Predicting car prices based on various factors

Josh Herr, Naveen Goyal, Dallas Hutchinson, Sai Kiran Butti

Introduction

Since they were invented cars have played a major role in global transportation, and every year the impact and market for cars grows even further. As a result there is a large market for used cars as an alternative to buying newer, generally more expensive cars. Within the used car market there is a clear incentive for both parties, buyer and seller, to know exactly what the car is worth so they can seek to earn the highest amount of money they can. However, there are a plethora of factors that contribute to the value of the car and the market can change quickly based on the supply or demand. For example, during the COVID-19 pandemic a major supply chain failure led the supply of new cars being produced to dramatically decrease. This decrease in supply dramatically increased the demand for new cars and therefore the value. In this paper we describe several models we built to try and predict the value of a used car based on some of the available factors. We made these models to assist both buyers and sellers to feel confident that they are paying or receiving the proper amount for the vehicle they are interested in.

Several other interesting questions can be answered based on how the factors of the model interact. For example, in a paper written by Sallee et al. the authors ask, what is the impact of a car's fuel economy on consumer demand. This was a question that we sought to analyze with our model as well. While in the paper by Sallee et al. the authors used a fairly similar dataset to the one we are using to build our models, it is important to note that the data is very different as well as how it was handled. Salle et al. used a dataset including millions of transactions at car auctions from 1993 to 2008. They also primarily focus on measuring the impact fuel economy has on consumer

choices, while in our model fuel economy will just be another factor to consider in pricing. In their paper Sallee et al. found that consumers fully value fuel economy, however they also note that their predictions for vehicle life is based on a single government study and that the consumers often undervalued fuel economy in extreme situations.

Another interesting question asked by other researchers studying the pricing for used cars is from Moresino who focuses on how quality labels for cars impacts prices in Switzerland. Consumers in Switzerland can reference three quality labels when purchasing a car, one for reliability, a second for fuel efficiency, and a third for safety (Moresino). Again it is important to consider the difference in scope between the model made in the paper and our model. Due to the author's focus on these labels offered exclusively in Switzerland the model would have to be extrapolated for use on cars in other countries, if those countries offer a similar service as the labels in Switzerland. This paper is very interesting to compare with ours because our model does not have factors to directly consider reliability and safety, which are some of the primary focuses for Moresino. In the paper Moresino chose to use a hedonic regression approach, and a stepwise regression to select his variables. As a result he found that the Swiss reliability label did have a significant impact on the price of the car. However, due to multicollinearity problems he could not conclude whether the fuel efficiency label or the safety label had a significant impact.

Another interesting paper to compare with our model is from Pal et al. who create a Random Forest model to predict used car prices. A Random Forest model is also one of the models we make to predict used car prices, so it is interesting to compare the

accuracies of both models, and to consider what influences and differences between the models. Pal et al. used a dataset from Kaggle which contained data about used car sales in Germany. Their dataset included more than three hundred seventy thousand data points with twenty factors, which they later reduced down to ten factors by removing the factors they felt were irrelevant. For their random forest model, Pal et al. used five hundred trees in the model by increasing in increments of fifty, and after training they were able to achieve a test accuracy of just under eighty-four percent.

Methodology:

Dataset Background:

https://www.kaggle.com/datasets/avikasliwal/used-cars-price-prediction?select=train-data.csv

This dataset talks about the prices of used cars being sold in India based on various factors like owner type, fuel type, location, kilometers driven etc. The target is to find out the best prediction for the car prices. The number of observations is 6019.

List of available variables:

Index	index
Name	brand and model of the car
Location	location of where the car is being sold
Year	year of the car
Kilometers_Driven	total KM driven in the car by previous owners
Fuel_Type	type of fuel used (Diesel/Petrol/Electric/LPG/CNG)
Transmission	type of transmission used (Automatic/Manual)
Owner_Type	whether ownership is first, second, etc
Mileage	standard mileage offered by the car in km/kg or kmpl

Engine	displacement volume of the engine in CC
Power	maximum power of the engine in bhp
Seats	number of seats in the car
New_Price	price of a new car of the same model
Price	price of the used car in INR Lakhs

Table 1. Variable list

Data exploration, Preprocessing and Cleaning:

The raw dataset includes 14 columns, few of them were taken out as being unrelated to the purpose of this project. Columns which were excluded: Name, New_Price, Index

After removing the unnecessary columns, we needed to prepare the dataset which includes replacing commas with decimals, dropping the rows which contain NA values, and converting data types of certain variables. For continuous variables, strings had to be stripped like “CC” and “bhp” and then converted to numeric type. Variables of categorical nature such as Location, Fuel Type, Transmission, and Owner Type were converted to factor types. The seats variable was initially a value from 2 to 10 but we noticed a high majority of 5 seats and converted the variable to a factor where group one is “<=5” and group two is “6+”. This decreased the influence the few cars with unusual seat numbers had. The mileage column contained units of both km/kg and kmpl, however, after being unable to identify a guaranteed conversion from km/kg to kmpl, the few rows which were km/kg were removed from the dataset. This likely had to do with the combustion type of the car.

Other methods used included checking for potential outliers and the distribution of each variable. Histograms were useful for identifying a normal distribution of

continuous variables and we used boxplots to observe differences in Price between categories of factor variables. The ggpairs() function created a useful pairplot for us to check the linear relationships between continuous variables. Figure 1 below is a heatmap-correlation matrix which reveals Pearson's correlation coefficient between all numeric variables, including our target variable. We did note the potential multicollinearity at play with a couple relationships, notably Power/Engine and Price/Power.

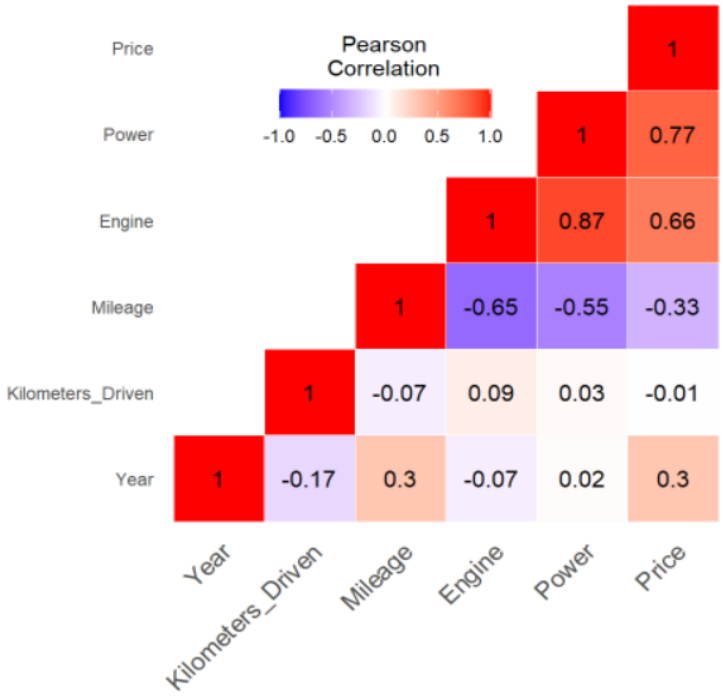

Fig. 1. Pearson's correlation matrix

Results:

Linear Regression Model

One of the models we implemented to understand the car price factors was the linear regression model. We used Price as a response variable and other few variables as predictors. The predictor variables are: Year, Mileage, Power, Transmission, Location

and Engine. We obtained a significantly high R-Squared value (0.721). Few plots were created in order to find out the linearity. Based on the VIF outcomes, there was multicollinearity present. Below are the few plots created to verify the linearity and confirm the multicollinearity based on VIF matrix:

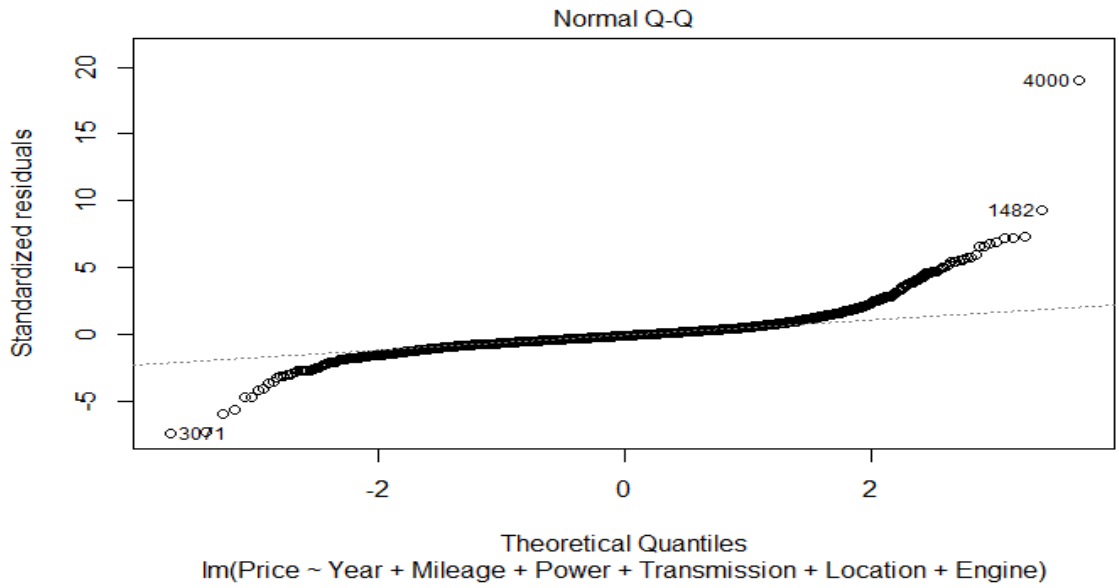

Fig 2. Normal Q-Q plot for price prediction

By looking at the Residuals vs Leverage all the observations were distributed symmetrical from the Cook's distance. So, it can be confirmed that based on these variables the linear regression model is overall a good fit.

Random Forest Model

Random Forest Regression is a supervised ensemble learning method which can be used for regression techniques. It constructs many decision trees in parallel based upon a random set of data points in the training set. Random forest models are generally more accurate since predictions are made by averaging across all constructed trees. One of the other main reasons our team wanted to utilize a random forest

technique is that the algorithm handles predictors of different types well. Since our predictors contain both continuous and factor/categorical types, random forest could lead to better accuracy.

In order to optimize our model, we used a for loop to iterate through and build a random forest model at levels of 'ntrees' from 2, 4, 8, etc to 2048 trees. This baseline is later optimized further by finding the lowest MSE for surrounding values of 'ntrees'. At each level of 'ntrees' a model is built using all predictors and the log transformation of Price as the target considering the variable's right-skewed distribution. An out-of-bag score is calculated at each step and stored to compare which model performed the best. A similar approach is taken to find the best value of 'mtry' which is the number of features to consider at each split point. After this, we are left with an optimal 'ntrees' of 250 trees and 'mtry' of 6 features.

Fitting the optimized model and predicting on our test set produces the following results. The obtained RMSE value is 1.02 INR Lakhs (approximately $1,300 USD) with a R^2 value of 0.9376. The random forest model seems to fit the data points very well, especially when the price falls below 30 INR Lakhs (~ $39,000).

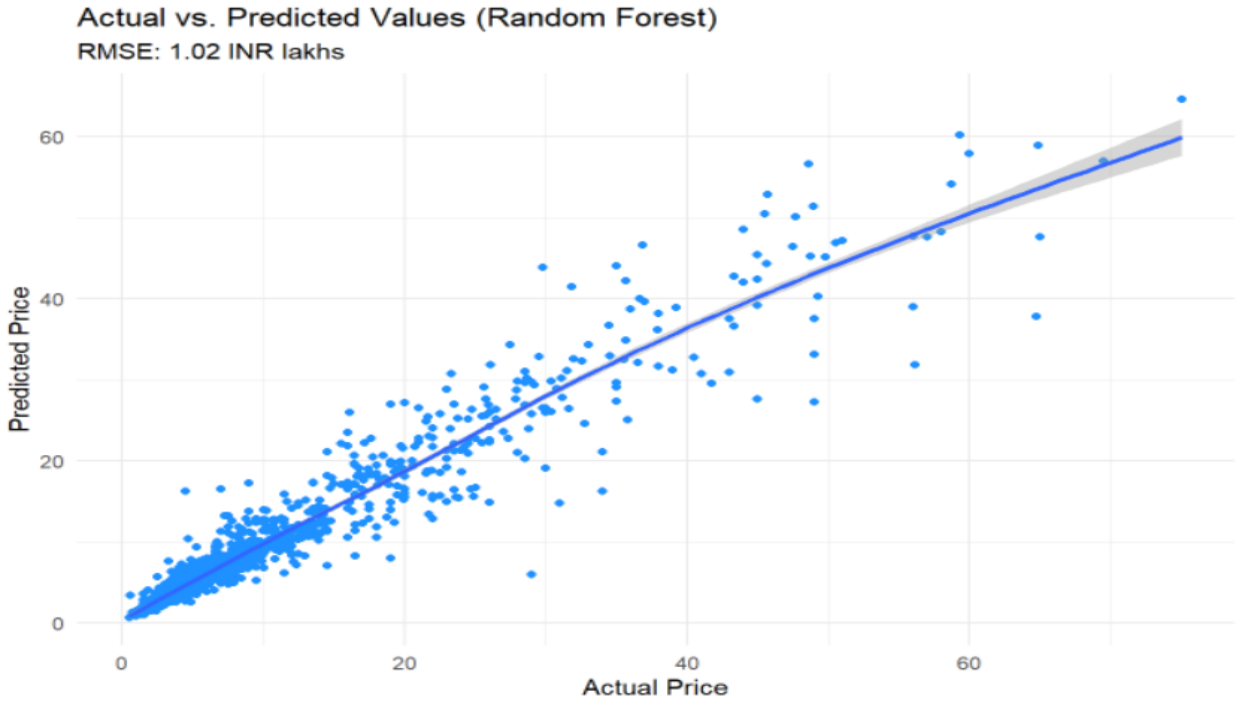

Fig. 3. Random Forest Actual vs. Predicted plot

Next, to find which variables were most influential in price determination we used the varImpPlot() function from the randomForest package to visualize feature importance. Specifically for %IncMSE, measuring mean decrease in accuracy as certain variables are removed, the Year and Location variables stand out as most impactful.

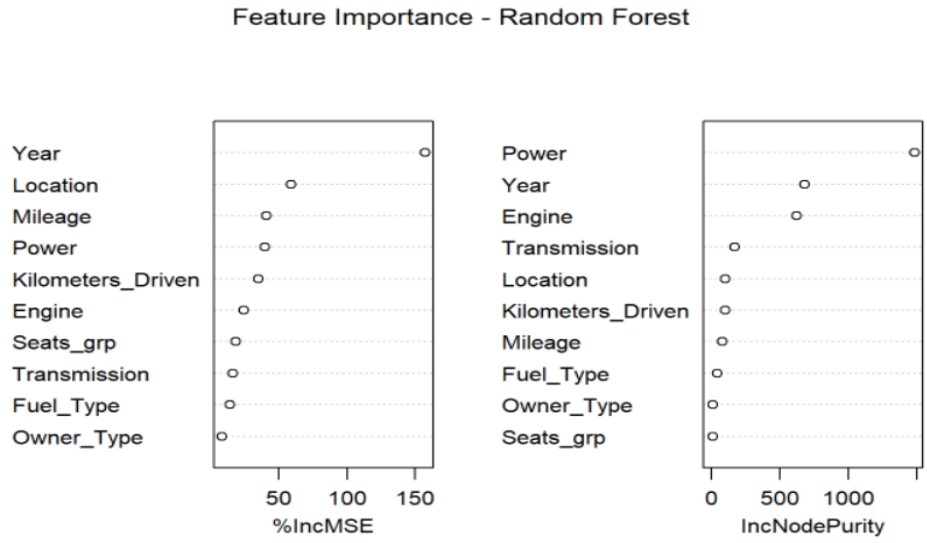

Fig. 4. Random Forest feature importance plot

Support Vector Regression Model

The third model we implemented to predict the used car price is the Support vector regression model. Support vector machine is a supervised learning method which can be used for both classification and regression. Support vector machine breakdown the data points into different parts in a n dimensional space using a boundary line known as the hyperplane. The points which are near to the hyperplane are known support vectors. The distance which separates the hyperplane and the support vector is margin. The best hyperplane will have the maximum margin. The major advantage of the Support vector machine is that it is memory efficient.

To improve the performance of the model we have encoded the following columns: Location, fuel type, transmission, owner type and seats group using one hot encoding approach. After encoding these columns, we have built the SVM model using these new columns and Year, Kilometers Driven, Mileage, Engine as the predictor and Price as the response. After building the model we predicted the values using the test data set and the r2 score for the predicted values came up to be 0.78 and RMSE was 5.65. Below is the plot between the actual and predicted values.

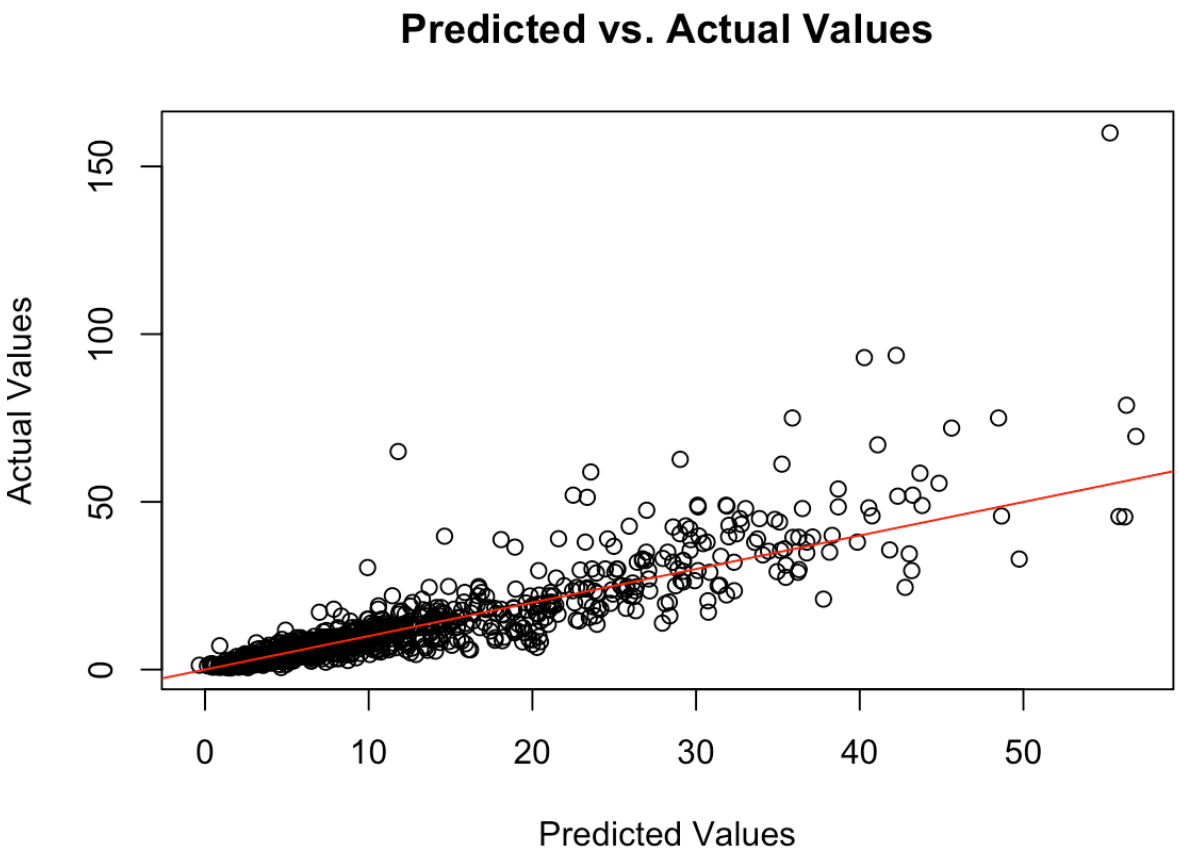

Fig. 5. SVM predicted vs. actual values

The table below defines the differences in performance of our three regression models on the test set. Linear regression seemed to perform best with less variables used while SVM used one hot encoding and had many more variables. The random forest regressor led to our highest R^2 value and lowest test error rate followed by SVM and linear regression.

Model	**R^2**	**RMSE**	**Num. variables used**
Linear Regression	0.72	6.433	6
Random Forest	0.9376	1.024	10
SVM	0.78	5.656	25

Table 2. Model comparison

Discussion

Our objective when making these models was to be able to accurately predict the value of used cars, and based on the results shown in Table 2 we were able to accomplish this task. The R^2 for the random forest model is by far the strongest at just under 0.94 but the linear regression model and the support vector machine model are both quite strong with a R^2 of 0.72 and 0.78 respectively. For users who only want to predict the value of a used car the random forest model will be the best, but for users who want to understand more about the contribution that factors have to the valuation the much more interpretable linear regression model may be the best.

Despite the strength of the models, there are still several ways that changes in the dataset could greatly improve each of the models. The most obvious change would be to add additional variables, for example a variable to account for vehicle reliability. This variable was shown to be significant for cars sold in Switzerland in the paper by Moresino. Another way the dataset can change to make models stronger is to broaden the scope of the dataset and increase the sample size. Our dataset contained just over six thousand observations from in and around India, therefore the predictions made with our models using this data may be unreliable when used in other countries and regions.

For future researchers our models prove that a strong model can be created to predict the value of used cars in India. In conjunction with the other research discussed at the beginning, which included models based on data from Germany and Switzerland, there is reason to believe an accurate global model can be made or a model for each region on Earth.

References

Moresino, Francesco. "A hedonic approach to estimate the price of reliability, energy efficiency and safety for new cars in Switzerland." American Journal of Industrial and Business Management 9.3 (2019): 468-481.

Pal, Nabarun, et al. "How much is my car worth? A methodology for predicting used cars' prices using random forest." Future of Information and Communication Conference. Springer, Cham, 2018.

Sallee, James M., Sarah E. West, and Wei Fan. "Do consumers recognize the value of fuel economy? Evidence from used car prices and gasoline price fluctuations." Journal of Public Economics 135 (2016): 61-73.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Greg Snow (2020). TeachingDemos: Demonstrations for Teaching and Learning. R package version 2.12. https://CRAN.R-project.org/package=TeachingDemos

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2022). dplyr: A Grammar of Data Manipulation. R package version 1.0.8. https://CRAN.R-project.org/package=dplyr

Hadley Wickham (2019). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.4.0. https://CRAN.R-project.org/package=stringr

Barret Schloerke, Di Cook, Joseph Larmarange, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg and Jason Crowley (2021). GGally: Extension to 'ggplot2'. R package version 2.1.2. https://CRAN.R-project.org/package=GGally

A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

Adam Petrie (2020). regclass: Tools for an Introductory Class in Regression and Modeling. R package version 1.6. https://CRAN.R-project.org/package=regclass

David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel and Friedrich Leisch (2021). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-9. https://CRAN.R-project.org/package=e1071

Ben Hamner and Michael Frasco (2018). Metrics: Evaluation Metrics for Machine Learning. R package version 0.1.4. https://CRAN.R-project.org/package=Metrics

Max Kuhn (2022). caret: Classification and Regression Training. R package version 6.0-91. https://CRAN.R-project.org/package=caret

Hadley Wickham (2007). Reshaping Data with the reshape Package. Journal of Statistical Software, 21(12), 1-20. URL http://www.jstatsoft.org/v21/i12/.

Lüdecke et al., (2021). performance: An R Package for Assessment, Comparison and Testing of Statistical Models. Journal of Open Source Software, 6(60), 3139. https://doi.org/10.21105/joss.03139

Chaitanya Sagar (2017). Building Regression Models in R using Support Vector Regression. KDnuggets. https://www.kdnuggets.com/2017/03/building-regression-models-support-vector-regression.html

Chaya Bakshi (2020). Random Forest Regression. Level Up Coding. https://levelup.gitconnected.com/random-forest-regression-209c0f354c84

ggcorrplot: Visualization of a correlation matrix using ggplot2. Statistical tools for high-throughput data analysis.

http://www.sthda.com/english/wiki/ggcorrplot-visualization-of-a-correlation-matrix-using-ggplot2#correlation-matrix-visualization

Barret Schloerke (2022). Pairs plot with ggpairs. R CHARTS

https://r-charts.com/correlation/ggpairs/

Noel Bambrick, AYLIEN (2016). Support Vector Machines: A Simple Explanation. KDnuggets.

https://www.kdnuggets.com/2016/07/support-vector-machines-simple-explanation.html

Zach (2021). How to Perform One-hot Encoding in R.

https://www.statology.org/one-hot-encoding-in-r/