**FINAL REPORT**

**CAPSTONE PROJECT**

**TIME SERIES ANALYSIS**

**OF**

**TRIPS IN COOK COUNTY, ILLINOIS USA**

**FROM 2019 - 2023**

**By:**

Nurul Hakimah Mohd Zaki

Date: 28 June 2024

Table of Contents

**Problem Statement3**

**Analysis3**

**Data4**

**Metrics5**

**Statistical Analysis5**

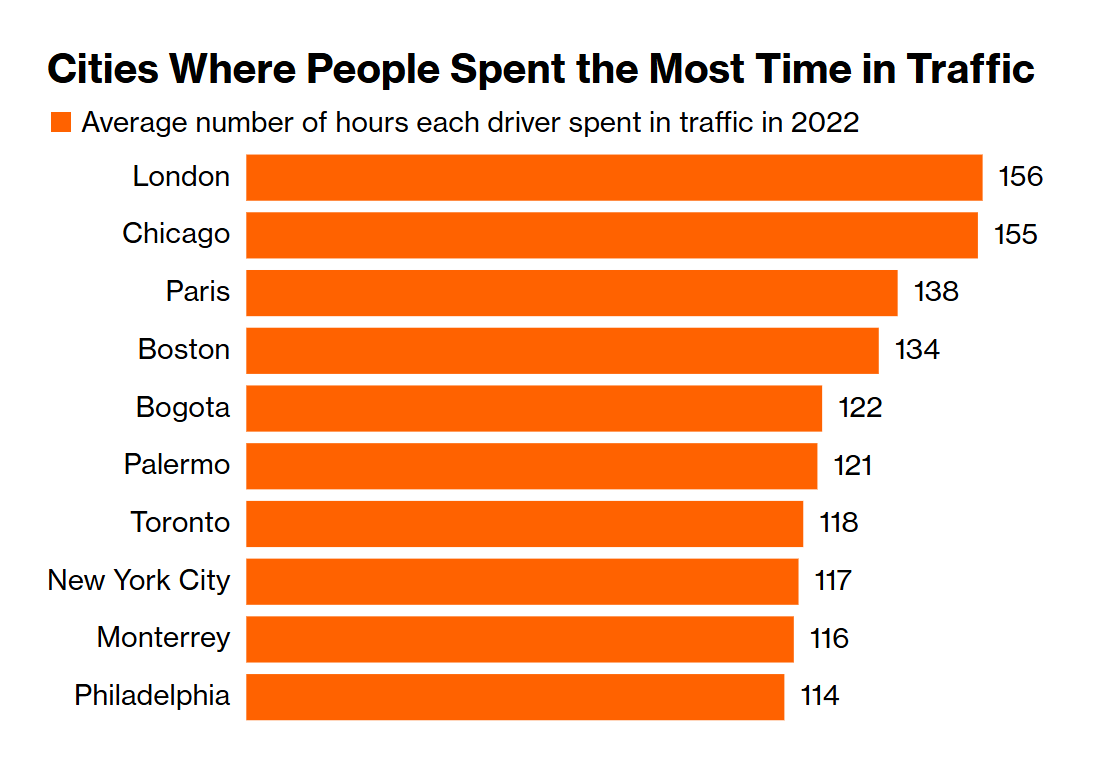
**Risk and Limitations10**

**Proposed Solution10**

**References11**

**PROBLEM STATEMENT**

Traffic congestion in the US has gotten worse over the past year. According to Inrix, the average driver in Chicago spent 155 hours, making it as the second most congested city in the world in 2022. The trend is also increasing rapidly to pre-Covid19 levels. This problem is expected to get worse as US population has steadily increased by 300 million people every year. By modelling the population movements, the pattern can be analyzed to determine the needs for significant changes that require attention by the town planner. Modelling and forecasting were done for Cook County, IL, where Chicago is located. The model was performed on data from Jan 2019 to May 2023. Forecasting was done 1 year ahead until June 2024.



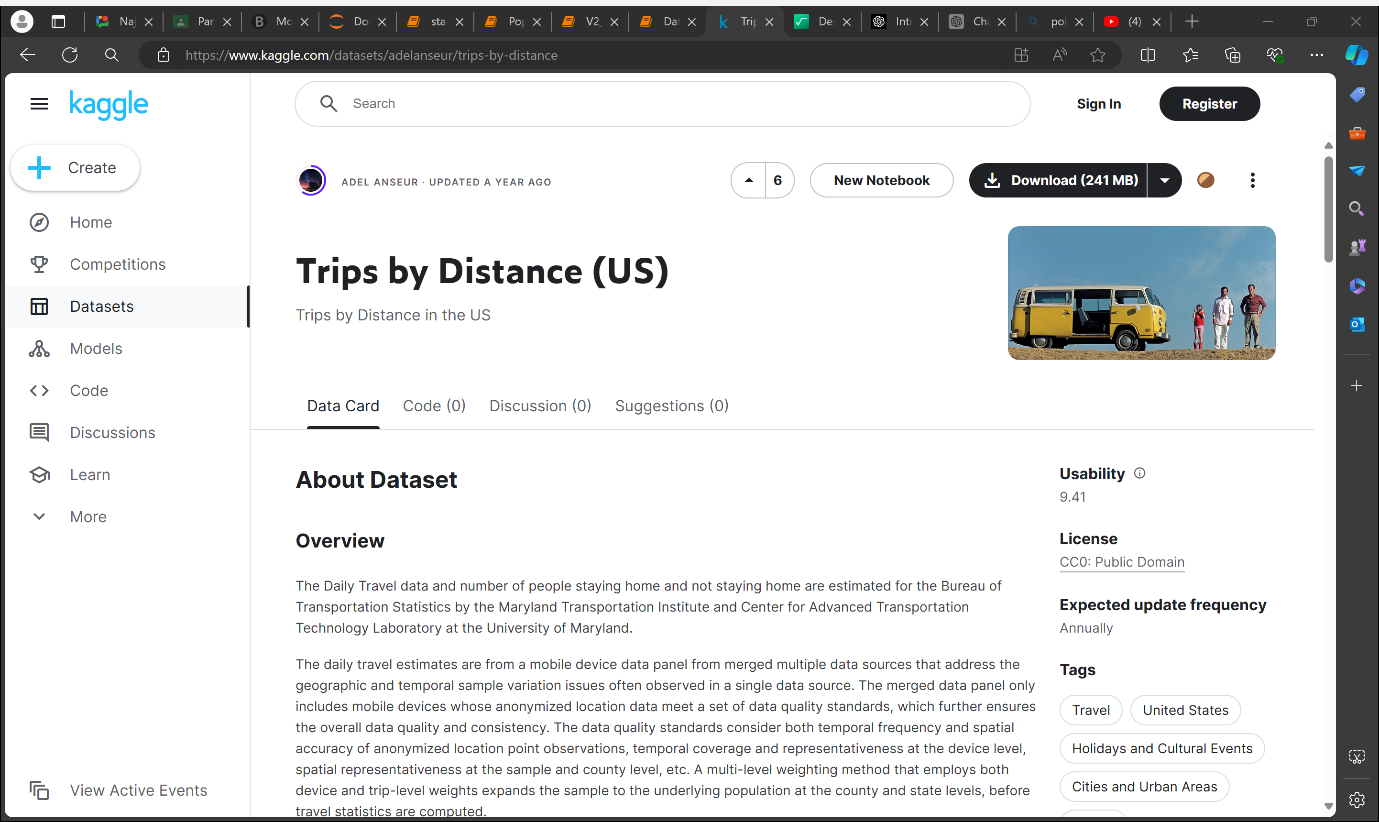
*Figure 1: Most Congested Cities Ranking in 2022 (Inrix,2023)*

**ANALYSIS**

This dataset is analyzed using time series analysis. ARIMA & SARIMAX model were used and compared to get better accuracy.

**DATA**

Data was obtained from Kaggle.com, and Bureau of Transportation Statistics by the Maryland Transportation Institute and Center for Advanced Transportation Technology Laboratory at the University of Maryland.



*Figure 2: Dataset Obtained from Kaggle.com*

The dataset has 22 columns of total number of trips and different frequency of trips taken in different counties in different states from 01 January 2019 until 13 May 2023. This data is recorded daily.

Upon checking on null values, there are 2% of null value, that are summation data in each county and each state. The data has combined total daily value into state level, and leaving county details as null. It is also true in country level, which it combines the values for state level and leaving states details as null. Hence, it is safe to drop null values as they are redundant.

Number of Trips column was chosen as the main dataset because I want to capture the whole

population movements, regardless of the distance.

After loading the time series data, it is apparent visually that the series is not stationary.

This condition determines the approaches needed to move further.

**METRICS**

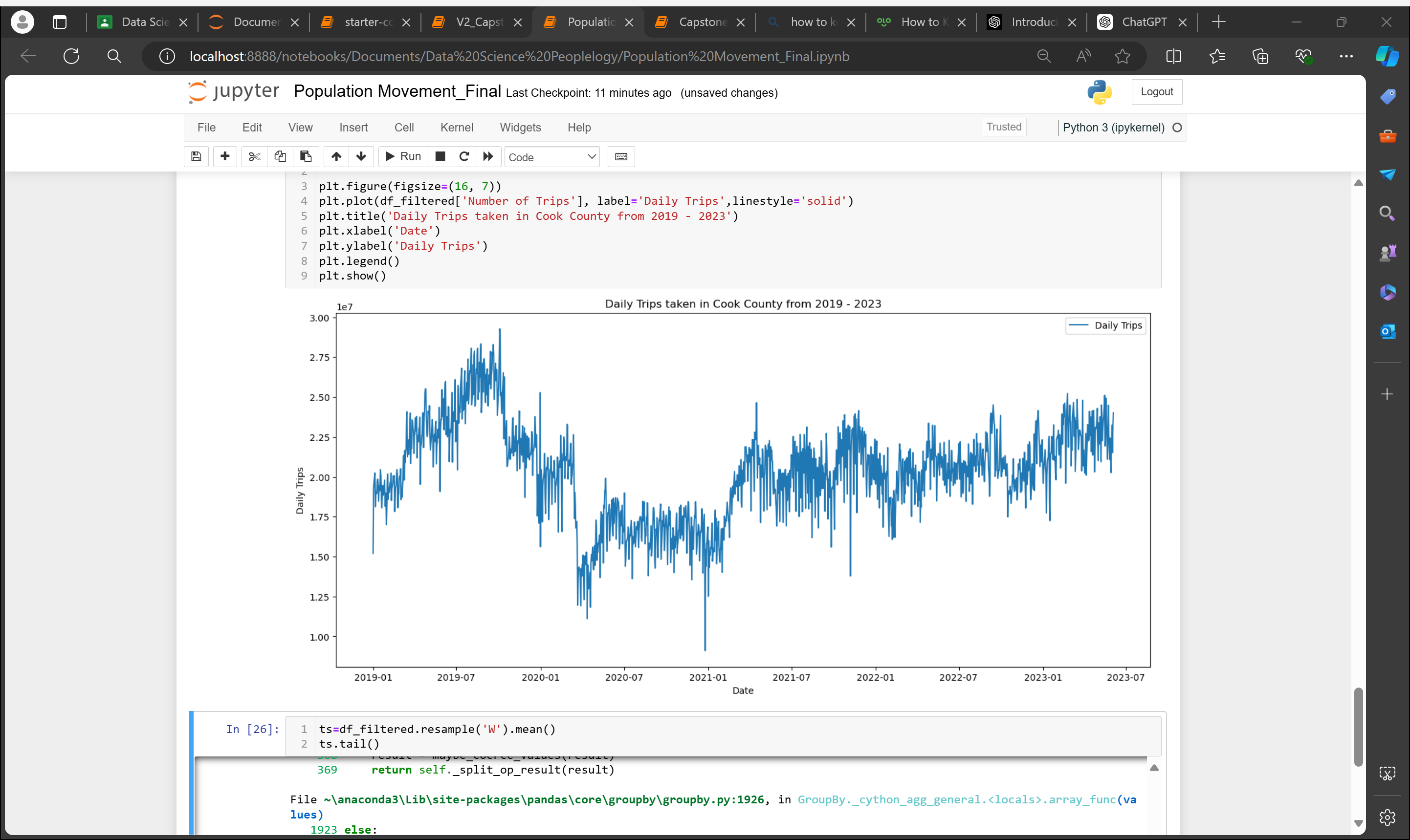
The stationarity is determined using Augmented Dicket Fuller (ADF) test using p-value.

The model accuracy is measured by RMSE.

**STATISTICAL ANALYSIS**

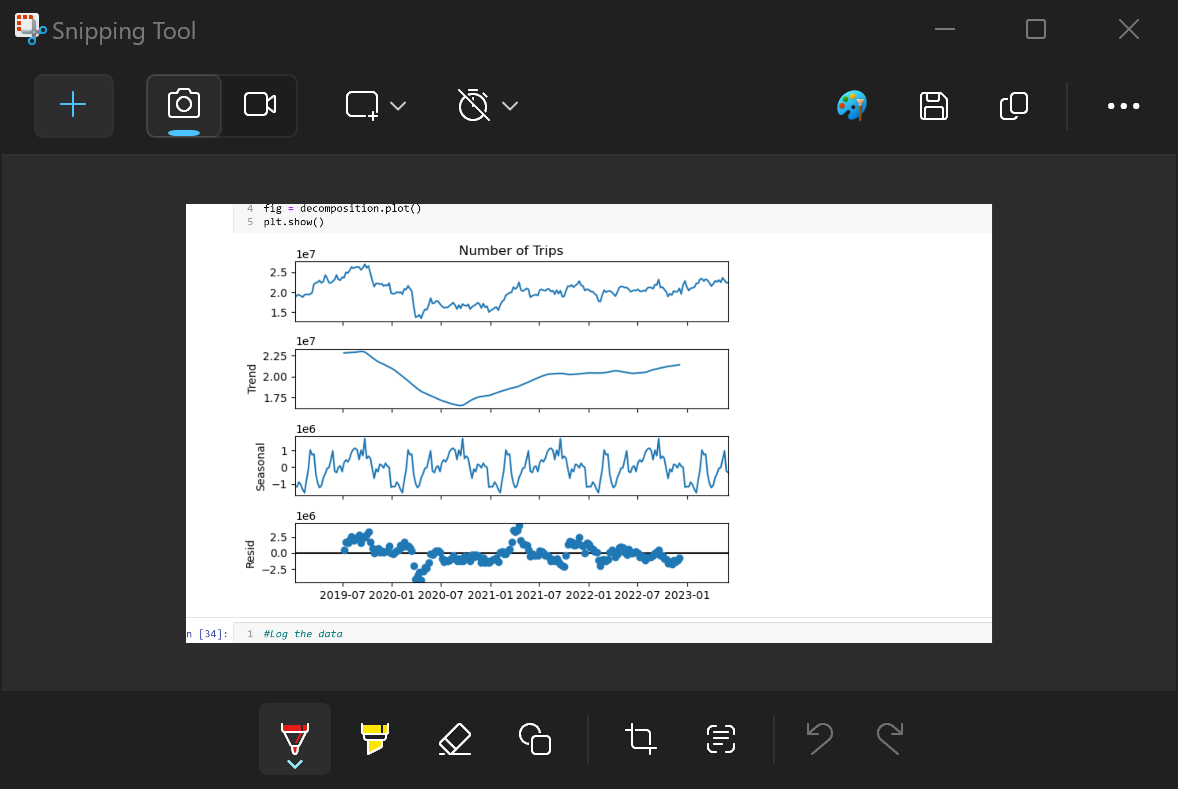
In the beginning, a routine data cleaning, and visualization is done to understand the whole data.

Then, Number of Trips column is chosen to be the main dataset and analyzed. Time series data is plotted. Data is not stationary based on visual observation.



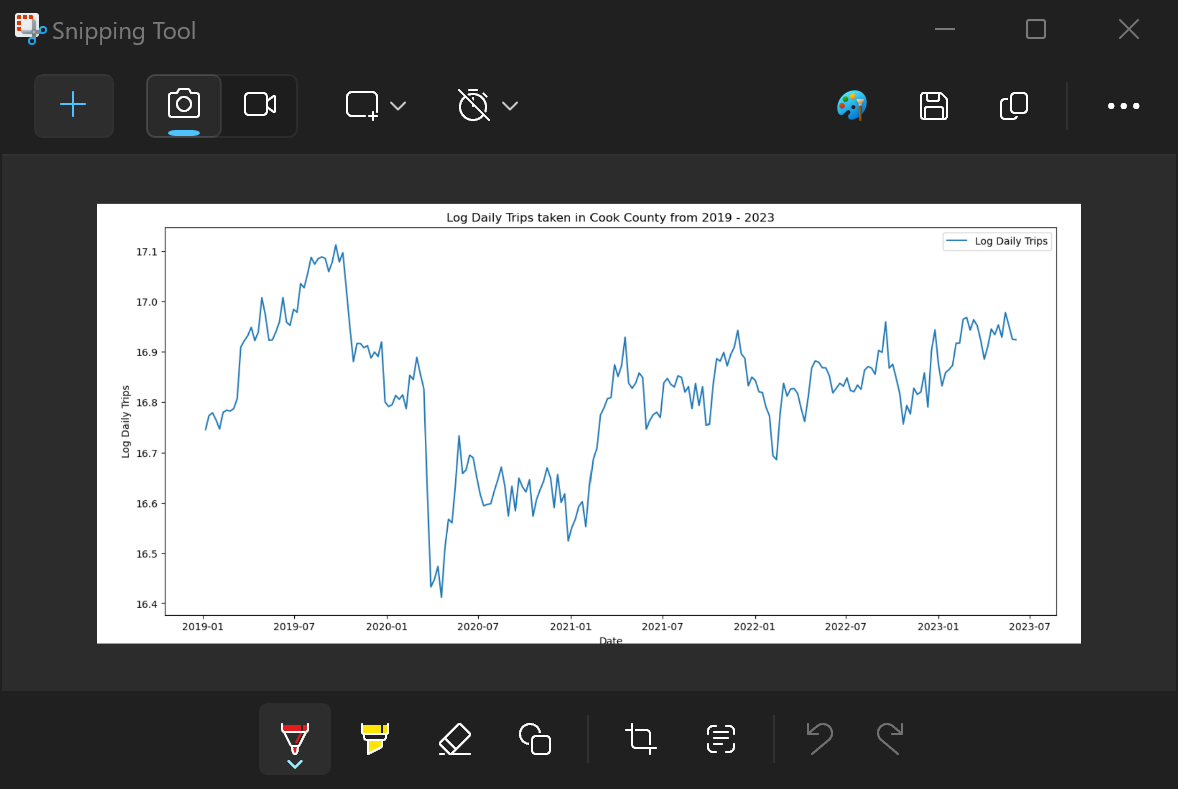
*Figure 3: Daily Trips in Cook County, IL from 2019 – 2023*

Dataset decomposition is done to check thoroughly on the trend, seasonality and residual components. It is apparent that all three components exist and data needs to be transformed, if it is not stationary.



*Figure 4: Dataset Decomposition showing trend, seasonality and residual error.*

The first step of transformation is weekly resampling and mean value is used throughout. Due to high magnitude and high residual in the dataset, log is necessarily needed. This is confirmed in the first round of modelling and forecasting without log, all RMSEs are extremely high. The dataset was reevaluated and found that log method works best.



*Figure 5: Transformed Dataset using Resampling, Log & Differencing Method.*

ADF test for weekly-resampled log data confirmed that data is not stationary, with p-value = 0.158.

Hence, 9 transformation methods is performed & found the best method that gives the lowest p-value in ADF test.

|  |  |
| --- | --- |
| **Method** | **ADF p-value** |
| Resampled weekly | 1.52E-01 |
| Differencing | 4.03E-28 |
| Double Differencing | 3.49E-09 |
| Log | 1.58E-01 |
| SQRT | 1.78E-01 |
| CBRT | 1.72E-01 |
| Log & SQRT | 1.57E-01 |
| Log & CBRT | 1.57E-01 |
| **Resampling 1 week & Log & Differencing** | **8.44E-28** |

*Table 1: ADF p-value with corresponding transformation method*.

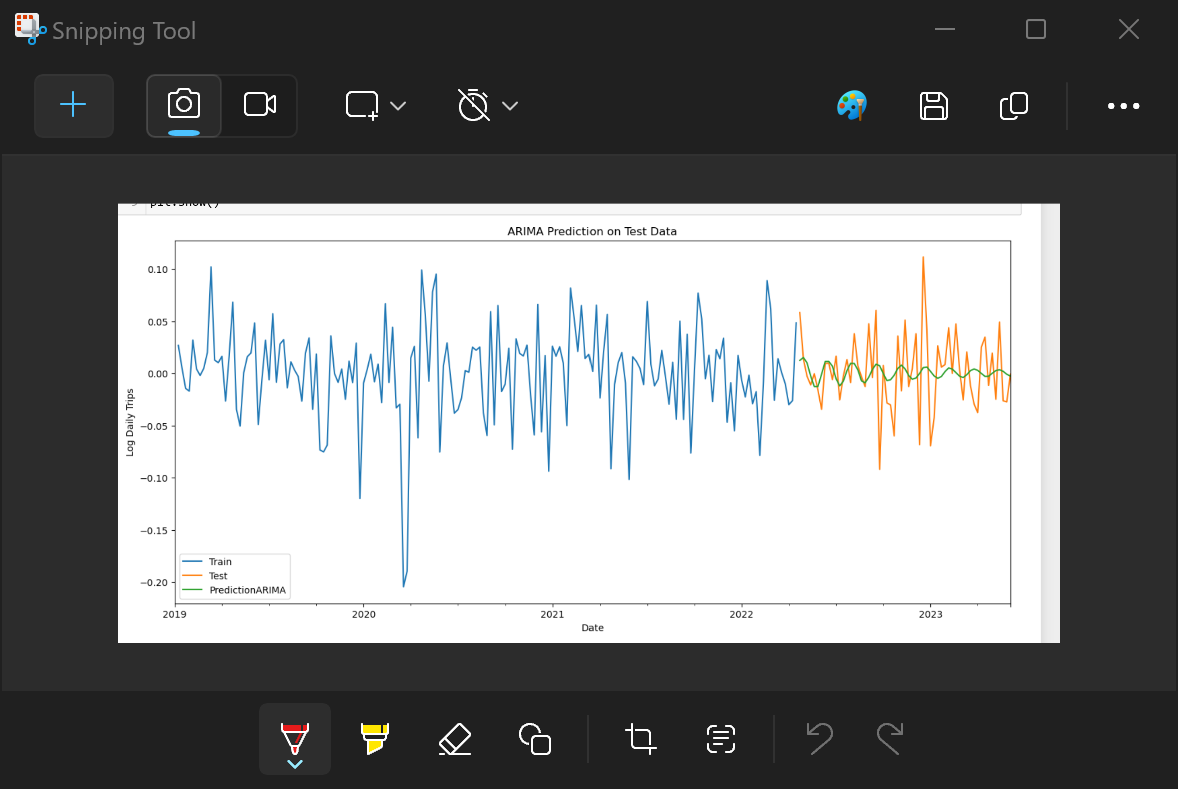
Train and test data is defined with 75:25 ratio.

ARIMA method is chosen because it is the most common model and has differencing component.

For loop was performed to find the best hyperparameter combination with the lowest RMSE. The best parameters for ARIMA is p=7, d=1, q=7.

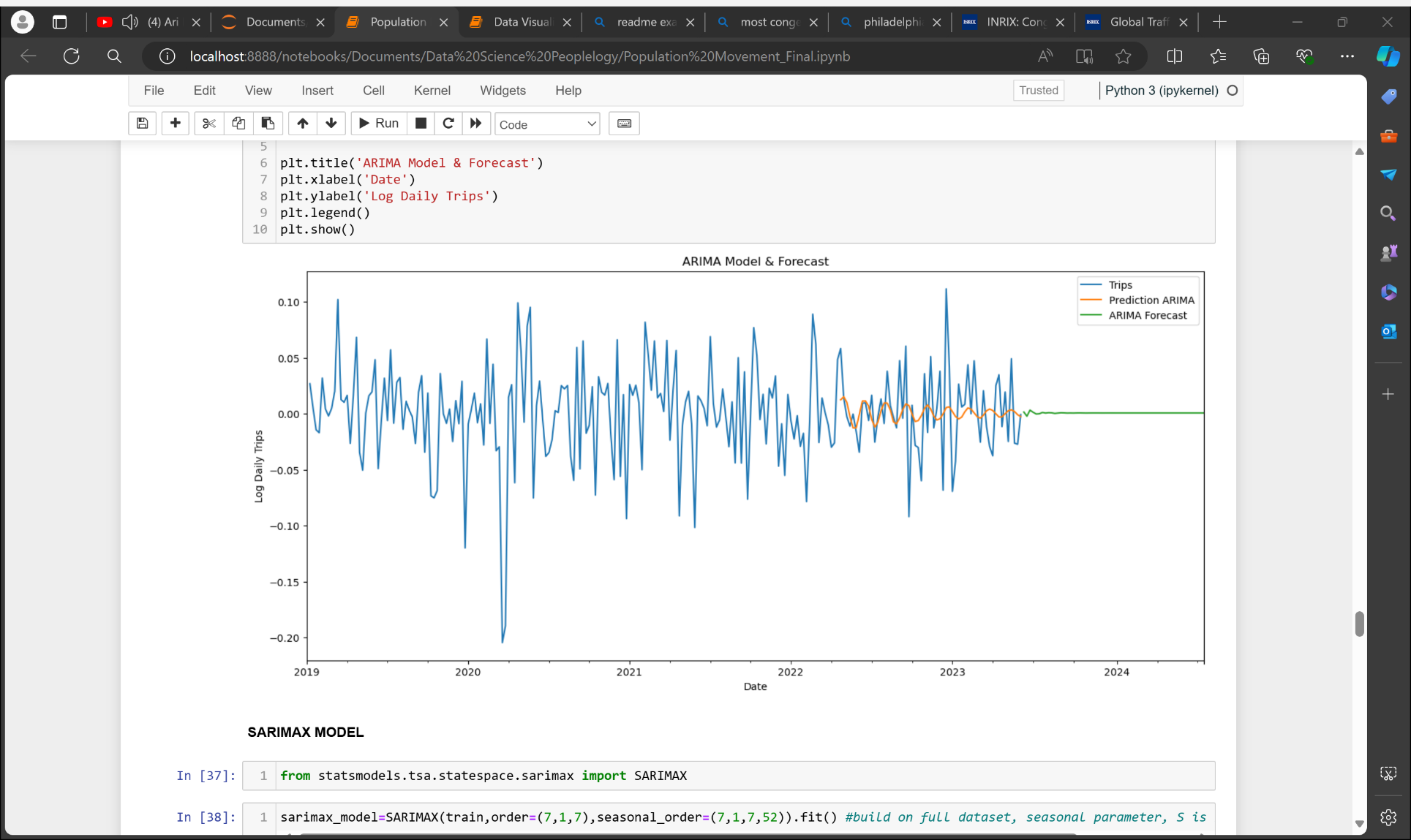
The first step is to use training data and model the prediction on test data.

The RSME for ARIMA Prediction on test data is 0.034. This shows that ARIMA is a good model to make prediction on top of test data.



*Figure 6: Prediction using ARIMA model on test data.*

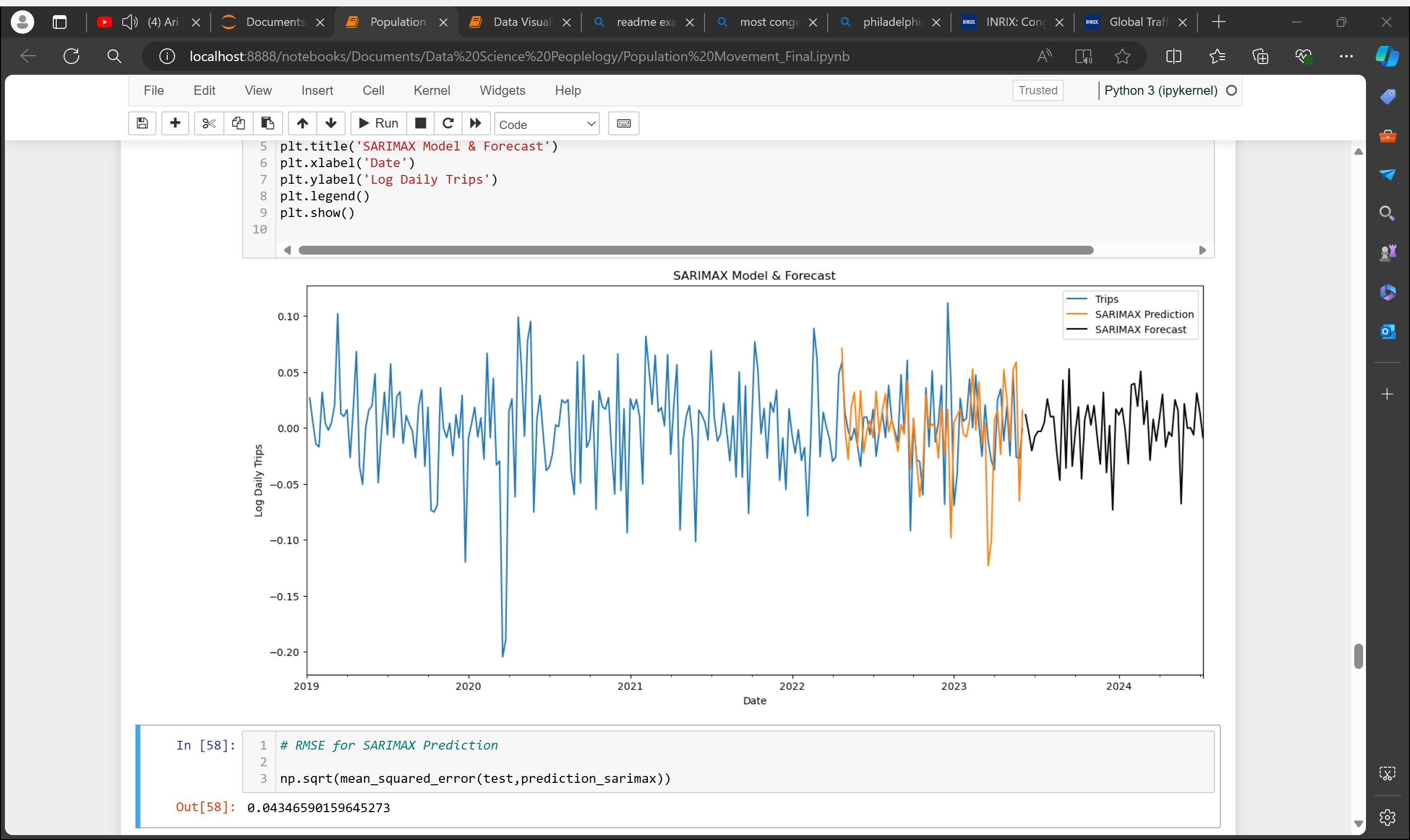
Next is to perform forecasting for 1 year ahead. The error is considered acceptable.



*Figure 7: Forecast using ARIMA model.*

In ARIMA model, forecast trend does not seem to capture seasonality. This could affect badly to the forecast because the original dataset has so much seasonality component.

Hence, SARIMAX model is introduced to improve existing model. Same train : test ratio, p, d, and q parameters were used, in addition to s=52 to account for 52 week cycle in a year. The RSMEs for SARIMAX prediction and forecast also showing 0.035 and 0.043 respectively, indicating good model.

****

*Figure 8: SARIMAX Model and forecast.*

In order to analyze the forecast data to propose relevant solutions, a linear regression line is plotted together. The increasing trend can be seen clearly visually and percentage increase can be calculated using forecasted data.

****

*Figure 9: Increasing Trend in SARIMAX Forecast.*

The linear regression line is used exclusively to show trend line, no analytics were done using the data due to huge error.

**RISK & LIMITATIONS**

This project commenced before I learnt Time Series Analysis formally, hence I have to do a lot of self-study & watching some tutorials throughout the project. The learning curve is steep yet informative.

I found another method which cross validation time series but given limited time and knowledge, I put that for future visit.

Running the model takes more than 30 minutes, depending on laptop/computers used, so time planning is also crucial.

Proposing solutions can be done on writing, no product can be developed with limited time and knowledge.

**PROPOSED SOLUTION**

Notice that there is significant upward trend before a decrease in trend end of 2019, due to COVID-19 spread. It seems like the population will move and reach the peak quite soon. Keep in mind that the number of population is also increasing, so the peak could be higher and could be reached faster.

An increasing trend is observed in SARIMAX forecast. This should raise an alarm to all local and federal authority in planning better transportation and traffic management system. Clearly in 2022, congestion was not managed really well, but changes need to be done to avoid it from happening again, or becoming worse.

|  |  |
| --- | --- |
| **Year** | **Average Daily Trips** |
| June 2023 | 22.3 Million |
| July 2024 | 24.0 Million |
| +1.72 Million (+7.7%) | |

*Table 2: Increasing Observation in SARIMAX Forecast*

In this data-oriented era, authorities should adapt, collect and apply to create intelligent transportation system.

Furthermore, with all the data the government has, data analytics can be integrated on mobility, traffic, traffic signals parking & population movement.

City planners and engineers should have stronger collaboration in making data-based decisions.

Necessary development such as increasing public transport coverage and frequencies, adding more highways or smart traffic lights shall be evaluated for future needs.

**REFERENCES:**

1. *Trips by Distance (US)*. (2023, August 3).

Kaggle. https://www.kaggle.com/datasets/adelanseur/trips-by-distance

1. Chandler-Wilde, H. (2023, January 10). *Bloomberg - These are the world’s most congested cities*. www.bloomberg.com. Retrieved May 31, 2024, from https://www.bloomberg.com/news/articles/2023-01-10/these-are-the-world-s-most-congested-cities
2. INRIX 2022 Global Traffic Scorecard: London Tops List as Most Congested City, U.S. Cities Inch Closer. (2023). In *https://inrix.com*. Retrieved May 25, 2024, from https://inrix.com/press-releases/2022-global-traffic-scorecard-uk/
3. Learnerea. (2022, December 8). *Time Series Analysis using Python| ARIMA & SARIMAX Model Implementation | Stationarity Handling* [Video]. YouTube. https://www.youtube.com/watch?v=O5pataOw33Y