# Capstone Car Dealership Analysis

![carsin](./img/carsin.jpg)

**## BUSINESS UNDERSTANDING**

Used Car Lot TM is a newly established local business in the United States. Requested some information about the **top make and model of cars** **that sold the most in the last year nationally.** From this analysis we intend on building a multiple linear regression model to predict what brand names will make the list.

The car lot intends on using this information to ensure their business objectives, marketing, and sales teams' concentrations are on point.

The information obtained will also be used to educate their staff and sales team.

**## DATA PREPARATION**

Exploratory Statistical analysis used:

* Missing data
* Null values
* Duplicates
* Creating data frame
* Creation of Dummy Variables
* Categorical Variables
* Conversion of datatypes
* Complex lists and dictionaries
* Statistical values to determine high correlation and significance
* Multicollinearity of Independent variables
* Reviewed for Normality of dataset

**## MODELLING**

The data science process used to build this model is the CRISPM process, visual diagram below.

![CRISPDM](./img/CRISPDM.jpg)

Modelling analysis used:

* Mean Standard Error – MSE
* Polynomial Regression using an r2 score from log transformation – (fit/overfit)
* Cross Validation
* Kernel Density Estimates
* Partition dataset to validate model using train-test splits or k-fold cross validation.
* Log Transformation

**Types of Graphical representations include:**

* Scatter Plots
* Histograms
* Heat Map

**## CAPSTONE PROECT KEY DELIVERABLES**

**##TWO major features sharing *Strong Relationships* with the Used Cars Dataset**

1.

2.

**##THREE Important Statistics**

1.

2.

3.

**## Final Evaluation**

**Points to review**:

* The data reflects homoscedasticity and the residuals **should** follow a normal distribution
* From this analysis we can infer that XYZ on the basis of
* The predictions are
* My non-bias review is:
* Strong connections observed are:
* (Chart type) maps the validity of ZYC
* I determine that the car sales in the USA for the following year are projected to be:

Make:

Model:

**## MULTIPLE LINEAR REGRESSION INSTRUCTIONS**

1. Import the required libraries and read the dataset
2. Data munging and processing actions
3. Exploratory Data Analysis (EDA)
   * Data Visualization
4. Check for normality
   * Density plots
   * Q-Q plots
5. Multiple linear regression model
   * Train test split
   * Train the model
   * Fit the model
   * Make predictions
   * Plot the results
6. Residual analysis
   * Remove autocorrelation with varying lag values
   * Check for the normality of the variables
   * Train and fit the model
   * Make predictions and plot the results
7. Symbolic regression model
   * Create a model
   * Train the model
   * Fit the model
   * Make predictions and plot the result

**## FOR MORE INFORMATION**

Please review our full analysis in the Jupyter Notebook[Jupyter Notebook] (./Capstone.ipynb) or our [Capstone](./presentation.pdf).

For more information contact \*\*Carla Kirby, [ra\_carlajoy@yahoo.com\*\*](mailto:ra_carlajoy@yahoo.com**)