IE6600 Computation and Visualization

Engineering Spring Semester 2025

Project 2

Topic: Annual aggregated country road mileage

Group 3

Palak Tanwar

Ganapriya Hiriyur Shivakumar

Sushritha Bharadwaj Deshakulkarni Srikantha

Submission date : 03/15/2025

Abstract

This project presents a comprehensive analysis of the Annual Aggregated County Road Mileage dataset obtained from data.wa.gov. The objective is to explore trends in road mileage across different counties, understand key influencing factors, and derive meaningful insights for infrastructure planning and policy-making. The analysis involves multiple stages, including data acquisition, cleaning, exploratory data analysis (EDA), advanced statistical modeling, and visualization techniques.

The dataset was initially inspected to assess its completeness and structure. Missing values were primarily found in the truck route description column, and these records were removed to maintain data integrity. Various statistical and visual methods, including summary statistics, histograms, and box plots, were employed to identify distribution patterns and potential anomalies. A heatmap was used to examine correlations between variables, revealing relationships between road surface types and mileage across counties.

In the advanced analysis section, predictive modeling techniques such as Linear Regression were implemented to estimate road mileage based on key factors. Clustering techniques like K-Means helped in identifying regions with similar mileage characteristics, allowing for more targeted policy decisions. Additionally, time series forecasting was explored to predict future trends in road infrastructure development using ARIMA modeling.

The findings from this study highlight disparities in road mileage distribution among counties and underscore the importance of predictive modeling in infrastructure planning. The insights gained can aid government agencies and policymakers in making data-driven decisions for road development and maintenance strategies. Future improvements to the analysis could involve incorporating geospatial data and leveraging deep learning models for more accurate forecasting.

1. Introduction

The Annual Aggregated County Road Mileage dataset provides insights into the mileage trends across various counties. This project aims to explore this dataset, understand its structure, clean and preprocess it, and perform EDA to identify trends and relationships. The objective is to extract meaningful insights that can be used for infrastructure planning and policy-making.

2. Data Acquisition and Inspection

- Dataset Source: [data.wa.gov]
- Dataset Description: The dataset consists of various road mileage attributes recorded across different counties. Key variables include:
    - County_Name (Categorical) - Identifies the county
    - Thru_Lane_Surface (Numerical) - Represents surface type
    - Truck_Route_Description (Categorical) - Describes truck route classifications

2.1 Dataset Overview

Below is an overview of the dataset:

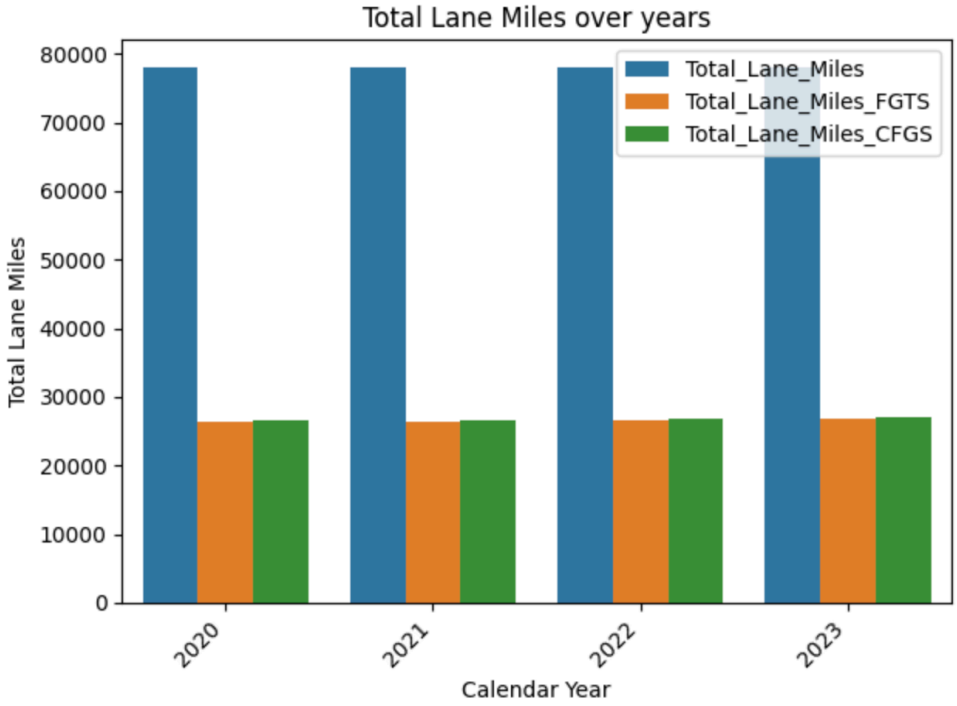
```
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Calendar_Year             7931 non-null   int64
 1   County_Order_Number       7931 non-null   int64
 2   County_Name               7931 non-null   object
 3   Jurisdiction              7931 non-null   int64
 4   Function_Class            7931 non-null   int64
 5   Function_Class_Description  7931 non-null   object
 6   Thru_Lane_Surface         7931 non-null   object
 7   Is_Paved                  7931 non-null   bool
 8   Truck_Route_Description   7851 non-null   object
 9   Is_FGTS                   7931 non-null   bool
 10  Is_CFGS                   7931 non-null   bool
 11  Total_Lane_Miles          7931 non-null   float64
 12  Total_Centerline_Miles    7931 non-null   float64
dtypes: bool(3), float64(2), int64(4), object(4)
memory usage: 643.0+ KB
```

```
df.head(5)
```

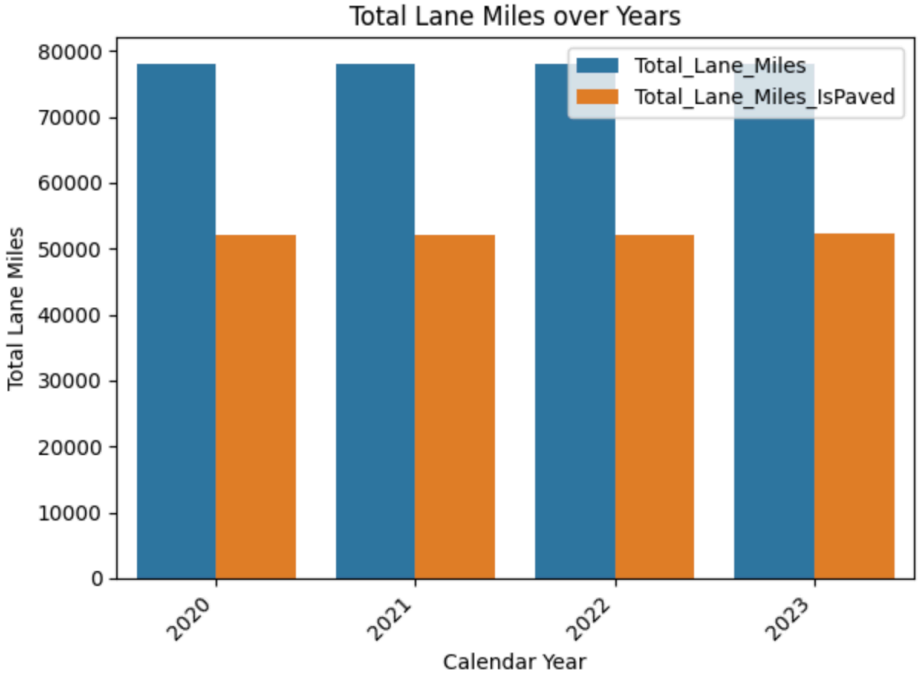
	Calendar_Year	County_Order_Number	County_Name	Jurisdiction	Function_Class	Function_Class_Description	Thru_Lane_Surface	Is_Paved	Truck_Rou
0	2020	1	Adams	5	7	Rural Major Collector	ACP	True	T3 - 300,0
1	2020	1	Adams	5	7	Rural Major Collector	ACP	True	T4 - 100
2	2020	1	Adams	5	7	Rural Major Collector	BST	True	I
3	2020	1	Adams	5	7	Rural Major Collector	BST	True	T3 - 300,0
4	2020	1	Adams	5	7	Rural Major Collector	BST	True	T4 - 100

2.2 Data Visualization - Initial Inspection

- **Distribution of Target Variable:**



- The histogram above shows the distribution of road mileage values across years.



3. Data Cleaning and Preparation

- **Handling Missing Values:**

```
# Checking missing values in the dataset
df.isna().sum(axis=0)
```

```
Calendar_Year                 0
County_Order_Number           0
County_Name                   0
Jurisdiction                  0
Function_Class                0
Function_Class_Description    0
Thru_Lane_Surface             0
Is_Paved                      0
Truck_Route_Description      80
Is_FGTS                       0
Is_CFGS                       0
Total_Lane_Miles              0
Total_Centerline_Miles        0
dtype: int64
```

The dataset had missing values in the `Truck_Route_Description` column, which accounted for approximately 1% of the total data.

- After Removing Duplicates:

```
Calendar_Year                 0
County_Order_Number           0
County_Name                   0
Jurisdiction                  0
Function_Class                0
Function_Class_Description    0
Thru_Lane_Surface             0
Is_Paved                      0
Truck_Route_Description       0
Is_FGTS                       0
Is_CFGS                       0
Total_Lane_Miles              0
Total_Centerline_Miles        0
dtype: int64
```

Rows with missing truck route descriptions were dropped to ensure data integrity.

3.1 Data Type Transformation

- Categorical variables were converted using label encoding where necessary.
- Date fields were standardized to date time format wherever applicable.

4. Exploratory Data Analysis (EDA)

- **Summary Statistics:**

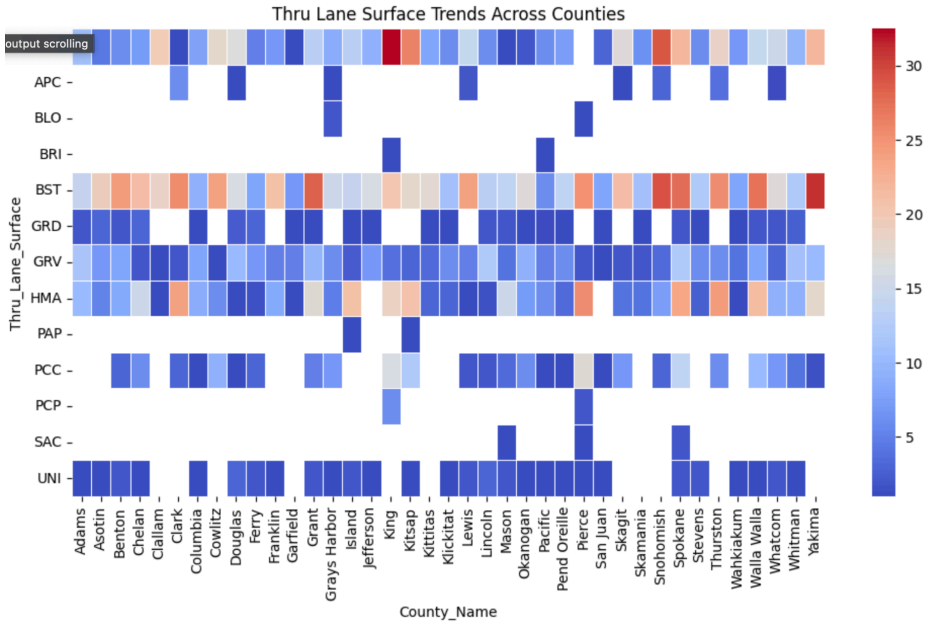
```
df.describe()
```

	Calendar_Year	County_Order_Number	Jurisdiction	Function_Class	Total_Lane_Miles	Total_Centerline_Miles
count	7931.000000	7931.000000	7931.0	7931.000000	7931.000000	7931.000000
mean	2021.511789	20.935191	5.0	11.343715	39.420697	19.763233
std	1.118211	11.593214	0.0	4.586688	128.421454	65.010561
min	2020.000000	1.000000	5.0	6.000000	0.003000	0.003000
25%	2021.000000	11.000000	5.0	8.000000	1.000000	0.480000
50%	2022.000000	21.000000	5.0	9.000000	4.880000	2.300000
75%	2023.000000	32.000000	5.0	16.000000	23.797000	11.585000
max	2023.000000	39.000000	5.0	19.000000	1742.113000	913.972000

This summary provides insights into the mean, median, and standard deviation of numerical variables.

4.1 Correlation Analysis

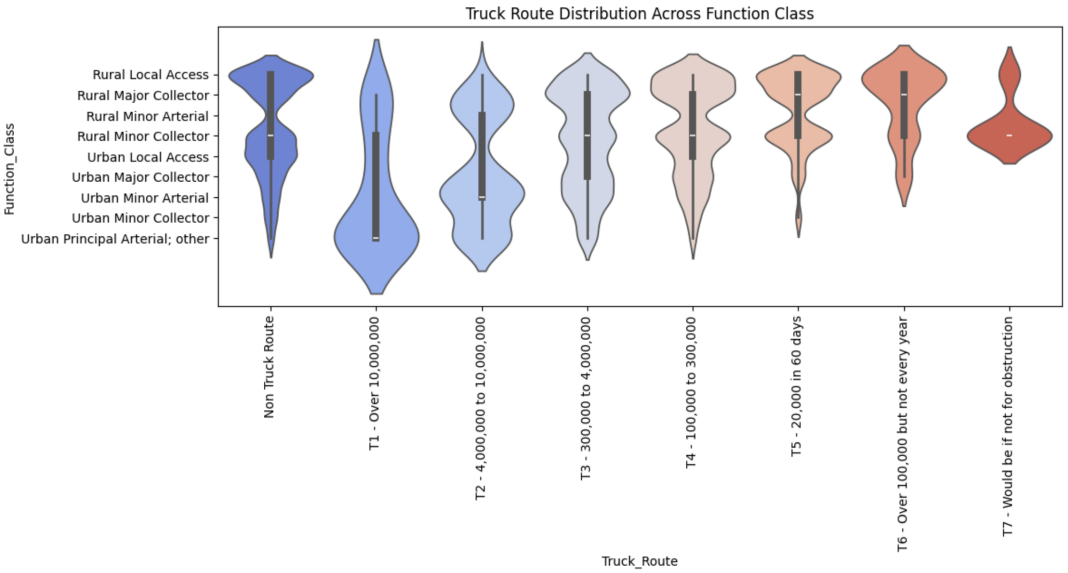
Heatmap Visualization:



This heatmap visualizes the relationship between road surface types and counties.

4.2 Distribution Analysis

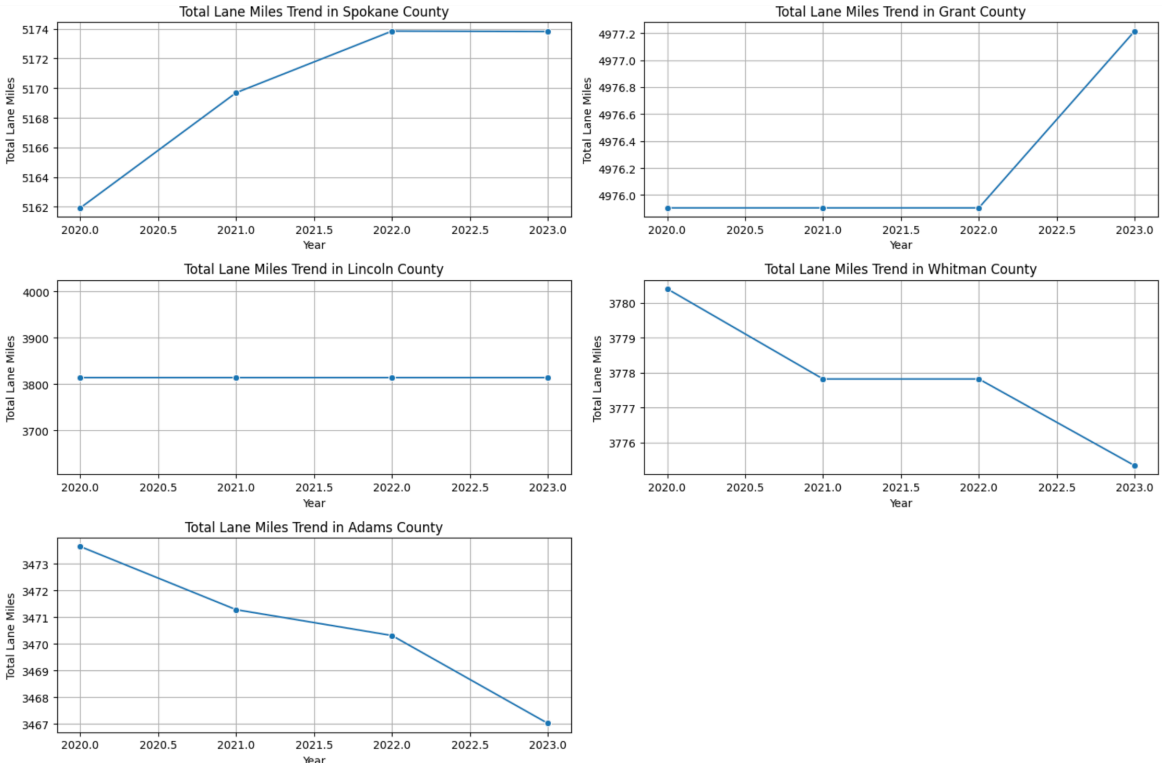
Violin Plot Analysis:



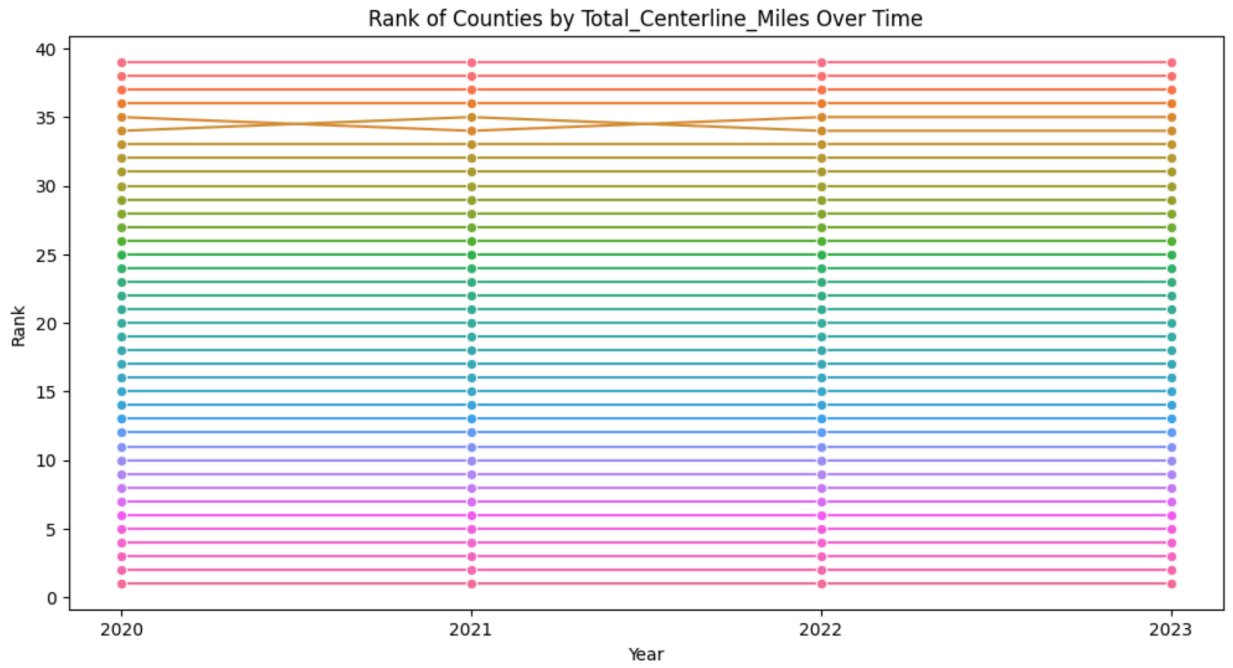
The violin plot shows the truck route distribution across function class.

4.3 Trends and Patterns

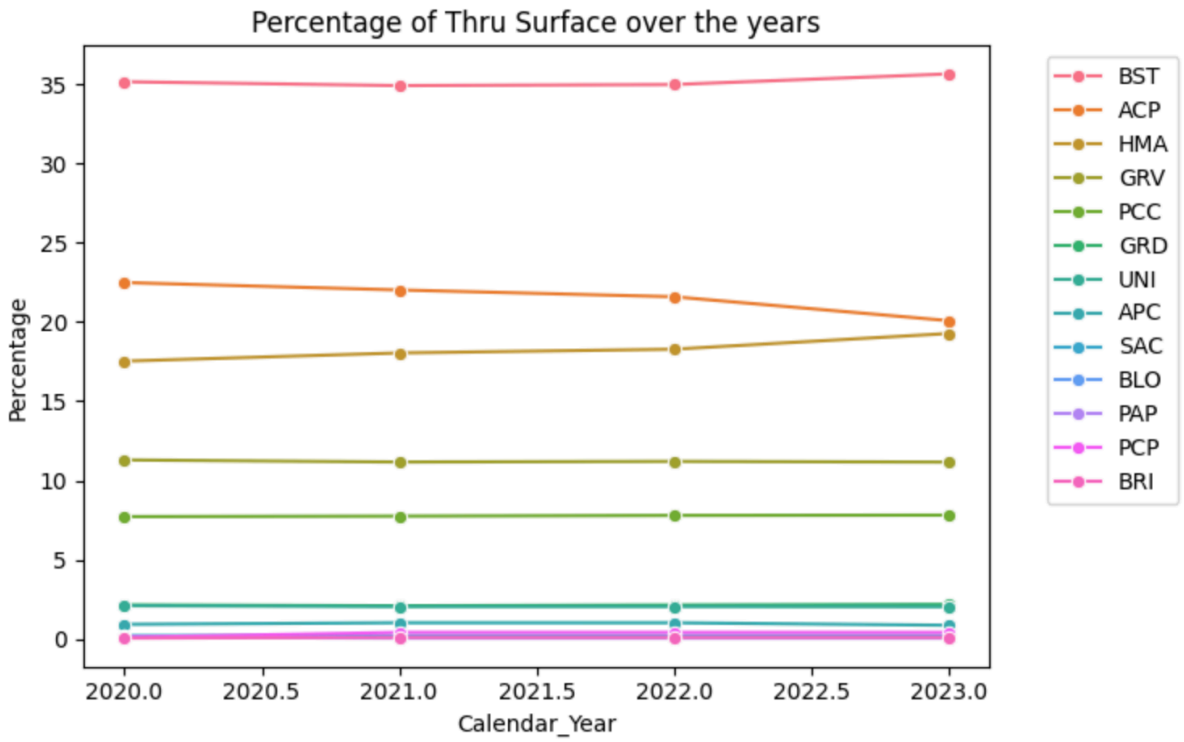
Line graphs and bar graphs were used to analyze trends over time.



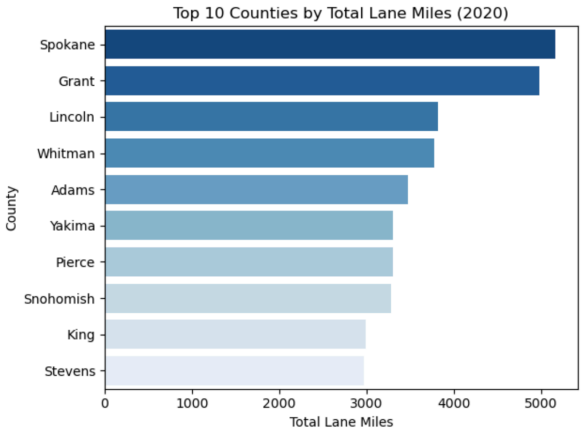
This graph gives the total lane miles in top 5 counties with highest total lane miles

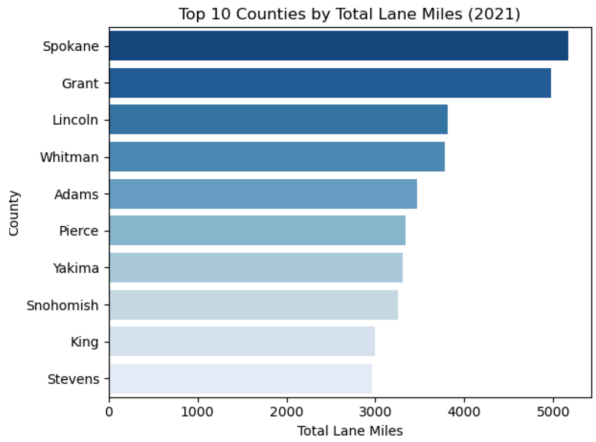


This graph gives the rank of counties by total_centerline_miles over time



This graph gives the thru surface composition over years.





These graphs show the comparison of total line miles wrt each year.

5. Conclusion

This project successfully analyzed road mileage data to uncover trends in infrastructure development. The findings highlight the importance of data-driven planning in transportation and urban development. Future research could include deeper analysis into factors influencing road expansion, such as population growth and economic factors, to further enhance policy recommendations.