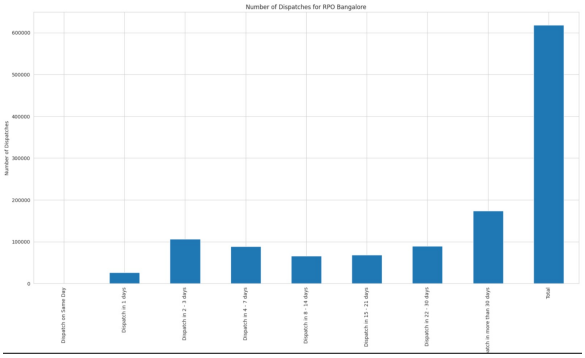


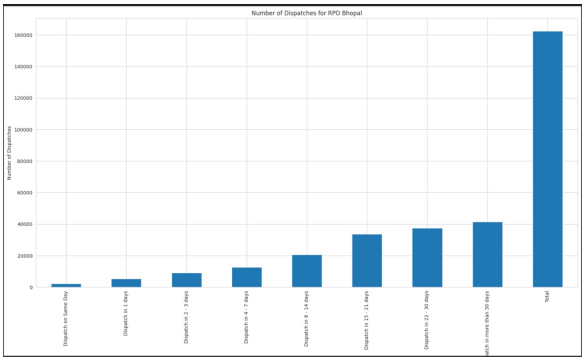
Figure 3.1 gives us a basic idea of dispatching the passports on the same day the applications were received and from which we figure out which RPOs are most efficient in terms of their speed of dispatching the passports of their customers. Thus, we conclude that the RPO of Cochin, Triandrum, and Jalandhar are the top three in service, respectively.

Similarly, the figure 3.2 provides us the information of the RPOs dispatching the passports after eight to fourteen days of registration by applicants. Here, it is evident that RPO Hyderabad gives out maximum dispatches after one week to two weeks.

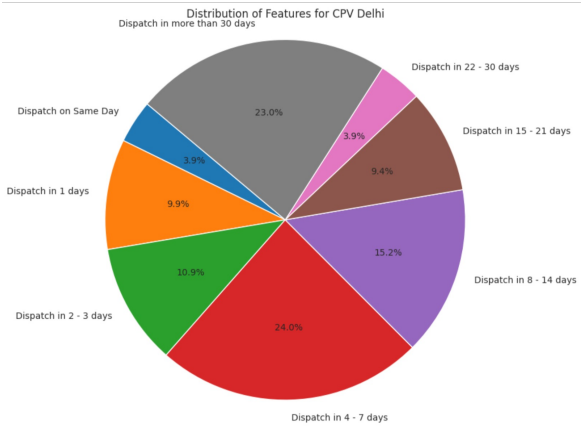
Now, the figure 3.3 gives us an idea of total passports dispatched over a year and we conclude from the graph that RPO Lucknow is having maximum passports being dispatched from its branch to its customers.

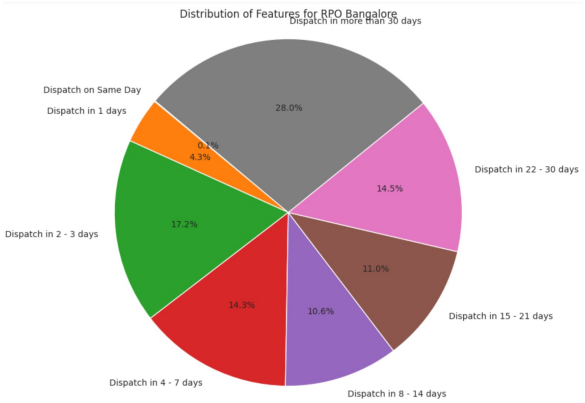


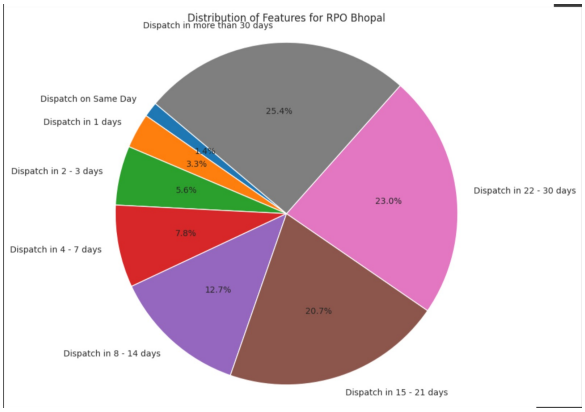




The above figures are the bar chart-type graph of the data given in the histogram data which is RPOs vs. passports Dispatched. The highest bar displays the highest number of passports dispatched in that given period mentioned in the bar chart.







The above figures are the pie chart-type graph of the data given in the bar chart data which is RPOs Vs Passports Dispatched. The sector area that occupies the maximum area of the pie chart has the most number of passports dispatched in that given period mentioned in the pie chart.

3.1.2 Statistical Observations

RPO Name	Highest Value	Feature
CPV Delhi	664	Dispatch in 4 – 7 days
RPO Ahmedabad	134725	Dispatch in 22 – 30 days
RPO Amritsar	41864	Dispatch in 15 – 21 days
RPO Bangalore	173448	Dispatch in more than 30 days
RPO Bareilly	83717	Dispatch in more than 30 days
RPO Bhopal	41364	Dispatch in more than 30 days
RPO Bhubaneswar	42745	Dispatch in more than 30 days
RPO Chandigarh	117309	Dispatch in 22 – 30 days
RPO Chennai	79564	Dispatch in 15 – 21 days
RPO Cochin	94989	Dispatch in more than 30 days
RPO Coimbatore	29489	Dispatch in 8 – 14 days
RPO Dehradun	20422	Dispatch in more than 30 days
RPO Delhi	105102	Dispatch in 15 – 21 days
RPO Ghaziabad	115661	Dispatch in more than 30 days
RPO Goa	16980	Dispatch in 8 – 14 days
RPO Guwahati	54242	Dispatch in more than 30 days
RPO Hyderabad	215722	Dispatch in 8 – 14 days
RPO Jaipur	104838	Dispatch in 22 – 30 days
RPO Jalandhar	110667	Dispatch in 15 – 21 days
RPO Jammu	36865	Dispatch in more than 30 days
RPO Kolkata	391706	Dispatch in more than 30 days
RPO Kozhikode	56444	Dispatch in 4 – 7 days
RPO Lucknow	301757	Dispatch in more than 30 days
RPO Madurai	56250	Dispatch in more than 30 days
RPO Malappuram	86286	Dispatch in more than 30 days
RPO Mumbai	95377	Dispatch in more than 30 days
RPO Nagpur	66150	Dispatch in more than 30 days
RPO Patna	248609	Dispatch in more than 30 days
RPO Pune	157330	Dispatch in more than 30 days
RPO Raipur	16934	Dispatch in more than 30 days
RPO Ranchi	35141	Dispatch in more than 30 days
RPO Shimla	29789	Dispatch in more than 30 days
RPO Srinagar	66020	Dispatch in more than 30 days
RPO Surat	54959	Dispatch in 8 – 14 days
RPO Thane	208028	Dispatch in more than 30 days
RPO Trichy	43153	Dispatch in more than 30 days
RPO Trivandrum	57775	Dispatch on Same Day
RPO Visakhapatnam	69776	Dispatch in 8 – 14 days

Figure 3.1: Highest Frequency

The above output shows us the highest frequency of a particular dispatch duration for every RPO. For example, CPV Delhi as most of their dispatches in 4-7 days. Similarly, you can analyse this information for each and every RPO.

	RPO Name	Dispatch on Same Day	Dispatch in 1 days	\
8	RPO Chennai	14405	58841	
14	RPO Goa	99	8416	
36	RPO Trivandrum	57775	34396	

	Dispatch in 2 – 3 days	Dispatch in 4 – 7 days	Dispatch in 8 – 14 days	\
8	53040	30203	64939	
14	5735	2158	16980	
36	8571	2623	33762	

	Dispatch in 15 – 21 days	Dispatch in 22 – 30 days	\
8	79564	48444	
14	9850	797	
36	43149	17472	

	Dispatch in more than 30 days	Total
8	54765	404201
14	2068	45303
36	11625	209373

Figure 3.2: atleast 80% in either same day or within 2-3 days

min80.png

Figure 3.3: atleast 80% in either 15-21 days or more than 30 days

On the left, we have the RPO's who have atleast 80% of their dispatches in either one the same day/in 1 day/in 2-3 days,i.e, these are the ones which can be classified as relatively effective and fast dispatchers.

On the contrast, on the right side, we have the RPO's who have atleast 80% of their dispatches in either 15-21 days/22-30 days/more than 30 days,i.e, these can be classified as the ones which are the slowest dispatchers out of all.

	RPO Name	Percentage
36	RPO Trivandrum	48.116042
14	RPO Goa	31.454871
8	RPO Chennai	31.243367
24	RPO Malappuram	29.445306
21	RPO Kozhikode	27.757295
35	RPO Trichy	26.924696
10	RPO Coimbatore	25.530159
0	CPV Delhi	24.620939
25	RPO Mumbai	23.878518
18	RPO Jalandhar	23.398283

Figure 3.4: Top-10 fastest dispatch RPOs

	RPO Name	Percentage
22	RPO Lucknow	5.944779e+07
20	RPO Kolkata	2.354090e+07
4	RPO Bareilly	1.845943e+07
1	RPO Ahmedabad	1.716852e+07
27	RPO Patna	1.418270e+07
7	RPO Chandigarh	9.226360e+06
13	RPO Ghaziabad	7.832975e+06
17	RPO Jaipur	7.027226e+06
16	RPO Hyderabad	6.206558e+06
19	RPO Jammu	6.135219e+06

Figure 3.5: 10 Slowest RPOs

Using the same logic and concept as the last observation, we have calculated the top-10 fastest dispatch RPOs and 10 slowest RPOs using the percentage of fast dispatches to slow dispatches.

The above figure 3.7 is box-plot. A box plot is a visual summary of the mean/median/mode i.e. central tendency, variance, and variability of a dataset, along with any potential outliers. It consists

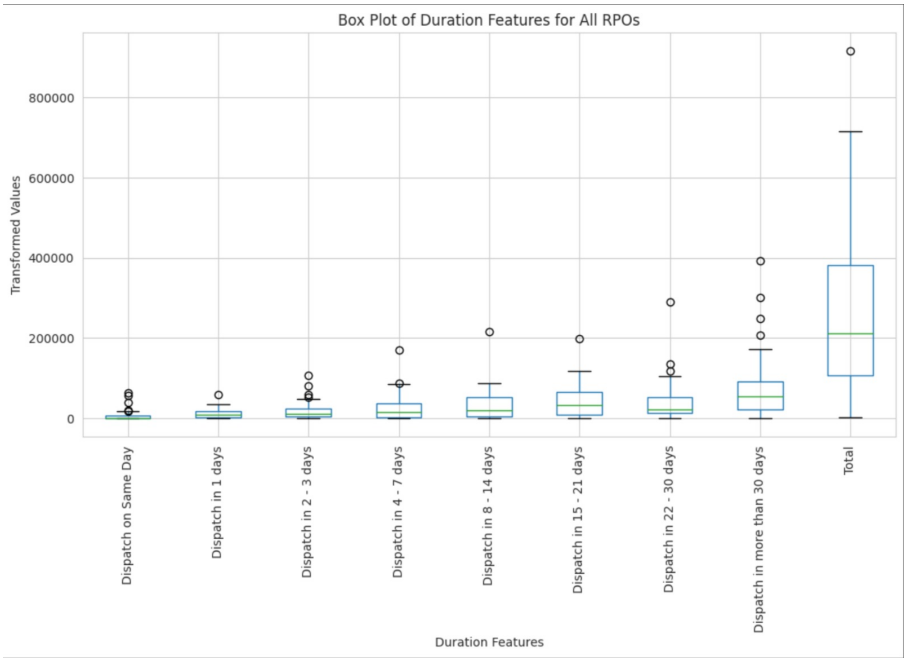


Figure 3.6: Box Plot

of several key elements:

3.1.3 Box Plot

Box

The interquartile range (IQR) of the data, which includes the middle 50%, or median, of the observations, is represented by the box plot. The box's top edge denotes the third quartile (Q3), and the bottom edge shows the first quartile (Q1). The dispersion of the data is indicated by the length of the box (Q3 - Q1).

Median:

The median (Q2) of the data, or the midway value when the dataset is sorted in ascending order, is represented by a horizontal line inside the box.

Whiskers:

The whiskers, which are usually 1.5 times the IQR from the first and third quartiles, reach the minimum and maximum values within a given range from the box's margins. Plotting of observations outside of this range is common, as they are regarded as possible outliers.

Outliers:

Data points that are outside of the whiskers' designated range are known as outliers. These could be anomalies or extreme levels in the data that need more research. When examining the distributions of several groups or variables within a dataset, box plots are especially helpful. They can discover any

potential outliers or skewed distributions, as well as differences in central tendency, spread, and variability between groups.

Here, we see that the size of the box grows as the passports take longer to be sent out. When the passports are dispatched later, the head length also grows. That occurs as a result of the number of passports shipped in the box before it being added to the box that comes after it, and so forth.

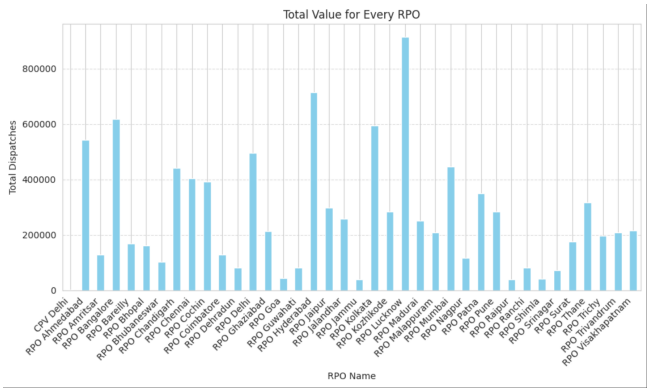


Figure 3.7: Total dispatches for every RPO

In this image, we have displayed the bar graph for Total Dispatches for every RPO showing us the activity levels of all the RPOs. The ones with the higher graphs are the ones that dispatch the most number of passports in general. Observation:- **RPO Lucknow, RPO Hyderabad and RPO Banglore** are the most active and high volume dispatchers.

RPO Name	Total
RPO Lucknow	915737
RPO Hyderabad	715228
RPO Bangalore	618415
RPO Kolkata	596090
RPO Ahmedabad	544213
RPO Delhi	496421
RPO Mumbai	447243
RPO Chandigarh	442494
RPO Chennai	404201
RPO Cochin	392547

Figure 3.8: Highest 10 RPOs

RPO Name	Total
CPV Delhi	2770
RPO Jammu	38334
RPO Raipur	38428
RPO Shimla	40757
RPO Goa	45303
RPO Srinagar	72358
RPO Dehradun	81555
RPO Ranchi	81594
RPO Guwahati	82258
RPO Bhubaneswar	103308

Figure 3.9: Lowest 10 RPOs

In the above outputs, we have calculated the highest 10 and lowest 10 RPO's based on the total dispatches they send out.
This displays their activity and volume on a large scale.

3.1.4 Correlation:-

Correlation analysis is used to identify the strength and direction of the linear relationship between variables in a collection. Values close to 0 imply little to no link, positive values indicate positive correlation, and negative values indicate negative correlation. It provides the correlation coefficient, a value with a range of -1 to 1.

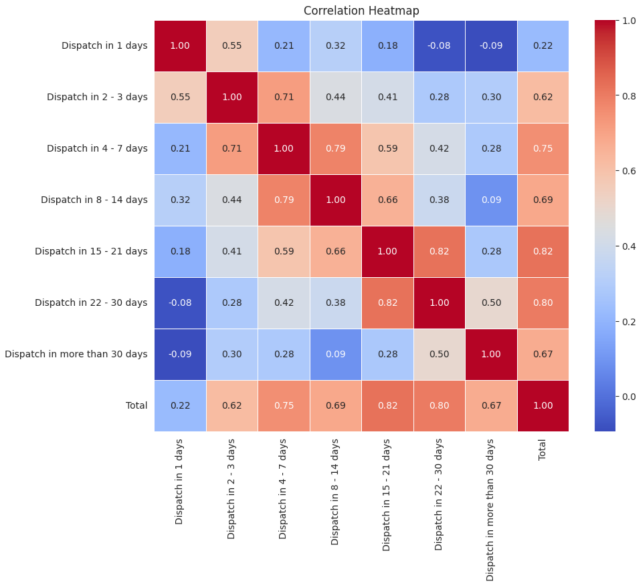
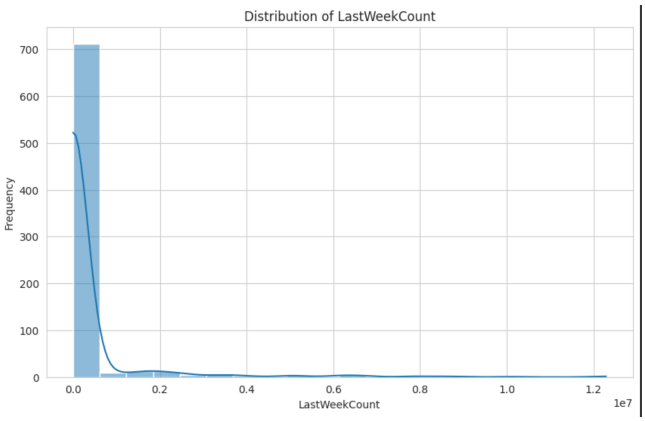


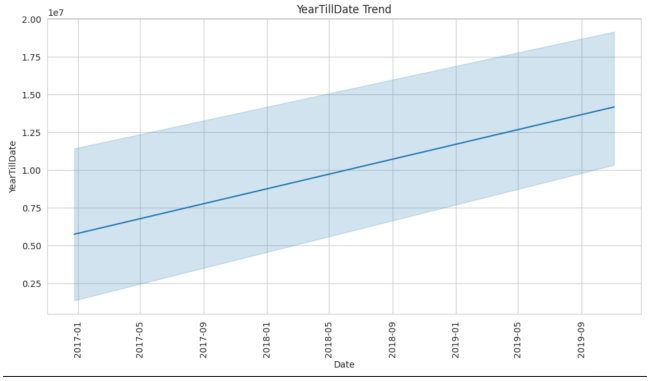
Figure 3.10: Correlation plot

3.2 Dataset-2

We will be looking at data for Dataset-2 and see what we could observe from all the visualisations.



There is not much to see here other than the fact that a minority of the values are outliers and their values are so significantly greater that the other normal values are relatively insignificant. It is almost impossible to infer things from the normal plots.



Here we can notice that the YearTillDay Values are linearly increasing which means the passport services observe constant growth every year on year.

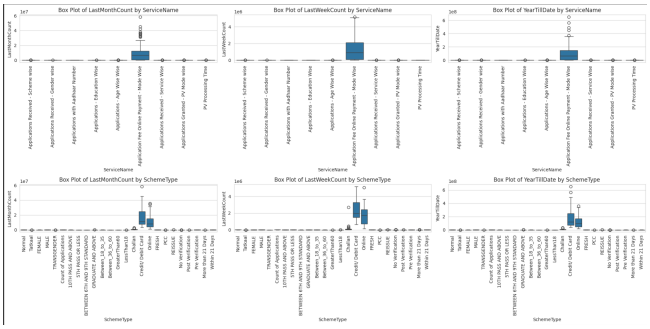


Figure 3.11: Box plots of scheme type and service name

In the above figure, we have plotted all the three numerical values with respect to its scheme type and service name in separate plots. We can observe that Mode Wise service is exponentially greater used than any of the other schemes. Also, the Credit/Debit Card and Online Scheme Types are significantly larger than other types.

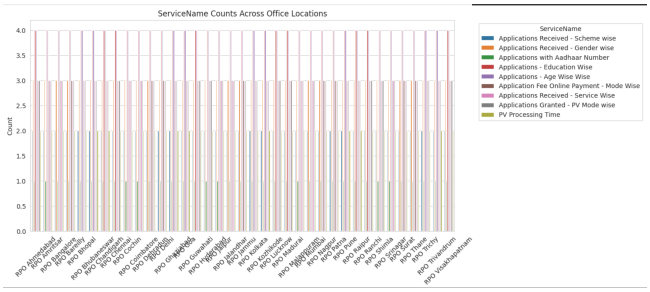


Figure 3.12: Comparison in the distributions

In this figure, we can see the comparision in the distribution in ServiceType for each RPO. Clearly we can observe that most of the RPOs majority Service provided is Age-Wise and Education-

Wise.

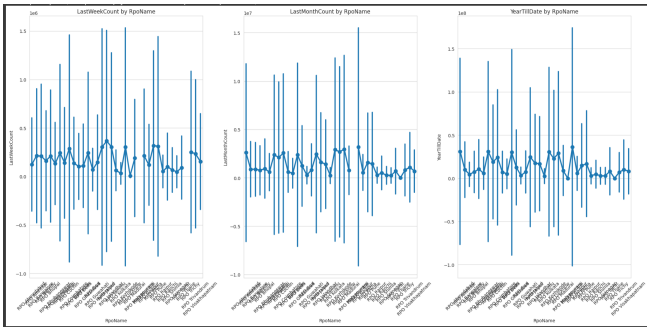


Figure 3.13: Trends in the data

In this figure, we can see the trend in the LastWeekCount, LastMonthCount and YearTillDate.

Chapter 4. Feature Engineering

Feature engineering is the process of transforming unstructured information into meaningful features that improve the performance of machine learning models. Tasks that come under this category include feature selection, new feature construction, numerical feature scaling, outlier management, missing value handling, and categorical variable encoding. These steps are crucial for extracting meaningful information from the data and getting it ready for modeling, which will lead to more precise predictions and insights in the end.

4.1 Standardization

Data normalization is a preprocessing approach used in feature engineering that ensures that all features have the same scale. The process of transforming numerical features to have a mean of 0 and a standard deviation of 1 is known as Z-score standardization. This process ensures that the algorithm treats each characteristic equally and prevents features with larger scales from taking over the model. Standardization is essential for models like neural networks and linear regression because feature sizes may have an impact on the model's performance. It usually happens after data cleansing and before model training, which contributes to the production of more accurate and dependable forecasts.

4.1.1 Dataset-1

	Dispatch in 1 days	Dispatch in 2 – 3 days	Dispatch in 4 – 7 days	Dispatch in 8 – 14 days	Dispatch in 15 – 21 days	Dispatch in 22 – 30 days	Dispatch in more than 30 days	Total
0	-0.926406	-0.827026	-0.781708	-0.825306	-0.981347	-0.821633	-0.936105	-1.270052
1	-0.887753	-0.067124	1.659983	1.316265	1.401092	1.702476	0.351064	1.335983
2	-0.308711	-0.438122	-0.590356	-0.101696	-0.025260	-0.541372	-0.732542	-0.663699
3	1.264644	3.665375	1.770859	0.791711	0.579507	0.854457	1.113186	1.693127
4	-0.794565	-0.800018	-0.718990	-0.620715	-0.639564	0.223396	0.049111	-0.469951

Figure 4.1: standardised dataset-1

The above data is the standardized format of data for the same set.

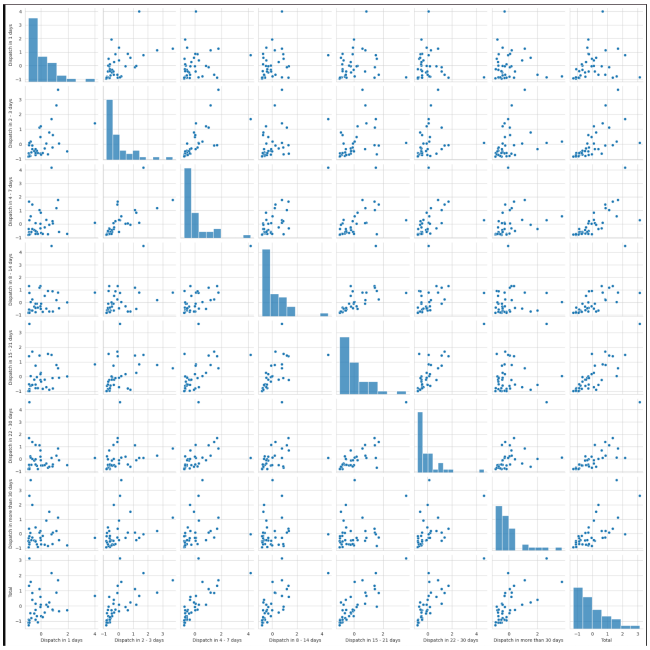


Figure 4.2: Pair plot

This pair plot represents the relationsip between each 2 variables take one by one.

4.1.2 Dataset-2

	ServiceName	RpoName	SchemeType	LastWeekCount	LastMonthCount	YearTillDate	Date
0	Applications Received - Scheme wise	RPO Ahmedabad	Normal	-0.247520	-0.230651	-0.224939	2019-11-03 04:29:09.831353
1	Applications Received - Scheme wise	RPO Ahmedabad	Tatkaal	-0.258003	-0.239743	-0.234417	2019-11-03 04:29:09.831353
2	Applications Received - Gender wise	RPO Ahmedabad	FEMALE	-0.253715	-0.236127	-0.230663	2019-11-03 04:29:09.831353
3	Applications Received - Gender wise	RPO Ahmedabad	MALE	-0.251808	-0.234269	-0.228695	2019-11-03 04:29:09.831353
4	Applications Received - Gender wise	RPO Ahmedabad	TRANSGENDER	NaN	-0.239992	-0.234673	2019-11-03 04:29:09.831353

Figure 4.3: Standardised dataset-2

We have standardized the data for better visulations and further feature engineering.

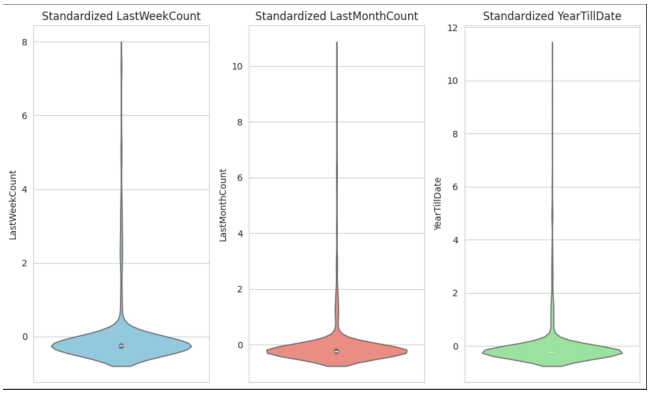


Figure 4.4: Violin plot

This is the violin plot for the following standardized data where the significant values are still overpowering but we can still get a better idea of the distribution.

4.2 Feature extraction

In feature extraction, unprocessed data is transformed into a new collection of features that are better suited for machine learning algorithms or more informative. Methods include domain-specific feature engineering, word embeddings, and dimensionality reduction (similar to PCA). It seeks to improve model performance and reduce dimensionality while capturing pertinent information.

4.2.1 Dataset-1

RPO Name	Total_to_Dispatch in 22 - 30 days_Ratio
CPV Delhi	25.648148
RPO Ahmedabad	4.039436
RPO Amritsar	8.551910
RPO Bangalore	6.909819
RPO Bareilly	3.026450
RPO Bhopal	4.342211
RPO Bhubaneswar	5.439267
RPO Chandigarh	3.772038
RPO Chennai	8.343675
RPO Cochin	4.811333
RPO Coimbatore	7.928502
RPO Dehradun	4.183810
RPO Delhi	4.756631
RPO Ghaziabad	5.012301
RPO Goa	56.841907
RPO Guwahati	26.846606
RPO Hyderabad	14.795474
RPO Jaipur	2.840735
RPO Jalandhar	40.569676
RPO Jammu	57.819005
RPO Kolkata	11.750246
RPO Kozhikode	7.260778
RPO Lucknow	3.155962
RPO Madurai	4.690617
RPO Malappuram	10.544298
RPO Mumbai	7.635782
RPO Nagpur	7.378097
RPO Patna	7.896563
RPO Pune	11.569562
RPO Raipur	6.250488
RPO Ranchi	4.117166
RPO Shimla	12.967547
RPO Srinagar	120.797997
RPO Surat	10.248840
RPO Thane	29.033962
RPO Trichy	5.061935
RPO Trivandrum	11.983345
RPO Visakhapatnam	17.706866

Figure 4.5: Engineered feat1

RPO Name	Total_to_Dispatch in more than 30 days_Ratio
CPV Delhi	4.355346
RPO Ahmedabad	4.984548
RPO Amritsar	7.232277
RPO Bangalore	3.565420
RPO Bareilly	2.018742
RPO Bhopal	3.929552
RPO Bhubaneswar	2.416844
RPO Chandigarh	7.617387
RPO Chennai	7.380645
RPO Cochin	4.132552
RPO Coimbatore	7.234680
RPO Dehradun	3.993487
RPO Delhi	5.812008
RPO Ghaziabad	1.842479
RPO Goa	21.906673
RPO Guwahati	1.516500
RPO Hyderabad	9.451686
RPO Jaipur	7.704090
RPO Jalandhar	29.354956
RPO Jammu	1.039848
RPO Kolkata	1.521779
RPO Kozhikode	6.719908
RPO Lucknow	3.034684
RPO Madurai	4.479893
RPO Malappuram	2.420694
RPO Mumbai	4.689212
RPO Nagpur	1.778216
RPO Patna	1.406719
RPO Pune	1.812464
RPO Raipur	2.269281
RPO Ranchi	2.321903
RPO Shimla	1.368190
RPO Srinagar	1.096001
RPO Surat	10.889471
RPO Thane	1.524636
RPO Trichy	4.570088
RPO Trivandrum	18.010581
RPO Visakhapatnam	11.505616

Figure 4.6: Engineered feat2

RPO Name	Duration_Difference_4	Duration_Difference_5
CPV Delhi	-162	-151
RPO Ahmedabad	17217	30792
RPO Amritsar	12285	-26809
RPO Bangalore	2602	21317
RPO Bareilly	6467	40710
RPO Bhopal	13056	3739
RPO Bhubaneswar	8672	2551
RPO Chandigarh	73627	200
RPO Chennai	14625	-31120
RPO Cochin	-53626	48262
RPO Coimbatore	-1505	-11634
RPO Dehradun	9881	-903
RPO Delhi	23839	-738
RPO Ghaziabad	2882	23357
RPO Goa	-7930	-8253
RPO Guwahati	-2806	2127
RPO Hyderabad	-107853	-59528
RPO Jaipur	45339	29308
RPO Jalandhar	40107	-104309
RPO Jammu	110	455
RPO Kolkata	10055	5821
RPO Kozhikode	40724	-14877
RPO Lucknow	134373	91108
RPO Madurai	29564	3108
RPO Malappuram	4002	11587
RPO Mumbai	35797	-18642
RPO Nagpur	3972	8716
RPO Patna	11077	25798
RPO Pune	-2238	12737
RPO Raipur	2154	653
RPO Ranchi	5548	11364
RPO Shimla	99	1679
RPO Srinagar	-3	250
RPO Surat	-21503	-16220
RPO Thane	-912	6885
RPO Trichy	25234	-855
RPO Trivandrum	9387	-25677
RPO Visakhapatnam	-31275	-26237

Figure 4.7: Engineered feat3

These are the features which were extracted from the original features because these new features are a lot more inferential the the original ones.
Now, because of these features having such a huge range, we will standardize these values too.

	Total_to_Dispatch in 15 – 21 days_Ratio
0	0.061141
1	-0.569629
2	-0.996367
3	-0.088909
4	0.100964
5	-0.638216
6	-0.413943
7	-0.836173
8	-0.595041
9	0.150064
10	-0.671797
11	-0.791542
12	-0.655736
13	0.097188
14	-0.607382
15	2.094625
16	-0.367207
17	-0.802752
18	-1.198763
19	2.832649
20	0.260999
21	-0.566248
22	-0.677487
23	-0.611938
24	0.878063
25	-0.483947
26	0.452548
27	0.595090
28	0.820940
29	-0.320617
30	-0.032610
31	0.966570
32	2.950173
33	-0.562600
34	1.984104
35	-0.616210
36	-0.633360
37	-0.506645

Figure 4.8: Engineered feat1 standardised

	Total_to_Dispatch in more than 30 days_Ratio	Total_to_Total_Ratio
0	0.035822	2.220446e-16
1	0.201229	2.220446e-16
2	0.676047	2.220446e-16
3	-0.201801	2.220446e-16
4	-0.817761	2.220446e-16
5	-0.087538	2.220446e-16
6	-0.633306	2.220446e-16
7	0.744124	2.220446e-16
8	0.702644	2.220446e-16
9	-0.027450	2.220446e-16
10	0.676482	2.220446e-16
11	-0.068350	2.220446e-16
12	0.394065	2.220446e-16
13	-0.907344	2.220446e-16
14	2.199848	2.220446e-16
15	-1.088717	2.220446e-16
16	1.031475	2.220446e-16
17	0.759030	2.220446e-16
18	2.619051	2.220446e-16
19	-1.401400	2.220446e-16
20	-1.085597	2.220446e-16
21	0.580363	2.220446e-16
22	-0.385817	2.220446e-16
23	0.070055	2.220446e-16
24	-0.631629	2.220446e-16
25	0.125872	2.220446e-16
26	-0.941395	2.220446e-16
27	-1.155134	2.220446e-16
28	-0.923151	2.220446e-16
29	-0.699042	2.220446e-16
30	-0.675266	2.220446e-16
31	-1.179164	2.220446e-16
32	-1.360964	2.220446e-16
33	1.223393	2.220446e-16
34	-1.083911	2.220446e-16
35	0.094363	2.220446e-16
36	1.922248	2.220446e-16
37	1.298625	2.220446e-16

Figure 4.9: Engineered feat2 standardised

	Duration_Difference_4	Duration_Difference_5	Duration_Difference_6
0	NaN	NaN	-1.998063
1	0.374571	0.957544	NaN
2	0.166763	NaN	-1.018778
3	-0.788736	0.734940	1.013540
4	-0.228290	1.126560	0.358282
5	0.204239	-0.318628	-0.805835
6	-0.047665	-0.549993	0.263151
7	1.269282	-1.885533	NaN
8	0.274111	NaN	-0.523583
9	NaN	1.229568	-0.077002
10	NaN	NaN	-1.351876
11	0.032689	NaN	-1.662729
12	0.574938	NaN	NaN
13	-0.725828	0.790260	0.931652
14	NaN	NaN	-1.476600
15	NaN	-0.659979	0.719391
16	NaN	NaN	0.346568
17	0.970750	0.927644	NaN
18	0.895252	NaN	-1.091873
19	-2.731322	-1.592467	0.513632
20	0.043436	-0.050731	1.846579
21	0.904652	NaN	-0.938953
22	1.639711	1.614194	-0.162980
23	0.707461	-0.430485	-1.068374
24	-0.523735	0.365938	0.874840
25	0.825253	NaN	0.523450
26	-0.528367	0.193601	0.708006
27	0.103035	0.850428	1.542198
28	NaN	0.423214	1.285599
29	-0.905033	-1.374174	-0.206014
30	-0.322652	0.354175	0.002651
31	-2.795580	-0.803074	0.331483
32	NaN	-1.953877	0.865323
33	NaN	NaN	NaN
34	NaN	0.050872	1.520825
35	0.609954	NaN	-0.767495
36	0.001112	NaN	NaN
37	NaN	NaN	-0.497015

Figure 4.10: Engineered feat3 standardised

Now that we have the standardized values for the extracted features, let us plot a boxplot to see the changes.

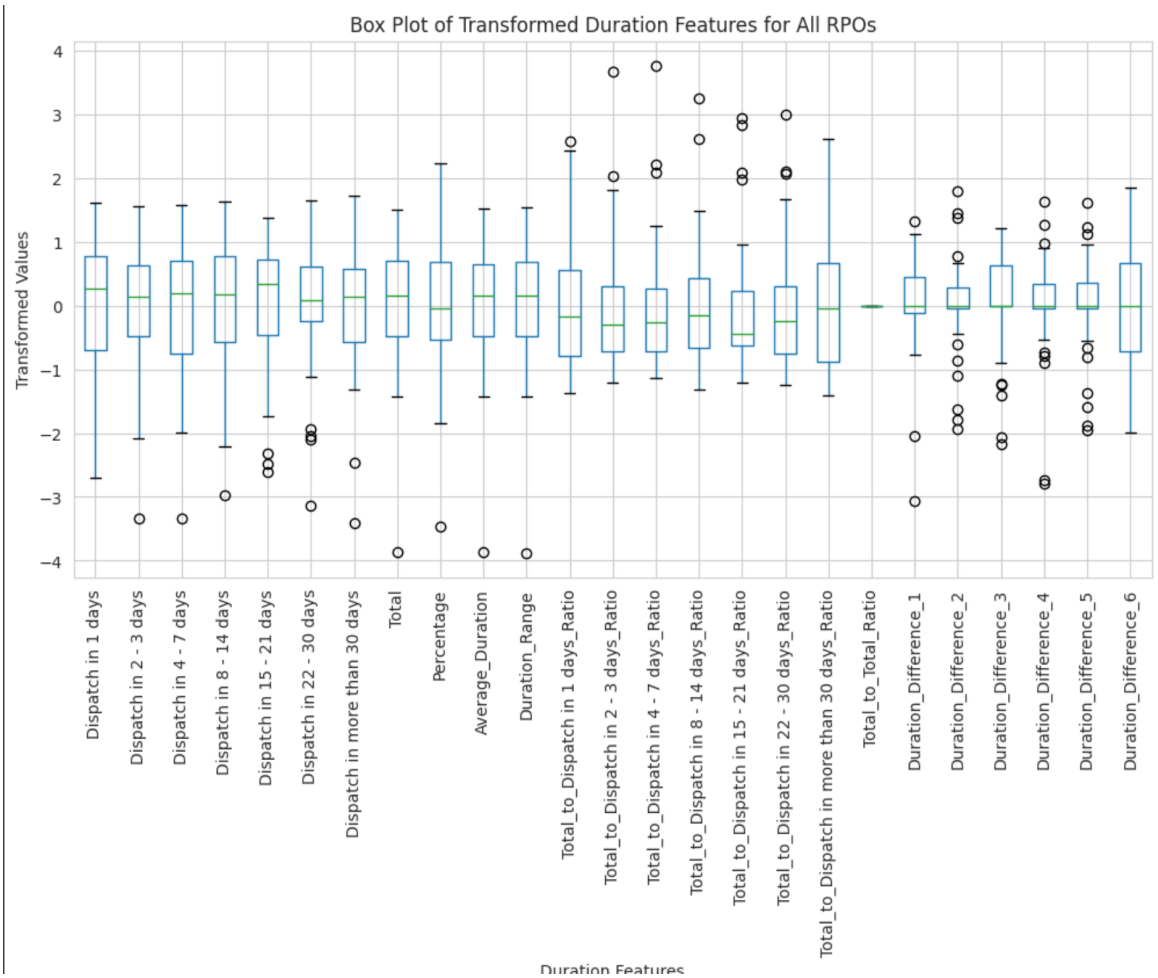


Figure 4.11: Box plot for standardised values

As we can observe, after standardizing these values, the box plots are significantly more inferential and clear to make observations and compare.

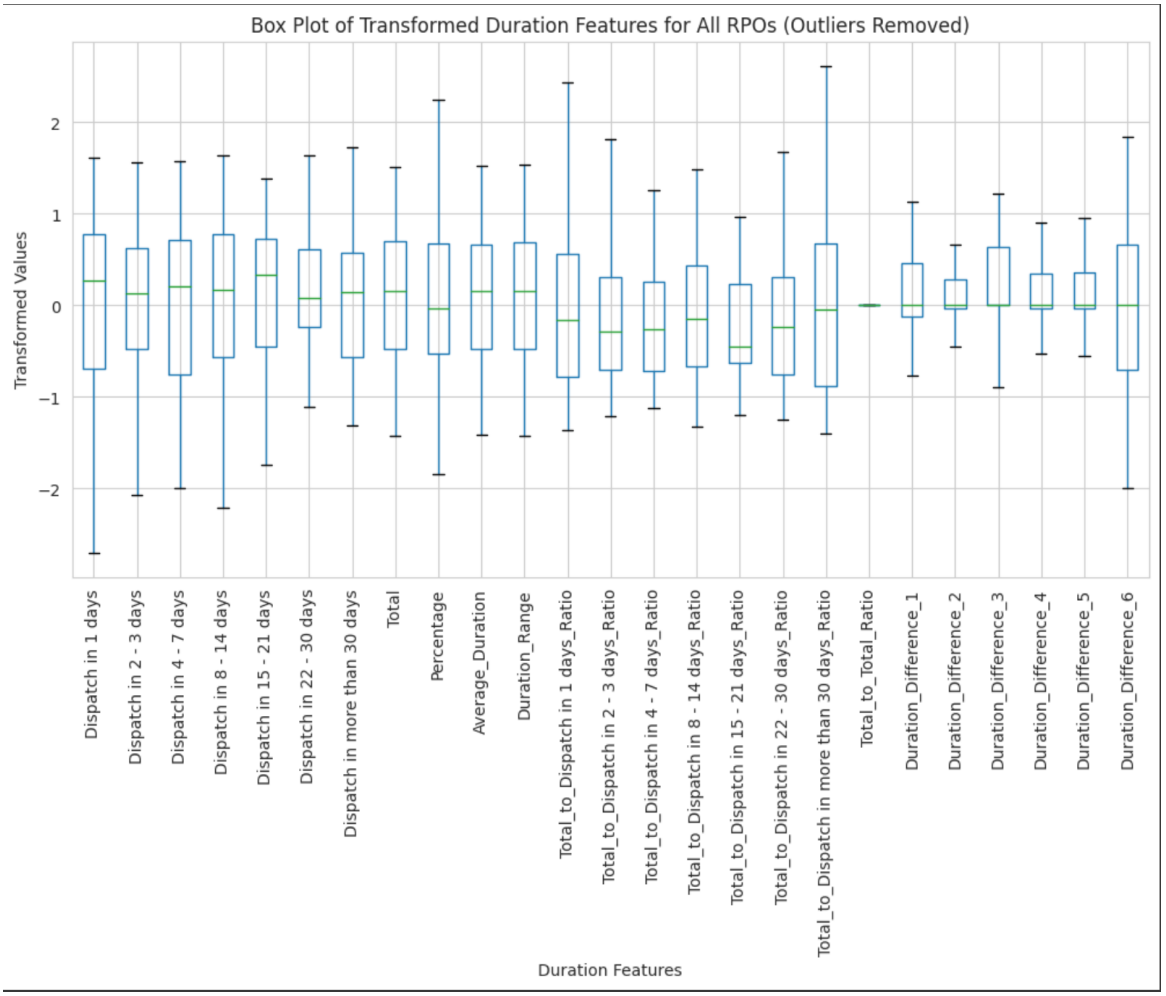


Figure 4.12: New box plot with no outliers

This is the boxplot after removing outliers using [-1.5*IQR,1.5*IQR] range to remove the outside values.

4.2.2 Dataset-2

Here, instead of relative values, we have taken the moving average features for better extraction as you can see below are the given features.

```
   LastMonthCount_RollingStd_7  YearTillDate_RollingMean_7  \
0                          NaN                    567081.0
1                35387.865971                    291005.5
2                25196.861650                    271882.0
3                20853.268335                    290982.0
4                21573.166033                    232788.6

   YearTillDate_RollingStd_7  LastWeekCount_RollingMean_14  \
0                        NaN                   7174.000000
1             390429.716339                   3667.000000
2             278055.401776                   3454.333333
3             230222.597294                   3667.000000
4             238084.422519                   3667.000000

   LastWeekCount_RollingStd_14  LastMonthCount_RollingMean_14  \
0                          NaN                       51414.00
1                 4959.646963                       26391.00
2                 3526.291300                       24685.00
3                 2910.451626                       26388.25
4                 2910.451626                       21110.80

   LastMonthCount_RollingStd_14  YearTillDate_RollingMean_14  \
0                           NaN                     567081.0
1                 35387.865971                     291005.5
2                 25196.861650                     271882.0
3                 20853.268335                     290982.0
4                 21573.166033                     232788.6

   YearTillDate_RollingStd_14
0                         NaN
1              390429.716339
2              278055.401776
3              230222.597294
4              238084.422519
```

Figure 4.13: feature engineering

Now, we have log-transformed all the new features for further feature selection.

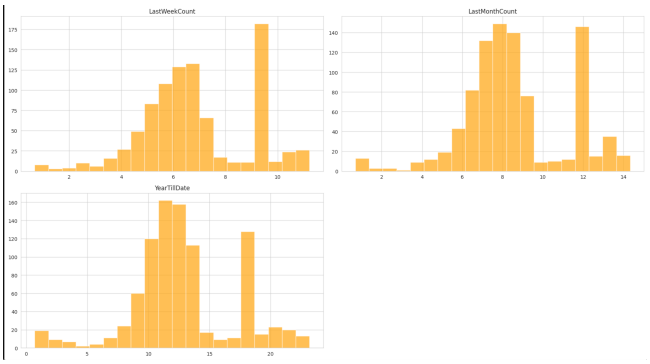


Figure 4.14: Log Transformed plot

Here, you can see the distribution is much more clearer than before.

4.2.3 Dataset-2

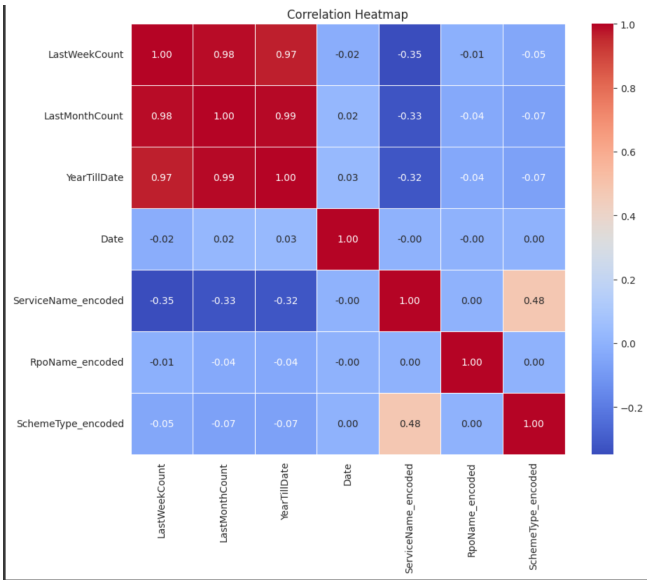


Figure 4.15: Correlation plot

Here also, we are using the correlation matrix to find out the top features.
As you can see below, we have selected the important features needed with high correlation.

```
Top 4 selected features:
                        Feature         Score
25  YearTillDate_RollingStd_14  5.261175e+08
19   YearTillDate_RollingStd_7  3.836048e+08
13   YearTillDate_RollingStd_3  2.580987e+08
2                 YearTillDate  2.243897e+08
```

Figure 4.16: Top 4 selected Features

4.3 Feature selection

Feature selection is the process of picking the most relevant attributes from the original collection to improve interpretability, reduce overfitting, and maximize model performance. methods include filter methods (which rely on statistical measurements), embedded methods (which are inserted into model training), and wrapper methods (which use specific ML algorithms). It streamlines models, prevents overfitting, and speeds up training.

4.3.1 Dataset-1

Now, we will be using this new correlation matrix to find out the important features by applying Feature Selection using Correlation.

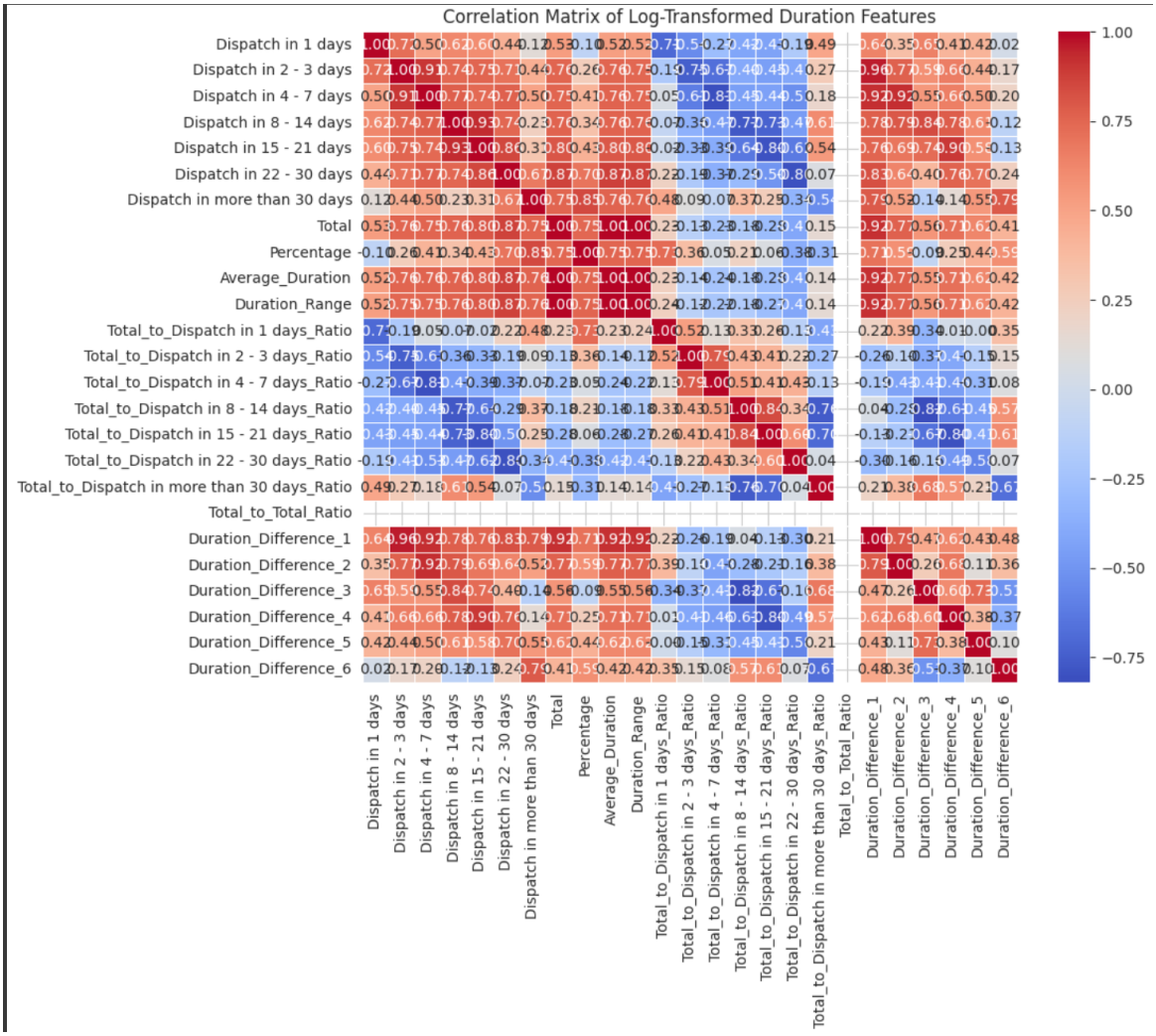


Figure 4.17: New Correlation Plot

Below is the given feature selection.

```
Selected Features based on correlation threshold:
                     level_0                          level_1  correlation
178                    Total                   Duration_Range     0.999823
177                    Total                Average_Duration     0.999611
226          Average_Duration                   Duration_Range     0.999383
42     Dispatch in 2 - 3 days             Duration_Difference_1     0.960741
76    Dispatch in 8 - 14 days          Dispatch in 15 - 21 days     0.931753
..                        ...                              ...          ...
237          Average_Duration             Duration_Difference_4     0.710743
210                Percentage             Duration_Difference_1     0.707162
29     Dispatch in 2 - 3 days          Dispatch in 22 - 30 days     0.706373
11         Dispatch in 1 days  Total_to_Dispatch in 1 days_Ratio     0.706302
142  Dispatch in 22 - 30 days             Duration_Difference_5     0.703392

[77 rows x 3 columns]
Top 5 selected features using SelectKBest:
Index(['Average_Duration', 'Duration_Range', 'Total_x_Duration_Range',
       'Average_Duration_x_Total', 'Duration_Range_x_Total'],
      dtype='object')
```

Figure 4.18: Feature Selection

These are the selected features having the highest correlation with the Total value of dispatches.

Chapter 5. Model fitting

In exploratory data analysis, model fitting comprises selecting an appropriate statistical model, estimating the model's parameters to fit the data, assessing the model's effectiveness, and analyzing the results to gain additional insight into the relationships between the variables. The aims of this iterative process are to comprehend the data and formulate theories for further investigation.
Many model fitting techniques are used in data analysis, each suitable for a certain set of data and set of research questions:

5.1 Regression

Linear Regression: With this approach, a continuous dependent variable is linearly fitted to one or more independent variables.

Decision Trees: non-linear models that separate the data into subgroups based on the values of predictor variables, forming a structure resembling a tree.

Random Forests:** a learning technique that, to boost accuracy and robustness, mixes the predictions of many decision trees fitted to arbitrary subsets of data.

Support Vector Machines (SVM):** supervised learning models that determine the best high-dimensional space hyperplane for class division.

5.2 ML algorithms

After applying certain linear and random forest classifiers, we have come up with the most optimal polynomial multi-variate regression which gives the following results:

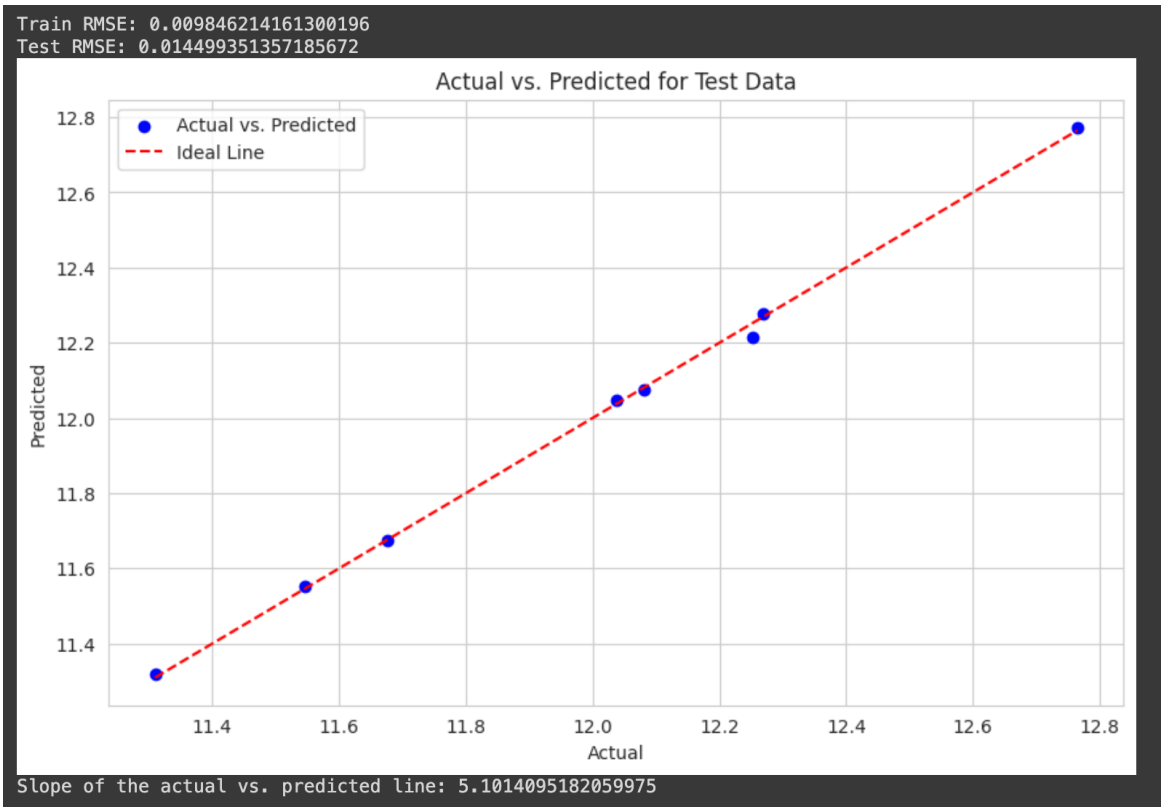


Figure 5.1: Actual vs. Predicted test data

Chapter 6. Conclusion & future scope

In this part, we concluded our project and applied every concept known to us about the Exploratory Data Analysis. We learned about applying the imputation of the missing data value in a column and dealing with such problems in the future as well. Then, we moved on to visualizing the data which was in the form of different plots such as box plots, scatter plots, violin plots, bar plots, histograms, line plots, and heat maps as well when it came to finding the correlation matrix between different features. After visualizing the data, we selected the features via the correlation matrix method. We selected those features that we observed a high correlation value among them. After feature selection, it was time that we log-transformed the features for further feature selection to make the distribution much clearer. Then again we solved the correlation matrix for these features and finally chose those that would work for the project. When we select the best features, it becomes easy for us to fit a machine learning model in the test data set as it decreases the overfitting and optimizes the ML algorithms. Selecting the features, now we performed various ML algorithms and models on the training data which was kept 80% of the data and the remaining was testing data. We found the root mean square error value for each of the models and opted for the model that had the least value of RMSE as that would be the one that would be the most optimized model of Machine learning. We found that the linear regression model was the most optimal model to go ahead with. So here was the overall viewpoint and conclusion that we had while doing the project. For the future scope, we would be performing a similar process on the other countries' dataset of passport services and would be running a function of intersection set that where Indians get located in the other countries and in the future what all countries (via regression model of Machine learning) are Indians targeting to go and live for. We would also be working on the number of Visas obtained by Indians and what type of visas (business, travel, work, etc.) Indians opting for in different countries. The final observation would then be compared and correlated with the data of Indian trades(in USD) with those countries Indians are opting for and then we would see the specifics of the trades that what all products are being traded by India with those countries and according to that we would find out the future trade of India adding to the GDP of India. Like this, we can optimize the trade routes and also the trade in terms of money with other countries and make a flow chart according to which if the trade is done, we can surely get a lot of output out of it after finding the reasons for the same.

6.1 Findings/observations

We observed from our study that passport services efficiency is highly dependent on the region of the Passport office. There needs to be a revolution in the passport service sector to make it an efficient and fast process for people all over the country. In the other dataset, we found out about the distribution of passport issuances over different types of Schemes and services.

6.2 Challenges

We faced a lot of challenges in actually finding the appropriate Dataset for Passport Services because there is very little data available for which you can do data analysis on Passport Service data. Also, visualization was a complicated part because the Passport Service Sector is very oddly distributed in terms of its services and schemes. We also faced a lot of challenges in fitting the accurate models of machine learning out of different machine learning algorithms. The main challenge was that the data that we found contained a lot of categorical values and hence model fitting according to the selected features was getting tough to regress the model upon.

Group Contribution

Dhyey Patel - 202103053

- Finding the Datasets
- Python Code and Data Analysis of Dataset-1 and a few parts of Dataset-2
- Latex Report (Visualisation, Feature Engineering, Model Fitting of Dataset-1)

Mit Desai - 202103013

- Python Code and Data Analysis of Dataset-2
- Latex Report (Visualisation, Feature Engineering, Model Fitting of Dataset-2)

Aarzoo Khambhoo - 202103026

- Latex Report Content about packages, observations and conclusion and overall latex report as well

Short Bio

1. **Dhyey Patel** is a passionate data analyst who likes working on Python.
2. **Mit Desai** is a passionate Machine learning enthusiast who prefers model fitting.
3. **Aarzoo Khambhoo** is also a passionate data analyst and also highly proficient in report writing and Latex.

References

[1] Data.gov.in database website of Indian Government. *URL:* https://data.gov.in/resource/weekly-data-passport-related-services

[2] Data.gov.in database website of Indian Government. *URL:* https://data.gov.in/resource/rpo-wise-time-taken-dispatch-passport-passport-dispatched-01-jan-2016-31-dec-2016-m