Analysis of PMGSY Scheme

Contents

1)Description of Data

1.1) Data Source

The data is collected from https://www.data.gov.in/resource/physical-financial-progress-pradhan-mantri-gram-sadak-yojna-pmgsy-date

The latest update is published by October 2024. It contains 2200+ records. In the current analysis, this data is well relatable as the information can be helpful to derive helpful insight in progress and success of this scheme. It is also a good collection for learning 'Data Transformation, Data Cleaning, and Data Visualization.'

1.2) Data Attributes

Dataframe name - df

S.NO	Attribute	Data Type	Example
1	STATE_NAME	Nominal Data	Kerala, Bihar, etc.
2	DISTRICT_NAME	Nominal Data	Vizianagar etc.
3	PMGSY_SCHEME	Ordinal Data	PMGSY – I etc.
4	NO_OF_ROAD_WORK_SANCTIONED	Continuous Data	32, 44, etc
5	NO_OF_BRIDGES_SANCTIONED	Continuous Data	0,2, etc
6	NO_OF_ROAD_WORKS_COMPLETED	Continuous Data	32, 44, etc
7	NO_OF_BRIDGES_COMPLETED	Continuous Data	0,2, etc
8	NO_OF_BRIDGES_BALANCE	Continuous Data	0,2, etc
9	LENGTH_OF_ROAD_WORK_SANCTIONED_KM	Continuous Data	151.5, 267.4, etc
10	COST_OF_WORKS_SANCTIONED_LAKHS	Continuous Data	196.57, 145,8, etc
11	LENGTH_OF_ROAD_WORK_COMPLETED_KM	Continuous Data	151.5, 267.4, etc
12	EXPENDITURE_OCCURED_LAKHS	Continuous Data	196.57, 145.8, etc
13	LENGTH_OF_ROAD_WORK_BALANCE_KM	Continuous Data	54.45, 92.8 etc.

Dataframe Name – area_df

S.NO	Attribute	Data Type	Example
1	STATE_NAME	Nominal Data	Kerala, Bihar, etc.
2	STATE_AREA	Continuous Data	41846, 63877 etc.

2) Data Transformation

Drop columns related to 'NO_OF_BRIDGES_SANCTIONED', 'NO_OF_BRIDGES_COMPLETED', 'NO_OF_BRIDGES_BALANCE' since most of the entries were null.

Merged two dataframes df and area_df on STATE_NAME and change data type of STATE_AREA column to float for easy conversion

Dropped null rows to create new columns named 'Sanctioned_expenditure_per_km', 'actual_expenditure_per_km', 'Sanctioned_expenditure_per_km_in_rupees', 'actual_expenditure_per_km_in_rupees', 'difference', 'Length Per SQ. KM'.

The below commands remove the null values and make the data clean and consistent.

S.No.	**Command**	**Purpose**
1	.drop(axis=1)	To drop columns with null values
2	[~__.isin()]	To drop rows of no use
3	.merge()	To merge dataframes
4	.replace()	To replace 0 with Nan
5	.dropna()	To drop nan rows
6	.astype()	To convert data type of the dataframe object

3) Data Normalisation

3.1) First Normal Form (1NF)

The cleaned dataframe df and area_df are already in 1NF as a combination of 'STATE_NAME', 'DISTRICT_NAME', 'BLOCK_NAME', and 'ROAD_NAME' uniquely identifies each road project.

The dataframe area_df satisfies 2NF as it has a single-column primary key and all other attributes are fully dependent on it and has STATE_NAME column has partial dependency in df datatframe.

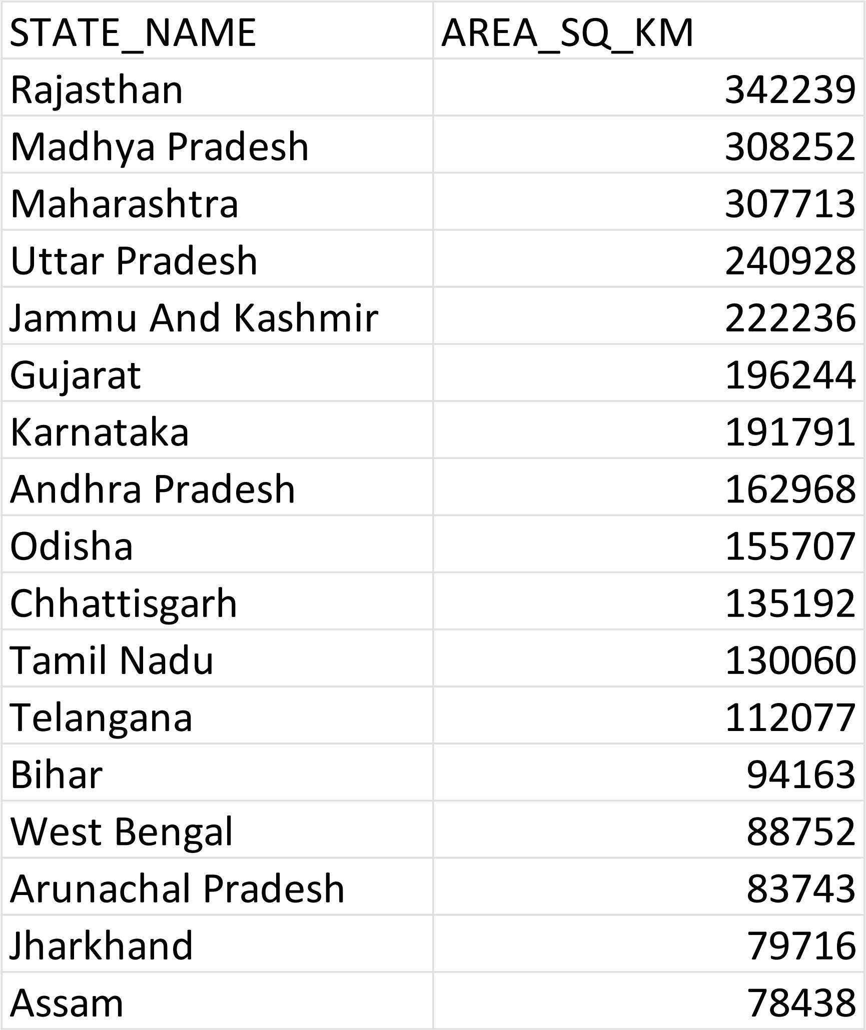
The cleaned data has no transitive dependencies. So, it is already in 3NF.

This table shows head of cleaned data frame to provide basic details regarding our data.

Dataframe df:

STATE_NAME	DISTRICT_NAME	PMGSY_SCHEME	NO_OF_ROAD_WORK_SANCTI	NO_OF_ROAD_WORKS_COMPLETED	NO_OF_ROAD_WORKS_BALANCE	LENGTH_OF_ROAD_WORK_SANCTIONED_KM	COST_OF_WORKS_SANCTIONED_LAKHS
Andhra Pradesh	Bapatla	PMGSY-II	6	6	0	53.43	28.7436
Andhra Pradesh	Chittoor	PMGSY-III	37	33	4	276.248	172.5041
Andhra Pradesh	Guntur	PMGSY-II	4	4	0	35.1	17.3794
Andhra Pradesh	Kakinada	RCPLWEA	3	3	0	19.47	12.19
Andhra Pradesh	Krishna	PMGSY-I	159	159	0	368.9	89.007

Dataframe area_df:



STATE_NAME	AREA_SQ_KM
Rajasthan	342239
Madhya Pradesh	308252
Maharashtra	307713
Uttar Pradesh	240928
Jammu And Kashmir	222236
Gujarat	196244
Karnataka	191791
Andhra Pradesh	162968
Odisha	155707
Chhattisgarh	135192
Tamil Nadu	130060
Telangana	112077
Bihar	94163
West Bengal	88752
Arunachal Pradesh	83743
Jharkhand	79716
Assam	78438

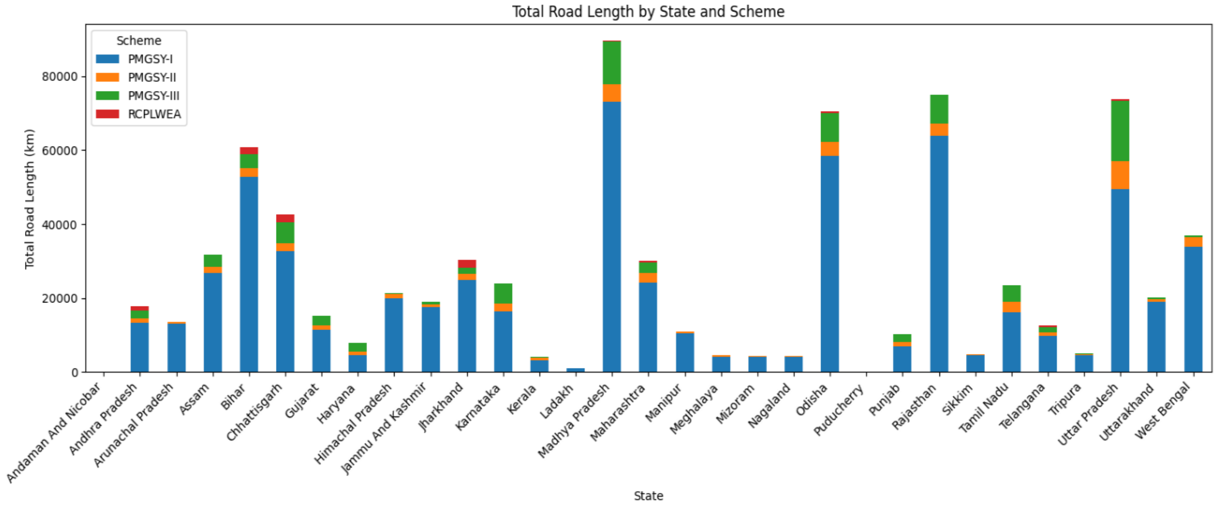
4) Data Visualisation

4.1) States/UTs wise analysis of PMGSY Scheme

i) **Objective:** To visualize the total length of roads constructed under scheme state-wise and phase-wise.

Number of variables	:	2
Type of Relation	:	Categorical
Type of Plot	:	Stacked Bar Chart

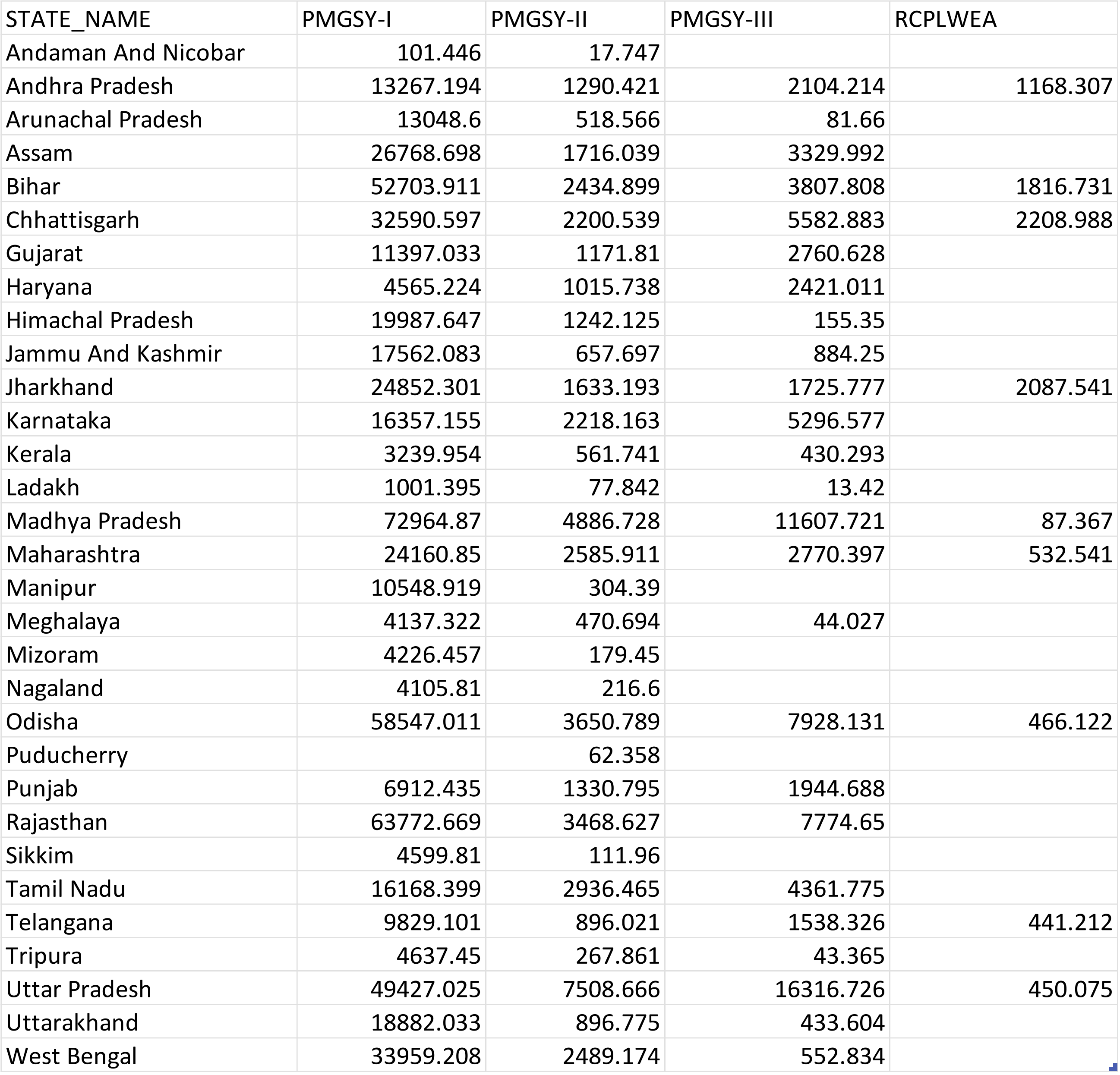
Plot: This stacked bar chart presents the total road length constructed under different PMGSY schemes (IA, II, III, etc.) for each state. The x-axis represents the states, and the y-axis represents the total road length in Kilometers. Each bar is divided into segments representing different schemes, and the height of each segment corresponds to the road length completed under that scheme in that state.



Inference: The bar chart shows the variation in the total length of road constructed under PMGSY Scheme across different states/UTs.

- Uttar Pradesh, Madhya Pradesh, and Bihar have the highest road lengths, indicating greater infrastructure development in these states.
- PMGSY-I and PMGSY-II are the most prominent schemes, with contributions varying across states, highlighting scheme-wise implementation focus.

Dataframe Generated:

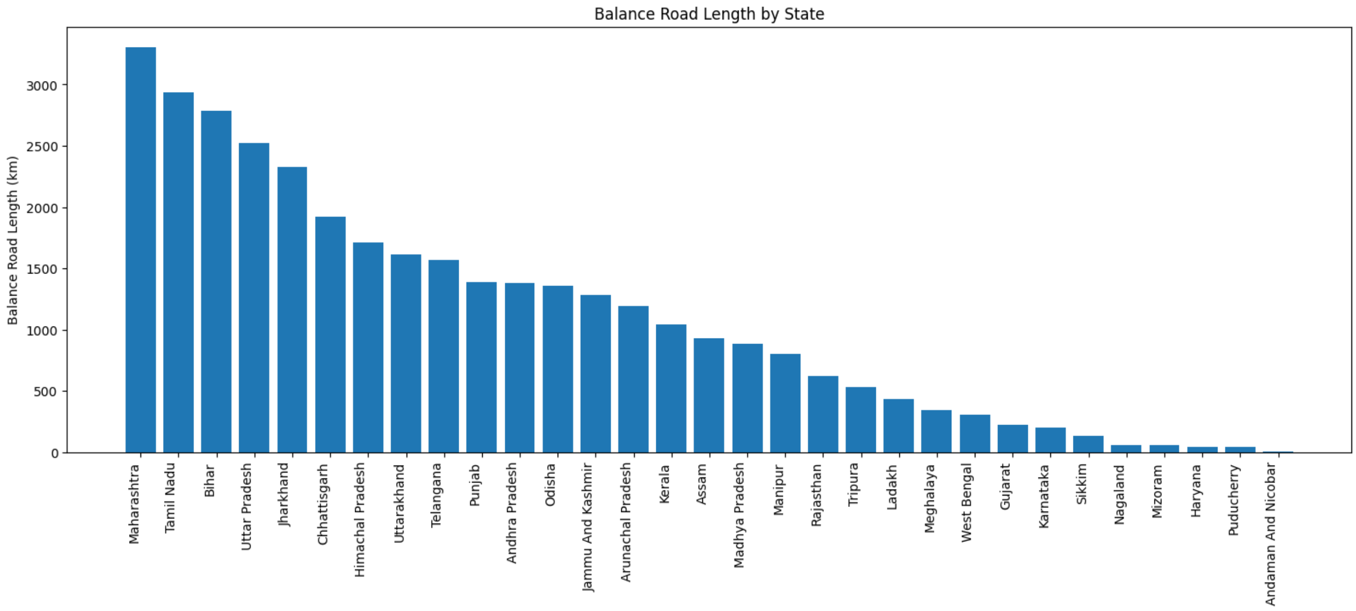


STATE_NAME	PMGSY-I	PMGSY-II	PMGSY-III	RCPLWEA
Andaman And Nicobar	101.446	17.747		
Andhra Pradesh	13267.194	1290.421	2104.214	1168.307
Arunachal Pradesh	13048.6	518.566	81.66	
Assam	26768.698	1716.039	3329.992	
Bihar	52703.911	2434.899	3807.808	1816.731
Chhattisgarh	32590.597	2200.539	5582.883	2208.988
Gujarat	11397.033	1171.81	2760.628	
Haryana	4565.224	1015.738	2421.011	
Himachal Pradesh	19987.647	1242.125	155.35	
Jammu And Kashmir	17562.083	657.697	884.25	
Jharkhand	24852.301	1633.193	1725.777	2087.541
Karnataka	16357.155	2218.163	5296.577	
Kerala	3239.954	561.741	430.293	
Ladakh	1001.395	77.842	13.42	
Madhya Pradesh	72964.87	4886.728	11607.721	87.367
Maharashtra	24160.85	2585.911	2770.397	532.541
Manipur	10548.919	304.39		
Meghalaya	4137.322	470.694	44.027	
Mizoram	4226.457	179.45		
Nagaland	4105.81	216.6		
Odisha	58547.011	3650.789	7928.131	466.122
Puducherry		62.358		
Punjab	6912.435	1330.795	1944.688	
Rajasthan	63772.669	3468.627	7774.65	
Sikkim	4599.81	111.96		
Tamil Nadu	16168.399	2936.465	4361.775	
Telangana	9829.101	896.021	1538.326	441.212
Tripura	4637.45	267.861	43.365	
Uttar Pradesh	49427.025	7508.666	16316.726	450.075
Uttarakhand	18882.033	896.775	433.604	
West Bengal	33959.208	2489.174	552.834	

ii) **Objective:** To compare balance of sanctioned roadwork remaining under PMGSY Scheme across different States/UTs

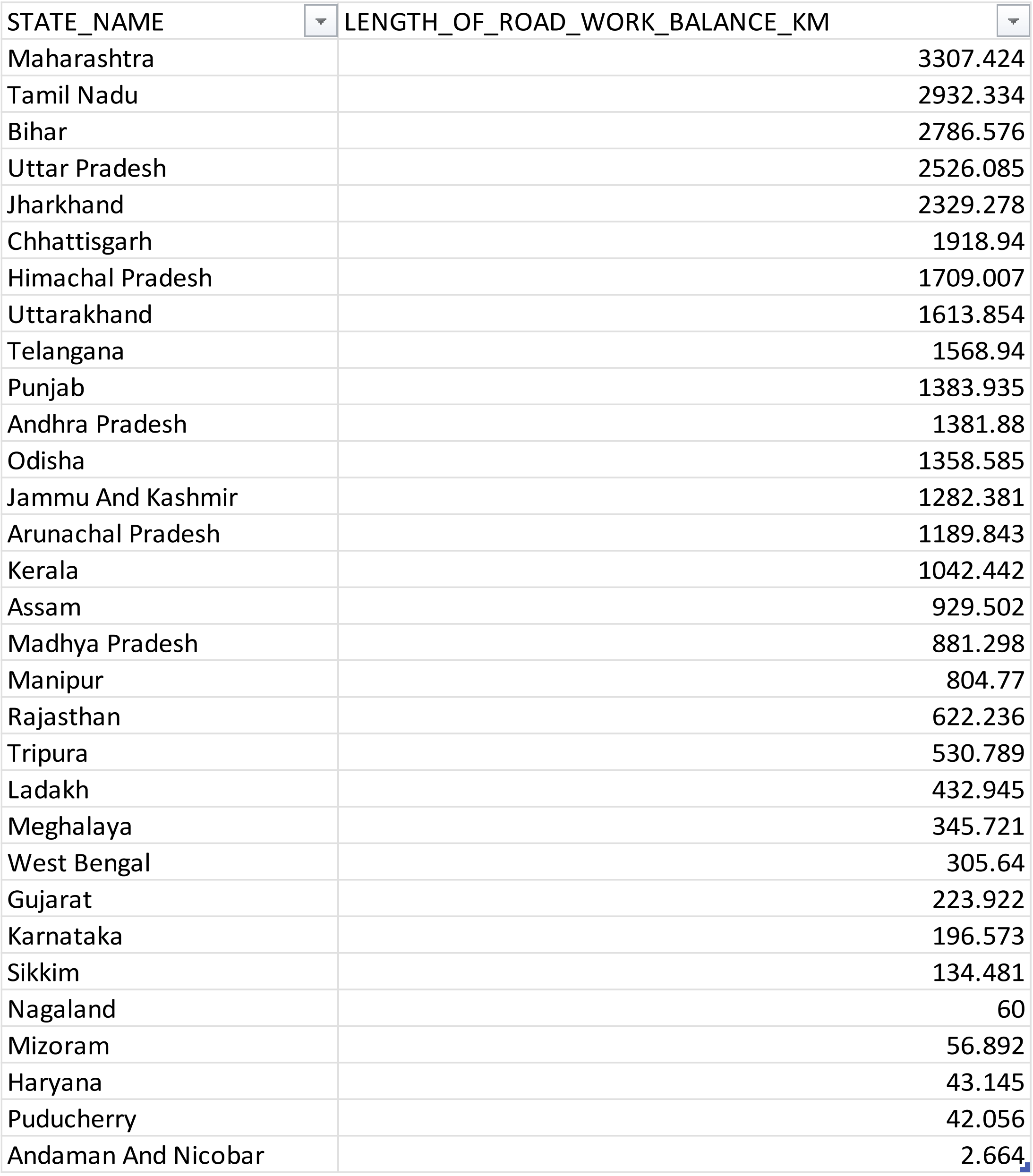
Number of variables	:	1
Type of Relation	:	Categorical
Type of Plot	:	Bar Chart

Plot: This bar chart displays the total length of roads remaining to be constructed (balance road length) under the PMGSY scheme for each state. The x-axis represents the states, and the y-axis represents the balance road length in kilometers. Higher bars indicate more unfinished road projects.



Inference: The primary inference is about the progress of road construction and the remaining backlog in different states. States with higher bars have a larger backlog of projects, potentially indicating slower progress or greater infrastructure needs. Bihar, Tamil Nadu, and Maharashtra have larger backlogs, indicating a greater need for future road construction in these states.

Dataframe Generated:

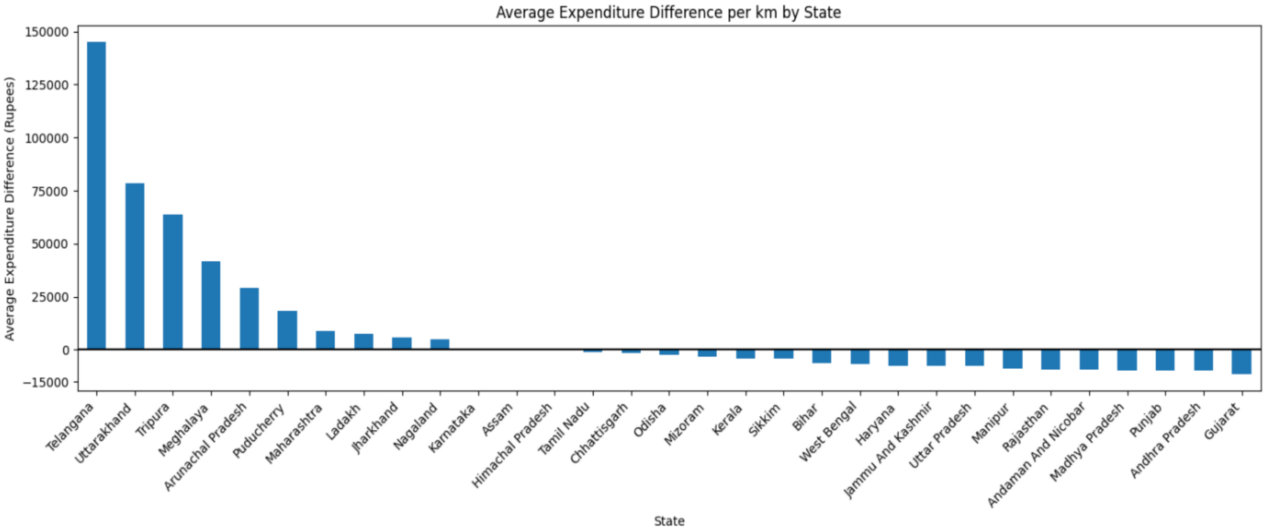


STATE_NAME	LENGTH_OF_ROAD_WORK_BALANCE_KM
Maharashtra	3307.424
Tamil Nadu	2932.334
Bihar	2786.576
Uttar Pradesh	2526.085
Jharkhand	2329.278
Chhattisgarh	1918.94
Himachal Pradesh	1709.007
Uttarakhand	1613.854
Telangana	1568.94
Punjab	1383.935
Andhra Pradesh	1381.88
Odisha	1358.585
Jammu And Kashmir	1282.381
Arunachal Pradesh	1189.843
Kerala	1042.442
Assam	929.502
Madhya Pradesh	881.298
Manipur	804.77
Rajasthan	622.236
Tripura	530.789
Ladakh	432.945
Meghalaya	345.721
West Bengal	305.64
Gujarat	223.922
Karnataka	196.573
Sikkim	134.481
Nagaland	60
Mizoram	56.892
Haryana	43.145
Puducherry	42.056
Andaman And Nicobar	2.664

iii) **Objective:** To compare difference in actual and allocated expenditure and sanctioned cost for different states

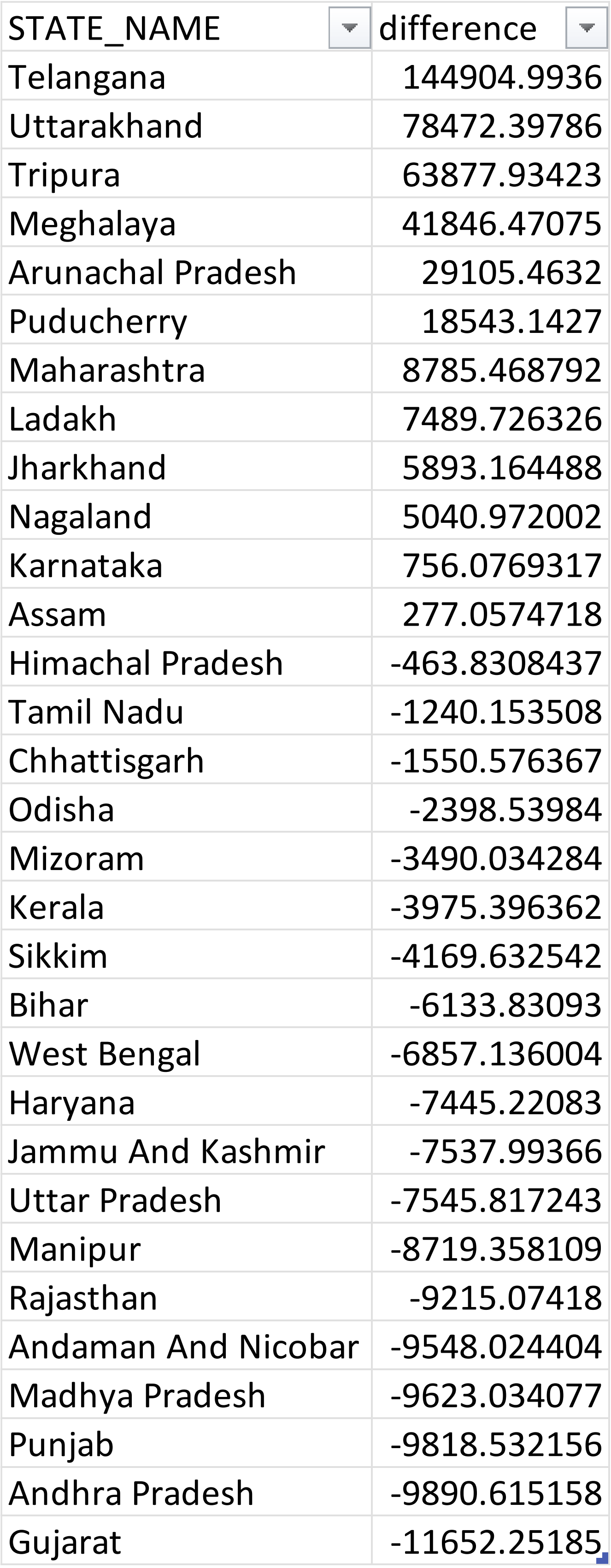
Number of variables	:	1
Type of Relation	:	Categorical
Type of Plot	:	Bar Chart

Plot: The bar chart compares the actual and sanctioned costs per Kilometer for road construction projects across different states. The x-axis represents the states, and the y-axis represents the average expenditure difference in rupees. Positive values indicate cost overruns, while negative values suggest cost savings.



Inference: The plot reveals how actual project costs compare to planned costs across different states. States with positive differences suggest potential cost overruns, while negative differences indicate cost savings. While the states like Telangana and Uttarakhand have quite high overrun costs while states like Gujrat and Andra Pradesh saved costs. This information highlights variations in cost efficiency and project management, potentially due to regional factors or implementation practices.

Dataframe Generated:

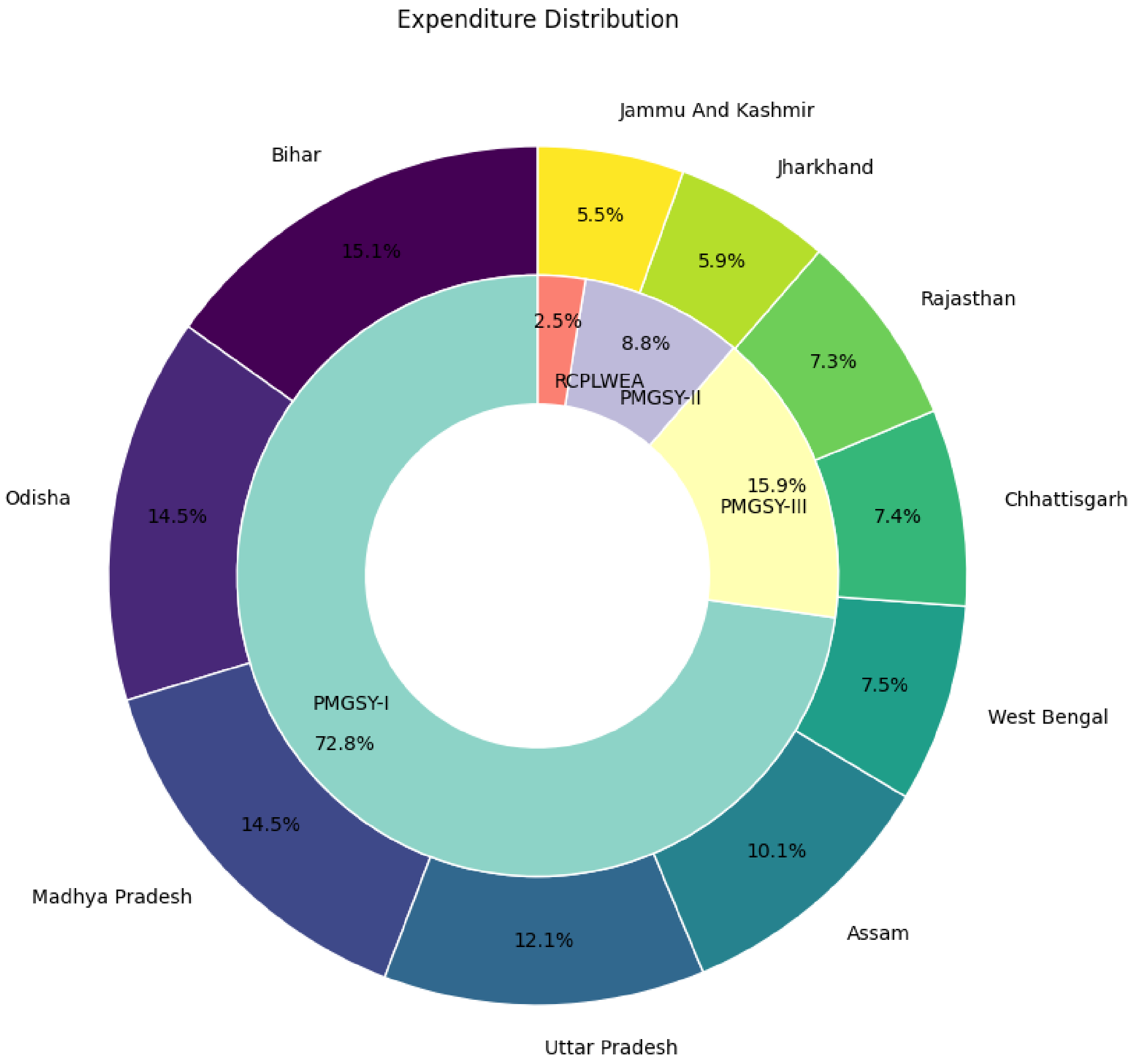


STATE_NAME	difference
Telangana	144904.9936
Uttarakhand	78472.39786
Tripura	63877.93423
Meghalaya	41846.47075
Arunachal Pradesh	29105.4632
Puducherry	18543.1427
Maharashtra	8785.468792
Ladakh	7489.726326
Jharkhand	5893.164488
Nagaland	5040.972002
Karnataka	756.0769317
Assam	277.0574718
Himachal Pradesh	-463.8308437
Tamil Nadu	-1240.153508
Chhattisgarh	-1550.576367
Odisha	-2398.53984
Mizoram	-3490.034284
Kerala	-3975.396362
Sikkim	-4169.632542
Bihar	-6133.83093
West Bengal	-6857.136004
Haryana	-7445.22083
Jammu And Kashmir	-7537.99366
Uttar Pradesh	-7545.817243
Manipur	-8719.358109
Rajasthan	-9215.07418
Andaman And Nicobar	-9548.024404
Madhya Pradesh	-9623.034077
Punjab	-9818.532156
Andhra Pradesh	-9890.615158
Gujarat	-11652.25185

iv) **Objective:** To show expenditure proportions in different Phases and top 10 states(by expenditure).

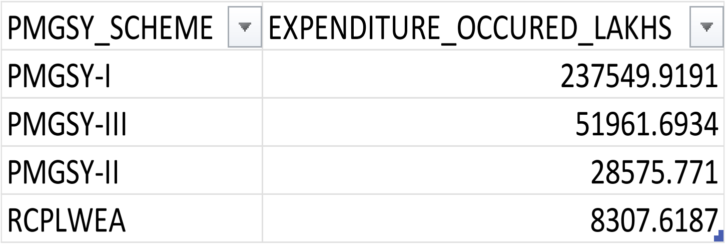
Number of variables	:	2
Type of Relation	:	Categorical
Type of Plot	:	Double Pie Chart

Plot: This double pie chart presents the distribution of expenditure under the PMGSY scheme. The outer pie chart shows the contribution of total expenditure in cost for the top 10 states. The inner pie chart shows the expenditure by scheme for the top 10 schemes.

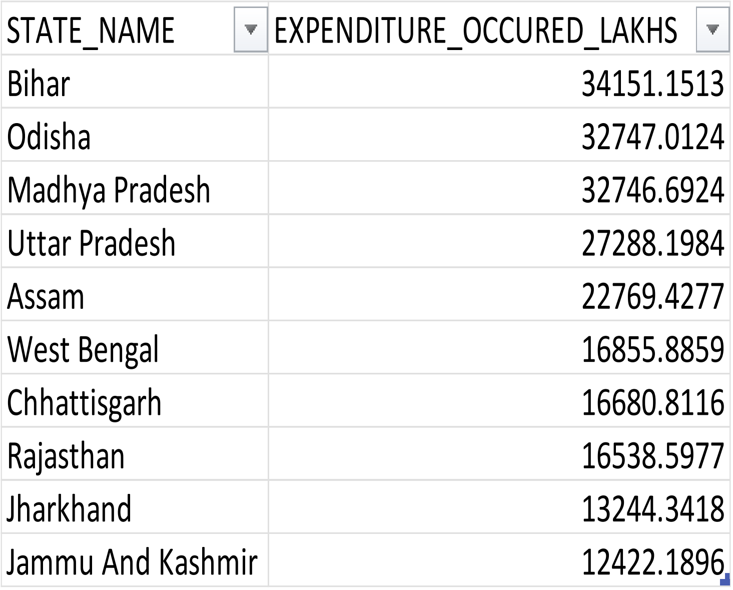


Inference: The double pie chart shows where PMGSY funding is being spent and on which schemes. The outer pie chart highlights states receiving the most funding, while the inner pie chart shows the distribution across different PMGSY schemes. Major expenditure is focused on Uttar Pradesh, Madhya Pradesh and Bihar under PMGSY-I and PMGSY-II schemes. By comparing the two, you can identify potential regional disparities and understand prioritization in resource allocation.

Dataframe Used:



PMGSY_SCHEME	EXPENDITURE_OCCURED_LAKHS
PMGSY-I	237549.9191
PMGSY-III	51961.6934
PMGSY-II	28575.771
RCPLWEA	8307.6187

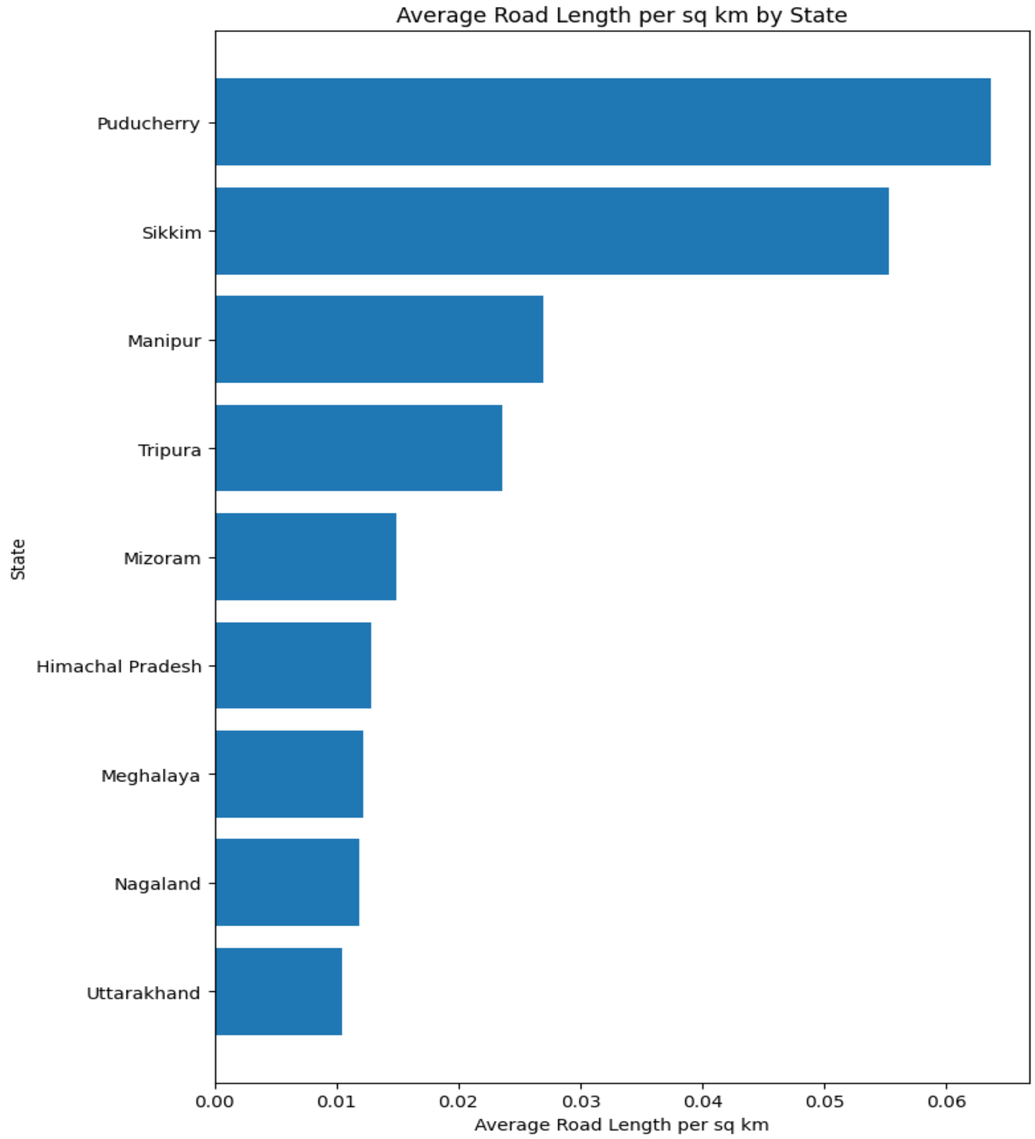


STATE_NAME	EXPENDITURE_OCCURED_LAKHS
Bihar	34151.1513
Odisha	32747.0124
Madhya Pradesh	32746.6924
Uttar Pradesh	27288.1984
Assam	22769.4277
West Bengal	16855.8859
Chhattisgarh	16680.8116
Rajasthan	16538.5977
Jharkhand	13244.3418
Jammu And Kashmir	12422.1896

v) **Objective:** To plot top 10 states with highest road length per sq. KM of state

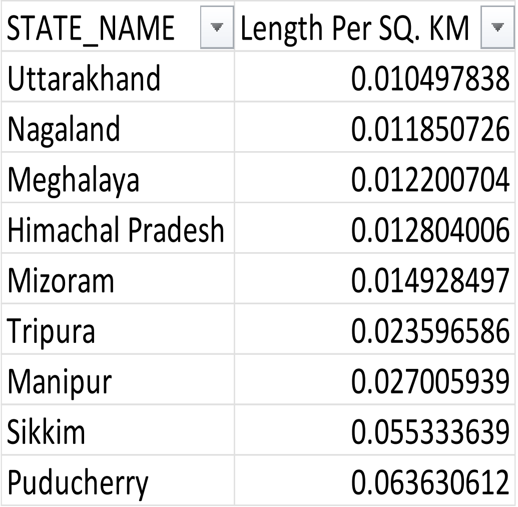
Number of variables	:	1
Type of Relation	:	Categorical
Type of Plot	:	Horizontal Bar Graph

Plot: This horizontal bar chart showcases the average road length per square Kilometer for each state, indicating road density. The y-axis represents the states, and the x-axis represents the average road length per sq. km. This chart compares road density across states. Longer bars indicate higher road density.



Inference: The bar chart reveals road density variations across different states, highlighting states with higher or lower road concentration relative to their area. This information can indicate differences in accessibility, connectivity, and infrastructure development. Puducherry and Sikkim have higher road density, suggesting better connectivity and potentially greater accessibility in those regions.

Dataframe Generated:

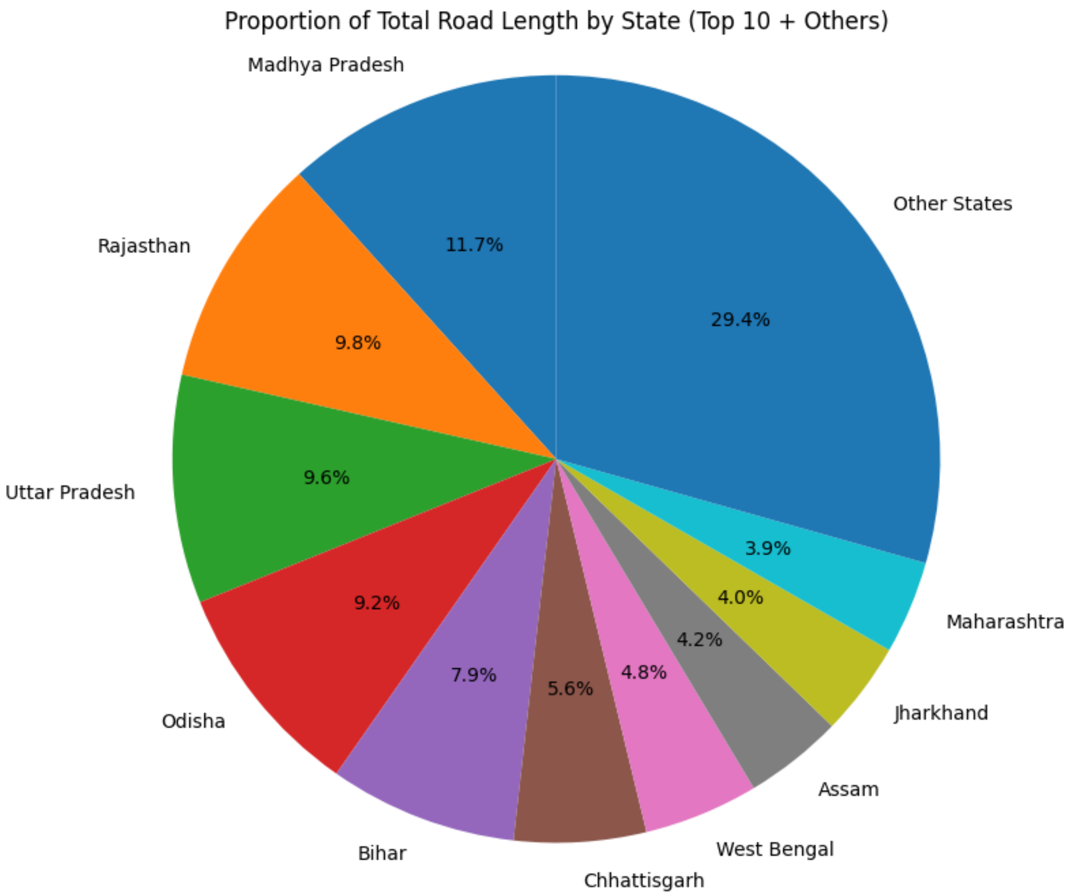


STATE_NAME ▾	Length Per SQ. KM ▾
Uttarakhand	0.010497838
Nagaland	0.011850726
Meghalaya	0.012200704
Himachal Pradesh	0.012804006
Mizoram	0.014928497
Tripura	0.023596586
Manipur	0.027005939
Sikkim	0.055333639
Puducherry	0.063630612

vi) **Objective:** To visualise proportion of top 10 and other states in total length of roadworks completed under PMGSY Scheme.

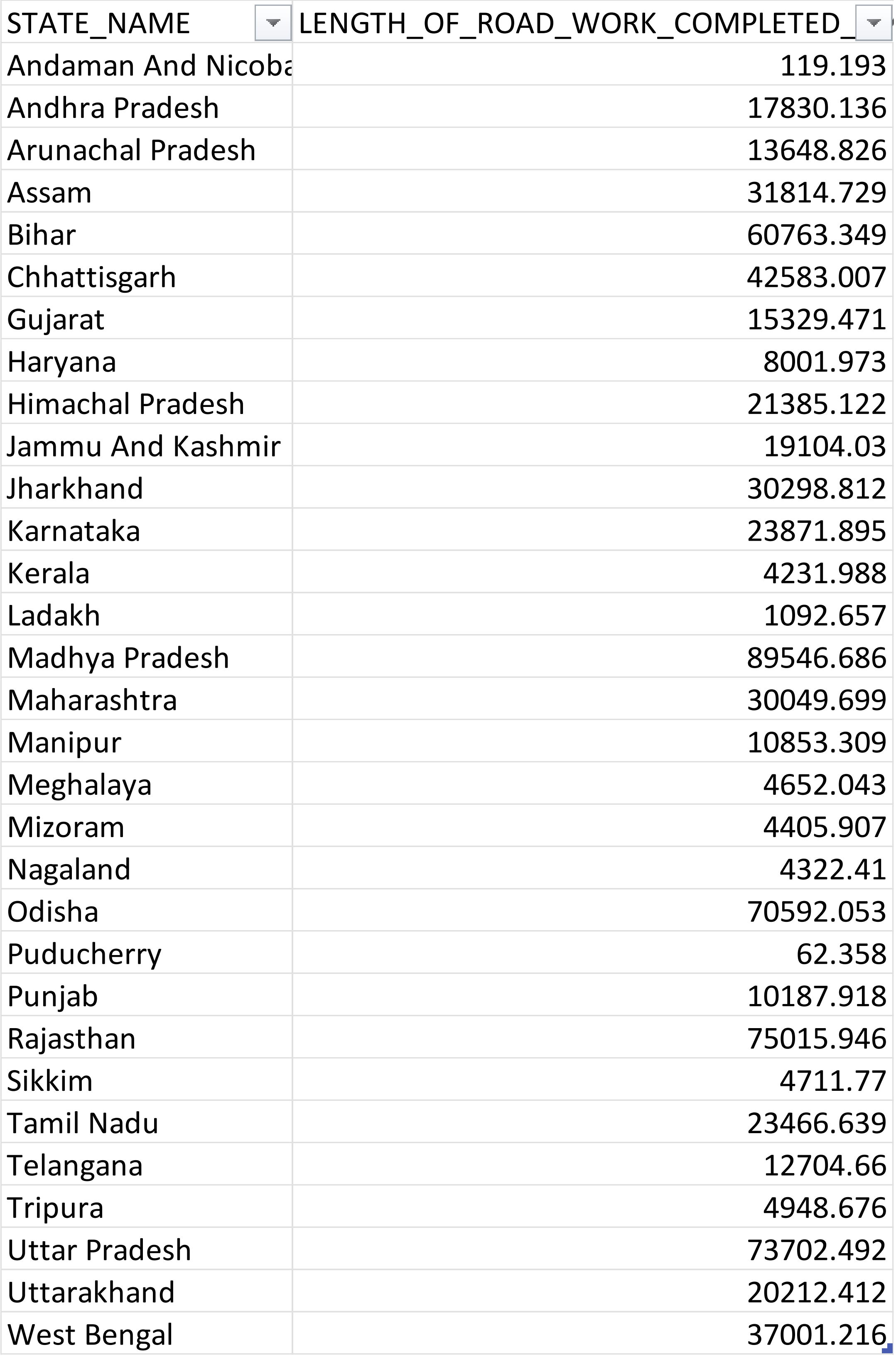
Number of variables	:	1
Type of Relation	:	Categorical
Type of Plot	:	Pie Chart

Plot: This pie chart represents the proportion of total road length contributed by the top 10 states and all other states combined under the PMGSY scheme. Each slice of the pie corresponds to a state or the "Other States" category. The size of each slice represents the percentage of total road length completed in that state or category.



Inference: The pie chart shows the distribution of road construction across different states under the PMGSY scheme. Larger slices represent states with a greater share of total road length completed, highlighting areas of concentrated infrastructure development. The top 10 states contribute a significant portion of total road length, indicating concentrated construction in those regions. This visualization provides a quick overview of the geographical distribution of road construction across India.

Dataframe Generated:



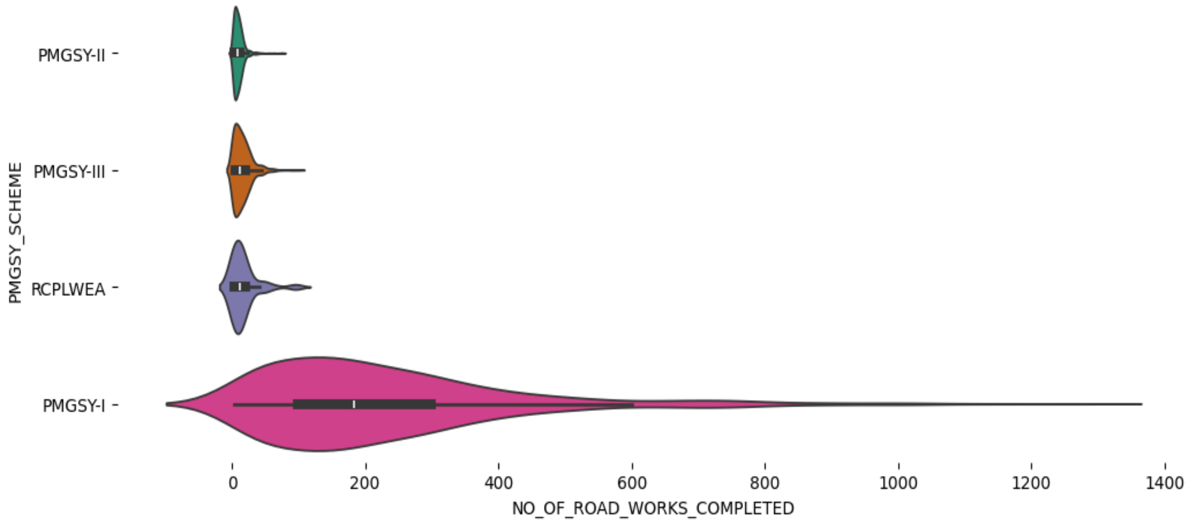
STATE_NAME	LENGTH_OF_ROAD_WORK_COMPLETED_
Andaman And Nicoba	119.193
Andhra Pradesh	17830.136
Arunachal Pradesh	13648.826
Assam	31814.729
Bihar	60763.349
Chhattisgarh	42583.007
Gujarat	15329.471
Haryana	8001.973
Himachal Pradesh	21385.122
Jammu And Kashmir	19104.03
Jharkhand	30298.812
Karnataka	23871.895
Kerala	4231.988
Ladakh	1092.657
Madhya Pradesh	89546.686
Maharashtra	30049.699
Manipur	10853.309
Meghalaya	4652.043
Mizoram	4405.907
Nagaland	4322.41
Odisha	70592.053
Puducherry	62.358
Punjab	10187.918
Rajasthan	75015.946
Sikkim	4711.77
Tamil Nadu	23466.639
Telangana	12704.66
Tripura	4948.676
Uttar Pradesh	73702.492
Uttarakhand	20212.412
West Bengal	37001.216

4.2) Phase Wise analysis of PMGSY Scheme.

Objective: To visualize no. of roadworks completed under different phases of PMGSY-Scheme.

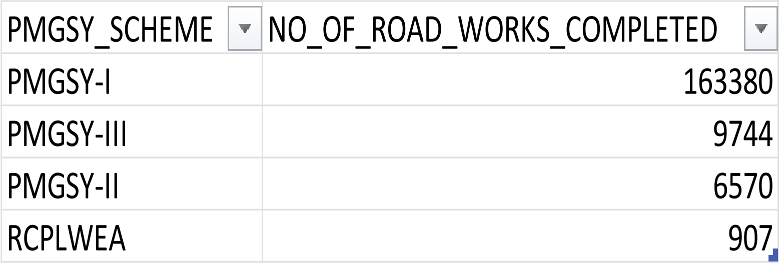
Number of variables	:	1
Type of Relation	:	Categorical
Type of Plot	:	Violen Plot

Plot: This violin plot compares the distribution of completed road works across different PMGSY schemes. Each violin represents a scheme, showing the range and frequency of completed road works. Wider sections indicate higher frequency, while the white dot marks the median.



Inference: The violin plot provides a detailed comparison of completed road works across PMGSY schemes, revealing variations in project size, frequency, and distribution. By analyzing violin shapes, you can identify schemes with larger or smaller projects and assess variability. It provides a comprehensive visual summary of how the number of completed road works varies across different PMGSY schemes, offering insights into project scale and implementation strategies.

Dataframe Generated:

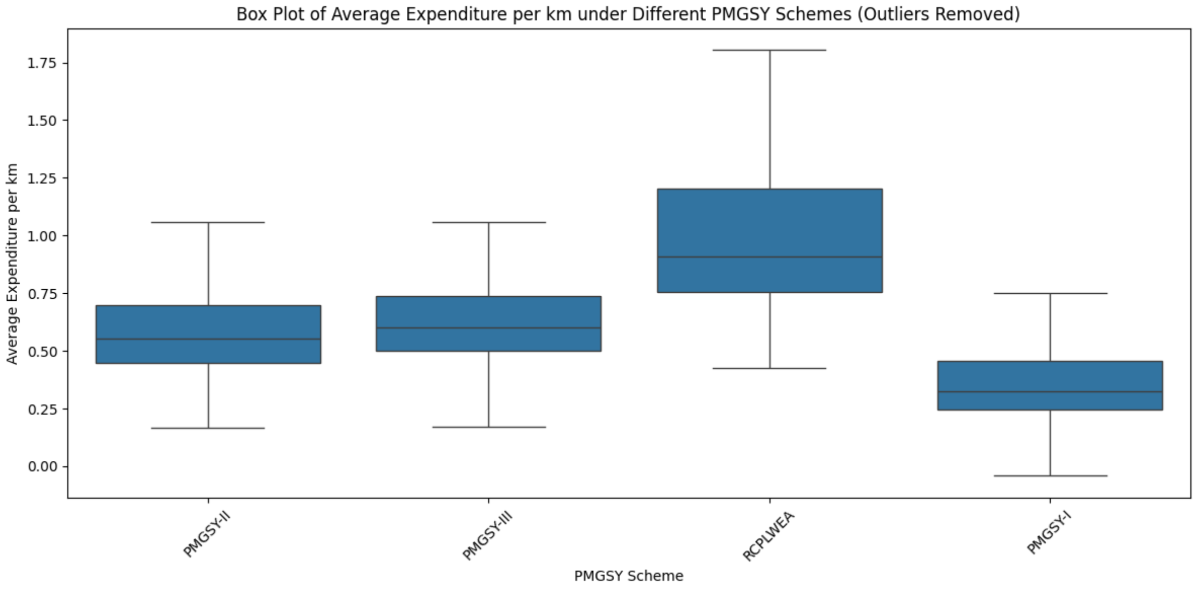


PMGSY_SCHEME	NO_OF_ROAD_WORKS_COMPLETED
PMGSY-I	163380
PMGSY-III	9744
PMGSY-II	6570
RCPLWEA	907

Objective: Analysis of Average cost per km construction of road under different PMGSY Phases.

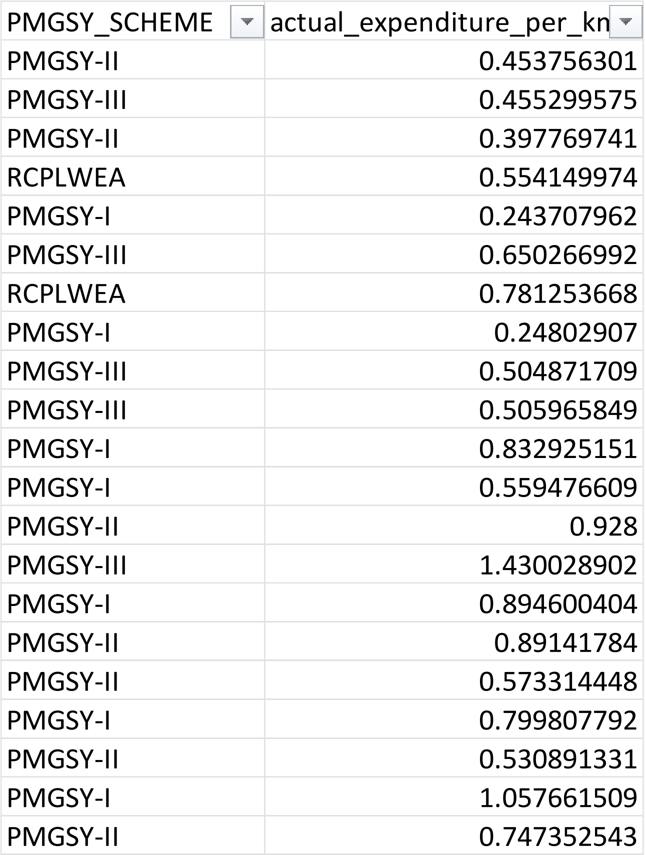
Number of variables	:	1
Type of Relation	:	Categorical
Type of Plot	:	Box Plot

Plot: This box plot displays the distribution of road lengths constructed under different PMGSY schemes. The x-axis represents the schemes, and the y-axis represents the road length in kilometers. Each box represents the interquartile range (IQR) of road lengths for a specific scheme, with the median marked by a line inside the box. Outliers have been hidden to show more average data.



Inference: The box plot of average expenditure per km under different PMGSY schemes provides a clear visual representation of cost distributions, allowing for comparisons of typical expenditures, variability, and potential outliers. By carefully analyzing the box plot elements and considering the context of different PMGSY schemes, you can gain insights into the cost structures and potential cost drivers for road construction projects under the program. This information can be valuable for evaluating the financial efficiency of the program and informing decision-making for future resource allocation and project planning.

Dataframe Used:



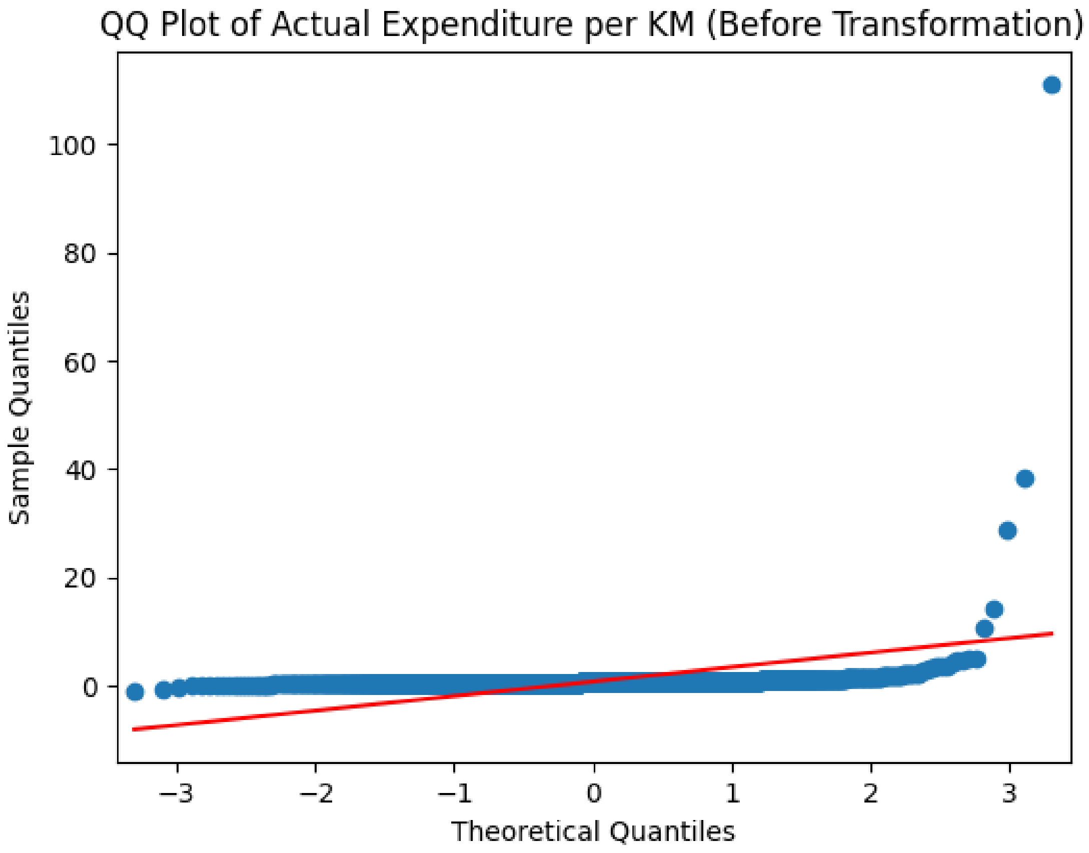
PMGSY_SCHEME	actual_expenditure_per_kn
PMGSY-II	0.453756301
PMGSY-III	0.455299575
PMGSY-II	0.397769741
RCPLWEA	0.554149974
PMGSY-I	0.243707962
PMGSY-III	0.650266992
RCPLWEA	0.781253668
PMGSY-I	0.24802907
PMGSY-III	0.504871709
PMGSY-III	0.505965849
PMGSY-I	0.832925151
PMGSY-I	0.559476609
PMGSY-II	0.928
PMGSY-III	1.430028902
PMGSY-I	0.894600404
PMGSY-II	0.89141784
PMGSY-II	0.573314448
PMGSY-I	0.799807792
PMGSY-II	0.530891331
PMGSY-I	1.057661509
PMGSY-II	0.747352543

Inferential Statistics:

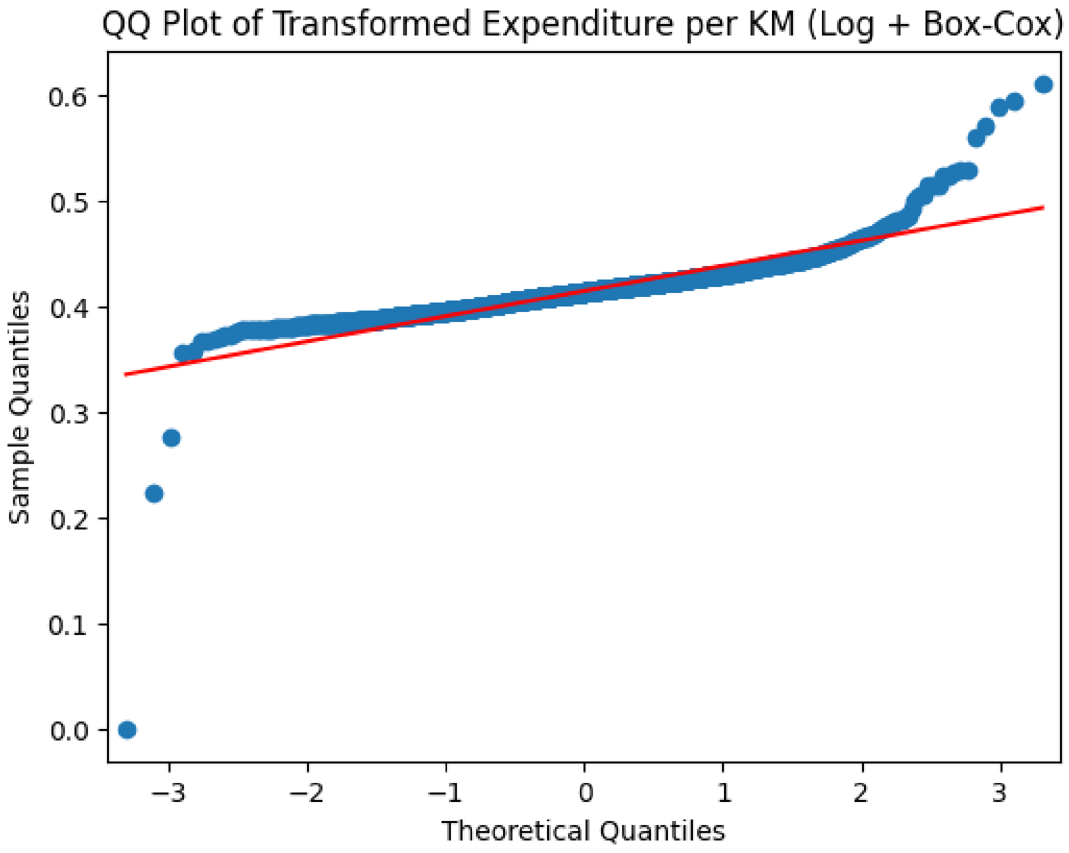
To check if the actual_expenditure_per_km data is normal or not if it is not We will apply normalisation to it.

Test – We are using Shapiro-Wilk test to see if the data is normal. Then we are using log and Box-Cox transformation to normalize it.

Result - Shapiro-Wilk Test (Before Transformation): Statistic=0.0541738978747911, p-value=9.201206107922576e-73



Result - Shapiro-Wilk Test (After Log + Box-Cox): Statistic=0.7885393309704485, p-value=5.59124070067253e-46



Conclusion - The log and Box-Cox transformations seem to have improved the normality of the 'actual_expenditure_per_km' data. This is indicated by:

- A better alignment of points with the diagonal line in the QQ plot after the transformations.
- A higher p-value from the Shapiro-Wilk test on the transformed data, suggesting reduced evidence against normality.

1st Hypothesis:

Null Hypothesis (H0): There is no significant difference in the average actual expenditure per km between the different PMGSY schemes.

Alternative Hypothesis (H1): There is a significant difference in the average actual expenditure per km between the different PMGSY schemes.

Test Used: ANOVA TEST for hypothesis, Shapiro wilk test for normalisation and for homogeneity we have used Levene's Test.

```
Results:              sum_sq     df       F    PR(>F)

PMGSY_SCHEME    160.194088     3.0  7.522492  0.000053

Residual     14849.956440  2092.0       NaN       NaN
```

```
Shapiro-Wilk Test for PMGSY-II: Statistic=0.8717022088165098, p-value=1.9862219935655772e-23

Shapiro-Wilk Test for PMGSY-III: Statistic=0.0717188690342726, p-value=1.069453111363139e-47

Shapiro-Wilk Test for RCPLWEA: Statistic=0.7375980003816217, p-value=1.757217906910643e-08

Shapiro-Wilk Test for PMGSY-I: Statistic=0.8994470330987803, p-value=2.8419742030582123e-21

Levene's Test: Statistic=3.2186130130316672, p-value=0.021928459665215252
```

Interpretation:

The p-value is less than our significance level (0.05), we are rejecting the null hypothesis and concluding that there is a significant difference in the average actual expenditure per km between the different PMGSY schemes.

2nd Hypothesis:

Null Hypothesis (H0): There is no significant relationship between the total road length completed and the state area.

Alternative Hypothesis (H1): There is a significant relationship between the total road length completed and the state area.

Test Used: Pearson correlation coefficient

Results: Pearson Correlation Coefficient: 0.7279355316540873, P-value: 5.141880524394789e-06

Interpretation: Based on the results, we can conclude that there is a statistically significant positive relationship between the total road length completed and the state area. In other words, larger states tend to have more roads completed under the PMGSY scheme.

Causality: Correlation does not imply causation. While there is a relationship between road length and state area, we cannot conclude that a larger state area directly causes more roads to be built. Other factors might be involved, such as population density, economic activity, or government policies. Strength of Relationship: The correlation coefficient of 0.727 suggests a moderate positive relationship. This means that the relationship is not extremely strong, but it is still noticeable and statistically significant.