**DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING**

**PROJECT REPORT**

(Project Semester January-April 2025)

***STREET NETWORK STUDY***

Submitted by

Anamika Sharma

Registration No.: 12319505

Programme and Section: Computer Science and Engineering, K23EU

Course Code: INT375

Under the Guidance of

**Dr. Tanima Thakur, UID: 23532**

**Discipline of CSE/IT**

**Lovely School of Computer Science and Engineering**

**Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that Anamika Sharma bearing Registration no. 12319505 has completed INT375 project titled, **“Street Network Study”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Signature and Name of the Supervisor**

**Designation of the Supervisor**

**School of Computer Science and Engineering**

Lovely Professional University

Phagwara, Punjab.

Date: 11-04-2025

**DECLARATION**

I, Anamika Sharma, student of Bachelor of Technology under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 11-04-2025 Signature

Registration No. 12319505 Anamika Sharma

**ACKNOWLEDGEMENT**

I would like to express my sincere gratitude to my teacher Dr. Tanima Thakur for their valuable guidance and support throughout the completion of this Excel assignment. Their encouragement and insights have been instrumental in helping me understand the practical applications of Excel.

I also extend my thanks to my classmates and friends who supported me during the preparation of this assignment.

Lastly, I am thankful to my family for their continuous support and motivation.

Date: 11-04-2025

Name: Anamika Sharma

Registration No. 12319505

**TABLE OF CONTENT**

* Introduction
* Source of Dataset
* EDA Process
* Analysis on Dataset
* Introduction
* General Description
* Specific Requirements, Functions and Formulas
* Analysis Results
* Visualization
* Conclusion
* Future Scope
* References

**1. INTRODUCTION**

The purpose of this Exploratory Data Analysis (EDA) report is to gain meaningful insights from the Street Centerline Dataset, which contains geospatial and descriptive information about road segments within a city. The dataset includes attributes such as street names, segment lengths, direction types, responsible authorities, and classifications, making it highly relevant for urban planning, traffic management, and infrastructure maintenance.

Through this analysis, we aim to answer key questions that can assist city planners, transportation departments, and decision-makers in understanding the underlying patterns and structure of the street network. Using powerful Python libraries like Pandas, NumPy, Matplotlib, and Seaborn, this report applies a combination of statistical methods and visualization techniques to explore trends, detect outliers, and uncover correlations.

This EDA focuses on the following primary objectives:

* Identifying the most frequently occurring street names.
* Analyzing the longest individual road segments.
* Investigating whether certain authorities are responsible for managing longer streets.
* Understanding the distribution of street directions.
* Exploring correlations between road length, shape length, and street classification.

By the end of this report, we expect to extract data-driven insights that support better city infrastructure design, improve transportation flow, and optimize road maintenance operations.

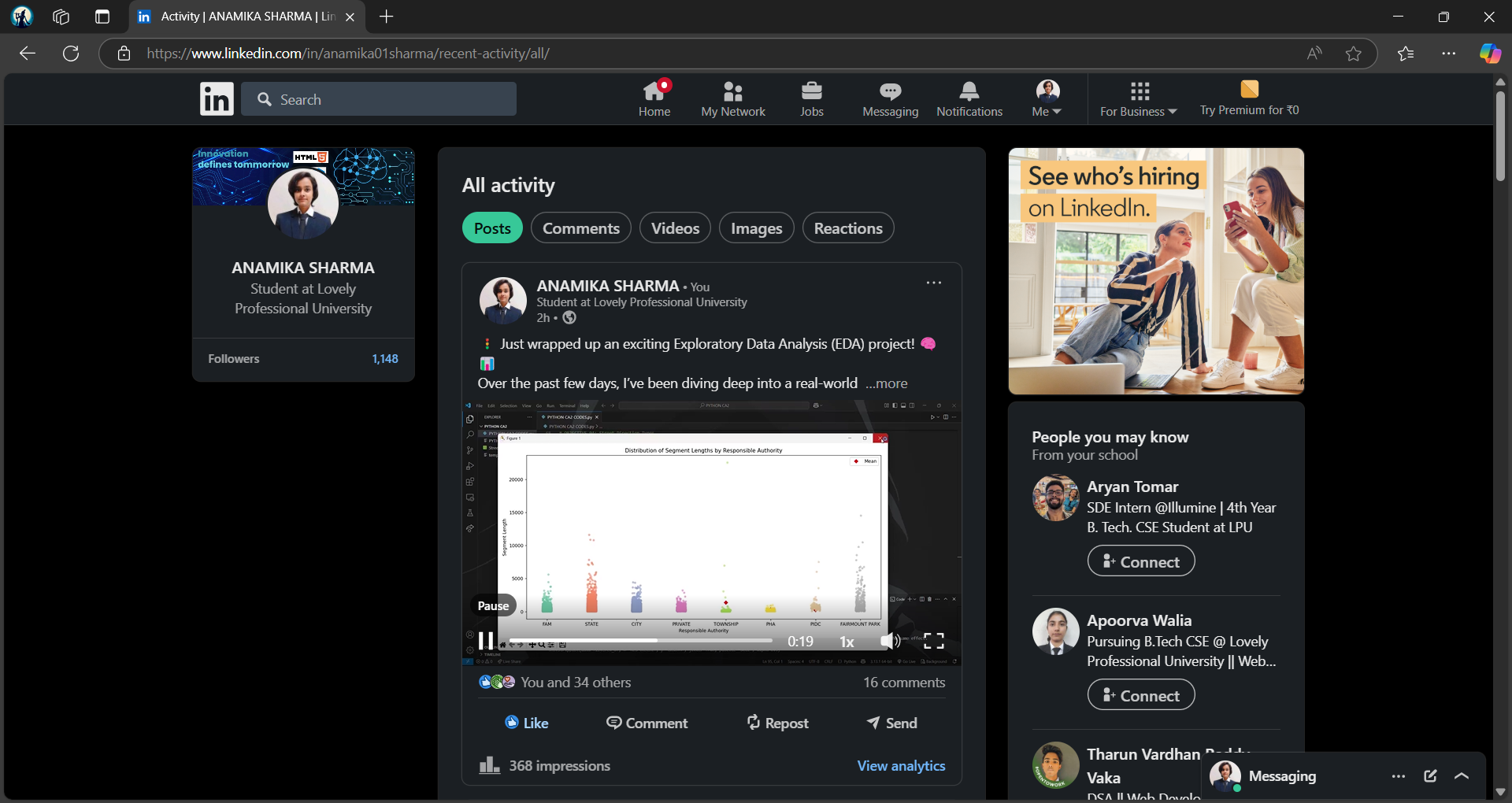
**2. SOURCE OF DATASET**

Link of dataset:

<https://catalog.data.gov/dataset/street-centerlines-70985>

Download Link: <https://hub.arcgis.com/api/v3/datasets/c36d828494cd44b5bd8b038be696c839_0/downloads/data?format=csv&spatialRefId=3857&where=1%3D1>

LinkedIn Post: <https://www.linkedin.com/posts/anamika01sharma_datascience-python-eda-activity-7317173373581570048-S_V9?utm_source=share&utm_medium=member_desktop&rcm=ACoAAEYNqrABcB6nOvMDcKme-OoeFYfAJ5rPLwc>



**3. EDA PROCESS AND ANALYSIS**

The Exploratory Data Analysis (EDA) process for the Street Centerline Dataset was carried out in a structured and methodical manner to ensure the extraction of reliable, insightful, and actionable information. The dataset contains rich geographic and administrative details of street segments, making it highly suitable for urban infrastructure analysis.

1. Data Loading and Initial Exploration

The dataset was first imported using the pandas library. A preliminary assessment was conducted to understand the dataset's structure, including its shape, column names, and the types of data present. This initial inspection helped identify key fields such as ST\_NAME, LENGTH, RESPONSIBL, ONEWAY, and Shape\_Length, which were crucial for further analysis.

1. Handling Missing Values

Before diving into visual exploration, data cleaning was performed. Missing or null values were carefully dropped or handled, depending on the objective, to ensure accuracy in visualizations and statistics. Columns such as LENGTH and RESPONSIBL were checked and filtered to exclude incomplete entries for meaningful analysis.

1. Univariate and Bivariate Analysis

A mix of univariate and bivariate techniques was applied:

Univariate Analysis involved assessing the distribution of single features, such as the frequency of street names and the count of one-way versus both-way streets.

Bivariate Analysis explored relationships between two variables — for example, the link between street length and responsible authority, and correlations between numeric features like LENGTH, Shape\_Length, and CLASS.

1. Visualization Techniques

Various types of visualizations were used to present the data clearly:

Bar plots were used to display the top 10 most common street names and direction counts.

Lollipop charts helped emphasize the longest road segments in a visually appealing way.

Strip plots with mean overlays showed how road lengths varied across different responsible authorities, allowing both distribution and average comparisons.

Heatmaps provided a compact view of correlations between road dimensions and classifications.

1. Insights and Trends

From the analysis:

Several street names appeared repeatedly, indicating a high concentration or repetition of commonly used names in the city’s layout.

A few streets had exceptionally long segments, possibly representing major arterial roads.

Certain responsible authorities were linked to consistently longer street segments, suggesting administrative zoning or jurisdiction over primary routes.

A majority of roads were bi-directional, with some specialized one-way streets, reflecting city design for smoother traffic.

The correlation matrix highlighted moderate relationships between length-based attributes, useful for predicting or cross-verifying spatial data.

Summary of EDA Approach:

* Emphasis on data integrity through filtering and cleaning.
* Balanced use of descriptive statistics and visual exploration.
* Selection of appropriate plot types for each objective to enhance interpretability.
* Clear breakdown of patterns relevant to urban planning and transport infrastructure.

**4. ANALYSIS ON DATASET**

* Objective 1: What are the most common street names in the dataset?

Introduction: Understanding which street names appear most frequently in a city’s dataset is valuable for identifying recurring naming conventions, high-density traffic zones, or even potential areas of confusion in navigation systems.

General Description: The dataset contains a column named `ST\_NAME` that records the name of each street segment. Since many streets may be broken into multiple segments for mapping or maintenance purposes, a single street name might appear several times.

Specific Requirements, Functions and Formulas:

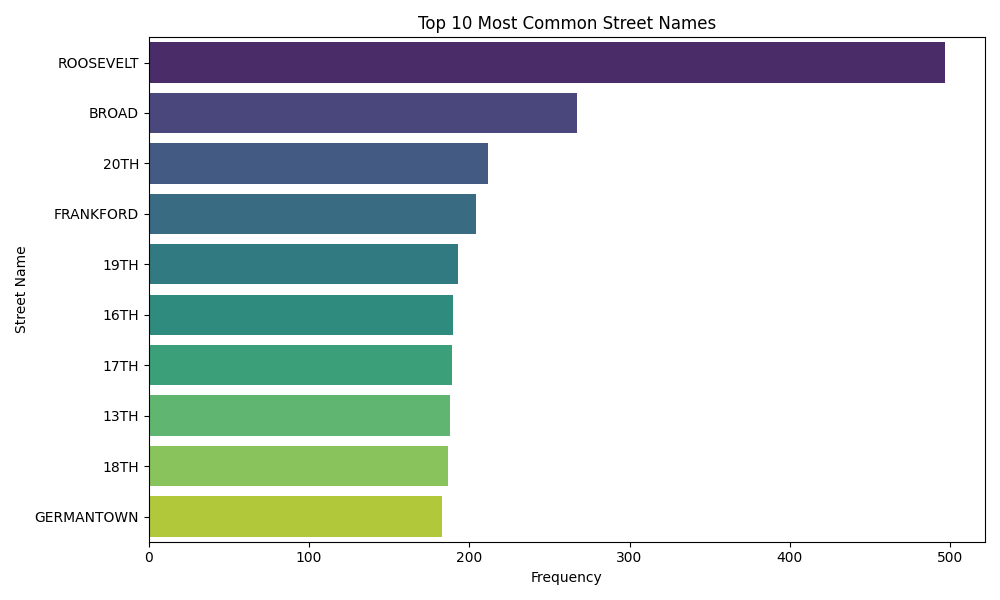
- Column used: ST\_NAME

- Functions Used: value\_counts(), head(), and visualization using seaborn.barplot()

- The value\_counts() function helps in calculating the frequency of each unique street name, while head(10) retrieves the top 10 most frequently occurring names.

Analysis Results: The analysis revealed the top 10 most commonly used street names in the dataset. These include frequently used generic names like "MAIN", "FIRST", and "PINE", indicating that these roads span multiple locations or districts. These high counts may suggest either the importance of these streets in the city’s structure or redundancy in naming conventions.

Visualization: A horizontal bar chart was used to present the top 10 street names based on their frequency. The use of a horizontal layout improved readability, especially for long street names. The viridis color palette enhanced visual appeal while maintaining clarity.



* Objective 2: What are the top 10 longest individual road segments in the city?

Introduction: Identifying the longest individual street segments is crucial for understanding road infrastructure. Longer segments may belong to highways, arterial roads, or major urban connectors.

General Description: The dataset contains a LENGTH column representing the physical length of each road segment. The associated ST\_NAME helps identify which street the segment belongs to.

Specific Requirements, Functions and Formulas:

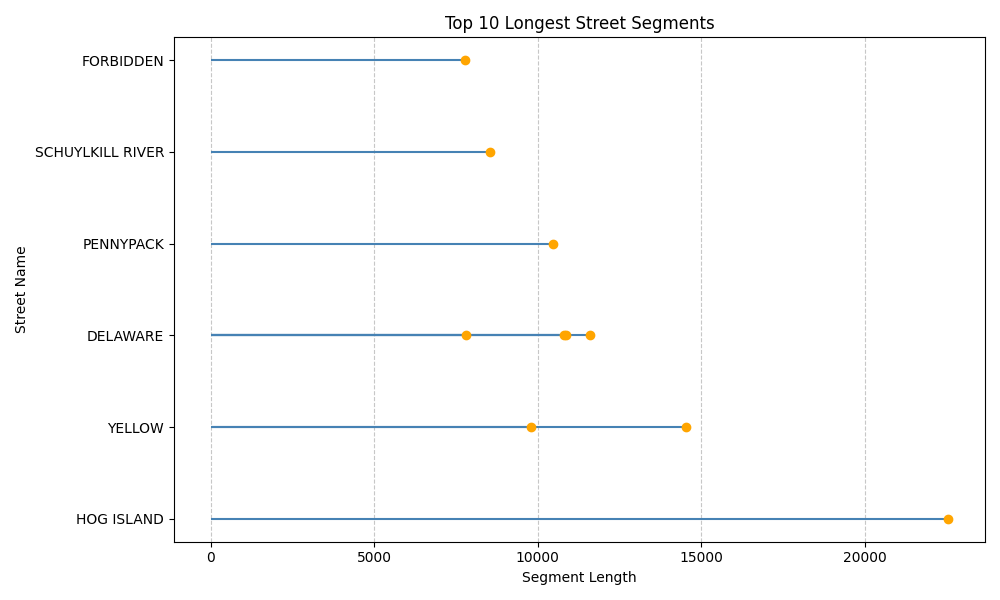
- Columns used: ST\_NAME, LENGTH

- Functions Used: dropna(), sort\_values(), head(), to\_numpy()

- Segments were sorted in descending order based on length, and the top 10 were selected. NumPy functions like max(), min(), and mean() were used to summarize the statistics.

Analysis Results: The top 10 longest segments were significantly longer than the average segment, with the maximum extending far beyond the mean. This suggests that a few exceptionally long roads may form the structural backbone of the city. These long segments are likely used for major commutes and might require higher priority in city maintenance and traffic management.

Visualization: A lollipop chart was employed, where each street name was aligned against its segment length. The dots represented exact lengths while horizontal lines showed their span. This approach offers a clean and modern alternative to bar charts, enhancing interpretability and drawing attention to the top contenders.



* Objective 3: Are longer streets assigned to specific responsible authorities?

Introduction: City infrastructure is typically managed by various government departments or agencies. This objective evaluates whether certain authorities are consistently assigned longer road segments.

General Description: The dataset includes a RESPONSIBL column indicating the department or authority responsible for the segment, and a LENGTH column representing the size of each segment.

Specific Requirements, Functions and Formulas:

- Columns used: RESPONSIBL, LENGTH

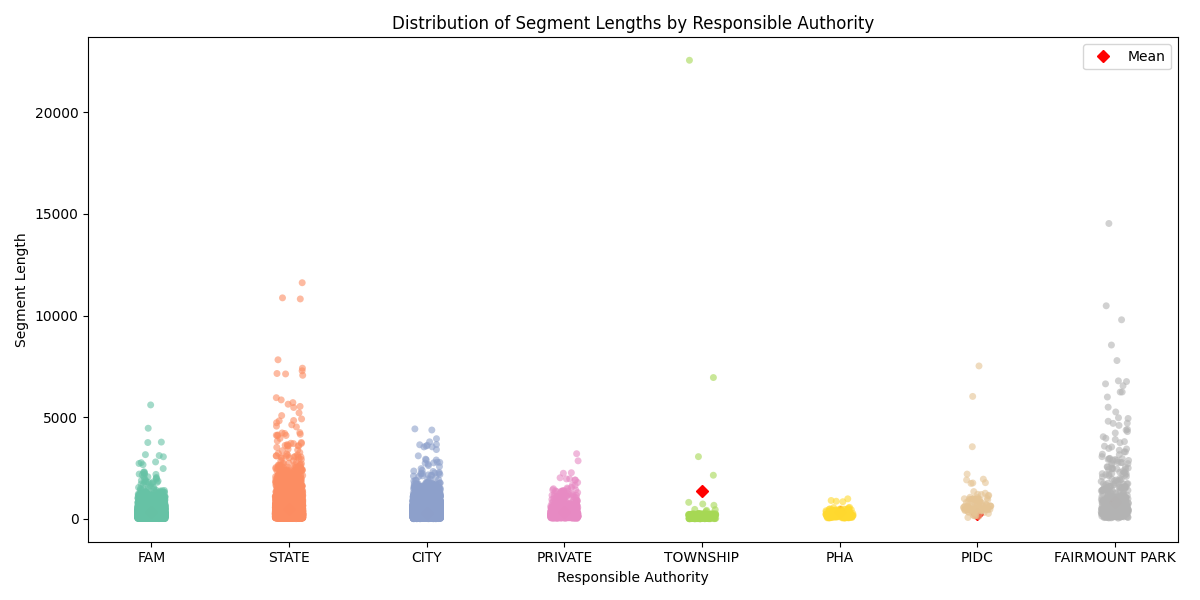
- Functions Used: dropna(), value\_counts(), groupby(), mean(), describe()

- Only the top 8 responsible authorities (by frequency) were included to maintain visual clarity and focus on major stakeholders.

- The stripplot() function from Seaborn with jitter was used to show distribution, and red diamonds represented mean values for comparison.

Analysis Results: The analysis revealed variation in the lengths of road segments managed by different authorities. Some agencies predominantly handled shorter local roads, while others managed significantly longer stretches, likely highways or expressways. This reflects role differentiation among departments, such as municipal bodies managing city roads and regional authorities overseeing intercity routes..

Visualization: A strip plot with mean overlay was selected for its ability to display both the distribution of values and their averages without aggregating the data into bins. This visual made it easy to compare agencies and understand how road lengths varied under each authority’s jurisdiction.



* Objective 4: What is the count of one-way vs both-way streets?

Introduction: Understanding the directionality of streets is essential for traffic modeling, navigation system design, and city planning. One-way streets may improve flow but require careful coordination.

General Description: The ONEWAY column in the dataset records the directionality of the street. Common values include B(both directions), T(toward), and TF(toward/from), representing different traffic flows.

Specific Requirements, Functions and Formulas:

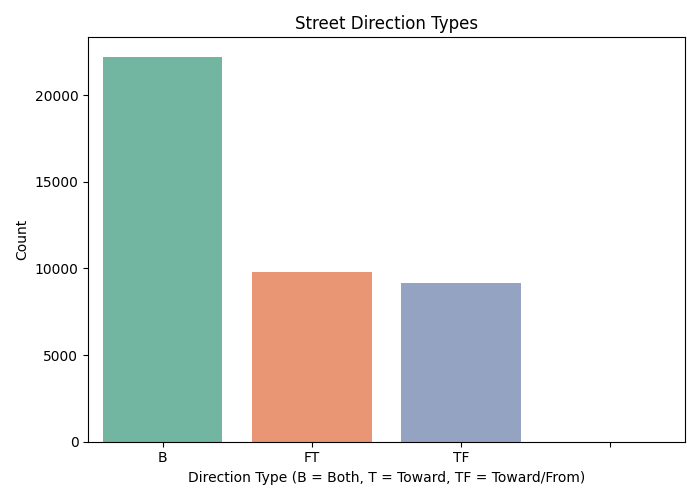
- Column used: ONEWAY

- Functions Used: value\_counts(), and visualization using barplot()

- The data was categorized based on direction type and counted using the value\_counts() method.

Analysis Results: The analysis indicated that the majority of streets in the dataset support two-way traffic. However, there are a notable number of one-way segments, likely found in high-traffic areas, business zones, or narrower streets. The proportion of each category aids urban planners in optimizing road usage and managing congestion.

Visualization: A vertical bar chart was used to represent the count of each direction type. The Set2 color palette was applied for distinction and clarity. Labels and titles ensured that the direction codes were clearly interpreted.



* Objective 5: Are there any correlations between length, shape length, and street class?

Introduction: Correlation analysis is essential to understand how variables are related. This objective examines the relationship between road segment length, its geometric shape length, and its classification.

General Description: The dataset includes three relevant columns:

- LENGTH: Represents the official segment length.

- Shape\_Length: Captures the geometric (possibly GIS-calculated) length.

- CLASS: Refers to the street classification code.

Specific Requirements, Functions and Formulas:

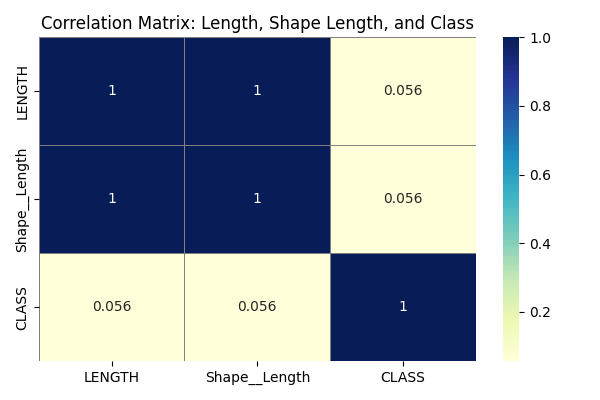
- Columns used: LENGTH, Shape\_\_Length, CLASS

- Functions Used: dropna(), describe(), corr(), and heatmap()

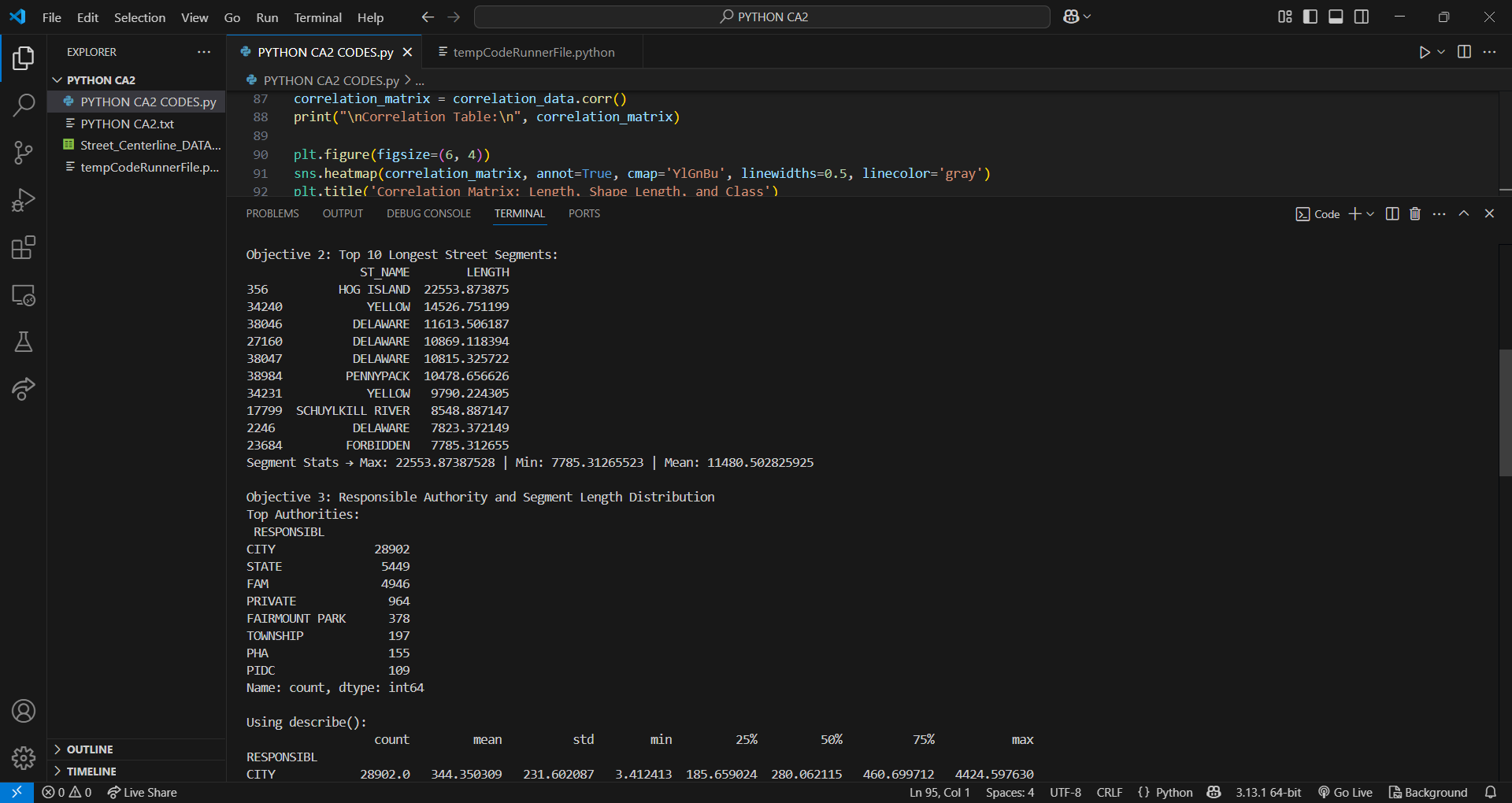
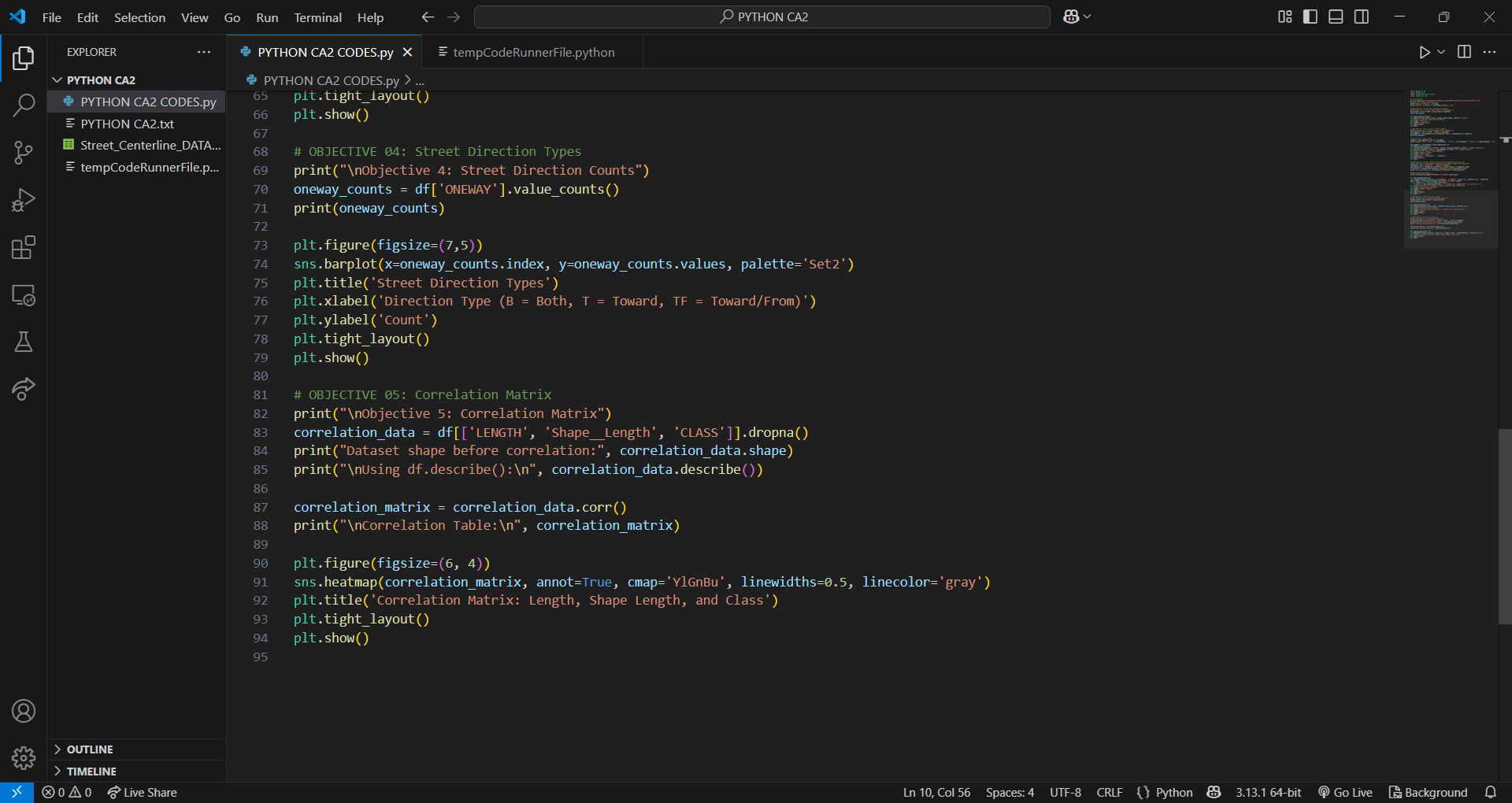
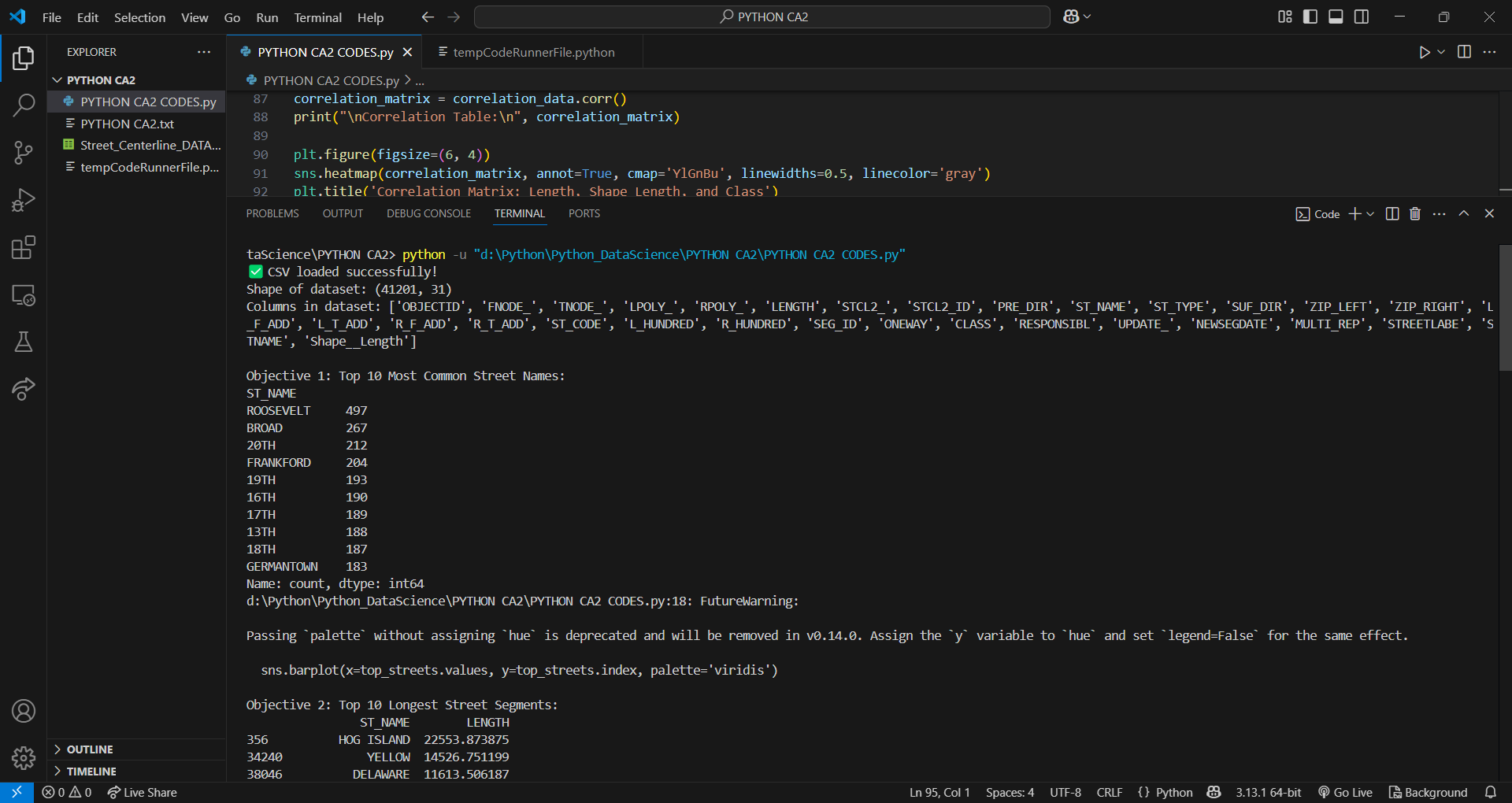
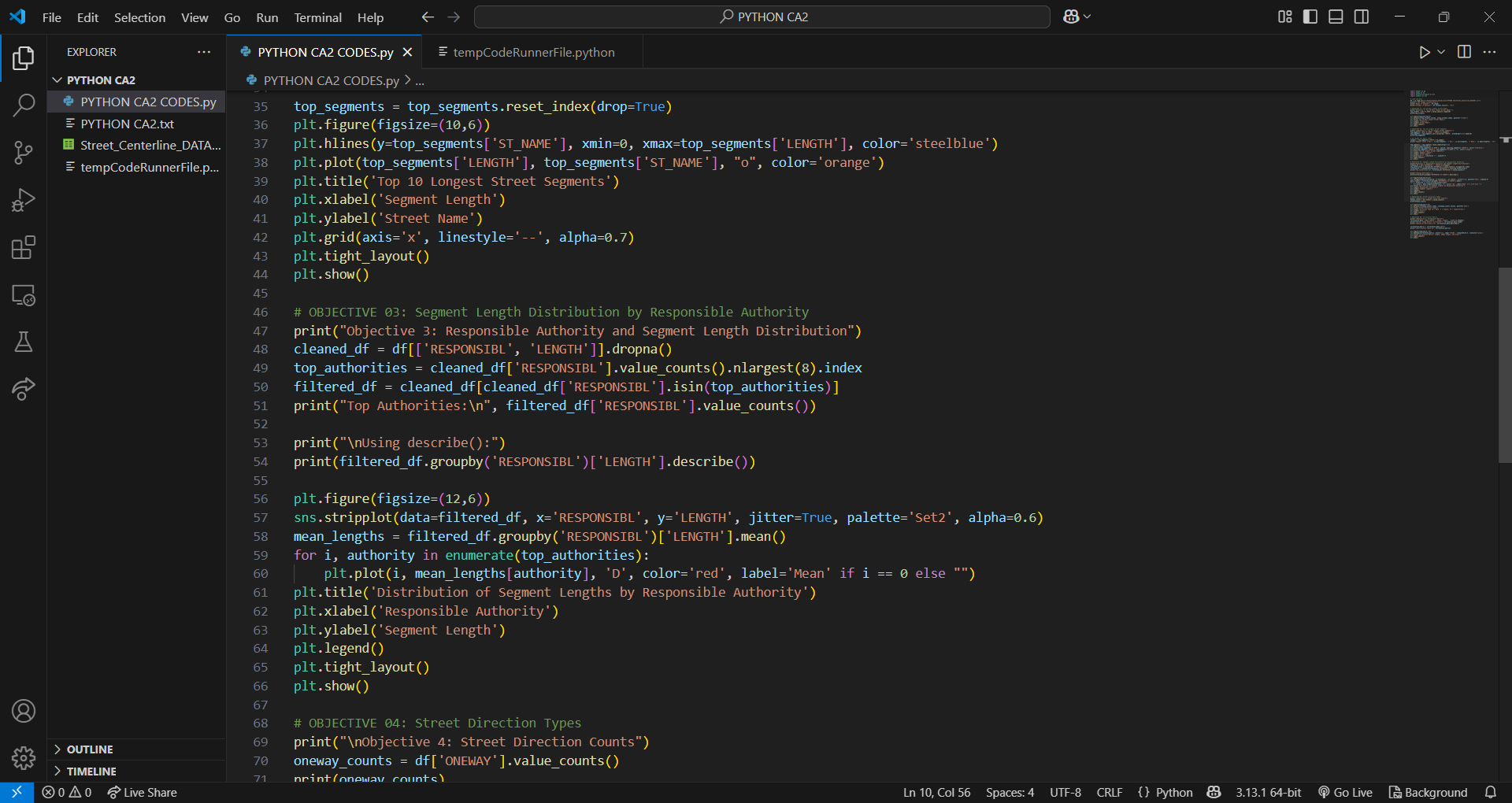
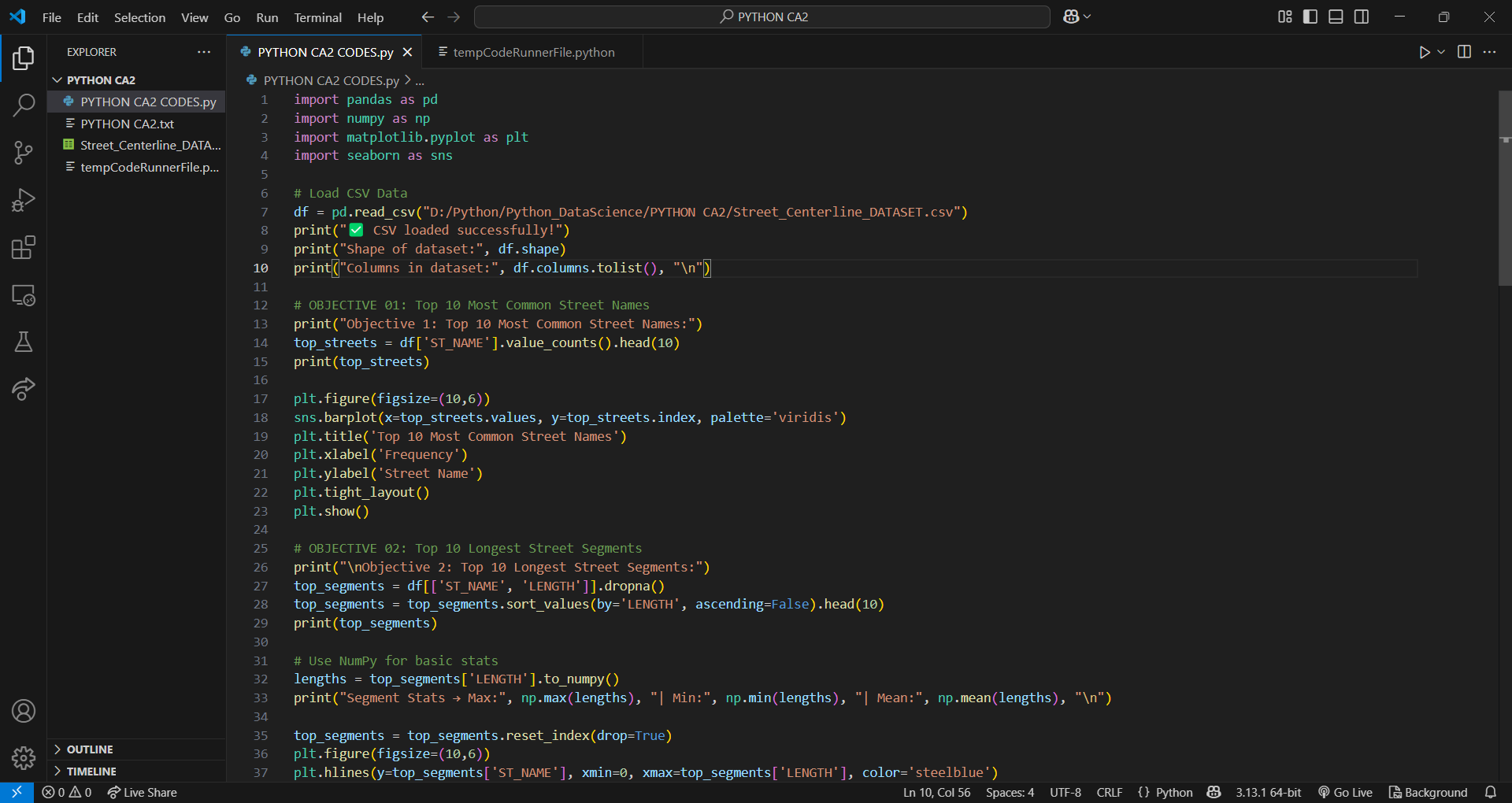
- The corr() function calculates pairwise Pearson correlation, and the matrix is visualized using Seaborn’s heatmap.

Analysis Results: The correlation matrix showed a strong positive relationship between LENGTH and Shape\_Length, which was expected as both represent similar measurements. The CLASS variable had a weaker correlation, indicating that classification codes may consider other factors beyond just physical size.

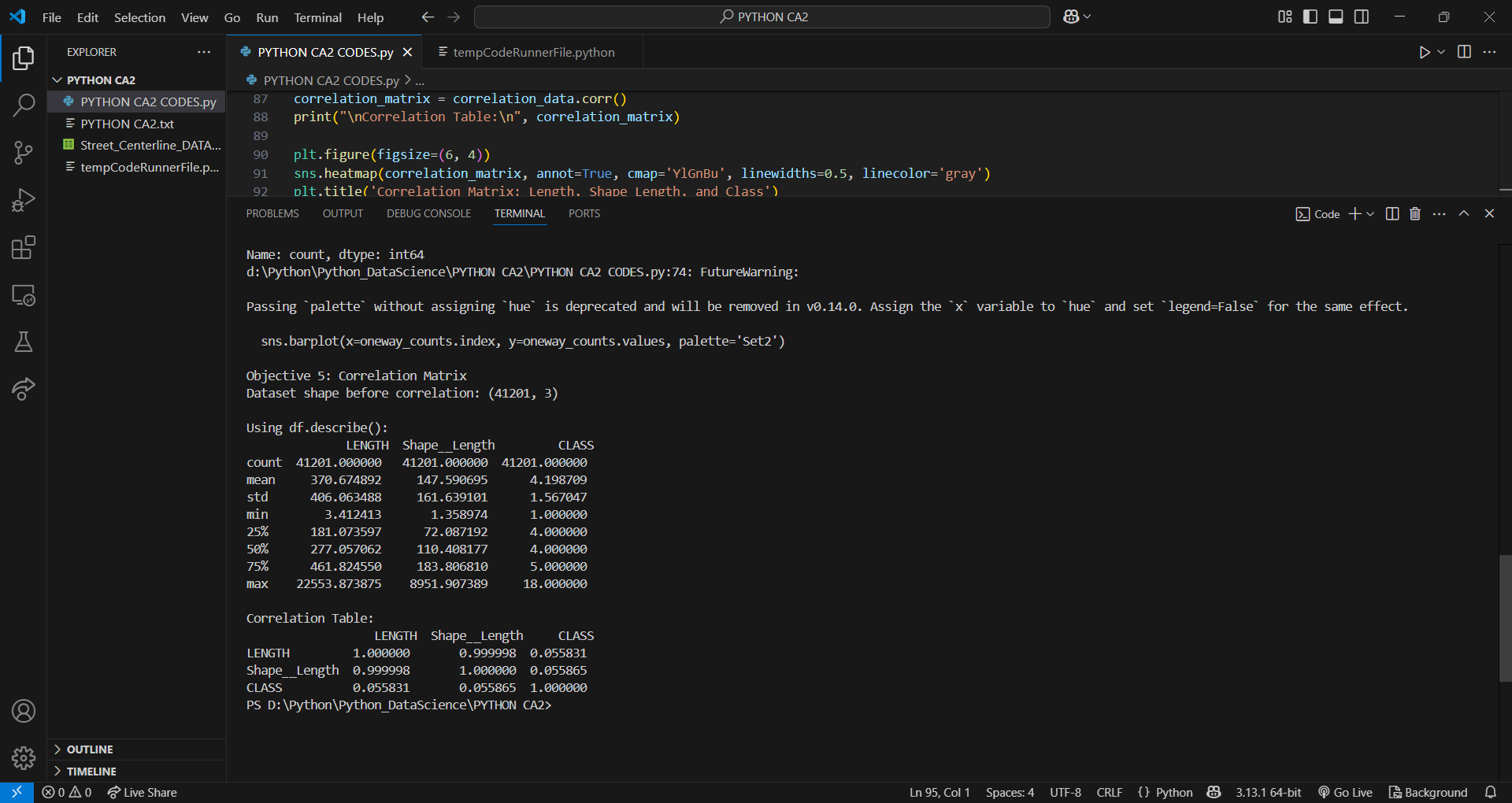
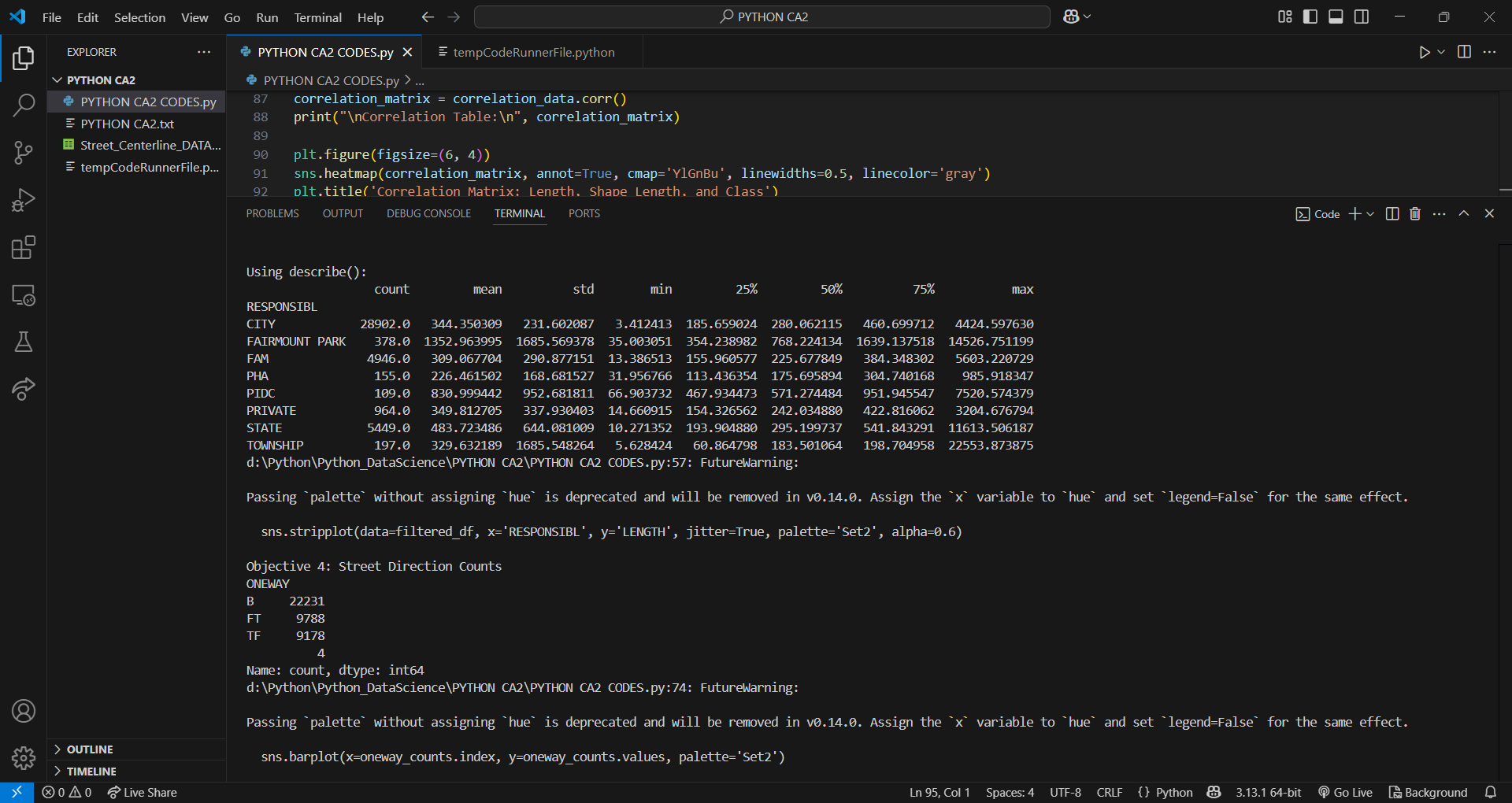
Visualization: A heatmap was chosen to present the correlation matrix. Annotated values and a diverging color scale (`YlGnBu`) made it easy to distinguish positive and negative correlations. This visual succinctly captured the interrelationships among the numerical features.



**CODE**



**6. CONCLUSION**



The Exploratory Data Analysis (EDA) of the Street Centerline dataset provided insightful revelations about the urban road structure, naming patterns, segment distributions, and administrative responsibilities. By breaking down the dataset through five specific objectives, we were able to identify key features such as the most frequently occurring street names, the longest road segments, and their corresponding managing authorities.

Furthermore, the directional flow of streets was analyzed, revealing the balance between one-way and two-way traffic infrastructure. A correlation matrix added depth to our analysis, highlighting how physical dimensions like length and shape length are interrelated, while also evaluating their connection with the classification of roads.

The visualizations used throughout the report helped convert raw data into easily interpretable formats, enhancing both comprehension and presentation of insights. Overall, this EDA not only laid the groundwork for data-driven decision-making in city planning but also highlighted the richness and complexity of urban road networks.

**7. FUTURE SCOPE**

While the current analysis covered foundational aspects of the dataset, several opportunities exist for deeper exploration in the future:

1. Temporal Trends: If timestamp data becomes available, seasonal or annual trends in road construction or modifications could be analyzed.

2. Geospatial Mapping: Incorporating GIS tools can provide map-based visualizations for identifying congested or underdeveloped areas.

3. Traffic Volume Analysis: By integrating traffic data, we can analyze which streets bear the most load and optimize them for efficiency.

4. Predictive Modeling: Machine learning models could be trained to predict street maintenance needs based on length, usage, and responsible authority.

5. Road Safety Metrics: Linking accident data to street segments could help in identifying high-risk roads and areas needing redesign.

6. Cross-City Comparisons: Analyzing similar datasets from other cities can help in benchmarking performance and learning best practices.

This future scope enables city planners, traffic engineers, and policymakers to use data as a tool for smarter and safer urban infrastructure development.

**8. REFERENCES**

* <https://seaborn.pydata.org/>
* <https://www.w3schools.com/python/matplotlib_intro.asp>
* <https://matplotlib.org/>
* <https://numpy.org/doc/stable/user/>
* <https://pandas.pydata.org/>