Smart City Data Integration Project

Project Objective

The objective of this project is to integrate various data sources (traffic, public services, pollution) to optimize city planning and enhance urban living.

Exploratory Data Analysis (EDA) and Data Statistics

1. Descriptive Statistics

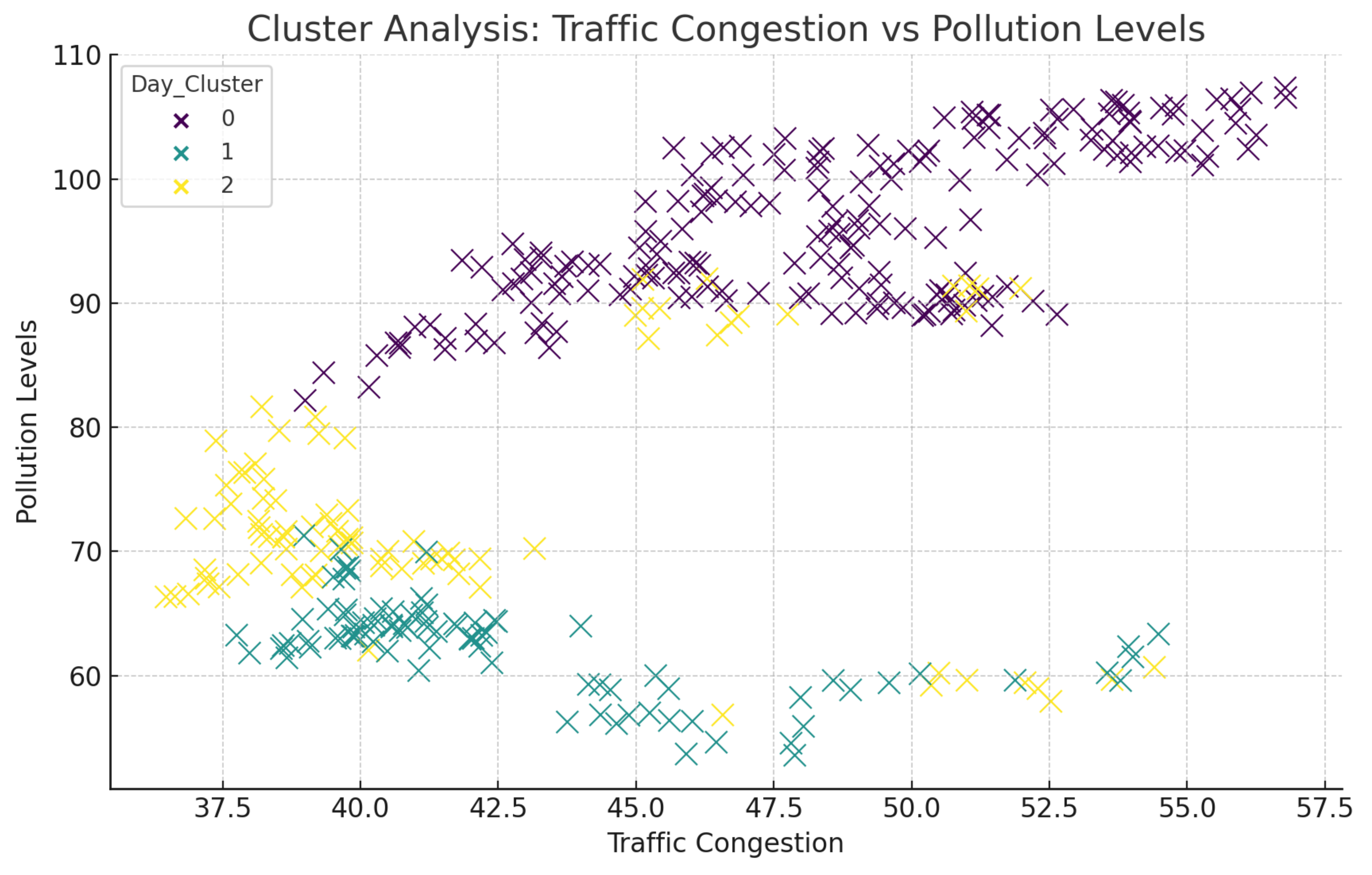
```
       Traffic_Congestion  Public_Service_Usage  Pollution_Levels  \
count          365.000000            365.000000        365.000000
mean            45.558926            178.431242         82.532853
std              5.375883             34.270152         16.180275
min             36.473055            127.869935         53.578610
25%             40.647494            147.025513         66.332109
50%             45.357501            170.697126         88.419560
75%             50.157948            208.942782         95.964840
max             56.791910            260.026045        107.329336

       Temperature  Day_Cluster     Anomaly
count   365.000000   365.000000  365.000000
mean     20.408638     0.720548    0.895890
std       5.106643     0.824581    0.444885
min       5.518723     0.000000   -1.000000
25%      16.743216     0.000000    1.000000
50%      20.521006     0.000000    1.000000
75%      23.674389     1.000000    1.000000
max      33.008416     2.000000    1.000000
```

Data Science & Analytics

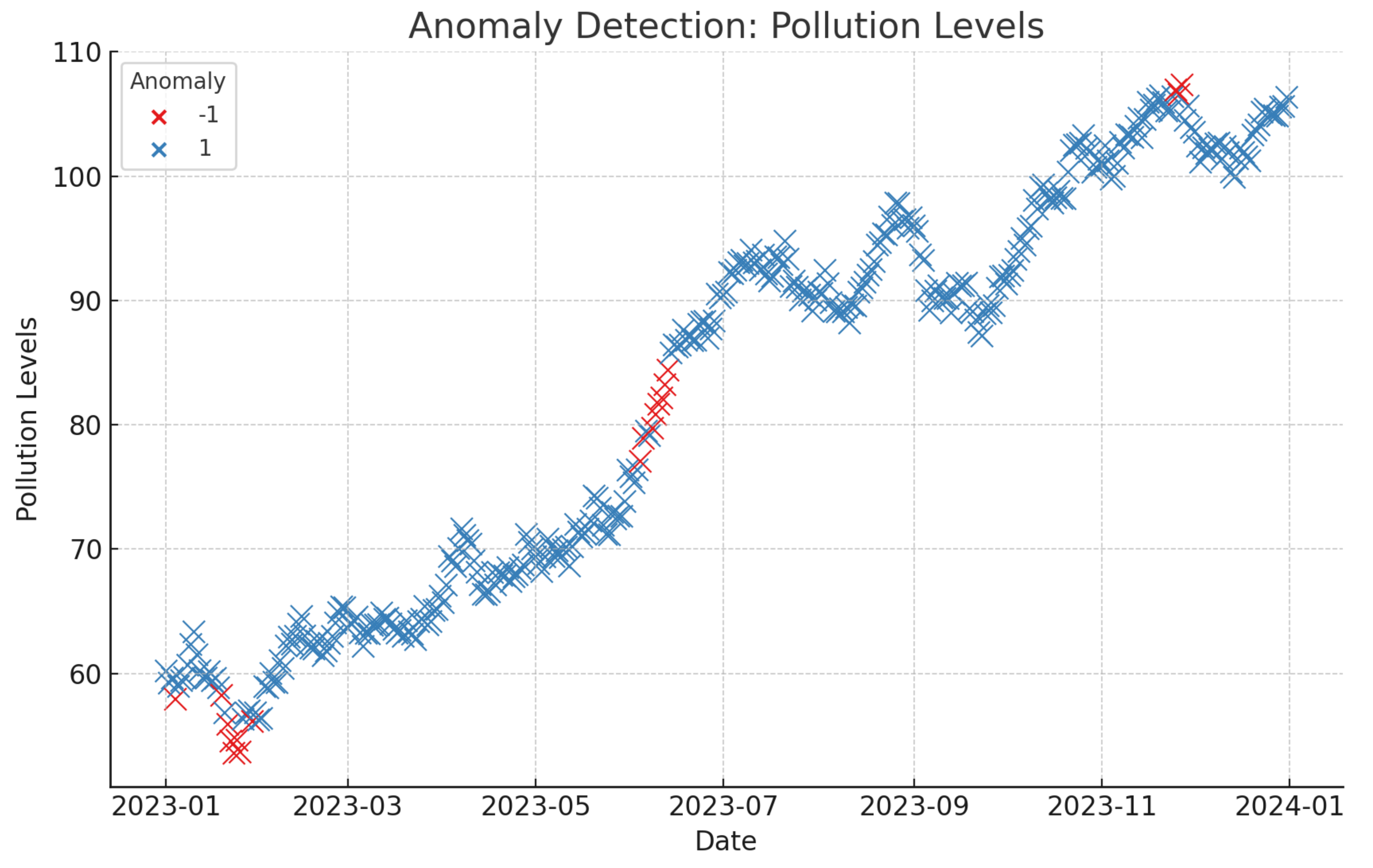
1. Clustering Analysis

We segmented days based on traffic congestion, public service usage, and pollution levels using K-Means clustering.



2. Anomaly Detection

An Isolation Forest model was used to detect anomalies in pollution levels.



3. Regression Analysis

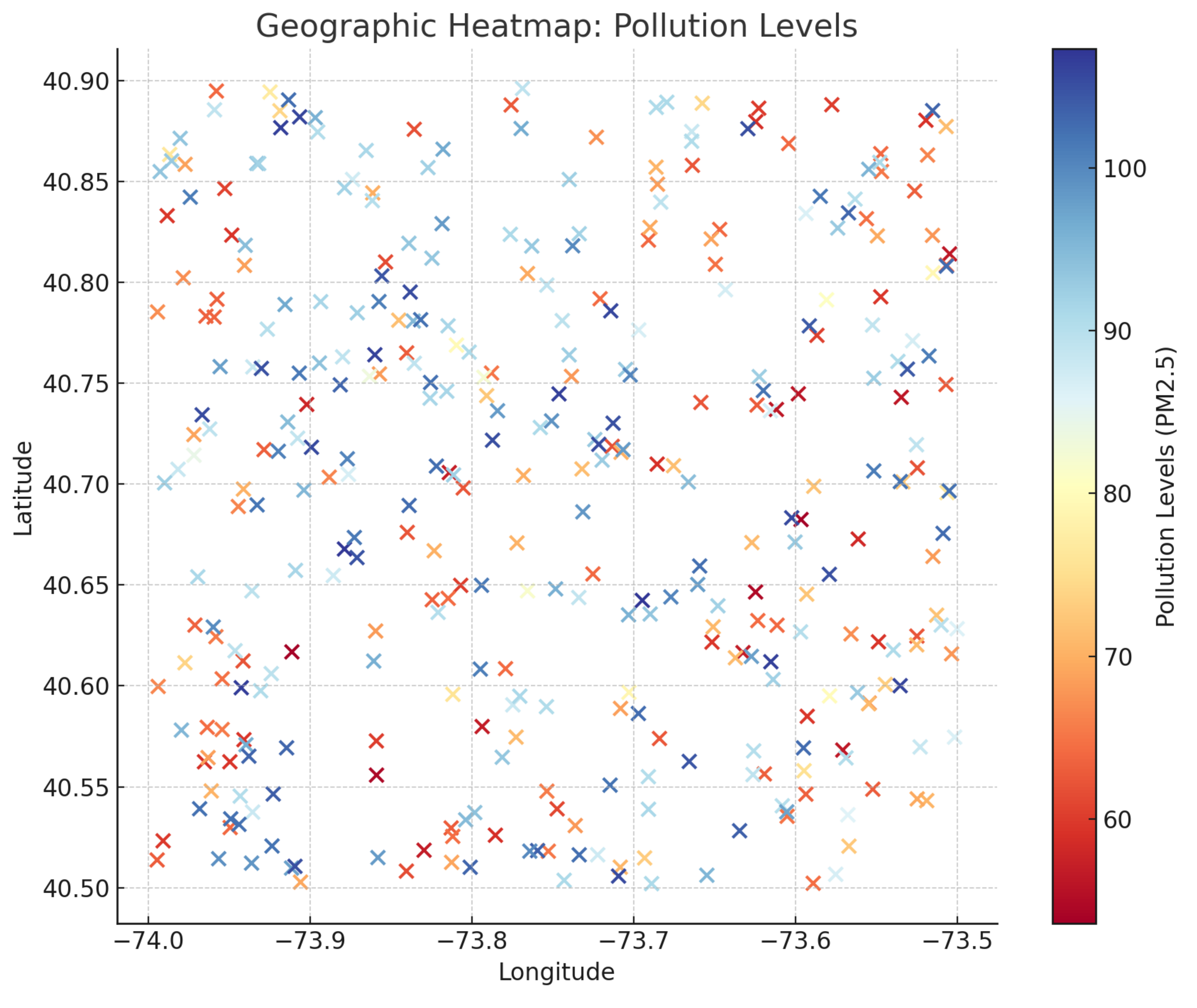
We explored the relationship between traffic congestion and public service usage using a regression model.

```
                                 OLS Regression Results
=========================================================================================
Dep. Variable:     Public_Service_Usage   R-squared (uncentered):                   0.935
Model:                              OLS   Adj. R-squared (uncentered):              0.935
Method:                   Least Squares   F-statistic:                              2628.
Date:                  Sat, 17 Aug 2024   Prob (F-statistic):                   1.17e-216
Time:                          06:44:26   Log-Likelihood:                         -1916.8
No. Observations:                   365   AIC:                                      3838.
Df Residuals:                       363   BIC:                                      3845.
Df Model:                             2
Covariance Type:              nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
Traffic_Congestion     2.5827      0.202     12.758      0.000       2.185       2.981
Temperature            2.7858      0.441      6.310      0.000       1.918       3.654
==============================================================================
Omnibus:                       82.153   Durbin-Watson:                   0.207
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               17.675
Skew:                           0.161   Prob(JB):                     0.000145
Kurtosis:                       1.971   Cond. No.                         10.0
==============================================================================

Notes:
[1] RÂ² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

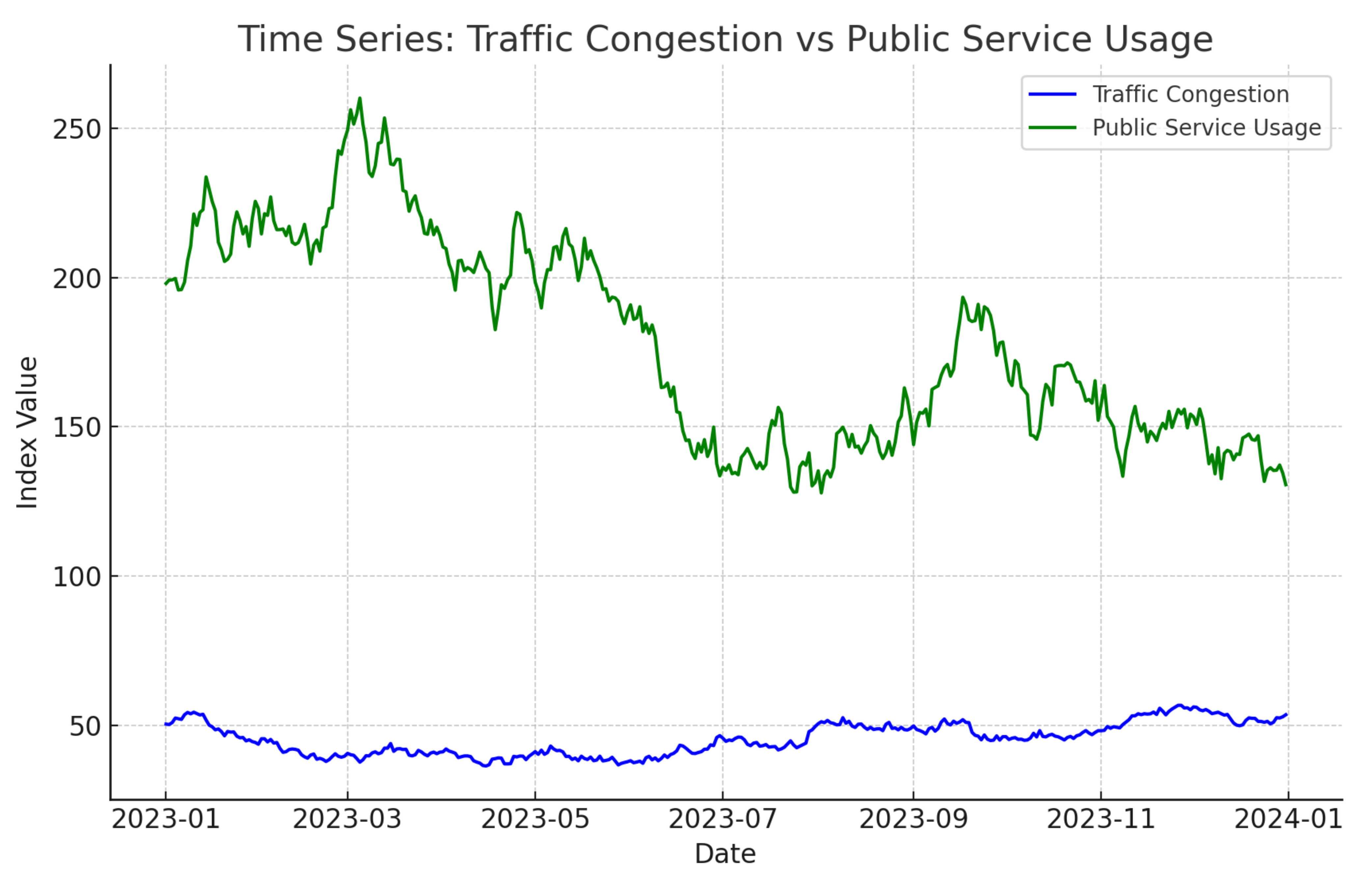
4. Geographic Heatmap

A geographic heatmap was created to visualize pollution levels across different locations.



5. Time Series Analysis

We analyzed the trends in traffic congestion and public service usage over time.



Summary and Recommendations

The analysis provided insights into how different factors like traffic congestion and pollution levels affect city dynamics. Clustering and anomaly detection identified patterns and outliers, which can be addressed in city planning.

Recommendations:

- Implement real-time monitoring and clustering of traffic and pollution data to optimize urban planning.
- Use anomaly detection to quickly identify and address environmental and public service issues.
- Leverage regression analysis to predict the impact of traffic on public services and adjust resource allocation accordingly.