1. Report on Data Extraction

1. Introduction:

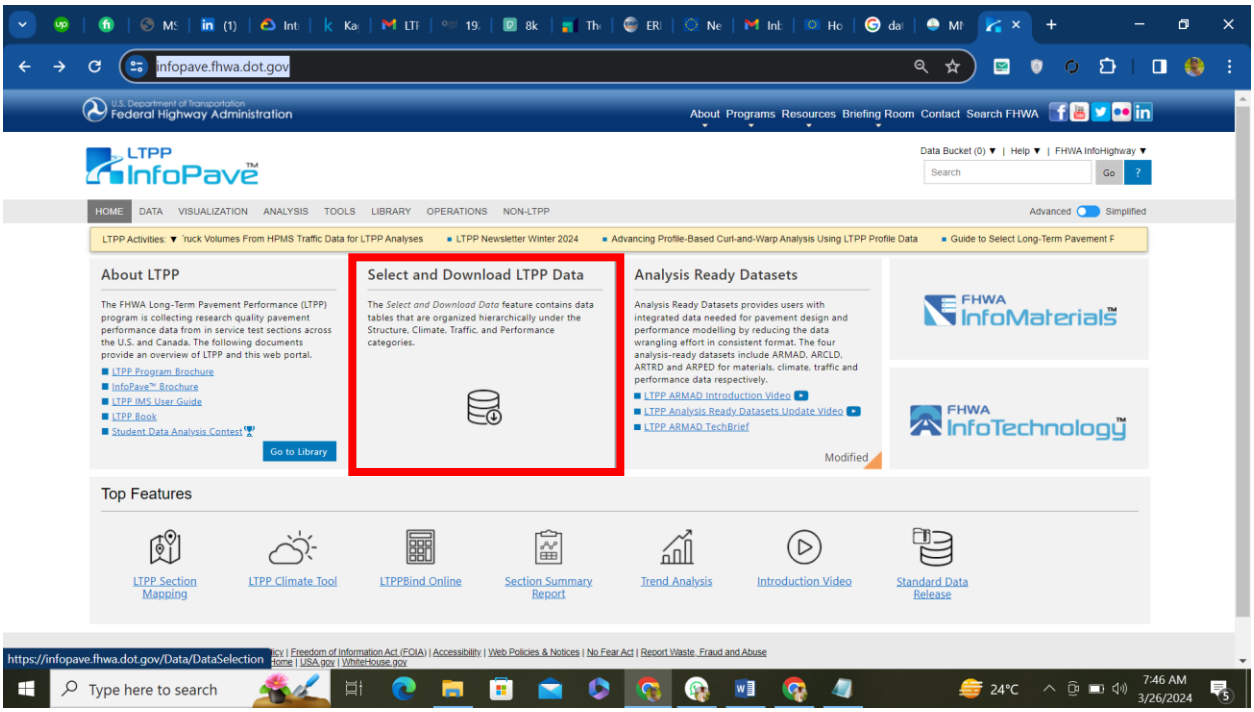
In this report, we detail the process of extracting and preprocessing data from the Long-Term Pavement Performance (LTPP) database for the purpose of developing a model to predict the International Roughness Index (IRI). The project aims to replicate a specific experiment outlined in a scholarly paper, focusing on data from a Canadian province or US state known for its comprehensive and high-quality data records.

2. Data Extraction:

We accessed the LTPP database and selected data from Canadian provinces. We focused on variables relevant to pavement condition and performance, including pavement structure, traffic, and performance metrics such as Mean IRI (MRI). Additionally, we utilized data from the Strategic Highway Research Program (SHRP) and NASA's Modern Era Retrospective-Analysis for Research and Applications (MERRA) to complement the dataset.

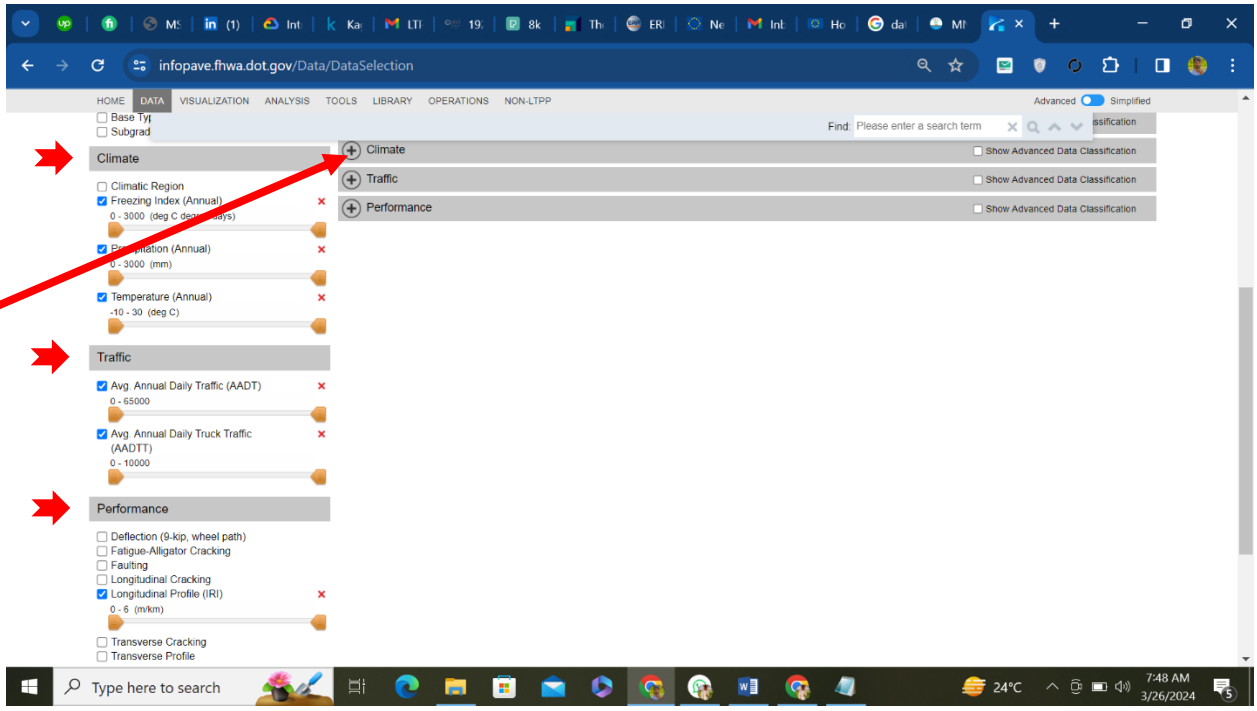
1. Accessing the LTPP Data Selection Page

- Go to the Long-Term Pavement Performance (LTPP) Data Selection page on the Federal Highway Administration (FHWA) website: https://infopave.fhwa.dot.gov/
- Then go to https://infopave.fhwa.dot.gov/Data/DataSelection
- Screenshot:



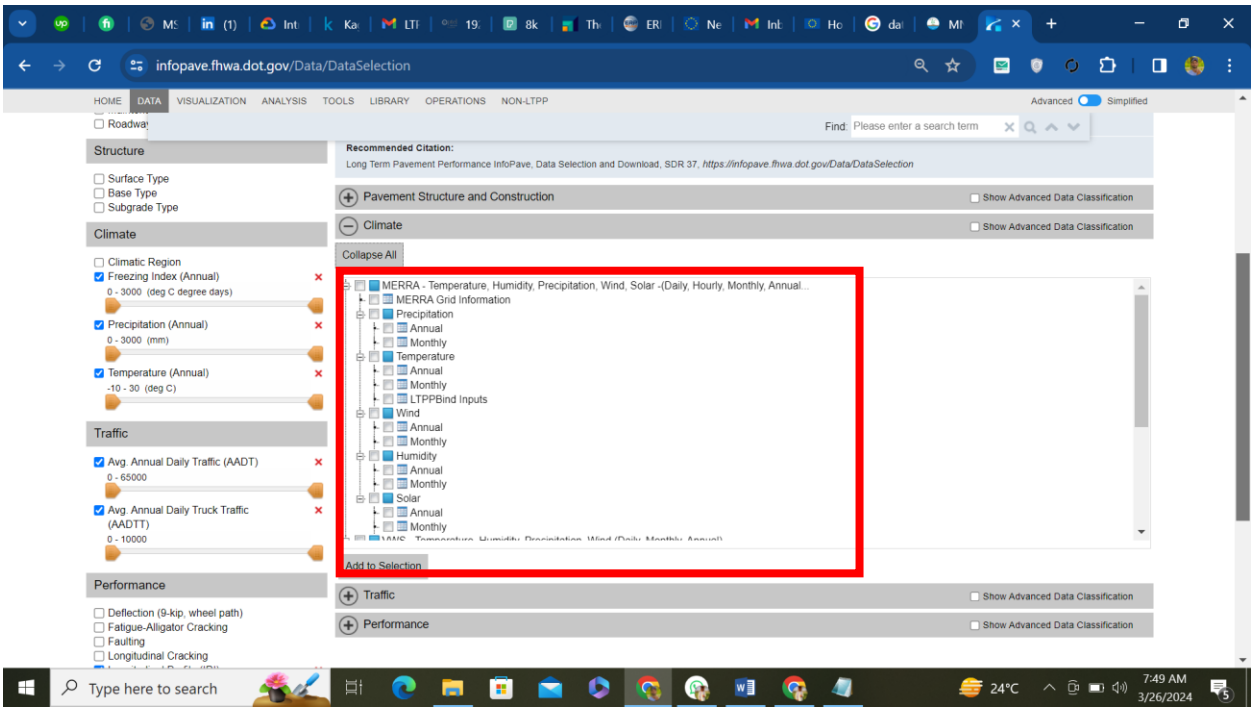
2. Filtering Required Table Categories

- Select the relevant table categories based on pavement structure, traffic, and performance variables.



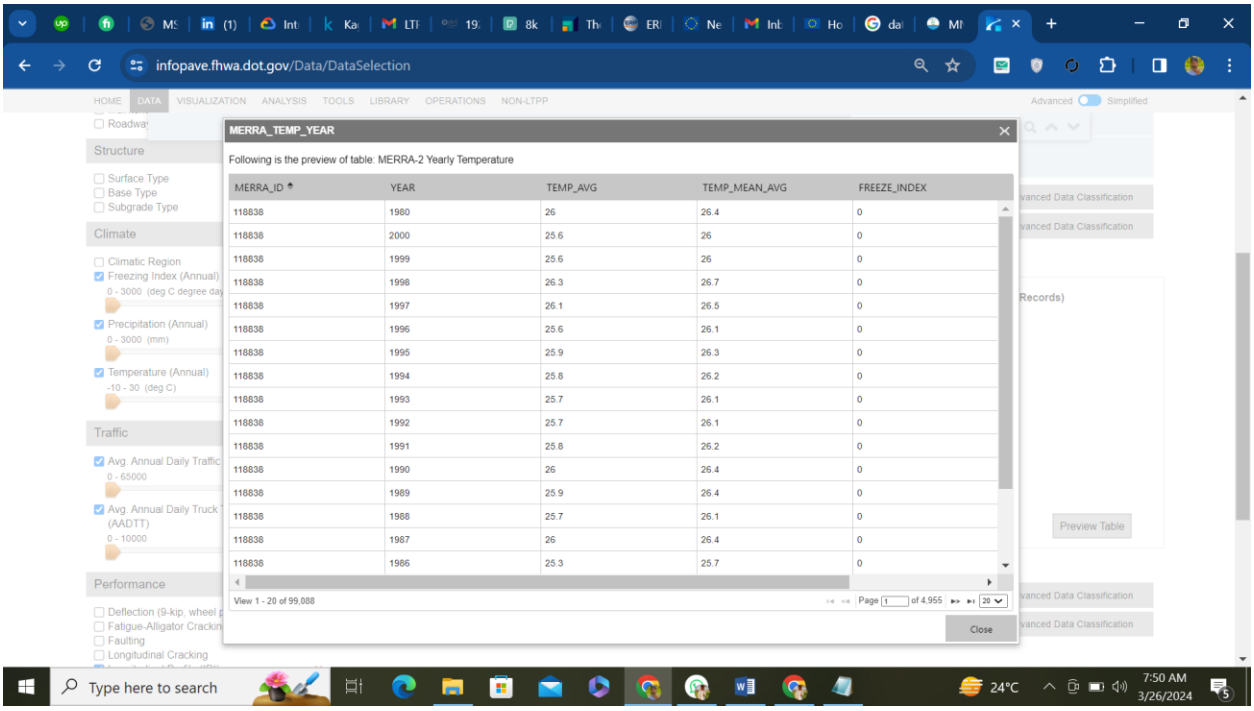
3. Selecting Related Columns

- Hover over the selected table categories to view available Table.
- Choose the table relevant to the project, such as climate zone, traffic volume, and Mean IRI (MRI).

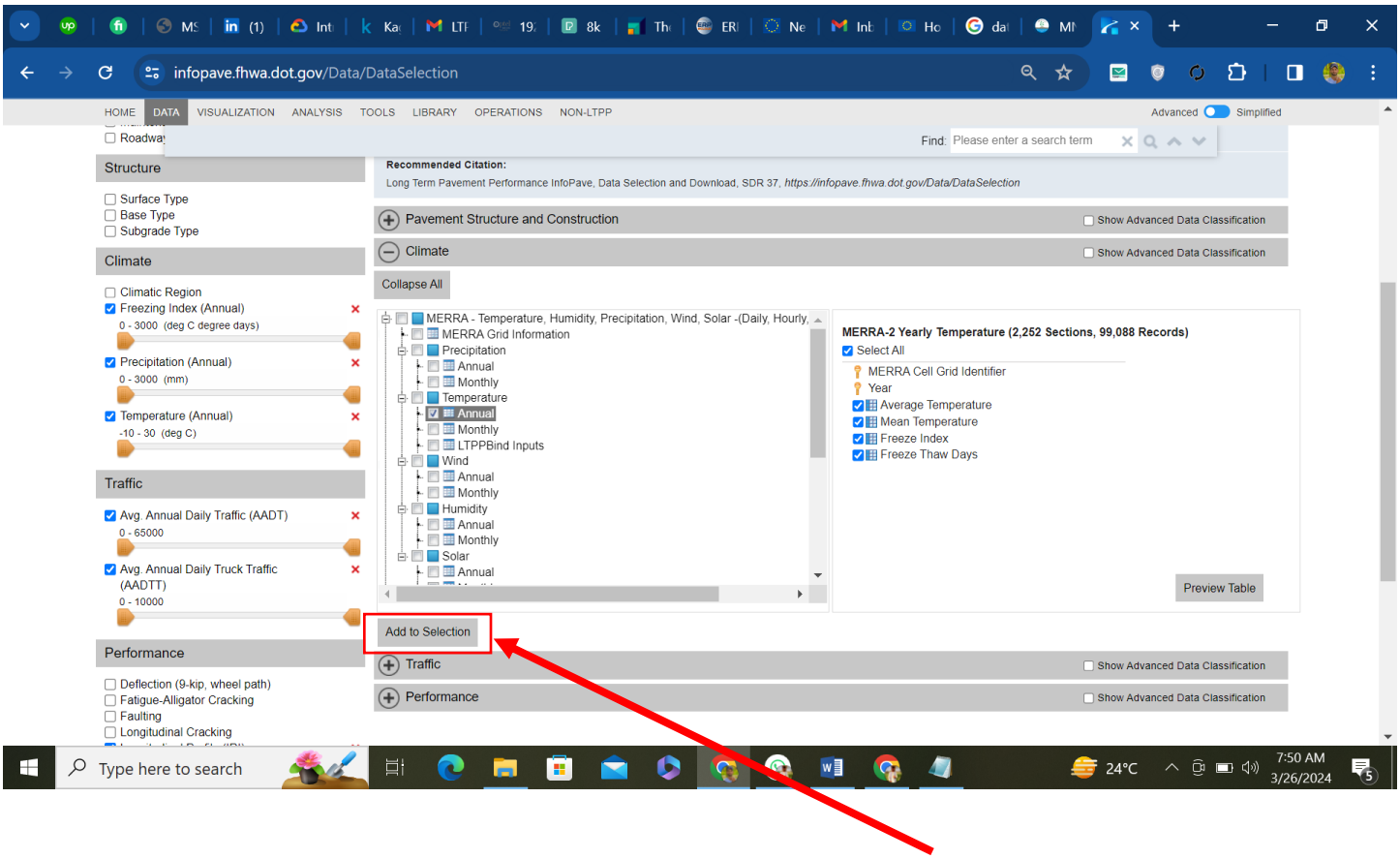


4. Previewing the Table

- Preview the selected data table to ensure it contains the desired variables.
- Verify that the columns and data entries align with the project requirements.

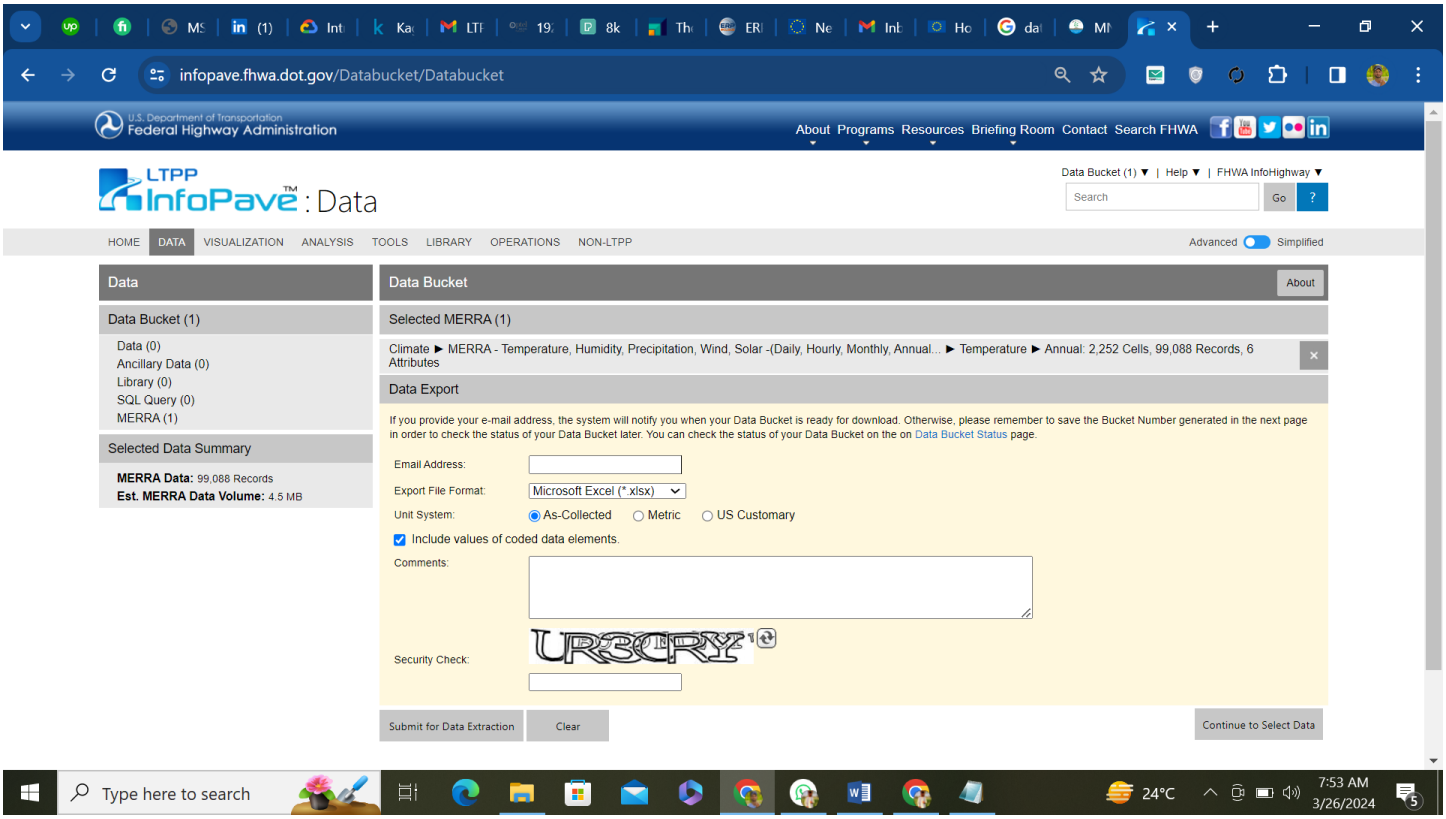


5. **Adding Data Sections**
   - Click on the "Add Section" button to include the selected data table in the download queue.
   - Repeat this step for each relevant data table until all required data sections are added.



6. **Selecting Download Format**
   - Choose the desired format for downloading the data (e.g., Excel).

- Ensure that all selected data sections are included in the download.



7. **Downloading Files**
   - Click on the SUBMIT FOR DATA EXTRACTION button to initiate the download process.
   - Wait for the files to be prepared and packaged for download.

8. **Finalizing Data Extraction**
   - Once the download is complete, verify that all necessary data files have been obtained.
   - Organize the downloaded files into a structured directory for further processing.

2. PRE PROCESSING and analysis:

1- ADDT PRE PROCESSING and analysis

1. Data Loading and Inspection: The initial step involves loading the dataset (`AADTT_MRI.csv`) into a Pandas DataFrame (`df_ADTT`) and inspecting the data using `describe()` method to get an overview of statistical summaries.

2. Data Cleaning:

- Conversion of `STATE_CODE` to string to maintain leading zeros.
- Sorting the DataFrame by `STATE_CODE` and `YEAR`.
- Identifying and handling null values:
- Determining the count and distribution of null values across the dataset.

- Filling null values with mean values based on the year and location.
- Dropping records with null values for `ANNUAL_TRUCK_VOLUME_TREND` after 2016 due to insufficient data for accurate imputation.

3. Data Analysis and Visualization:

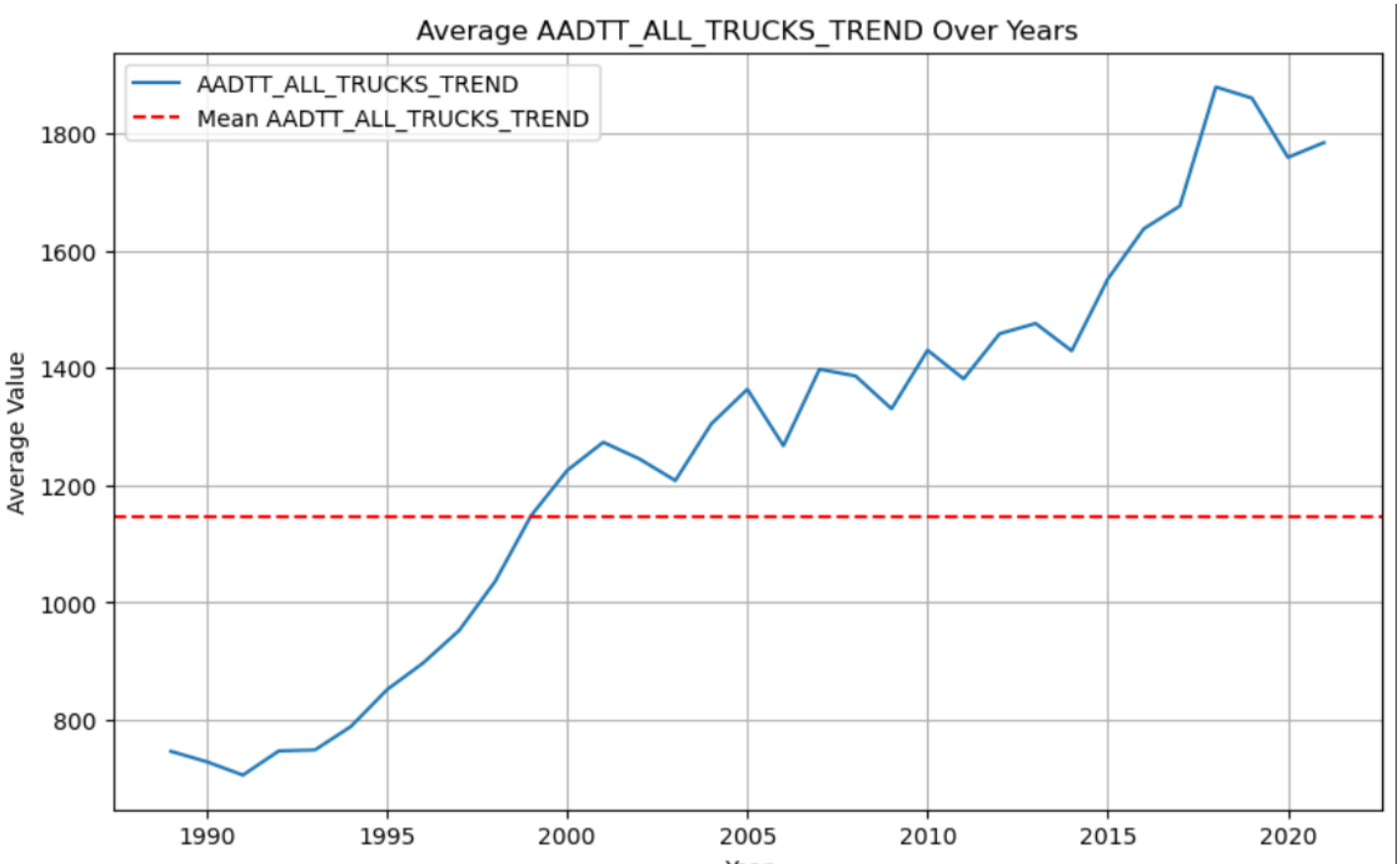
- Calculation of mean values for AADTT and annual truck volume.
- Grouping the dataset by year and calculating the mean AADTT and annual truck volume for each year.
- Visualization of the trends using matplotlib, plotting average AADTT and annual truck volume over the years, along with horizontal lines representing the mean values.
- Generating plots for visualizing the trends before and after data processing.

4. Final Summary Report:

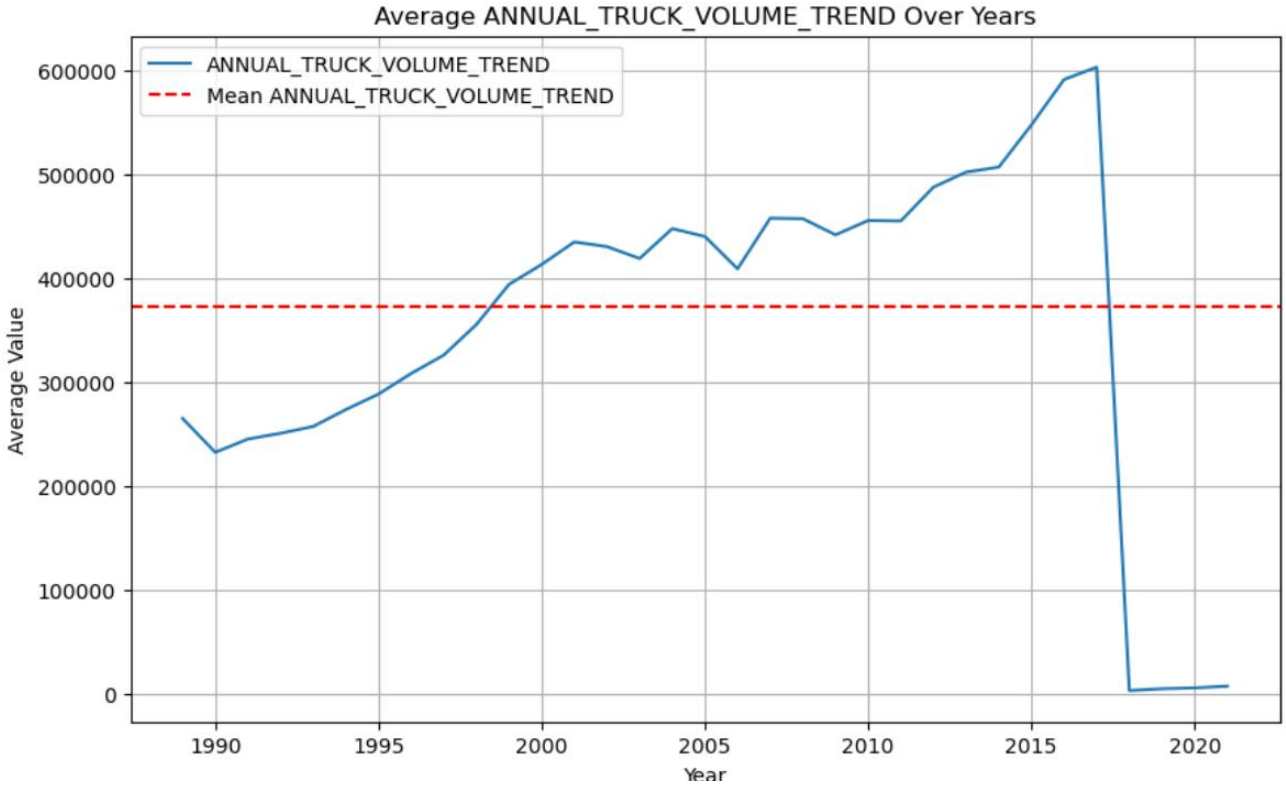
- After preprocessing and analysis, the dataset is cleaned and ready for further modeling or analysis.
- The null values have been handled appropriately, considering both statistical imputation and data quality.
- An explanation is provided for the drop in data after 2016, attributing it to fewer locations providing data, which affects the accuracy of mean calculation for subsequent years.
- The final processed dataset contains data after dropping 28 records with null values.
- We highlight the observation of a substantial proportion of null values after 2014 in the dataset. Recognizing the potential impact on data analysis, especially for ANN (Artificial Neural Network) models, we propose a strategy to address this issue effectively.

5. Plotting Results

AADTT_ALL_TRUCKS_TREND: This variable represents the trend in Average Annual Daily Truck Traffic.



ANNUAL_TRUCK_VOLUME_TREND: This variable reflects the trend in annual truck volume.



2- Hum preprocessing and analyssis

1. Data Loading and Sorting:

- The dataset (`MERRA_HUMID_YEAR.csv`) is loaded into a Pandas DataFrame (`df_hum`).
- The DataFrame is sorted by the 'YEAR' column to ensure data consistency and facilitate further analysis.

2. Exploratory Data Analysis (EDA):

- Null values in the dataset are identified and counted using the `isnull().sum()` function.
- The structure of the dataset is examined, with each record containing columns: 'MERRA_ID', 'YEAR', and 'REL_HUM_AVG_AVG'.
- It is noted that each 'MERRA_ID' represents the relative humidity values under the same condition for several years.

3. Data Visualization:

- A random subset of 'MERRA_ID's is selected to illustrate the variation of relative humidity over the years for different locations or conditions.
- The selected IDs' relative humidity values are plotted against the years to visualize the trends.
- The plot shows the stability of relative humidity over the years for some locations/conditions and volatility for others.

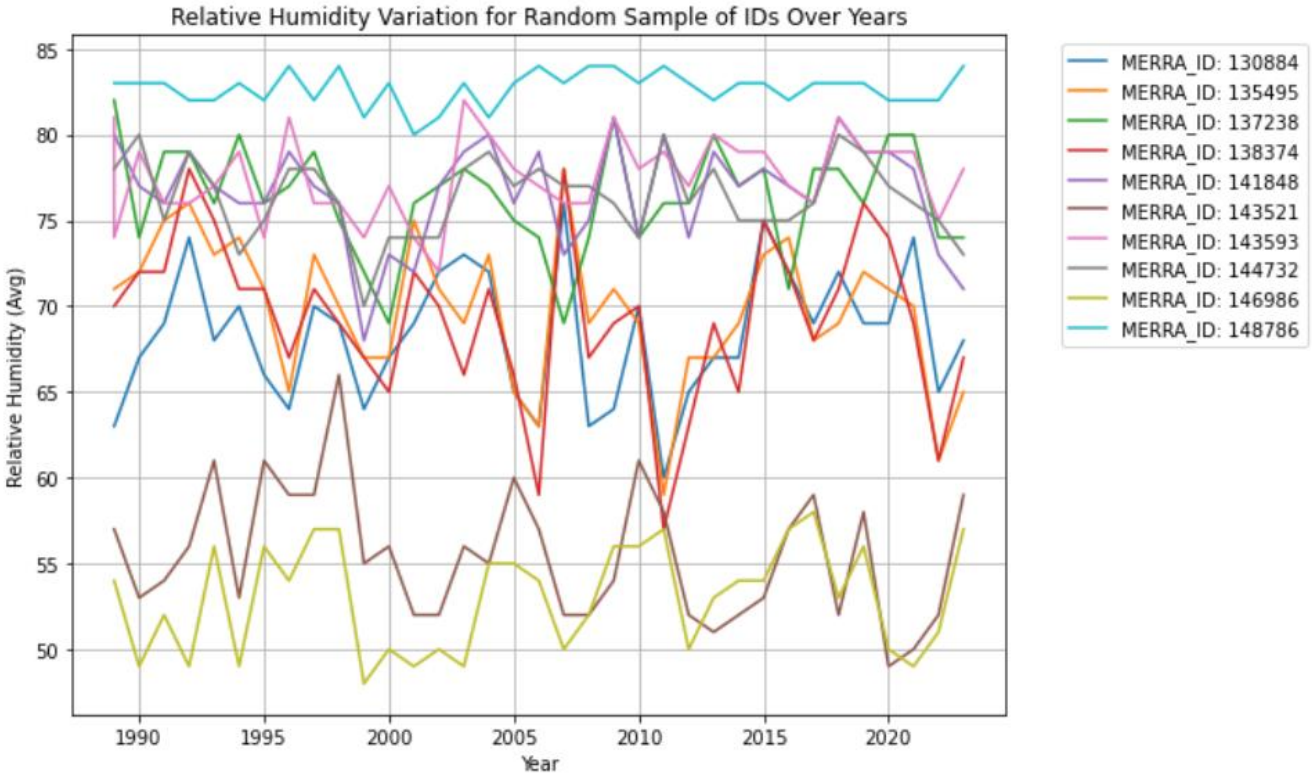
4. Interpretation:

- The analysis reveals that relative humidity exhibits different patterns over the years, depending on the location or condition represented by each 'MERRA_ID'.
- The stability or volatility of relative humidity over time indicates potential impacts on other meteorological variables or models that utilize humidity data, such as MRI (Magnetic Resonance Imaging).

5. Final Summary Report:

- The dataset consists of yearly average relative humidity values collected over multiple years from various locations or conditions.
- Initial data inspection identifies null values, which may require handling depending on the analysis requirements.
- The visualization highlights the variation of relative humidity over the years for different locations or conditions, providing insights into the dataset's characteristics and potential implications for further analysis.
- Understanding the dynamics of relative humidity is crucial for accurate modeling and interpretation of related phenomena, such as MRI data, where humidity can significantly influence outcomes.

Relative Humidity Variation for Random Sample of IDs Over Years



3- PRECIPITATION preprocessing and analyssis

1. Data Loading and Filtering:

- The dataset (`MERRA_PRECIP_YEAR.csv`) is loaded into a Pandas DataFrame (`df_pre`).
- Data from years earlier than 1989 are filtered out to focus on a specific time period.

2. Identifying Null Values:

- Null values in the dataset are identified and counted using the `isnull().sum()` function to understand data completeness.

3. Exploratory Data Analysis (EDA):

- The structure of the dataset is examined, with each record containing columns: 'MERRA_ID', 'YEAR', and 'PRECIPITATION'.
- Similar to the humidity dataset, each 'MERRA_ID' represents precipitation values under the same condition for several years.

4. Data Visualization:

- A random subset of 'MERRA_ID's is selected to illustrate the variation of precipitation levels over the years for different locations or conditions.
- The selected IDs' precipitation values are plotted against the years to visualize the trends.

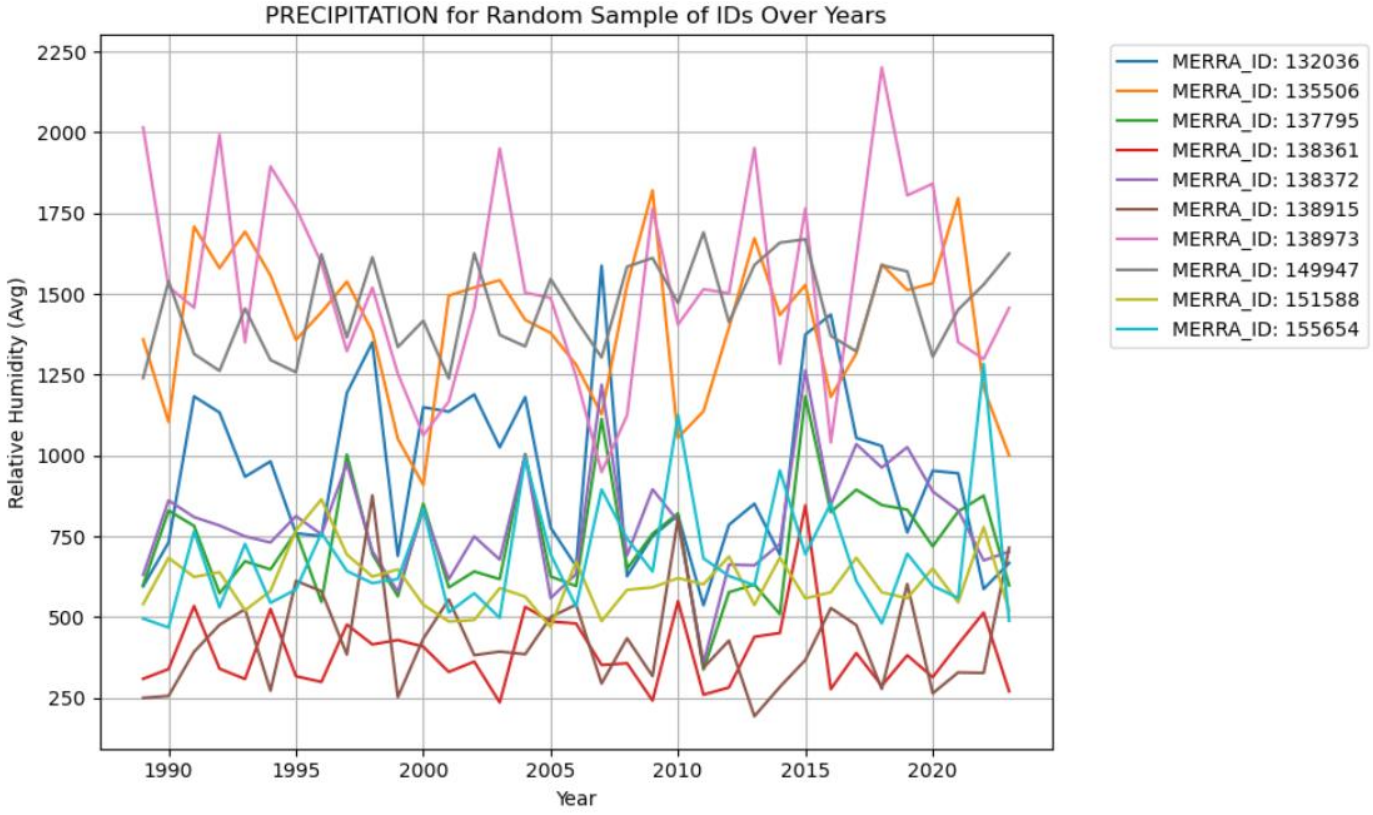
- The plot shows the variation in precipitation levels over time for different locations or conditions, providing insights into the dataset's characteristics.

5. Interpretation:

- The analysis reveals the variability in precipitation levels over the years for different locations or conditions, indicating potential impacts on various environmental factors and phenomena.

6. Final Summary Report:

- The dataset consists of yearly precipitation levels collected over multiple years from various locations or conditions.
- Null values are identified and may require handling based on the analysis requirements.
- Visualization highlights the variation in precipitation levels over the years for different locations or conditions, providing insights into the dataset's characteristics and potential implications for further analysis.
- Understanding the dynamics of precipitation is crucial for various applications, including climate modeling, water resource management, and environmental impact assessments.



4- Freez Data Preprocessing and analysis

1. Data Loading and Sorting:

- The dataset is loaded into a Pandas DataFrame (`df_tmp`).

- The DataFrame is sorted by the 'YEAR' column to ensure the data is arranged chronologically.

2. Identifying Null Values:

- Null values in the dataset are identified and counted using the `isnull().sum()` function to assess data completeness.
- Null records in the DataFrame are displayed to understand the extent of missing data.

3. Handling Null Values:

- It is observed that all null values are in the year 2023, indicating that the last available full data was taken in 2022.
- To address this, the dataset is filtered to exclude data for the year 2023, ensuring that only complete data up to 2022 is retained.

4. Final Summary and Key Insights:

- The analysis ensures that the dataset contains complete and consistent data up to the year 2022.
- It is noted that 'MERRA_ID' serves as the key for these tables, implying that data scientists can combine this dataset with others using 'MERRA_ID' as a common identifier.
- This information is crucial for data scientists and modelers who may need to integrate temperature data with other environmental variables or datasets for comprehensive analysis or modeling tasks.

5- MRI PREPROCESSING and analyssis

1. Data Loading and Sorting:

- The dataset (`MON_HSS_PROFILE_SECTION.csv`) is loaded into a Pandas DataFrame (`df_mri`).
- The 'STATE_CODE' column is converted to string type to maintain leading zeros.
- The DataFrame is sorted by 'STATE_CODE' and 'YEAR' to ensure data consistency.

2. Identifying Null Values:

- Null values in the dataset are identified and counted using the `isnull().sum()` function.
- Null records are displayed to assess the extent of missing data.

3. Handling Null Values:

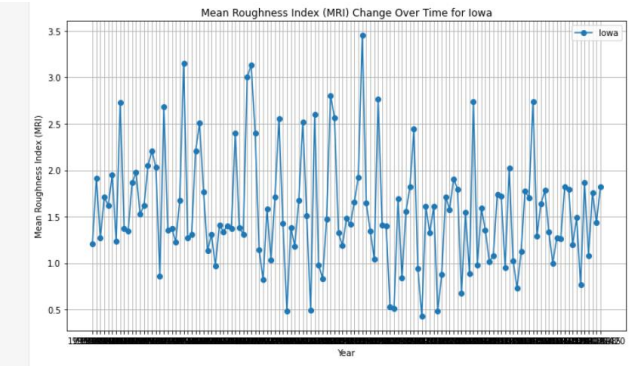
- It is observed that only 2 values are missing in the 'MRI' column, which are subsequently dropped from the sorted DataFrame.

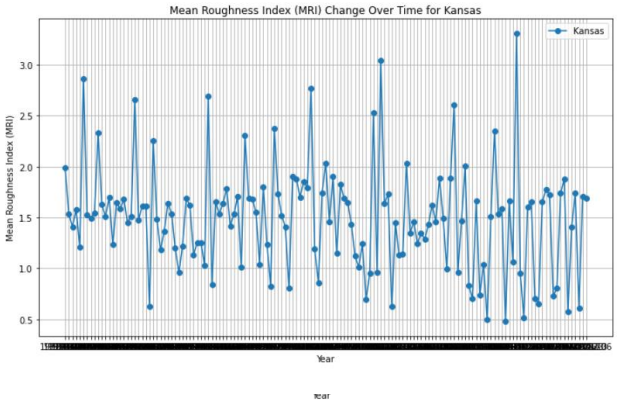
4. MRI Change Over Time Visualization:

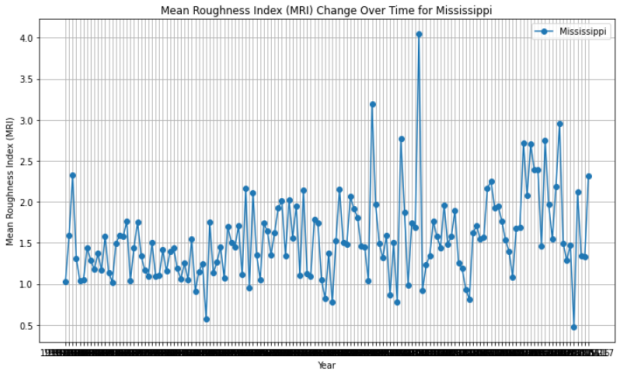
- The sorted DataFrame is grouped by 'YEAR' and 'STATE_CODE_EXP' to calculate the mean MRI for each location over time.
- Unique locations are identified, and MRI change over time is plotted separately for each location.
- For each location, a line plot is generated to visualize the mean MRI change over the years.

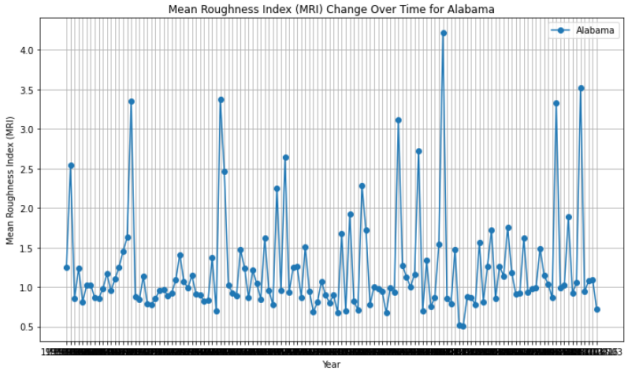
5. Final Summary:

- The preprocessing ensures data integrity by handling null values appropriately.
- Visualization allows for the examination of MRI trends over time for different locations, providing insights into the variation of road surface roughness across various areas.
- Understanding MRI change over time is crucial for road maintenance and planning, as it helps identify areas requiring attention or improvement.









After data extraction and analysis, the variables will be mapped to specific identifiers and dates as follows:

For AADTT, ADDT, MRI, and ESAL:

- Location: This will be represented by the 'STATE_CODE_EXP' or 'STATE_CODE' depending on the dataset.

- Date: This will be represented by the 'YEAR' column.

For HUM, PRE, and TMP:

- MERRA_ID: This will serve as the identifier for specific environmental conditions or locations.
- Date: This will be represented by the 'YEAR' column.

6- AADTT_VEH_CLASS_9_TREND PREPROCESSING and analyssis

1. Data Loading and Sorting:

- The dataset (`TRF_TREND_1.csv`) is loaded into a Pandas DataFrame (`df_trend9`).
- The 'STATE_CODE' column is converted to string type to maintain leading zeros.
- The DataFrame is sorted by 'STATE_CODE' and 'YEAR' to ensure data consistency.

2. Identifying Null Values:

- Null values in the dataset are identified and counted using the `isnull().sum()` function.
- Null records are displayed to assess the extent of missing data.

3. Handling Null Values:

- It is observed that the null values are present only in the 'AADTT_VEH_CLASS_9_TREND' column, amounting to less than 0.5% of the total data.
- Due to the low percentage of null values, the decision is made to drop these records from the sorted DataFrame.

4. Final Summary:

- The preprocessing ensures data integrity by handling the small percentage of null values appropriately.
- With less than 0.5% of the data affected, dropping the null records is a reasonable approach to maintain the quality of the dataset.
- The dataset is now ready for further analysis or modeling tasks related to the trend of AADTT for vehicle class 9.

7- ANNUAL_ESAL_TREND PREPROCESSING and analysis

1. Data Loading and Sorting:

- The dataset (`TRF_TREND.csv`) is loaded into a Pandas DataFrame (`df_esal`).
- The 'STATE_CODE' column is converted to string type to maintain leading zeros.
- The DataFrame is sorted by 'STATE_CODE' and 'YEAR' to ensure data consistency.

2. Identifying Null Values:

- Null values in the dataset are identified and counted using the `isnull().sum()` function.
- It's observed that there are no null values present in any of the columns.

3. Grouping Data and Calculating Mean:

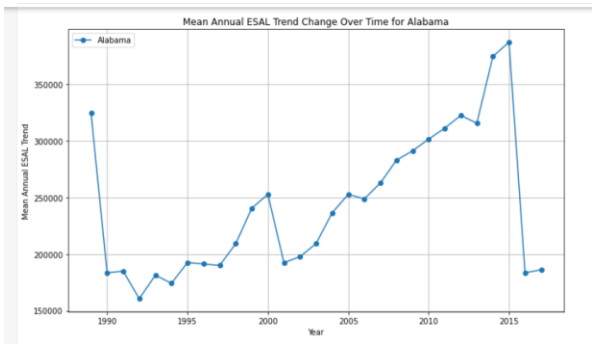
- The data is grouped by 'YEAR' and 'STATE_CODE_EXP', and the mean of 'ANNUAL_ESAL_TREND' is calculated for each group.
- This step aggregates the annual ESAL trends by location, providing insight into how ESAL values vary over time for different areas.

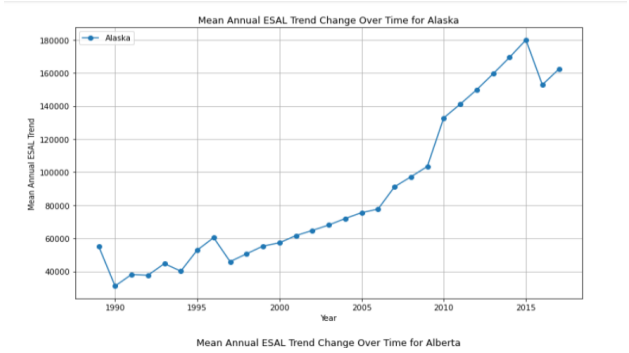
4. Visualization:

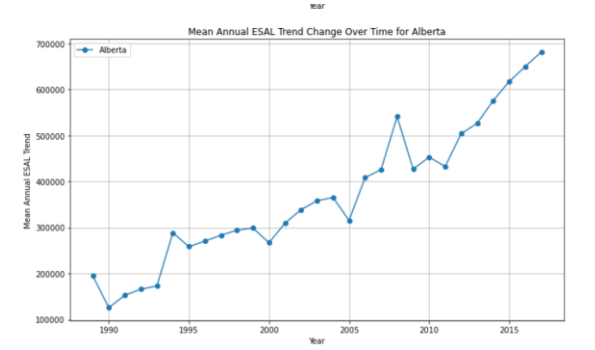
- Unique locations are identified based on 'STATE_CODE_EXP'.
- For each location, a line plot is generated to visualize the mean annual ESAL trend change over the years.
- The plot helps in understanding the temporal variations of ESAL trends across different locations.

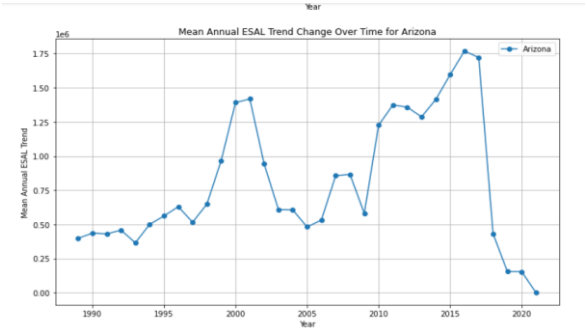
5. Final Summary:

- The preprocessing ensures data integrity by sorting and checking for null values, confirming that the dataset is ready for analysis.
- Grouping the data and calculating the mean allows for the aggregation of ESAL trends by location and year, facilitating insights into long-term trends.
- Visualization of the annual ESAL trend change provides a clear understanding of how ESAL values evolve over time for each location.









8. References:

Long-Term Pavement Performance (LTPP) database: https://infopave.fhwa.dot.gov/Data/DataSelection

Strategic Highway Research Program (SHRP): https://www.fhwa.dot.gov/research/

Modern Era Retrospective-Analysis for Research and Applications (MERRA): https://gmao.gsfc.nasa.gov/reanalysis/MERRA/