BIRZEIT UNIVERSITY

Faculty of Engineering and Technology

Electrical and Computer Engineering Department

ENCS5341

MACHINE LEARNING AND DATA SCIENCE

Assignment1 - Report

Prepared by:

Manar Shawahni, 1201086.

Layan Abuershaid, 1200098.

Instructor's Name: Dr. Yazan Abu Farha

Section: 1

30 October, 2024

Contents

Brief description, including the number of examples, number and type of features, and context.

```
Dataset Preview:
   VIN (1-10)     County       City State  Postal Code  Model Year     Make  \
0  5UXTA6C0XM     Kitsap    Seabeck    WA      98380.0        2021      BMW
1  5YJ3E1EB1J     Kitsap    Poulsbo    WA      98370.0        2018    TESLA
2  WP0AD2A73G  Snohomish    Bothell    WA      98012.0        2016  PORSCHE
3  5YJ3E1EB5J     Kitsap  Bremerton    WA      98310.0        2018    TESLA
4  1N4AZ1CP3K       King    Redmond    WA      98052.0        2019   NISSAN

      Model                   Electric Vehicle Type  \
0        X5  Plug-in Hybrid Electric Vehicle (PHEV)
1   MODEL 3          Battery Electric Vehicle (BEV)
2  PANAMERA  Plug-in Hybrid Electric Vehicle (PHEV)
3   MODEL 3          Battery Electric Vehicle (BEV)
4      LEAF          Battery Electric Vehicle (BEV)

  Clean Alternative Fuel Vehicle (CAFV) Eligibility  Electric Range  \
0           Clean Alternative Fuel Vehicle Eligible            30.0
1           Clean Alternative Fuel Vehicle Eligible           215.0
2             Not eligible due to low battery range            15.0
3           Clean Alternative Fuel Vehicle Eligible           215.0
4           Clean Alternative Fuel Vehicle Eligible           150.0

   Base MSRP  Legislative District  DOL Vehicle ID  \
0        0.0                  35.0       267929112
1        0.0                  23.0       475911439
2        0.0                   1.0       101971278
3        0.0                  23.0       474363746
4        0.0                  45.0       476346482

              Vehicle Location  \
0  POINT (-122.8728334 47.5798304)
1  POINT (-122.6368884 47.7469547)
2    POINT (-122.206146 47.839957)
3  POINT (-122.6231895 47.5930874)
4      POINT (-122.13158 47.67858)
```

```
                               Electric Utility  2020 Census Tract
0                        PUGET SOUND ENERGY INC       5.303509e+10
1                        PUGET SOUND ENERGY INC       5.303509e+10
2                        PUGET SOUND ENERGY INC       5.306105e+10
3                        PUGET SOUND ENERGY INC       5.303508e+10
4  PUGET SOUND ENERGY INC||CITY OF TACOMA - (WA)       5.303303e+10

Number of examples: 210165
Number of features: 17

Feature Types:
VIN (1-10)                                            object
County                                                object
City                                                  object
State                                                 object
Postal Code                                          float64
Model Year                                             int64
Make                                                  object
Model                                                 object
Electric Vehicle Type                                 object
Clean Alternative Fuel Vehicle (CAFV) Eligibility     object
Electric Range                                       float64
Base MSRP                                            float64
Legislative District                                 float64
DOL Vehicle ID                                         int64
Vehicle Location                                      object
Electric Utility                                      object
2020 Census Tract                                    float64
dtype: object
```

This dataset contains records of electric vehicles registered in Washington State. It includes 210,165 entries across 17 features, detailing vehicle specifics like make, model, year, and electric range, along with clean fuel eligibility, geographic data, and legislative district information. The dataset features various types of data: numerical (like model year, electric range), categorical (like make, vehicle type), and geographic (latitude and longitude). This information supports analysis of electric vehicle trends, popular models, and fuel eligibility across different regions, which can be valuable for government agencies and research studies focused on EV adoption and clean energy initiatives.

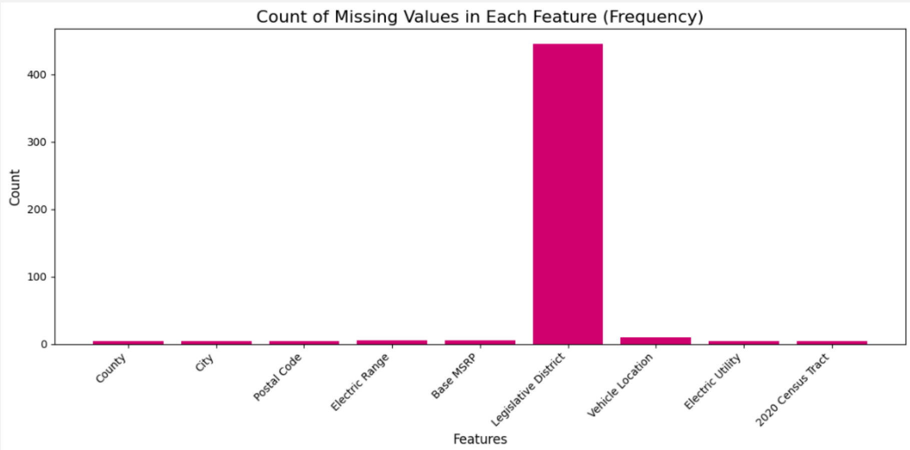
1. Data Cleaning and Feature Engineering:

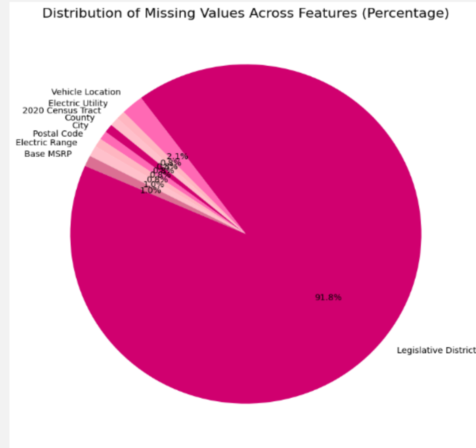
1.1. Document Missing Values:

Check for missing values and document their frequency and distribution across features.

Missing Data Summary:

	Feature	Count	Percentage
0	VIN (1-10)	0	0.000000
1	County	4	0.001903
2	City	4	0.001903
3	State	0	0.000000
4	Postal Code	4	0.001903
5	Model Year	0	0.000000
6	Make	0	0.000000
7	Model	0	0.000000
8	Electric Vehicle Type	0	0.000000
9	Clean Alternative Fuel Vehicle (CAFV) Eligibility	0	0.000000
10	Electric Range	5	0.002379
11	Base MSRP	5	0.002379
12	Legislative District	445	0.211738
13	DOL Vehicle ID	0	0.000000
14	Vehicle Location	10	0.004758
15	Electric Utility	4	0.001903
16	2020 Census Tract	4	0.001903





As shown, the dataset's missing values are mostly in the Legislative District feature, comprising 91.8% of all gaps. The bar chart shows counts per feature, while the pie chart highlights the dominance of missing data in Legislative District.

1.2. Missing Value Strategies: If missing values are present, apply multiple strategies (e.g., mean/median imputation, dropping rows) and compare their impact on the analysis.

```
Missing Values After Mean Imputation (Numerical Features):
VIN (1-10)                                            0
County                                                4
City                                                  4
State                                                 0
Postal Code                                           0
Model Year                                            0
Make                                                  0
Model                                                 0
Electric Vehicle Type                                 0
Clean Alternative Fuel Vehicle (CAFV) Eligibility     0
Electric Range                                        0
Base MSRP                                             0
Legislative District                                  0
DOL Vehicle ID                                        0
Vehicle Location                                     10
Electric Utility                                      4
2020 Census Tract                                     0
dtype: int64
```

```
Missing Values After Median Imputation (Numerical Features):
VIN (1-10)                                            0
County                                                4
City                                                  4
State                                                 0
Postal Code                                           0
Model Year                                            0
Make                                                  0
Model                                                 0
Electric Vehicle Type                                 0
Clean Alternative Fuel Vehicle (CAFV) Eligibility     0
Electric Range                                        0
Base MSRP                                             0
Legislative District                                  0
DOL Vehicle ID                                        0
Vehicle Location                                     10
Electric Utility                                      4
2020 Census Tract                                     0
dtype: int64
```

```
Missing Values After Mode Imputation (Categorical Features):
VIN (1-10)                                            0
County                                                0
City                                                  0
State                                                 0
Postal Code                                           4
Model Year                                            0
Make                                                  0
Model                                                 0
Electric Vehicle Type                                 0
Clean Alternative Fuel Vehicle (CAFV) Eligibility     0
Electric Range                                        5
Base MSRP                                             5
Legislative District                                445
DOL Vehicle ID                                        0
Vehicle Location                                      0
Electric Utility                                      0
2020 Census Tract                                     4
dtype: int64
```

```
Missing Values After Dropping Rows:
VIN (1-10)                                            0
County                                                0
City                                                  0
State                                                 0
Postal Code                                           0
Model Year                                            0
Make                                                  0
Model                                                 0
Electric Vehicle Type                                 0
Clean Alternative Fuel Vehicle (CAFV) Eligibility     0
Electric Range                                        0
Base MSRP                                             0
Legislative District                                  0
DOL Vehicle ID                                        0
Vehicle Location                                      0
Electric Utility                                      0
2020 Census Tract                                     0
dtype: int64
```

Mean and median imputations filled gaps in numerical features like Electric Range, keeping the dataset size. Mode imputation worked for categorical data like Electric Utility, while dropping rows removed all gaps but slightly reduced the dataset size.

Fill missing values (Imputation vs. Dropping Rows):

```
Missing Values After imputation:
VIN (1-10)                                            0
County                                                0
City                                                  0
State                                                 0
Postal Code                                           0
Model Year                                            0
Make                                                  0
Model                                                 0
Electric Vehicle Type                                 0
Clean Alternative Fuel Vehicle (CAFV) Eligibility     0
Electric Range                                        0
Base MSRP                                             0
Legislative District                                  0
DOL Vehicle ID                                        0
Vehicle Location                                      0
Electric Utility                                      0
2020 Census Tract                                     0
dtype: int64
```

```
Missing Values After Dropping Rows:
VIN (1-10)                                            0
County                                                0
City                                                  0
State                                                 0
Postal Code                                           0
Model Year                                            0
Make                                                  0
Model                                                 0
Electric Vehicle Type                                 0
Clean Alternative Fuel Vehicle (CAFV) Eligibility     0
Electric Range                                        0
Base MSRP                                             0
Legislative District                                  0
DOL Vehicle ID                                        0
Vehicle Location                                      0
Electric Utility                                      0
2020 Census Tract                                     0
dtype: int64
```

Imputation keeps the dataset size (210,165 entries) by filling gaps, while dropping rows creates a smaller complete dataset (209,709 entries).

1.3. Feature Encoding: Encode categorical features (e.g., Make, Model) using techniques like one-hot encoding.

First: imputation strategy

```
Low Cardinality Columns: ['State', 'Make', 'Electric Vehicle Type', 'Clean Alternative Fuel Vehicle (CAFV) Eligibility']
High Cardinality Columns: ['VIN (1-10)', 'County', 'City', 'Model', 'Vehicle Location', 'Electric Utility']

Shape of Encoded Dataset (Mean + Mode Imputed): (210165, 104)

Encoded Dataset (Mean + Mode Imputed) Preview:
   VIN (1-10)  County  City  Postal Code  Model Year  Model  Electric Range  \
0        3754      87   595      98380.0        2021    147            30.0
1        3875      87   524      98370.0        2018     88           215.0
2       10664     169    61      98012.0        2016    100            15.0
3        3903      87    64      98310.0        2018     88           215.0
4        1781      85   546      98052.0        2019     86           150.0

   Base MSRP  Legislative District  DOL Vehicle ID  ...  Make_TESLA  \
0        0.0                  35.0       267929112  ...           0
1        0.0                  23.0       475911439  ...         1.0
2        0.0                   1.0       101971278  ...           0
3        0.0                  23.0       474363746  ...         1.0
4        0.0                  45.0       476346482  ...           0

   Make_TH!NK  Make_TOYOTA  Make_VINFAST  Make_VOLKSWAGEN  Make_VOLVO  \
0           0            0             0                0           0
1           0            0             0                0           0
2           0            0             0                0           0
3           0            0             0                0           0
4           0            0             0                0           0

   Electric Vehicle Type_Plug-in Hybrid Electric Vehicle (PHEV)  \
0                                                1.0
1                                                  0
2                                                1.0
3                                                  0
4                                                  0

   Clean Alternative Fuel Vehicle (CAFV) Eligibility_Eligibility unknown as battery range has not been researched  \
0                                                  0
1                                                  0
2                                                  0
3                                                  0
4                                                  0

   Clean Alternative Fuel Vehicle (CAFV) Eligibility_Not eligible due to low battery range
0                                                  0
1                                                  0
2                                                1.0
3                                                  0
4                                                  0
```

Second: Drop rows strategy

```
Shape of Encoded Dataset (Dropped Rows): (209709, 58)

Encoded Dataset (Dropped Rows) Preview:
   VIN (1-10)  County  City  Postal Code  Model Year  Model  Electric Range  \
0        3752      17   371      98380.0        2021    147            30.0
1        3873      17   329      98370.0        2018     88           215.0
2       10659      30    35      98012.0        2016    100            15.0
3        3901      17    37      98310.0        2018     88           215.0
4        1780      16   346      98052.0        2019     86           150.0

   Base MSRP  Legislative District  DOL Vehicle ID  ...  Make_TESLA  \
0        0.0                  35.0       267929112  ...           0
1        0.0                  23.0       475911439  ...         1.0
2        0.0                   1.0       101971278  ...           0
3        0.0                  23.0       474363746  ...         1.0
4        0.0                  45.0       476346482  ...           0

   Make_TH!NK  Make_TOYOTA  Make_VINFAST  Make_VOLKSWAGEN  Make_VOLVO  \
0           0            0             0                0           0
1           0            0             0                0           0
2           0            0             0                0           0
3           0            0             0                0           0
4           0            0             0                0           0

   Electric Vehicle Type_Plug-in Hybrid Electric Vehicle (PHEV)  \
0                                                1.0
1                                                  0
2                                                1.0
3                                                  0
4                                                  0

   Clean Alternative Fuel Vehicle (CAFV) Eligibility_Eligibility unknown as battery range has not been researched
0                                                  0
1                                                  0
2                                                  0
3                                                  0
4                                                  0

   Clean Alternative Fuel Vehicle (CAFV) Eligibility_Not eligible due to low battery range
0                                                  0
1                                                  0
2                                                1.0
3                                                  0
4                                                  0
```

One-hot encoding was applied to low-unique features like 'Make', while label encoding handled high-unique ones like 'VIN'. Imputation kept all rows and expanded to 104 columns, preserving data but adding complexity. Dropping rows reduced it to 58 columns and fewer rows, simplifying the dataset but losing some information.

1.4. Normalization: Normalize numerical features if necessary for chosen analysis methods.

```
Min-Max Scaled Data:
   Postal Code  Model Year  Electric Range  Base MSRP  Legislative District  \
0     0.987766    0.846154        0.089021        0.0              0.708333
1     0.987664    0.730769        0.637982        0.0              0.458333
2     0.984005    0.653846        0.044510        0.0              0.000000
3     0.987051    0.730769        0.637982        0.0              0.458333
4     0.984414    0.769231        0.445104        0.0              0.916667

   DOL Vehicle ID  2020 Census Tract
0        0.559050           0.945730
1        0.993024           0.945730
2        0.212763           0.946202
3        0.989794           0.945730
4        0.993932           0.945693

Z-Score Normalized Data:
   Postal Code  Model Year  Electric Range  Base MSRP  Legislative District  \
0     0.082518   -0.016279       -0.236881  -0.117289              0.407157
1     0.078428   -1.019981        1.890216  -0.117289             -0.397760
2    -0.067968   -1.689116       -0.409348  -0.117289             -1.873443
3     0.053893   -1.019981        1.890216  -0.117289             -0.397760
4    -0.051610   -0.685414        1.142858  -0.117289              1.077922

   DOL Vehicle ID  2020 Census Tract
0        0.546015           0.035964
1        3.468961           0.035964
2       -1.786327           0.052697
3        3.447210           0.035957
4        3.475075           0.034637
Custom Scaled Data (divided by 10^j for each column):
   Postal Code  Model Year  Electric Range  Base MSRP  Legislative District  \
0      0.98380      0.2021           0.030        0.0                  0.35
1      0.98370      0.2018           0.215        0.0                  0.23
2      0.98012      0.2016           0.015        0.0                  0.01
3      0.98310      0.2018           0.215        0.0                  0.23
4      0.98052      0.2019           0.150        0.0                  0.45

   DOL Vehicle ID  2020 Census Tract
0        0.267929           0.530351
1        0.475911           0.530351
2        0.101971           0.530611
3        0.474364           0.530351
4        0.476346           0.530330
```

The normalization adjusts data for model performance: Min-Max Scaling bounds data between 0 and 1, Z-score centers around 0 for normal distribution assumptions, and Custom Scaling reduces large values without centering.

Apply Z-score on drop rows strategy

```
Normalized Dataset (Dropped Rows Dataset - Continuous Numerical Columns) Preview:
   Postal Code  Model Year  Electric Range  Base MSRP  Legislative District  \
0     0.369996   -0.016552       -0.236797  -0.117191              0.407127
1     0.337505   -1.020142        1.890604  -0.117191             -0.397789
2    -0.825672   -1.689202       -0.409289  -0.117191             -1.873469
3     0.142559   -1.020142        1.890604  -0.117191             -0.397789
4    -0.695709   -0.685612        1.143139  -0.117191              1.077890

   DOL Vehicle ID  2020 Census Tract
0        0.545747          -0.299420
1        3.468452          -0.299432
2       -1.786402           1.285967
3        3.446703          -0.300098
4        3.474566          -0.425159
```

Apply Z-score on imputation strategy

```
Normalized Dataset (Imputed Dataset) Preview:
   Postal Code  Model Year  Electric Range  Base MSRP  Legislative District  \
0     0.083517    0.682369       -0.115408  -0.118877              0.404805
1     0.055406   -0.990402        1.122189  -0.118877              0.404805
2    -0.031820   -0.321294        2.427910  -0.118877              0.941935
3    -0.021898   -2.328619        0.361464  -0.118877              1.143359
4     0.173223   -1.324956        2.109995  -0.118877             -0.602314

   DOL Vehicle ID  2020 Census Tract
0        0.179974           0.036545
1        3.420452           0.036539
2       -1.590342           0.035233
3       -1.658778           0.035229
4       -0.711472           0.056641
```

Applying Z-score normalization centers data around 0, ideal for models that assume normal distribution. Using it with imputation keeps all data, capturing patterns, while with drop rows, it reduces dataset size and potential noise.

---

2. Exploratory Data Analysis:

2.1. Descriptive Statistics: Calculate summary statistics (mean, median, standard deviation) for numerical features.

```
Descriptive Statistics (Imputed Dataset):
                      Mean (Imputed)  Median (Imputed)  \
Postal Code            -4.104795e-15         -0.021759
Model Year             -1.736501e-14          0.318288
Electric Range          5.977397e-17         -0.581822
Base MSRP              -2.238143e-17         -0.117290
Legislative District   -4.132788e-16          0.206146
DOL Vehicle ID          2.894034e-17          0.160761
2020 Census Tract       4.632077e-15          0.034636

                      Standard Deviation (Imputed)
Postal Code                               1.000002
Model Year                                1.000002
Electric Range                            1.000002
Base MSRP                                 1.000002
Legislative District                      1.000002
DOL Vehicle ID                            1.000002
2020 Census Tract                         1.000002
```

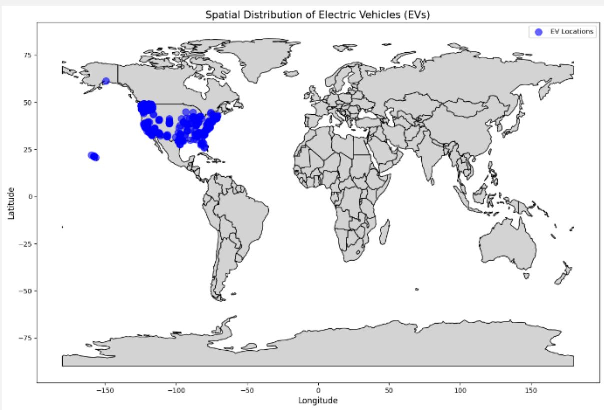
```
Descriptive Statistics (Dropped Rows Dataset):
                      Mean (Dropped Rows)  Median (Dropped Rows)  \
Postal Code                  5.054700e-15              -0.458524
Model Year                   1.476510e-14               0.317978
Electric Range               4.635101e-17              -0.581781
Base MSRP                   -1.456940e-17              -0.117191
Legislative District        -2.439527e-17               0.205898
DOL Vehicle ID               1.843198e-17               0.160690
2020 Census Tract           -5.286618e-14              -0.425295

                      Standard Deviation (Dropped Rows)
Postal Code                                    1.000002
Model Year                                     1.000002
Electric Range                                 1.000002
Base MSRP                                      1.000002
Legislative District                           1.000002
DOL Vehicle ID                                 1.000002
2020 Census Tract                              1.000002
```

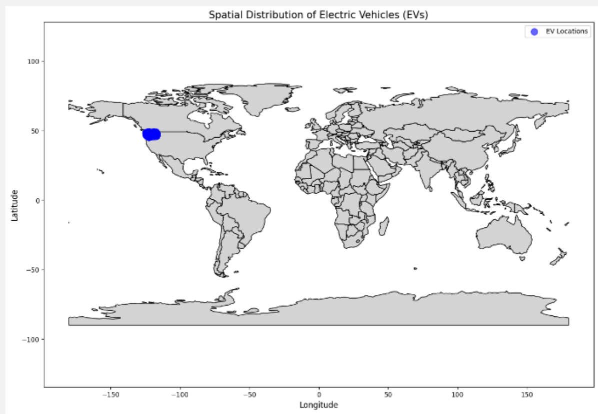
The descriptive statistics for the imputed and dropped rows datasets provide key insights into numerical features. Both datasets have been standardized which resulting in a mean close to 0 and a standard deviation of 1 for each feature. The slight differences in mean and median between the imputed and dropped rows datasets suggest minor variations in central tendency due to the different handling of missing data. These statistics confirm that both datasets are normalized, ensuring comparability across features.

2.2. Spatial Distribution: Visualize the spatial distribution of EVs across locations.

Spatial distribution on imputation strategy

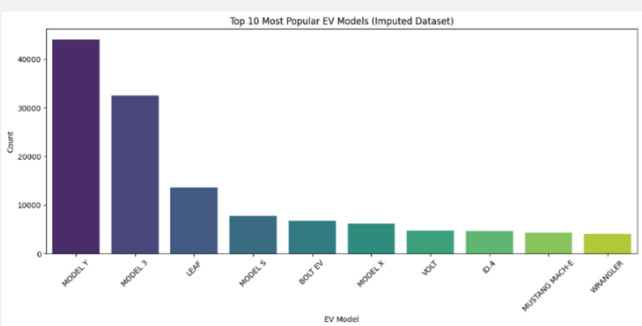


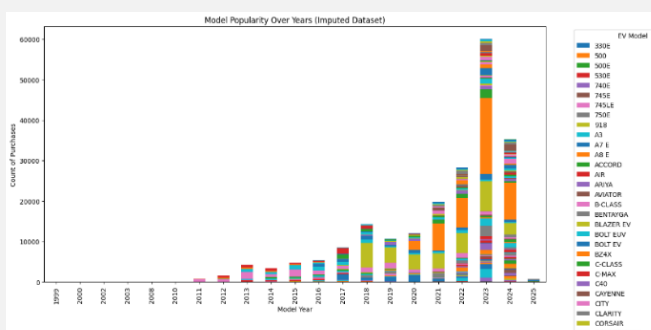
Spatial distribution on drop rows strategy:



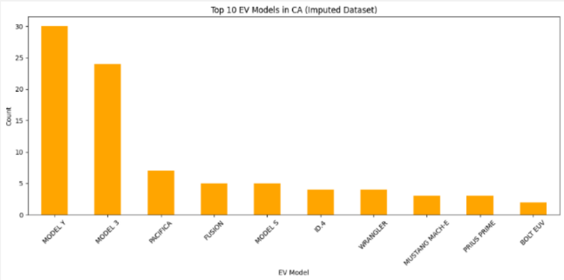
The spatial distribution maps show the impact of imputation and drop rows strategies on EV data locations. The imputed dataset displays a wider spread of EVs across regions, preserving data completeness. In contrast, the drop rows strategy results in a concentrated cluster in a limited area, indicating reduced geographic diversity. This shows how imputation maintains a broader spatial perspective, while dropping rows can limit regional analysis insights.

2.3. Model Popularity: Analyze the popularity of different EV models (categorical data) and identify any trends.



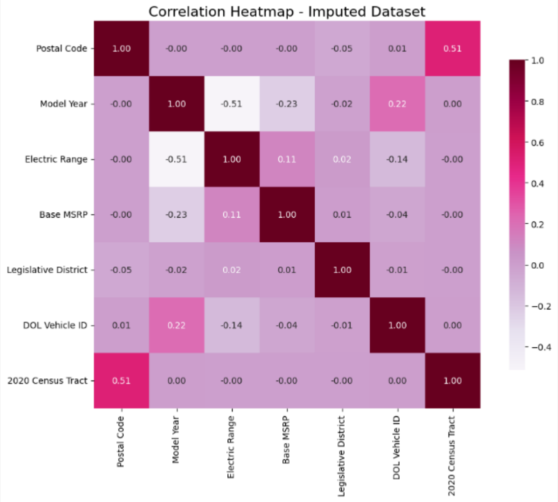


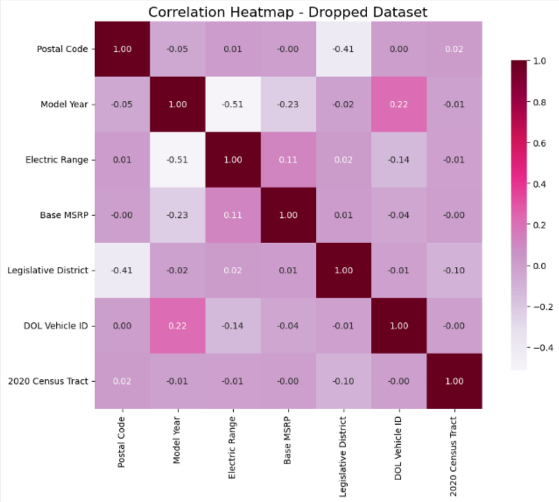
The bar charts display the top 10 EV models, while stacked bar charts show model popularity from 2015 onward, comparing the dropped rows and imputed datasets. Both strategies yield similar trends, with Model Y and Model 3 as the most popular.



We noticed that CA data is missing in the dropped rows dataset but is preserved with imputation, as shown in the chart. Imputation retains important regional data for a fuller analysis.

2.4. Investigate the relationship between every pair of numeric features. Are there any correlations? Explain the results.





The correlation heatmaps for the imputed and dropped datasets show only a few relationships between numeric features. Both heatmaps show a moderate negative correlation (-0.51) between Model Year and Electric Range, which suggests that newer models may have a greater electric range. In the imputed dataset, Postal Code has a moderate positive correlation (0.51) with 2020 Census Tract, but this is less visible in the dropped dataset because there are fewer data points. Overall, most feature pairs have low correlation values, which means the numeric features are mostly independent.

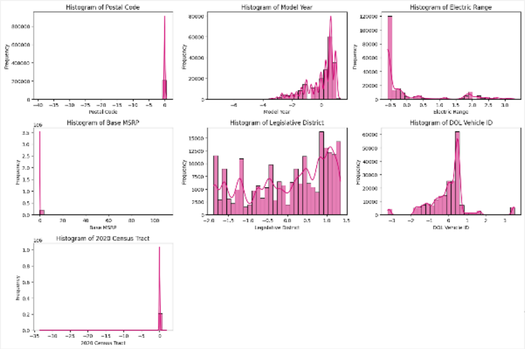
---

3. Visualization

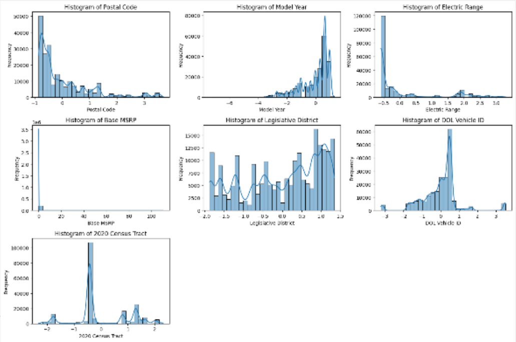
3.1. Data Exploration Visualizations: Create various visualizations (e.g., histograms, scatter plots, boxplots) to explore the relationships between features.

- **Histograms**

Histogram for imputed dataset:



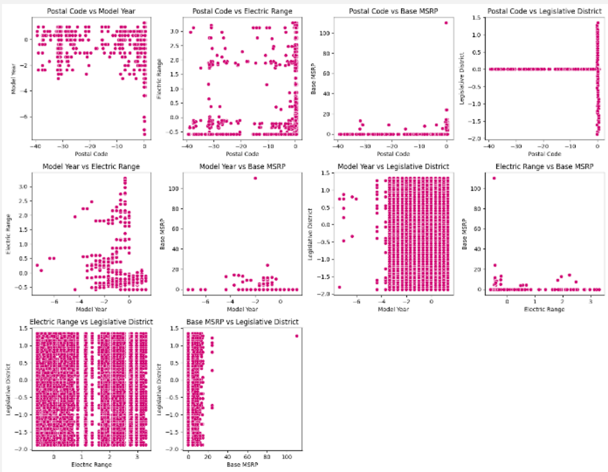
Histogram for dropped dataset:



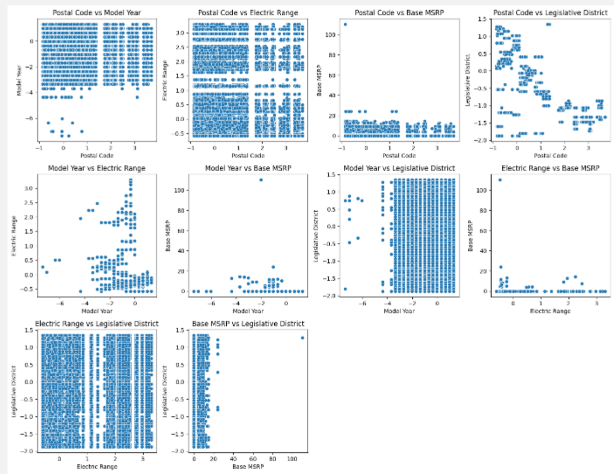
The histograms are similar overall, but the dropped dataset shows more bars in Postal Code and 2020 Census Tract because it retains more unique values. In contrast, the imputed dataset has fewer bars, as filling missing values reduces variation.

- **Scatter plots**

Scatter plots for imputed dataset:



Scatter plots for dropped dataset:

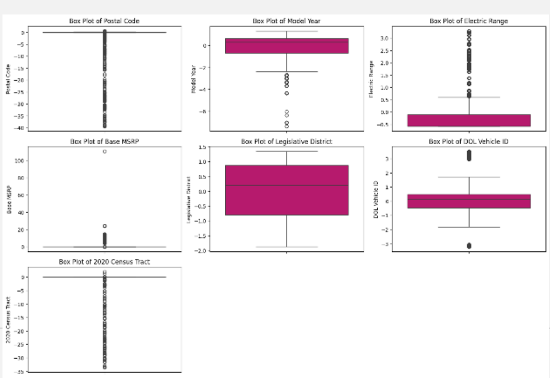


The scatter plots for both datasets show similar patterns, but the dropped rows dataset has more scattered points, especially in pairs like Postal Code vs. Model Year and Electric Range vs. Postal Code. This difference because imputation fills gaps creating a more organized distribution, while dropping leaves more variation due to removed entries.

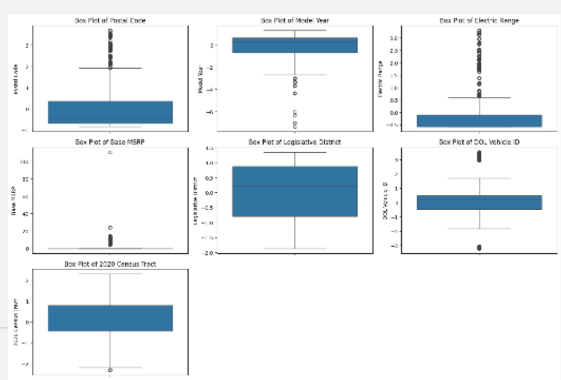
The scatter plot analysis shows that EV registrations concentrate in certain regions, evident from the clustering in Postal Code vs. Legislative District, which suggests regional preferences or policy influences. A positive relationship between Model Year and Electric Range indicates that newer models tend to offer better ranges. Conversely, Electric Range vs. Legislative District reveals no notable variation across districts. Outliers in the Electric Range vs. Base MSRP plot point to high-end EVs with premium prices, highlighting vehicles with extended ranges at higher costs.

- **Boxplots**

Boxplots for imputed dataset:

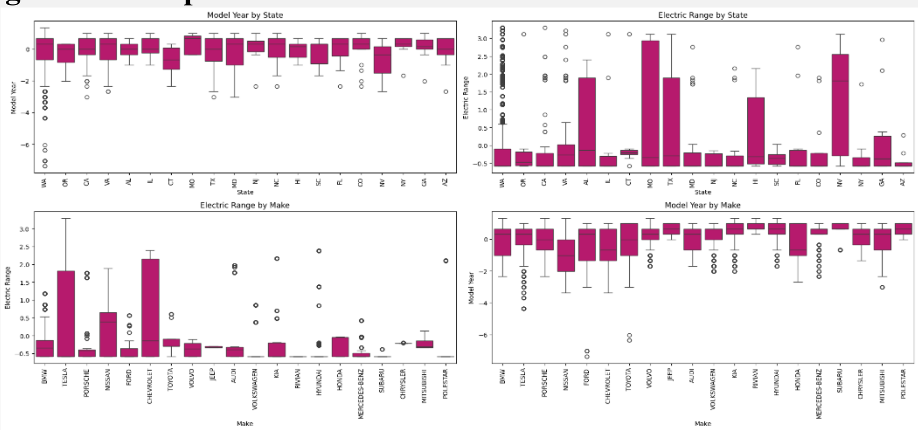


Boxplots for dropped dataset:



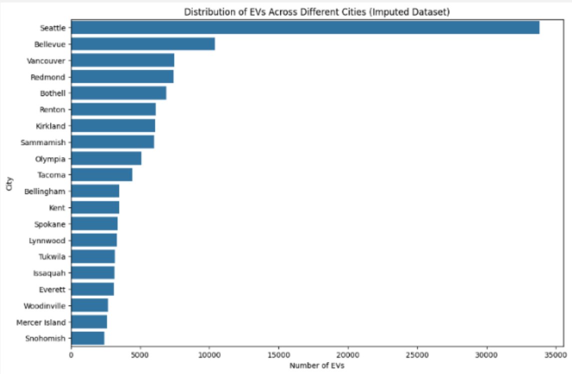
The box plots show tighter distributions in the imputed dataset for features like Postal Code and 2020 Census Tract, as imputation fills missing values for consistency. Dropping rows increases variation by removing incomplete entries, widening the spread in some features. This difference shows how imputation can create a more uniform dataset, while dropping rows preserves natural variation as we saw before.

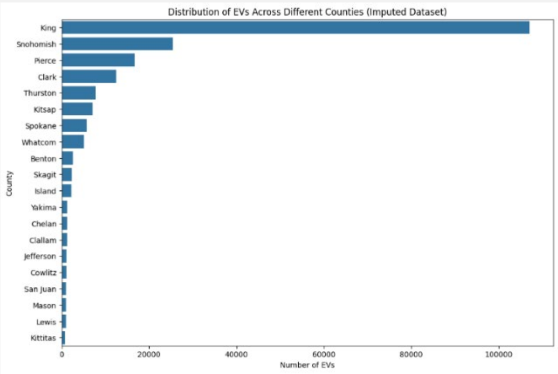
Exploring Relationships Between Vehicle Features:



In this analysis, boxplots help show the relationships between key features of electric vehicles. We compared metrics like Model Year and Electric Range by State and Make. For example, the range varies widely across states and makes, with some states and brands having vehicles with much higher ranges. Similarly, Model Year distributions reveal differences in how recent vehicles are across various states and brands. These visualizations provide a clearer view of how vehicle characteristics vary by location and manufacturer.

3.2. Comparative Visualization: Compare the distribution of EVs across different locations (cities, counties) using bar charts or stacked bar charts.

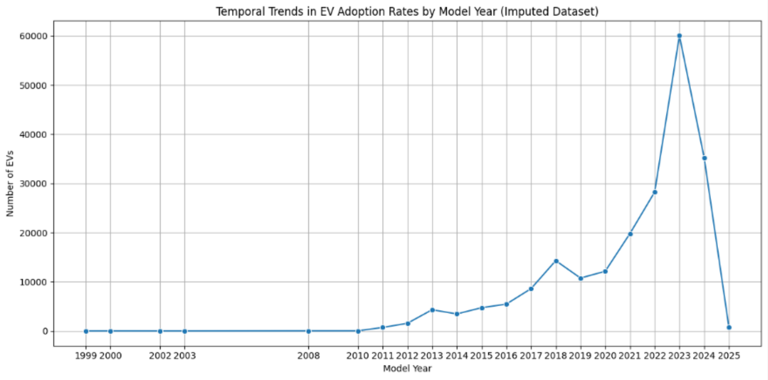


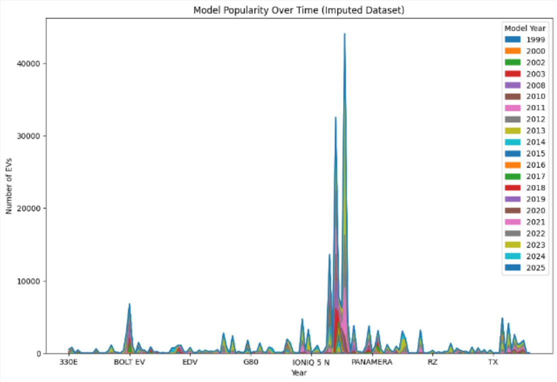


The bar charts show similar EV distribution patterns across cities and counties in both imputed and dropped datasets, with no notable differences. Seattle leads in EV counts among cities, followed by Bellevue and Vancouver, highlighting strong urban adoption. At the county level, King County dominates, with Snohomish and Pierce counties next, suggesting higher EV adoption in urban and suburban areas.

4. Additional Analysis:

4.1. Temporal Analysis: If the dataset includes data across multiple time points, analyze the temporal trends in EV adoption rates and model popularity.





The temporal analysis graphs show similar trends for both imputed and dropped datasets, with no noticeable differences. The first chart reveals a sharp increase in EV adoption starting around 2018, peaking in 2023, reflecting rapid growth in recent years. The second chart displays model popularity over time, showing a few standout models dominating at certain points, particularly from 2020 to 2023.