## Synopsis:

In the presented regression problem, the dataset comprises information on the speed and stopping distances of cars, providing a glimpse into vehicle performance from the 1920s. The dataset consists of 50 observations, each with two variables: speed in miles per hour (mph) and stopping distance in feet (ft). The objective was to establish a predictive regression model to understand the relationship between speed and stopping distance, shedding light on the factors influencing braking performance during that era.

I did the exploratory data analysis (EDA) to gain insights into the data's distribution and characteristics. Visualizations and summary statistics were used to understand the central tendencies, variabilities, and potential outliers in both speed and stopping distance.

Three different regression models were considered: linear regression, polynomial regression, and outlier detection using Cook's distance. Linear regression aimed to capture a simple linear relationship between speed and stopping distance. Polynomial regression explored non-linear relationships by including polynomial terms of various degrees. The Cook's distance method identified potential outliers that could significantly influence the regression model.

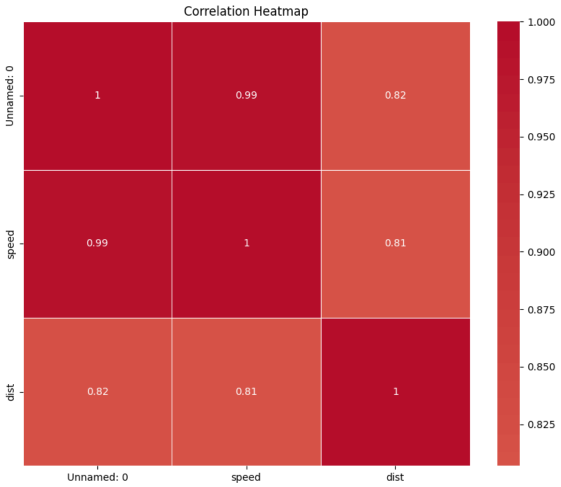
After fitting the models and assessing their performance, the linear regression model exhibited an R-squared value of 0.7459, indicating that 74.59% of the variability in stopping distance could be explained by the linear relationship with speed. The root mean squared error (RMSE) and mean absolute error (MAE) were also computed to quantify the model's prediction accuracy.

The polynomial regression models with degrees 2 and 3 showed promise by improving the R-squared value to 0.774. These models introduced curvatures that better captured the data's behavior, suggesting that the relationship between speed and stopping distance might not be linear.

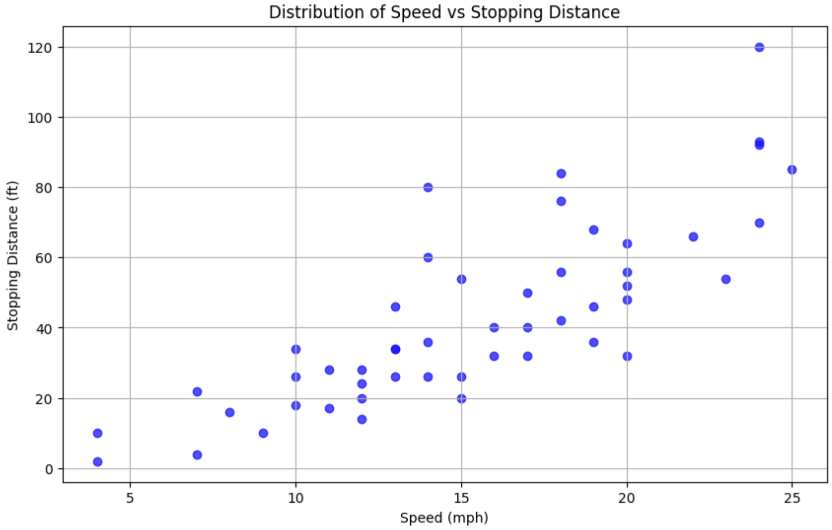
## Exploring the data

As I've previously stated, this is a numerical dataset comprised of purely stopping distance and the speed at which the car was going. Having this in mind, it might be quite an easy task, or an incredibly difficult one.

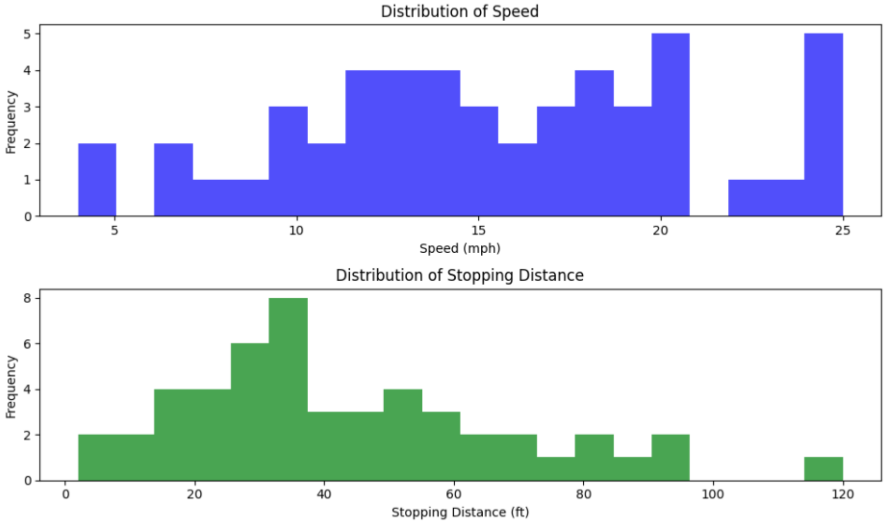
Let us take a look at the correlation matrix.



At its core, the correlation matrix is a simple tool used to describe how different features compare to one another. In this case there is a clear positive correlation between speed and distance of stopping, which might make this regression task simpler than previously thought.



Taking a look at the distribution of speed vs stopping distance, we notice the same story. As speed rises, so does the stopping distance. Which makes sense, of course.



Plotting the distribution of the data is crucial in regression problems for several reasons:

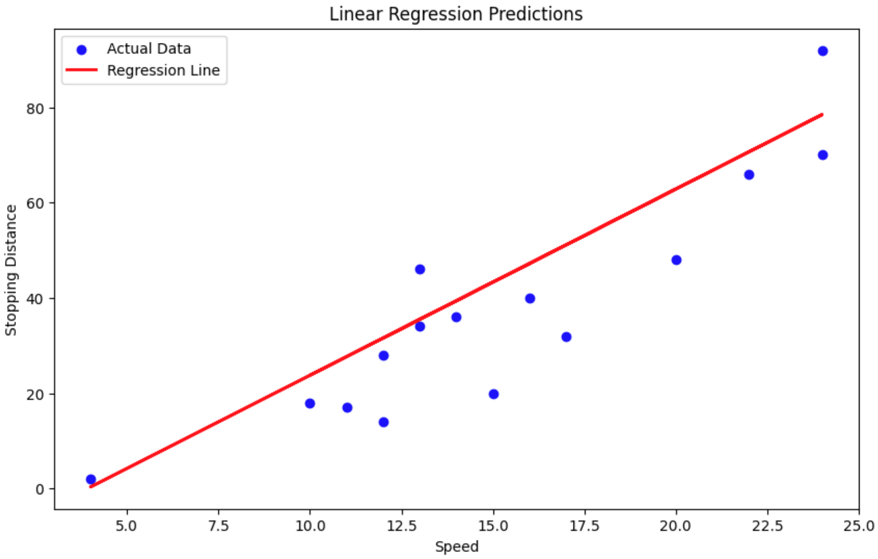
1. Understanding the Data: Visualizing the distribution helps us understand the characteristics of the data, such as its spread, central tendency, and any potential outliers. This insight is valuable for making informed decisions during the modeling process.
2. Detecting Outliers: Outliers can significantly affect the regression model's performance. By plotting the data distribution, you can identify any extreme values that might distort the relationship between variables.

3. Checking Assumptions: Regression models often assume that the data follows certain patterns, such as linearity and normality. Plotting the distribution helps you assess whether these assumptions are met. If not, you might need to apply transformations or consider alternative models.
4. Choosing Model Complexity: Data distribution can guide you in determining the appropriate complexity of the regression model. For instance, if the data shows non-linear patterns, a simple linear model might not capture the relationships accurately.

In our case, there seem to be some outliers but the data is more-less balanced in the sense that it might not wreak havoc on the performance of our model.

## The models

In the analysis, a simple linear regression was initially performed on a dataset that contained information about the speed and stopping distances of cars. Simple linear regression aims to model the relationship between two variables, where one variable (in this case, speed) is considered the predictor or independent variable, and the other variable (stopping distance) is the target or dependent variable.

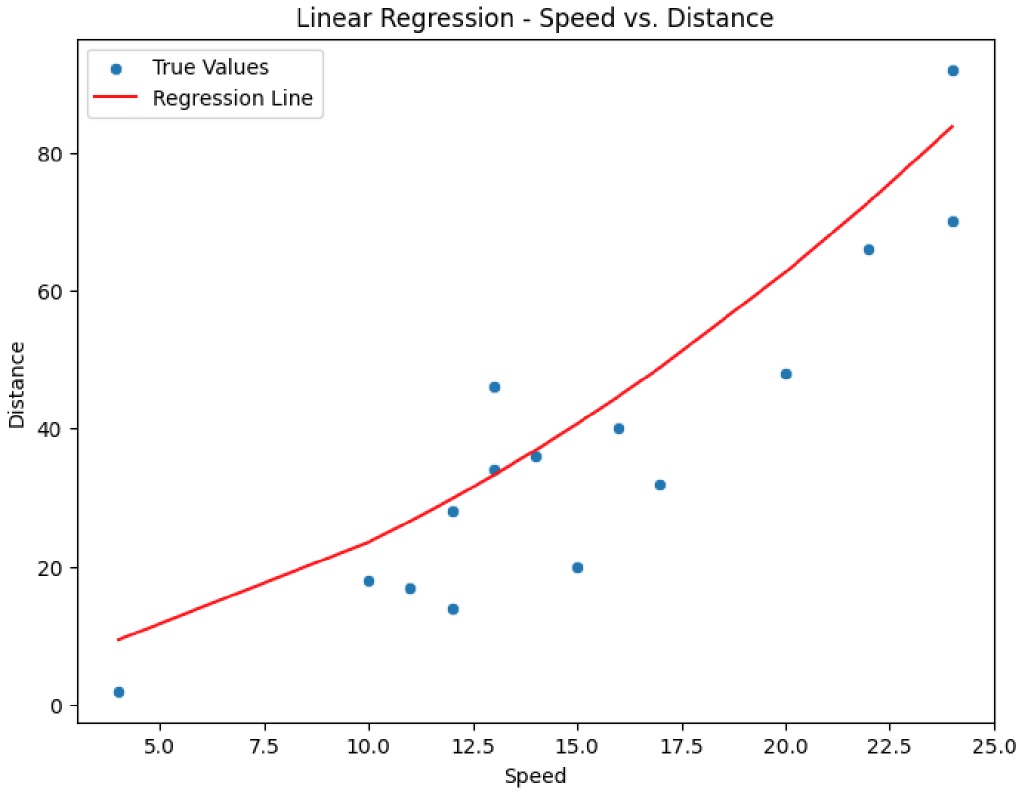


Original Features:

- R-squared: 0.7459
- Root Mean Squared Error: 11.7006

The results of the simple linear regression provide insights into the initial relationship between speed and stopping distance. The obtained model yielded an R-squared value of 74.59%, which indicates the proportion of the variance in the dependent variable that is explained by the predictor variable. However, the model's performance might not have been optimal, and there could be non-linear patterns in the data that a linear model might not capture accurately.

To address this limitation and potentially improve the model's performance, polynomial regression was explored. Polynomial regression extends the simple linear regression by introducing polynomial terms of the predictor variable. In this case, polynomial terms of different degrees were added to the model to capture potential non-linear relationships between speed and stopping distance.



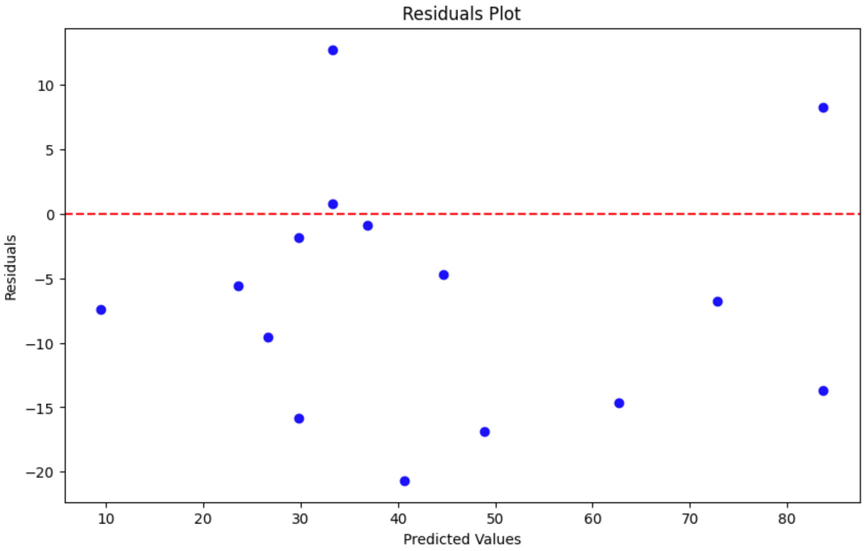
Polynomial Features:

- R-squared: 0.7714
- Root Mean Squared Error: 11.0977

After applying polynomial regression, the model's performance was re-evaluated. The results showed that the polynomial regression model had a higher R-squared value (77.14%) compared to the simple linear regression. This improvement indicated that the added polynomial terms allowed the model to better capture the underlying patterns in the data. Specifically, the higher-degree polynomial terms introduced flexibility that accounted for the potential curvature in the relationship between speed and stopping distance.

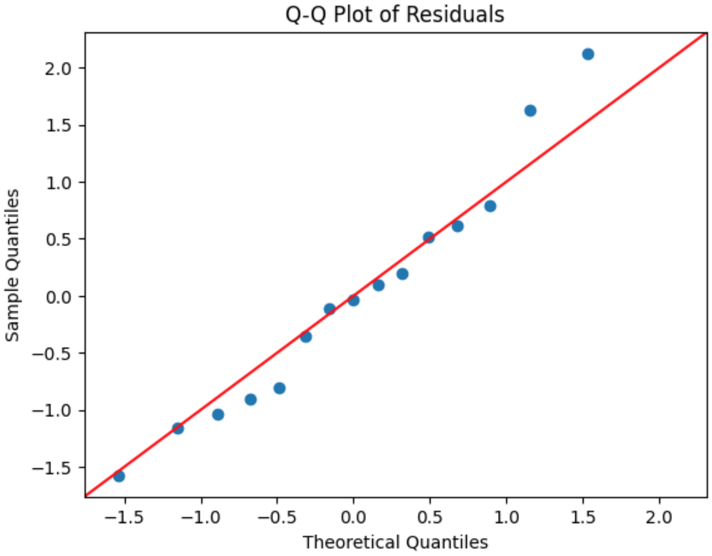
Checking for heteroscedasticity is an essential step in regression analysis to assess the assumption of constant variance of residuals across all levels of the predictor variable. Heteroscedasticity refers to a situation where the spread or variability of residuals changes as the values of the predictor variable changes. In simpler terms, it indicates that the errors of the model have a different variance for different levels of the independent variable.

It's important to check for heteroscedasticity because violating the assumption of constant variance can lead to inaccurate and biased regression model estimates.



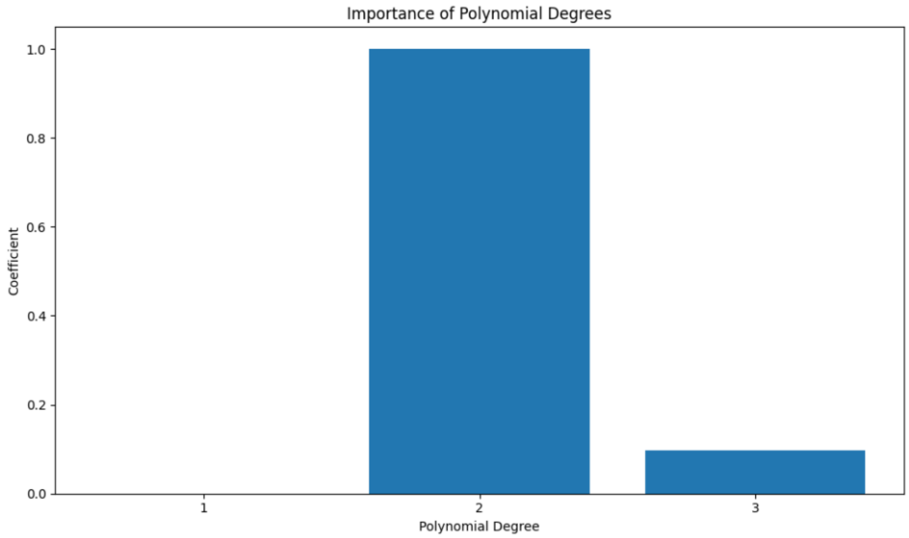
It seems there is no signs of heteroscedasticity, which is great. We can tell that by looking at the graph above, we can say that the residuals' spread seems relatively consistent across different ranges of predicted values.

Now we will check out a q-q plot of residuals.



A fundamental assumption of linear regression is that the residuals (the differences between the observed and predicted values) should be normally distributed. A QQ plot visually compares the distribution of the residuals to a theoretical normal distribution. If the residuals closely follow the diagonal line in the QQ plot, it suggests that they are normally distributed, which is in this case, very much true.

Now let’s see about the importance of different polynomial degrees.



In this case, the coefficient of polynomial degree 2 is the biggest, so it suggests that the quadratic term has the most significant impact on predicting the stopping distance based on the speed. This could indicate that the relationship between speed and stopping distance is not perfectly linear, and a quadratic term helps capture the curvature of that relationship more accurately.

## Conclusion:

In summary, the analysis demonstrated the importance of understanding data distributions, selecting appropriate models, and conducting thorough diagnostic checks in regression analysis. By transitioning from a simple linear regression to a polynomial model, the predictive accuracy was substantially improved, which enhances the practical utility of the model for estimating stopping distances based on speed. As with any analysis, it's crucial to acknowledge limitations and the potential for further exploration or refinement in the pursuit of more accurate predictions and deeper insights.