Electric Vehicle Data Analysis Assignment - Questions & Answers

Section 1: Data Cleaning Questions

Q: How many missing values exist in the dataset, and in which columns?

A: Use df.isnull().sum() to count missing values per column. For numeric columns where 0 represents missing (like Base MSRP or Electric Range), also check (df['col'] == 0).sum().

Q: How should missing or zero values in the Base MSRP and Electric Range columns be handled?

A: Options: Replace with mean/median (imputation), drop rows if too many are missing, or use external data. For MSRP, external pricing sources may be helpful.

Q: Are there duplicate records in the dataset? If so, how should they be managed?

A: Check with df.duplicated().sum(). If duplicates are exact matches, remove them using df.drop_duplicates().

Q: How can VINs be anonymized while maintaining uniqueness?

A: Apply a hashing function (e.g., SHA-256) or generate surrogate unique IDs. This preserves uniqueness while hiding sensitive info.

Q: How can Vehicle Location (GPS coordinates) be cleaned or converted for better readability?

A: Round to 3–4 decimal places, or use reverse geocoding to map coordinates to county/city names for better interpretation.

Section 2: Data Exploration Questions

Q: What are the top 5 most common EV makes and models in the dataset?

A: Group by Make and Model, then count occurrences. Use df.groupby(['Make','Model']).size().sort_values(ascending=False).head(5).

Q: What is the distribution of EVs by county? Which county has the most registrations?

A: Group by 'County' and count entries. Visualize with a bar chart to show the distribution. The county with the maximum count has the most EVs.

Q: How has EV adoption changed over different model years?

A: Group data by Model Year and count vehicles. Plot as a line chart to show growth trends over time.

Q: What is the average electric range of EVs in the dataset?

A: Calculate df['Electric Range'].mean(). Exclude zeros or missing values for accuracy.

Q: What percentage of EVs are eligible for Clean Alternative Fuel Vehicle (CAFV) incentives?

*A: Calculate: (eligible_count / total_count) * 100. Use df['CAFV Eligibility'].value_counts(normalize=True).*

Q: How does the electric range vary across different makes and models?

A: Group by Make and Model, then calculate mean and median range. Compare using boxplots or bar charts.

Q: What is the average Base MSRP for each EV model?

A: Group by Model and compute df.groupby('Model')['Base MSRP'].mean().

Q: Are there any regional trends in EV adoption (e.g., urban vs. rural areas)?

A: Compare counts across counties or zip codes. Urban areas usually show higher adoption rates than rural areas.

Section 3: Data Visualization Questions

Q: Create a bar chart showing the top 5 EV makes and models by count.

A: Use matplotlib or seaborn barplot after grouping data by Make & Model.

Q: Use a heatmap or choropleth map to visualize EV distribution by county.

A: Map counts per county using geopandas or plotly choropleth maps.

Q: Create a line graph showing the trend of EV adoption by model year.

A: Plot year vs count of EVs with plt.plot(). This shows adoption growth over time.

Q: Generate a scatter plot comparing electric range vs. base MSRP to see pricing trends.

A: Scatter plot with x=Base MSRP, y=Electric Range. Add regression line to see relationship.

Q: Plot a pie chart showing the proportion of CAFV-eligible vs. non-eligible EVs.

A: Use df['CAFV Eligibility'].value_counts().plot.pie().

Q: Use a geospatial map to display EV registrations based on vehicle location.

A: Use folium or geopandas to plot latitude/longitude on a map for spatial distribution.

Section 4: Linear Regression Model Questions

Q: How can we use Linear Regression to predict the Electric Range of a vehicle?

A: Fit a regression model with Electric Range as dependent variable, using features like Model Year, Base MSRP, Make.

Q: What independent variables (features) can be used to predict Electric Range?

A: Features include Model Year, Base MSRP, Vehicle Make/Model (encoded), and possibly CAFV eligibility.

Q: How do we handle categorical variables like Make and Model in regression analysis?

A: Convert them into numerical form using one-hot encoding or label encoding.

Q: What is the R^2 score of the model, and what does it indicate about prediction accuracy?

A: R^2 shows how much variance in Electric Range is explained by the features. Closer to 1 means better fit.

Q: How does the Base MSRP influence the Electric Range according to the regression model?

A: Check regression coefficients. A positive coefficient indicates higher MSRP correlates with longer range.

Q: What steps are needed to improve the accuracy of the Linear Regression model?

A: Feature scaling, adding interaction terms, removing outliers, or trying advanced models like Random Forest or Gradient Boosting.

Q: Can we use this model to predict the range of new EV models based on their specifications?

A: Yes, but predictions are only reliable if new EV specs are similar to training data. Extrapolation to unseen types may be inaccurate.