Castillo, Anjelica M.
PSMDSSC 103-PSMDS12S1 - Applied Machine Learning

ML TASK 3

Instructions: Implement the requirements in Python and submit both the HTML and IPYNB files. Use Markdown to incorporate the questions below, and provide a PDF file containing answers to some of the questions.

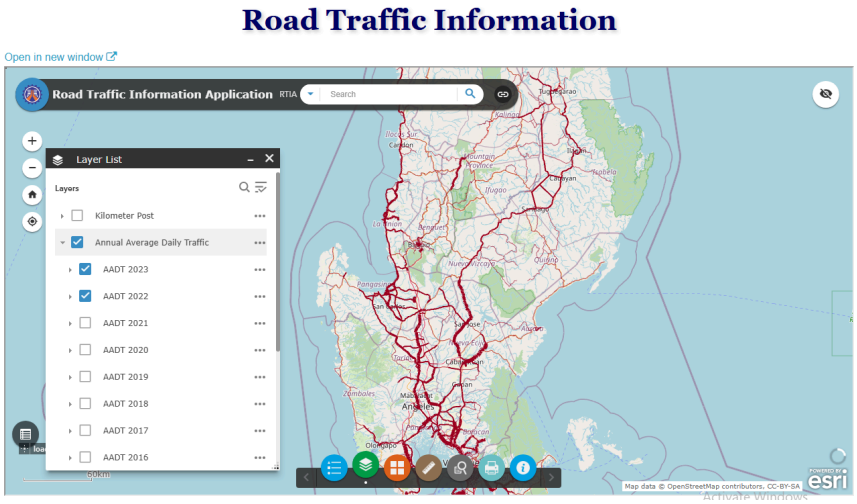
1. Define the problem you are trying to solve using machine learning.

Analyzing traffic flow in NCR in 2023 using Annual Average Daily Traffic data to optimize signal timings, reduce congestion, and improve transportation efficiency.

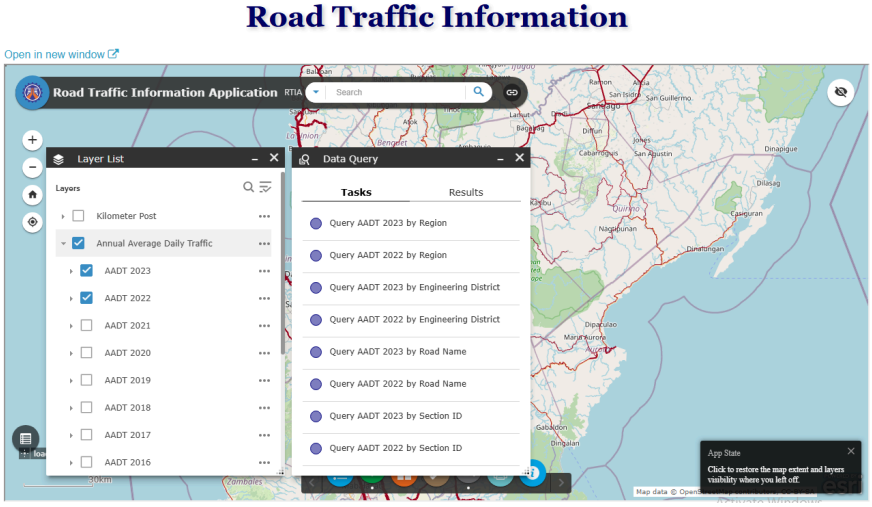
2. Identify the data sources and data collection methodologies.

Data source is: https://www.dpwh.gov.ph/dpwh/gis/rti

1. Go to Layer List and select Annual Average Daily Traffic, check AADT(Annual Average Daily Traffic) 2022 and AADT 2023.



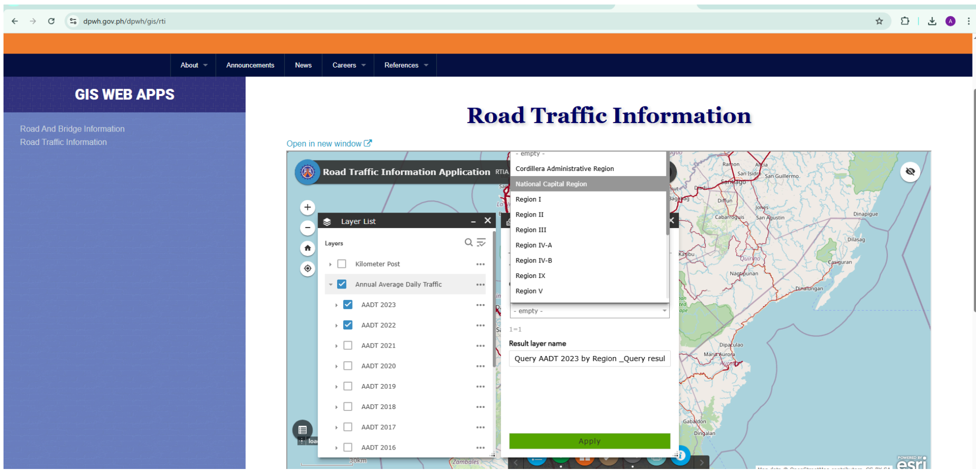
2. Then select Data Query and this will pop-out. Select Query AADT 2023 by Region and Query AADT 2022 by Region.



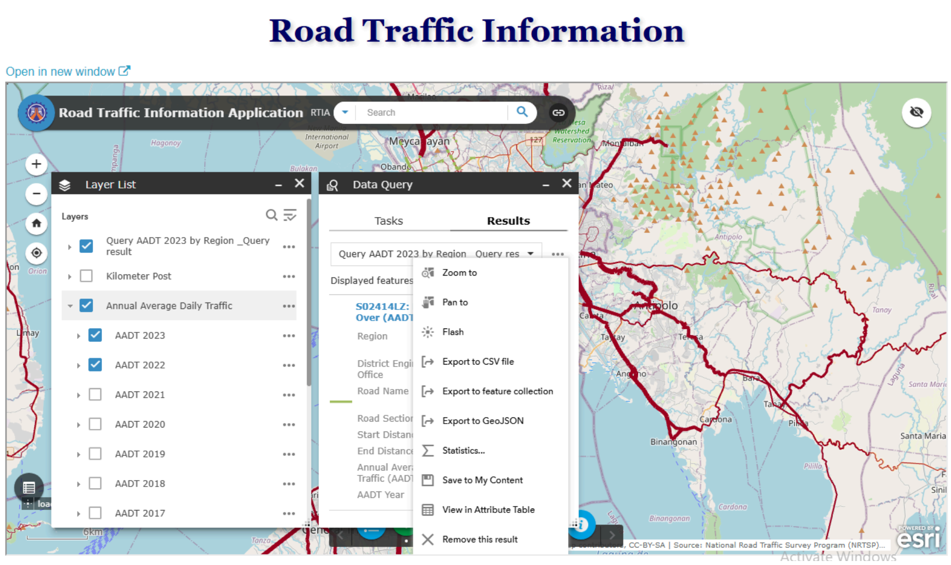
\

Castillo, Anjelica M.
PSMDSSC 103-PSMDS12S1 - Applied Machine Learning

3. Select **National Capital Region** then click Apply.



4. Then export to csv.



You may also access the 2022 and 2023 aadt here.

https://mmda.gov.ph/efoi-mmda/2-uncategorised/3345-freedom-of-information-foi.html?appgw_azwaf_jsc=4GiLsS6UL4au7PM0fMq53aA_dhyepabtW64mpAHrlBI

2. Determine what type of data is needed (structured or unstructured).

Structure of data is structure.

3. Define the features (variables) relevant to the problem.

4. Explain why this data is critical for solving the problem.

This data is critical as it provides insights into the number of vehicles on each road segment. By analyzing this data, one can predict traffic congestion, optimize traffic flow, and improve city planning.

5. Assess the data quality and relevance by verifying whether the dataset is complete, unbiased and represented.

This data is critical as it provides insights into the number of vehicles on each road segment. By analyzing this data, one can predict traffic congestion, optimize traffic flow, plan for infrastructure development, and improve city planning.

10. Given a dataset with multiple features, how would you determine which features are most relevant for your machine learning model? Describe at least two feature selection techniques and their advantages.

- SelectKBest ranks features based on their correlation with AADT and selects the most relevant ones.
- RandomForest provides an importance score, allowing us to see which features contribute the most to predictions.

11. You detected several extreme values in your dataset using the IQR method and Z-score analysis. How would you decide whether to remove, transform, or retain these outliers? Provide examples of when each approach is appropriate.

The decision to remove, transform, or retain outliers depends on their cause and impact on analysis. Outliers should be removed if they result from data entry errors, sensor malfunctions, or temporary anomalies (e.g., traffic spikes due to accidents). Transformation methods (such as log, square root, or Box-Cox transformations) are suitable when outliers distort normality but still contain valuable information, such as consistently high traffic volumes on major roads. Outliers should be retained if they represent natural variations in traffic patterns, as removing them may lead to loss of important insights. A careful analysis of the data context ensures the best approach for maintaining accuracy and relevance in traffic flow predictions.

```
Mean Absolute Error: 11533.855
Mean Squared Error: 315956281.86382484
R-Squared: 0.9659360962713641
```

Castillo, Anjelica M.
PSMDSSC 103-PSMDS12S1 - Applied Machine Learning

12. You created a boxplot for a dataset and observed that one feature has a long right whisker, indicating a positive skew. What does this tell you about the data, and how would you address this issue before training your machine learning model?

This shows that the effect of log transformation on skewness reduction. The original skewness of 1.30 indicates a right-skewed (positively skewed) distribution, meaning a few large values are pulling the tail to the right. After applying log transformation:

- The skewness is reduced to 0.59, showing better symmetry.

- The boxplot of AADT_log shows fewer extreme outliers compared to the original AADT.

- This transformation helps improve the performance of machine learning models by ensuring a more normal-like distribution, leading to better generalization. If further normalization is needed, Box-Cox transformation (which reduced skewness to 0.11) could be a more effective alternative.