Assignment 1

Luis Dale Gascon
Computer Science
Towson University
lgascon1@students.towson.edu

I. Introduction

Gasoline is both expensive and environmentally damaging, making fuel efficiency an important concern for consumers and manufacturers. Understanding how vehicle features such as horsepower, weight, acceleration, and displacement can influence miles per gallon (mpg) will allow for informed decision-making and promote sustainability. This paper aims to identify the key predictors of mpg by constructing a linear regression model while experimenting with different pre-processing methods and evaluating its performance using established metrics.

II. Dataset

The dataset is a modified version from Carnegie Mellon University's StatLib dataset [1]. Each row contains the following vehicle information:

- horsepower: Enging power output
- weight: Vehicle mass
- acceleration: Rate at which velocity changes over time
- displacement: Total volume of air and fuel an engine can displace
- model_year: A vehicle's production period

These features are used to predict a vehicle's miles per gallong (mpg), which is the focus of our regression model and is the prediction question.

We split the training and testing dataset via 70/30 where we take 70% of the dataset as the training data and 30% as the testing data to avoid an overfitted model.

III. Exploratory Data Analysis

A. Missing Values

6 rows from the dataset were missing horsepower values (about 1% of 398 total rows). We opted to remove these rows to maintain data quality with minimal loss.

B. Scatter Plots

Scatter plots were generated to explore relationships between the target variable: mpg and individual features to visualize their patterns.

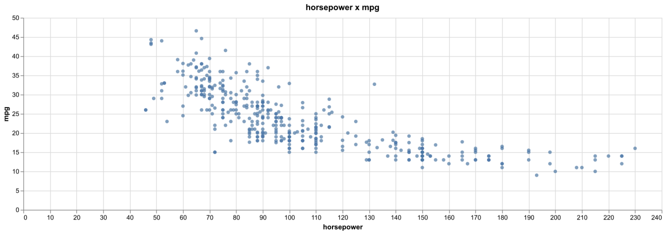

Fig. 1. Mpg decreases as horsepower increasese

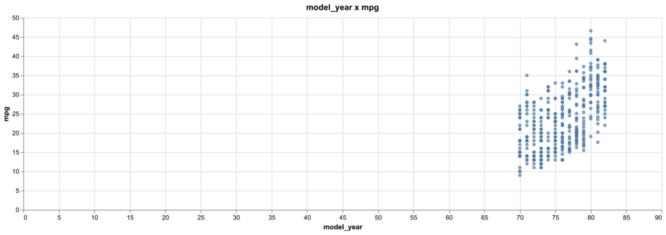

Fig. 2. Mpg increase as model year increases

IV. Data Preprocessing

A. Encoding

The number of cylinders, the model year and the origin of the car are all integer variables but they weren't continuous and had little cardinality, except for the model year, so we figured we could apply ordinal encoding for these features to preseve the meaningful order in the data. Transforming these features should improve the model's performance.

B. Normalizing numerical features

We want to set a standard value for numerical values to allow features with different magnitudes to contribute the same impact. We standardized the numerical variables: displacement, horsepower, and acceleration using the z-score method.

C. Data Binning

We plan to experiment with 2 binning techniques: Equal Width and K-Means Clustering

For the inital binning technique, we want to find the optimal number of bins via Sturges' Rule: $\lceil \log_2 n + 1 \rceil$, where n is the height of our dataset. Our training set has 274 rows, so we plug that value into Sturges' rule and we get a value of 9.

For the second binning technique. We tested a range of clusters from 2 to 10. We plotted each cluster's Within

Cluster Sum of Squares or WCSS. The equation for WCSS is as follows:

$$\sum_{i=1}^{k} \sum_{x\in C_i} \| x - \mu_i \|^2$$

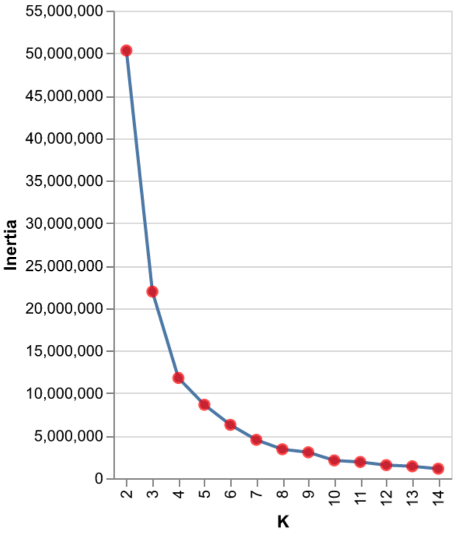

Fig. 3. Elbow plot

Looking at the plot, we can estimate that 4 would be the optimal value of clusters to set for K-means. We utilized Sklearn's KBinsDiscretizer to perform both types of binning by setting the keyword argument strategy to uniform or kmeans.

The data was binned using one-hot encoding, which resulted in 8 extra features added to the dataset using the equal width binning technique and 3 extra features added using the clustering binning technique.

D. *Transforming a feature with non-normal distribution*

Looking at the histogram for the feature of displacement, the data is positively skewed. To mitigate the impact of outliers and improve the performance of our regression model, we apply 3 different kinds of data transformation independently to compare their results.

I applied natural log, square root and inverse square root transformations to the dataset invidually. To evaluate how each transformation affected normality, I used a normal Quantile-Quantil (QQ) plot to visually compare the quantiles of the transformed data to those of a normal distribution.

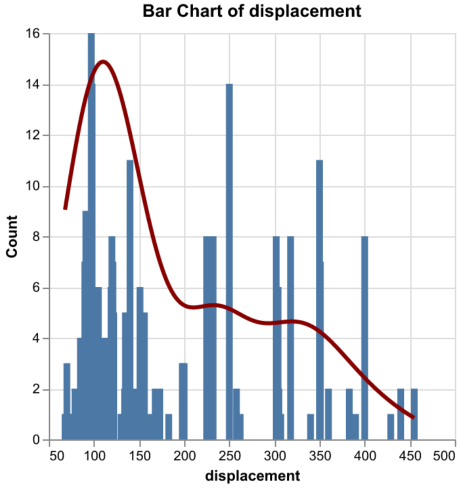

Fig. 4. Displacement shape via kernel density estimation

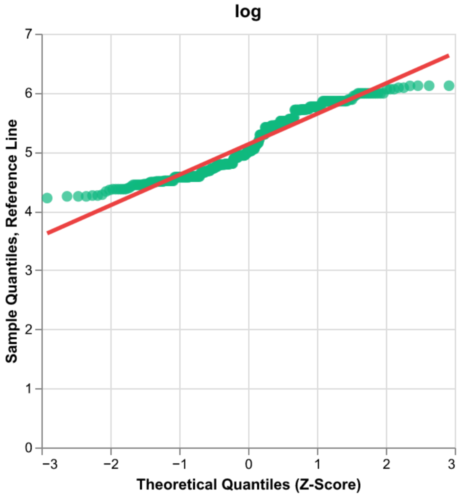

Fig. 5. Log transformation normal QQ plot

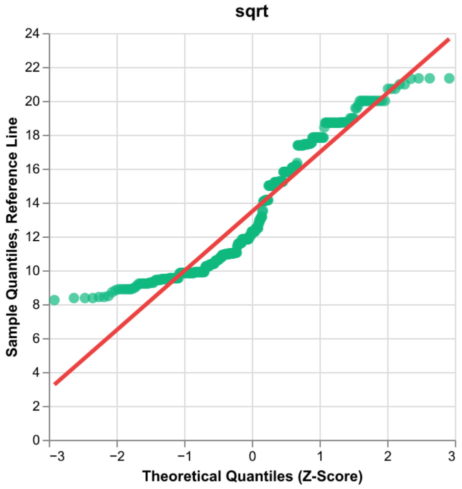

Fig. 6. Square root transformation normal QQ plot

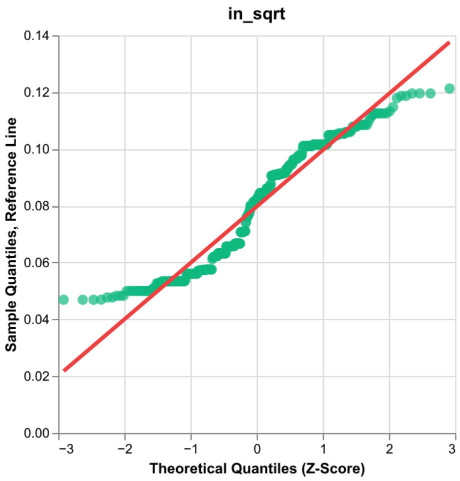

Fig. 7. Inverse square root transformation normal QQ plot

Visually, the normal QQ plot from a log transformation is the closest to matching the normal distribution, so for our data pre-processing step, we'll apply log transformation to the displacement feature.

V. Regression Analysis

To evaluate our model, we look at 3 metrics: R^2 score, and mean absolute error (MAE)

Metric	Score
R^2	0.81
MAE	2.56

R^2 score, also known as coefficient of determination is a measure of how predictable the target variable is based on the features of the dataset. We can say that 81% of the variation of the target variable, mpg, can be explained by the features of the training data. As for MAE, on average, the model's prediction are off by 2.56 mpg.

We wanted to see if the binning method chosen made a difference in the model's performance, and it did. Equal width binning gave us the R^2 score that we see in the table, while binning by clustering gave us an R^2 score of .79

VI. Conclusion

Our analysis shows how a vehicle's features are significant predictors of miles per gallon. By experimenting with different data preprocessing methodologies such as normalization, binning and distribution transformation to improve model performance, we achieved a linear regression model with a respectable R^2 score and low mean absolute error, which indicates our model can predict a vehicle's mpg fairly well.

In comparing distribution transformation, we found that log transformation gave us a result that closely resembles a normal distribution through comparing the normal QQ plot of each transformation. This improvements in normality is a benefit to the regression model since linear models perform much better when the features more closely resemeble a normal distribution and with outliers mitigated. It made sense that log transformation was the most effective as displacement is a postively skewed feature.

References

[1] R. Quinlan, "Auto MPG." 1993.