Machine Learning and Data Science-Assignment 1

Department of Electrical and Computer Engineering, Faculty of Engineering and Technology Birzeit University

Instructor: Dr.Ismail Khater, Section: 3.

Jana Qutosa-1210331, Shiyar Dar-Mousa-1210766

October 29, 2024

Dataset Overview:

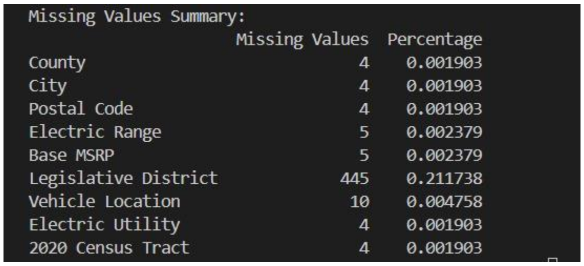
This dataset, provided by the State of Washington, offers details about battery electric vehicles (BEVs) and plug-in hybrid electric vehicles (PHEVs) registered with the Washington State Department of Licensing. It in**cludes 17 columns of information, featuring** each vehicle's VIN, registration county and city, make and model, electric type, and electric range. The dataset covers vehicle model years from 2013 to the present, with metadata updated regularly by the Washington government.

Requirements:

Data Cleaning and Feature Engineering:

1. Document Missing Values:

Documenting missing values is an important part of preparing data for analysis. This step involves finding out which columns have missing data, counting how many entries are missing, and calculating what percentage of the total each missing value represents. By organizing this information into a summary table, we easily see which features might need more attention. Understanding the extent of missing data helps you decide the best ways to handle it, like filling in missing values or removing incomplete entries.

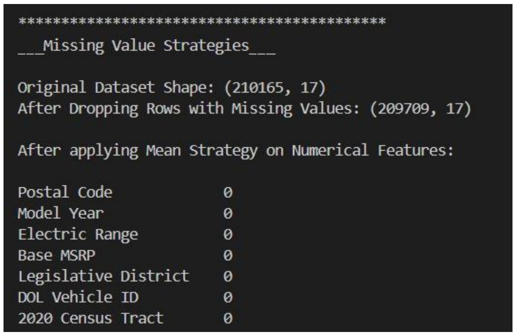


```
Missing Values Summary:
                      Missing Values  Percentage
County                             4    0.001903
City                               4    0.001903
Postal Code                        4    0.001903
Electric Range                     5    0.002379
Base MSRP                          5    0.002379
Legislative District             445    0.211738
Vehicle Location                  10    0.004758
Electric Utility                   4    0.001903
2020 Census Tract                  4    0.001903
```

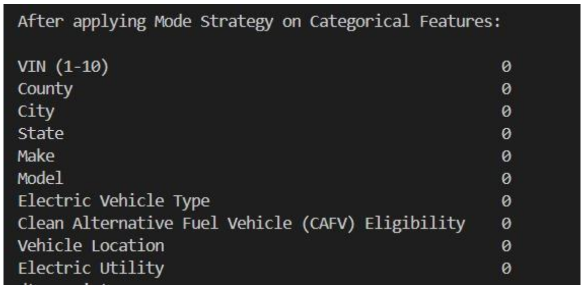
The summary of the missing values in the dataset shows that only some columns contain missing entries. Among such, the frequency and percentage vary. **The Legislative District column contains more missing values throughout the column, 445 entries missing,** which is approximately 21.17% of the whole column. Above, the high percentage indicates that the Legislative District feature might need to be given special attention, probably imputed, or maybe removed at all, depending on how important it is in this analysis. Other columns, like Vehicle Location with 10 missing values, have much more insignificant percent's. As can be seen from here, such a simple imputation technique as mean or mode imputation could work and not strongly affect the overall quality of the dataset because of these lower percentages. Overall, **most features have less than 0.1% missing values.** Hence, the dataset may be considered relatively complete with only a few columns requiring some special handling of the missing values.

2. Missing Value Strategies:

There are several ways to approach missing values of a variable within a dataset; each methodology has associated advantages and/or costs. For example, one common way to handle missing values for numerical columns is **mean imputation**: a missing value gets filled in with the mean of the column. This approach is rather good for normally distributed data since it maintains the size of the dataset and doesn't introduce extreme values, but it may weaken the variability in the data. It could also result in biased estimates if the data that are missing are not random. **Mode imputation** works best on categorical features and fills missing values with the most frequent category, keeping nearly all the patterns in nominal data intact without major distortions. Finally, the easiest approach **is dropping the rows which contain missing values**. It retains the data's authenticity, but the size of the dataset goes down, and sometimes it also leads to a loss of important information, specifically when the data is highly missing. A comparison of these strategies allows the analyst to choose the most appropriate approach given data characteristics and goals of the analysis.



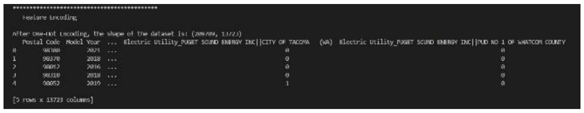
```
******************************************
___Missing Value Strategies___

Original Dataset Shape: (210165, 17)
After Dropping Rows with Missing Values: (209709, 17)

After applying Mean Strategy on Numerical Features:

Postal Code            0
Model Year             0
Electric Range         0
Base MSRP              0
Legislative District   0
DOL Vehicle ID         0
2020 Census Tract      0
```



```
After applying Mode Strategy on Categorical Features:

VIN (1-10)                                         0
County                                             0
City                                               0
State                                              0
Make                                               0
Model                                              0
Electric Vehicle Type                              0
Clean Alternative Fuel Vehicle (CAFV) Eligibility  0
Vehicle Location                                   0
Electric Utility                                   0
```

3. Feature Encoding:

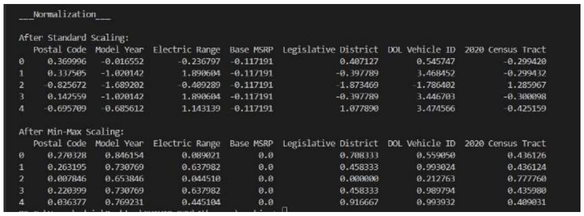
Feature encoding has become one of the most integral parts of a data preparation pipeline in machine learning, as most algorithms require numerical inputs. Common encoding techniques for categorical variables **include one-hot encoding**, which, for example, can be used for the categorical variables "Make" and "Model" of electric vehicles. In this process, **each categorical value is converted to a new categorical column and a binary value of 0 or 1 indicates whether a feature is present**. One-hot encoding might look like the following for three unique values of the feature "Make": "Tesla," "Nissan," and "Ford". Three new columns would be created, "Make_Tesla", "Make_Nissan", and "Make_Ford". Each row would then fill in the column corresponding to its make with a 1 and the other columns with 0s. In this way, no ordinal relationship between categories is assumed, and each category is dealt with independently by the model. One-hot encoding is particularly useful when categorical variables have no natural order. This will enable the machine learning model to make a more accurate prediction or insight.



```
******************************************
   Feature Encoding

After One-Hot Encoding, the shape of the dataset is: (209709, 13723)
   Postal Code  Model Year  ...  Electric Utility_PUGET SOUND ENERGY INC||CITY OF TACOMA   (WA)  Electric Utility_PUGET SOUND ENERGY INC||PUD NO 1 OF WHATCOM COUNTY
0        98380        2021  ...                                                 0                                                                     0
1        98370        2018  ...                                                 0                                                                     0
2        98012        2016  ...                                                 0                                                                     0
3        98310        2018  ...                                                 0                                                                     0
4        98052        2019  ...                                                 1                                                                     0

[5 rows x 13723 columns]
```

4. Normalization:

Any technique that scales numerical features into a common range, which is often necessary for some sensitive analysis methods. Common ones include **Min-Max Scaling and Standard Scaling.** Min-Max Scaling rescales values into a defined range so that comparisons may be done without biases toward features with larger numerical ranges. On the other hand, Standard Scaling centers values on an average of 0 with a standard deviation of 1. This is especially helpful in cases when data distribution approaches normal distribution. The process ensures that no one feature gets to disproportionately influence the model, especially when features vary widely in value ranges. This could be the case with the EV data, where "Electric Range" and "Base MSRP" span very different ranges; hence, normalization brings balance into play, supporting more accurate model performance and insightful analyses.

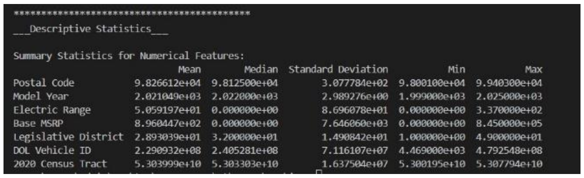


```
___Normalization___

After Standard Scaling:
   Postal Code  Model Year  Electric Range  Base MSRP  Legislative District  DOL Vehicle ID  2020 Census Tract
0     0.369996   -0.016552       -0.236797  -0.117191              0.407127        0.545747          -0.299420
1     0.337505   -1.020142        1.890604  -0.117191             -0.397789        3.468452          -0.299432
2    -0.825672   -1.689202       -0.409289  -0.117191             -1.873469       -1.786402           1.285967
3     0.142559   -1.020142        1.890604  -0.117191             -0.397789        3.446703          -0.300098
4    -0.695709   -0.685612        1.143139  -0.117191              1.077890        3.474566          -0.425159

After Min-Max Scaling:
   Postal Code  Model Year  Electric Range  Base MSRP  Legislative District  DOL Vehicle ID  2020 Census Tract
0     0.270328    0.846154        0.089021        0.0              0.708333        0.559050           0.436126
1     0.263195    0.730769        0.637982        0.0              0.458333        0.993024           0.436124
2     0.007846    0.653846        0.044510        0.0              0.000000        0.212763           0.777760
3     0.220399    0.730769        0.637982        0.0              0.458333        0.989794           0.435980
4     0.036377    0.769231        0.445104        0.0              0.916667        0.993932           0.409031
```

Exploratory Data Analysis:

5. Descriptive Statistics

Descriptive statistics summarize key characteristics of numerical data, including the distribution and the central tendencies of each feature. In computing for measures like **mean, median, and standard deviation,** we obtain notions about the average values, middle points, and variability in data. The mean gives the average value a feature takes on, while the median is a measure of the middle value that reduces effects from extreme values or outliers. Standard deviation is the measure of how values are spread around the mean and gives a measure of variability or consistency of a feature. Together, these metrics provide a succinct summary of the structure of the dataset and make patterns and anomalies visible in the data.

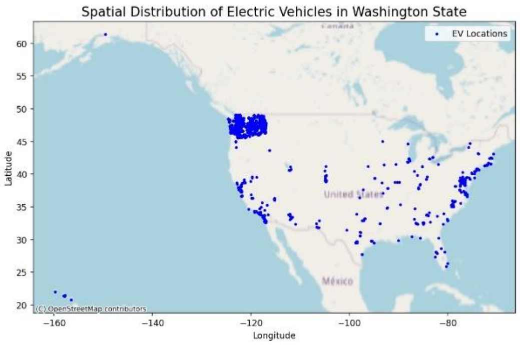


```
*******************************************
___Descriptive Statistics___

Summary Statistics for Numerical Features:
                             Mean        Median  Standard Deviation           Min           Max
Postal Code          9.826612e+04  9.812500e+04        3.077784e+02  9.800100e+04  9.940300e+04
Model Year           2.021049e+03  2.022000e+03        2.989276e+00  1.999000e+03  2.025000e+03
Electric Range       5.059197e+01  0.000000e+00        8.696078e+01  0.000000e+00  3.370000e+02
Base MSRP            8.960447e+02  0.000000e+00        7.646060e+03  0.000000e+00  8.450000e+05
Legislative District 2.893039e+01  3.200000e+01        1.490842e+01  1.000000e+00  4.900000e+01
DOL Vehicle ID       2.290932e+08  2.405281e+08        7.116107e+07  4.469000e+03  4.792548e+08
2020 Census Tract    5.303999e+10  5.303303e+10        1.637504e+07  5.300195e+10  5.307794e+10
```

Descriptive statistics for the numerical features in this dataset expose a few patterns: Mean and median values of Postal Code are at 98261.26, reflecting nearly symmetrical distribution. Model Year is one of the low variability variables since most of the entries fall around a central value; probably, this is because of coding. Electric Range and Base MSRP are highly variable, with standard deviations of 76.66 and 74608.65, respectively, showing large differences in electric capabilities and the price of vehicles. On the other hand, Legislative District and DOL Vehicle ID have small variability with values clustering around their means. Finally, 2020 Census Tract has a very large range and a large mean, which shows big geographic diversity. These are the models that represent important variables with great variation and can be pivotal in further analysis, especially when looking at the trend across vehicle types or regions.

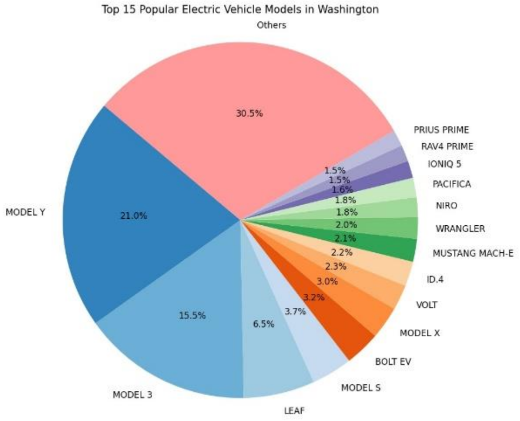
6. Spatial Distribution

Visualizing the spatial distribution of EVs across various locations offers a geographical perspective on adoption patterns. **Using maps**, such as choropleth or dot density maps, to represent the density or number of EVs in different areas (e.g., counties, cities) provides a clear view of where EV adoption is highest. By mapping EVs according to their registered locations, we can easily observe patterns and hotspots, which can inform infrastructure needs like the placement of charging stations. This spatial analysis can also reveal regional adoption trends, identifying areas with high or low EV uptake, and can support policymakers and businesses in addressing geographic disparities in EV accessibility and infrastructure.


Spatial Distribution of Electric Vehicles in Washington State
EV Locations
Canada
United States
México
(C) OpenStreetMap contributors
Latitude
Longitude
60
55
50
45
40
35
30
25
20
−160
−140
−120
−100
−80

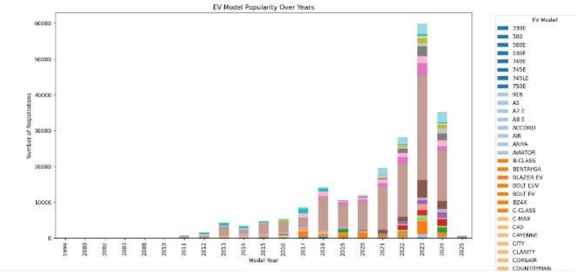
7. Model Popularity:

Analyzing the popularity of different EV models involves examining the frequency distribution of each model within the dataset to identify which models are most commonly adopted. By counting occurrences of each model in categorical data, we can visualize this information using bar charts, where the length or height of each bar represents the popularity of specific models. Further, grouping models by characteristics like year of manufacture or vehicle type can reveal trends in preferences over time or by category. Identifying these trends may uncover which features (e.g., range, price, brand reputation) contribute most to a model’s popularity, providing insights into consumer behavior and industry shifts.



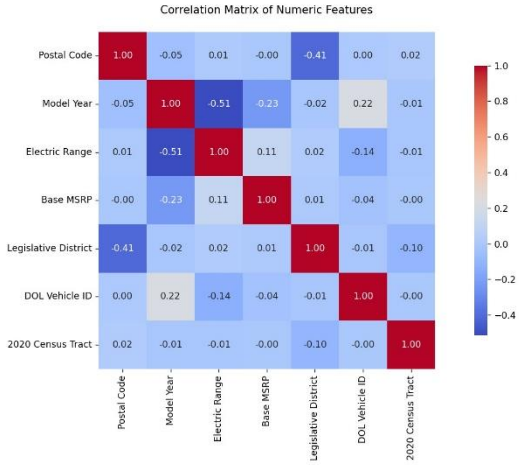
The pie chart shows the popularity of the top 15 models of electric vehicles in Washington. Model Y, by far, takes the biggest piece of the pie with 21% of the total market share, followed by Model 3 at a total of 15.5%, and Leaf with 6.5%. These three top models together highlight very strong preferences of

Tesla and Nissan models among EV users in the region. While the rest, like Bolt EV, Model S, and Mustang Mach-E, represent a market share ranging from 1.5% to 3.7%, indicating their existence in the market but at a more modest rate. Obviously, 30.5% is from "Others," which suggests there is a varied lot in those lesser-known models that together have quite a significant chunk of the electric vehicle market landscape. This distribution indicates that there are just a few models, largely by Tesla, that dominate popularity, while a variety of other models meet specific niche preferences in Washington's EV market.



8. Investigate the relationship between every pair of numeric features.

Examining the relationship of each pair of numeric features involves computation of their correlations to understand how a change in one feature may be related to a change in another. Using a correlation matrix, we are able to calculate values for each numeric feature pair; values near +1 indicate a strong positive relationship, values near -1 a strong negative relationship, and values near 0 suggest no clear relationship. A heatmap will help see these correlations. For example, high positive correlation between "Electric Range" and "Base MSRP" will hint that the pricey vehicles tend to have a longer range presumably because of better battery technology. This kind of insight would let us understand which features interact with each other in a significant way and may drive further analysis or model design.



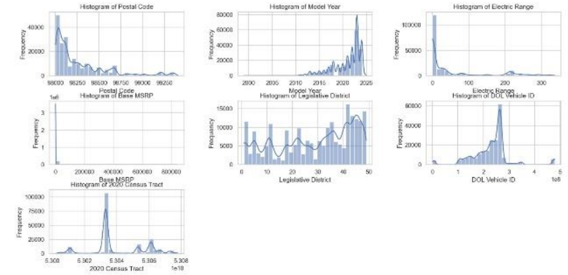
A correlation matrix in numeric features shows many relationships, which include both positive and negative values. The important thing to note is that Electric Range and Model Year are moderately positively correlated (.51), which makes sense since electric range is greater for more recent models, showing there is some technological advancement in electric vehicle technology. Postal Code and Legislative District are highly negatively correlated (-0.41), which would suggest there is a good deal of geographic clustering where each legislative district has a range of postal codes within it.

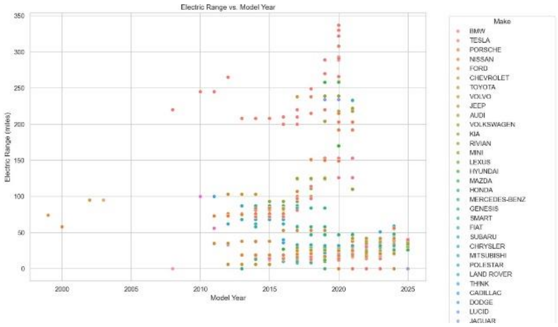
The Base MSRP is very lowly correlated with most features; however, it does share a somewhat positive correlation of 0.23 with Electric Range, which suggests that those vehicles that have higher electric ranges may have a slightly higher MSRP. There is a minor positive correlation of 0.22 between DOL Vehicle ID and Model Year, perhaps due to some kind of chronological ordering of vehicle registration. Overall, most numeric features in the matrix are very weakly correlated, with only a few moderate relationships influenced perhaps by geography or model year advancements. This is overall generally uncorrelated, which can be beneficial for modeling because it reduces multicollinearity concerns in those instances.

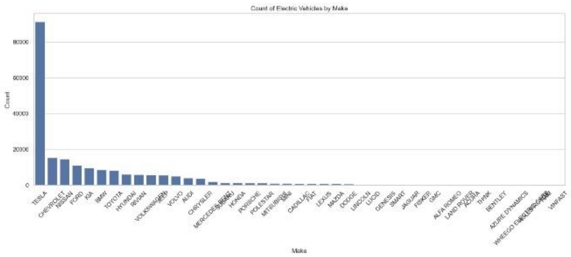
Visualization:

9. Data Exploration Visualizations:

Data exploration visualizations help **to uncover patterns, distributions, and relationships within the dataset**, making it easier to understand feature interactions. **Histograms** are useful for examining the distribution of individual numerical features, such as "Electric Range" or "Base MSRP," showing how frequently different ranges or prices appear. Scatter plots can visualize the relationship between two continuous features, like "Electric Range" and "Base MSRP," where clustering patterns or trends indicate possible correlations. Boxplots are particularly helpful for spotting outliers and comparing distributions across categorical features (e.g., "Electric Vehicle Type" and "Base MSRP"), showing the spread and central tendencies of prices across different EV types. These visualizations provide an intuitive understanding of the data's structure and highlight key areas for further investigation.

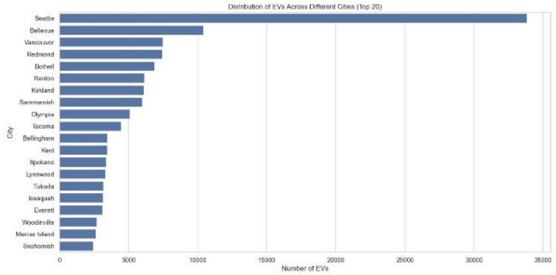


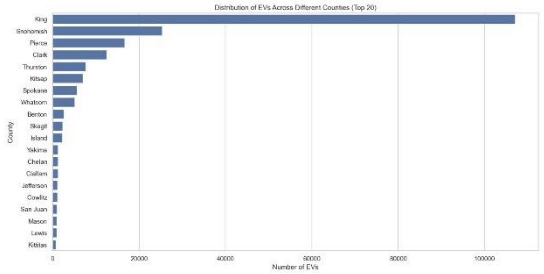




10. Comparative Visualization:

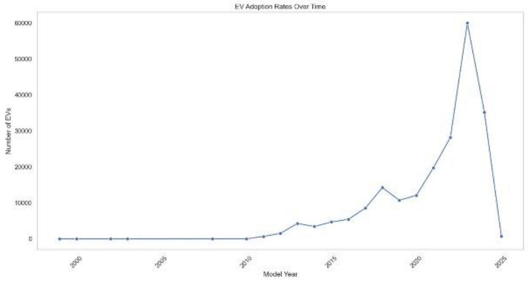
Comparative visualizations using bar charts or stacked bar charts provide a clear view of how electric vehicles (EVs) are distributed across different locations, such as cities or counties. A simple bar chart could display the count of EVs per city, showing which areas have higher or lower EV adoption rates. Stacked bar charts add an extra layer by allowing comparison within categories, such as EV types or model years, within each location. These visualizations can highlight regional trends and identify areas with significant adoption or gaps, providing valuable insights for policymakers or businesses focusing on regional EV deployment and infrastructure needs.

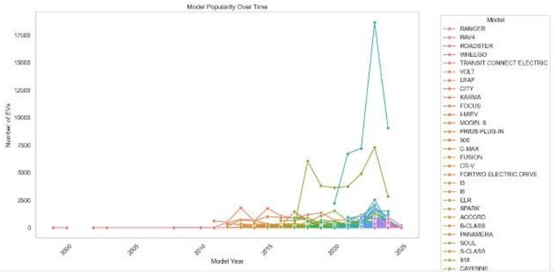




11. Temporal Analysis :

By working out the trends over time, one will gain insight into how the adoption, and consequently the model popularity, of electric vehicles have changed. If the dataset has data which relates to time, like dates of registration or model release years, this might disclose a pattern showing spikes in EV adoption rates, seasonal variations, or shifts in consumer preferences for certain models. For example, temporal trends in EV adoption can be up during some years as a reflection of the effectiveness of policy interventions or changes in consumer behavior. Tracking the temporal popularity of specific models could therefore indicate drivers behind consumer choices concerning improvements in technology, price, or range capability. These trends are of much importance to analysts, as they might help project the future growth of the EV market by informing strategic decisions for manufacturers, policymakers, and other stakeholders.


EV Adoption Rates Over Time
Number of EVs
Model Year


Model Popularity Over Time
Number of EVs
Model Year
Model
RANGER
RAV4
ROADSTER
WHEEGO
TRANSIT CONNECT ELECTRIC
VOLT
LEAF
CITY
KARMA
FOCUS
I-MIEV
MODEL S
PRIUS PLUG-IN
C-MAX
FUSION
CR-V
FORTWO ELECTRIC DRIVE
I3
I8
ELR
SPARK
ACCORD
B-CLASS
PANAMERA
SOUL
S-CLASS
CAYENNE