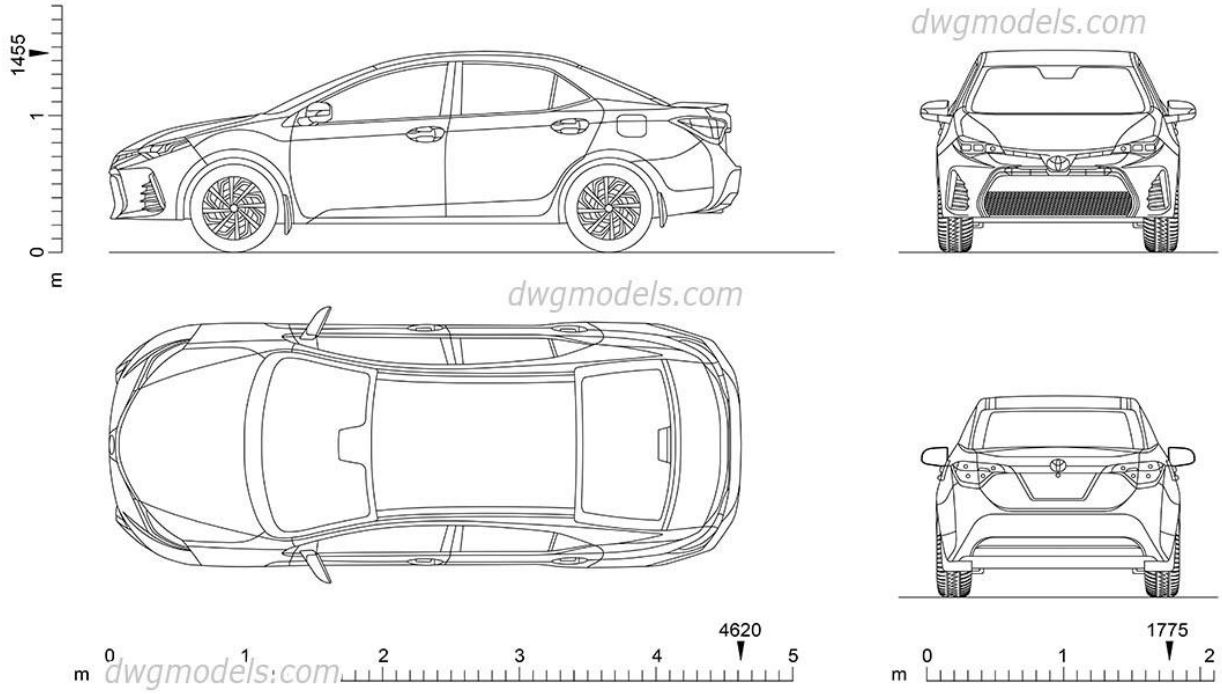
Making Toyota Corolla Resale Values Transparent:
CSDA 1000 Group 4 Project Proposal



Vaibhav Rastogi

Table of Contents

Abstract

As data becomes readily available, many avenues of research allow for consumers to understand the fair value of their purchases. Aiding in alleviating the 'market for lemons', our aim was to determine a new model for second-hand car buyers to derive an unbiased fair resale value of Toyota Corollas. Our initial research concluded that we would be able to conduct this research incorporating factors which differentiate us from other vehicle price prediction companies within the market. Adopting the CRISP-DM methodology, we framed our research to answer whether variables such as the car's model year, the number of kilometers, engine capacity, and total weight of the car have a correlation that is strong enough to predict a fair price for any used Toyota Corolla vehicle. Our approach will complement existing services and tools that are publicly available by offering an unbiased perspective on fair pricing. The aim is to deploy a code representation of the model into an operating system to score or categorize new unseen data and to create a mechanism for the use of that new information in the solution of the original business problem. The model will treat new raw data in the same manner as during model development. Data prep and estimated association measures have been prepared for each variable for the fair pricing model. The results of which will be further analyzed for our final paper.

Introduction

The purchase of a car is an important life decision individuals make at least once in their lifetime. The large price tag often has consumers questioning whether the price of the used car is a true reflection of its value.

The main objective of our study is to provide an unbiased opinion of the price of different models of used TOYOTA Corollas using a linear regression model that incorporates four independent variables in the determination of price. We will be solving this problem using data analysis.

Initially, we started off our analysis by reviewing an array of websites with the same objective of providing a price tag or price range for the value of used cars based on a given list of variables. Our value proposition is based on these factors:

1) We offer an unbiased view as the team members have no affiliations with any Toyota Car Dealership for new or used cars;
2) Our methodology is premised on the CRISP-DM approach; and,
3) Our model is simple and transparent and includes an evaluation of our results/findings.

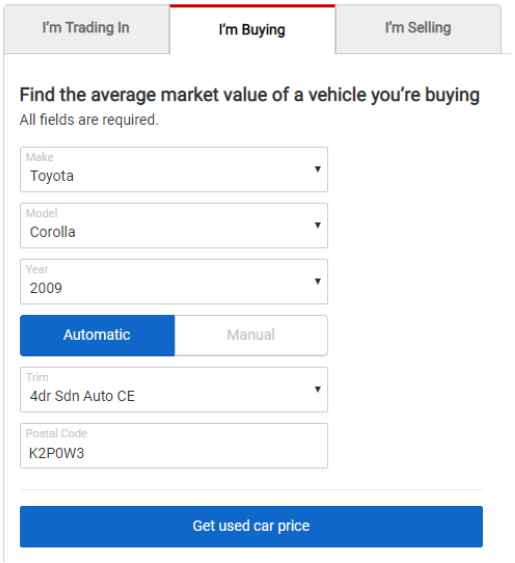
The purpose of our study is to provide a new model for second-hand car buyers to find an unbiased opinion on the fair resale value of Toyota Corollas.

Our approach in this study is to build a model to predict the price of a Toyota Corolla using a set of variables from the Toyota Corolla Sales Data Set from America, for cars manufactured from 1998 to 2004.

Competitive Analysis

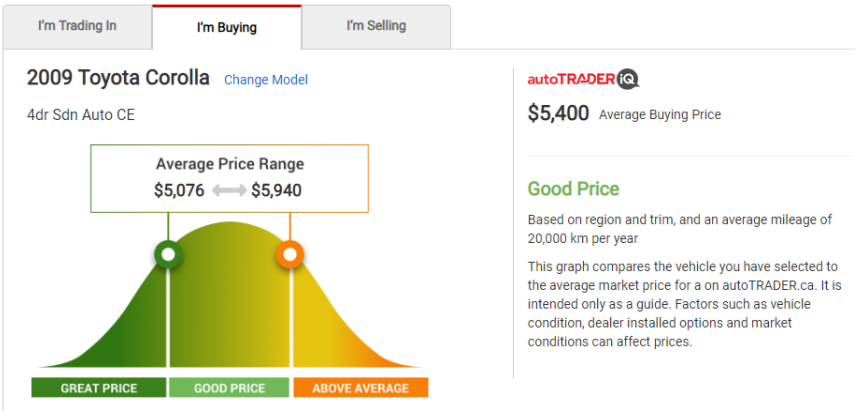
Several online companies, such as autoTrader.ca, Black Book Canada and Kelly's Blue Book in the United States offer online services that prospective car buyers and sellers can consult to assess fair purchase and resale values of used cars. Common parameters used to assess prices are: Make, Model, Year, Trim, Postal Code (location). Some sites also consider colour, condition, number of doors and other features. The viewer can select a particular combination of these features, as illustrated in Figure 1 below, generated by the autoTrader.ca website.

Figure 1: User interface and input variables for autoTrader.ca's online used car valuation tool.



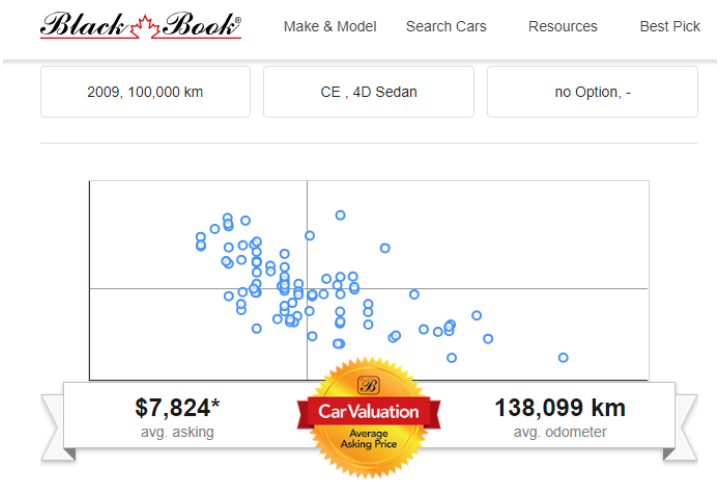
The models underlying the user interface on the site will then generate an average price and price range as illustrated in Figure 2 below.

Figure 2: AutoTrader.ca used car valuation tool output - average price and price range - based on selected input variables (AutoTrade, 2019).



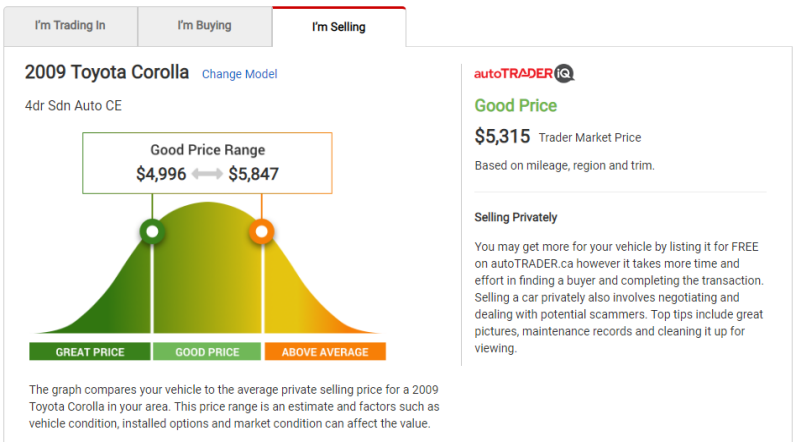
Comparable website Canadian Black Book generates a distribution of sale prices for vehicles with similar parameters, as illustrated below in Figure 3 generated Black Book's website. Importantly, the two sites' valuations for the average sale price of 10-year old Toyota Corolla based on selecting the same features (4-door sedan, automatic, same location) differ considerably from $5,400 (Auto Trader) to $7,824 (Black Book). In fact, Black Book's estimated average falls outside of Auto Trader's above average range.

Figure 3: Black Book Canada used car valuation output - average asking price and odometer reading - based on selected input variables (Book, 2019).



The Black Book site takes additional parameters into consideration (mileage, colour). The site also generates an additional output: average odometer reading. To test whether different input parameters explain the variation between site results, we can input Black Book's average odometer reading (138,099 km) as a parameter in the Auto Trader site by using a different view created for car sellers. The model output, an average price of $5,315 is similar to the Auto Trader price estimate for buyers but remains far below the Canadian Black Book estimate.

Figure 4: Auto Trader used car valuation tool output - average price and price range - based on selected input variables including average mileage.



A third site, CarFax Canada, estimates the resale value of a 2009 Toyota Corolla in Ontario as in a range between $4,315 and $7,302. Again, the Black Book average valuation falls outside of this range (CARFAX CANADA, 2019).)

Correlation between different features and price is implicit in these models. However, the workings of the models are not transparent to prospective buyers and sellers. Moreover, most of the sites that offer these tools also sell and/or purchase cars; and therefore may have an interest in the sale or purchase price.

Our approach will build on the valuation tools that are already publicly available by providing an unbiased model to assess a fair price based on different sets of features, and by making transparent the relationship between price and different features.

Proposed Plan

CRISP-DM or cross-industry data mining is a methodology used to provide a structured approach to data mining projects. It's a step-wise approach to solve any business problem that involves data mining and analysis.

Our team decided to adopt the CRISP-DM methodology to develop our predictive model. We used this data set to answer two simple questions:

1) Which predictor variables are to be used and why?
2) Which model (linear, exponential) are we going to use?

Business Understanding

This step involves understanding the project objectives and requirements from a business perspective and then converting this knowledge into a data mining problem definition and a preliminary plan.

After some preliminary research, our team observed that multiple websites use a variety of variables to determine the fair price of used vehicles. These variables include the car's model year, the number of kilometers the car was driven, engine capacity, and the total weight of the car. Our team's objective is to use the data set to test whether the variables mentioned above have a correlation that is strong enough to predict a fair price for any used Toyota Corolla vehicle.

Data Understanding

First, our team scanned the data set and identified the dependent and independent variables. Second, we attempted to clean the dataset of any observations that are outliers. We created

box plots for every variable to check how many outliers were observed and to determine if these outliers were statistically significant.

Finally, the team discussed how to determine the strength of correlation between the dependent and independent variables.

Data Preparation

The team prepared the data set to make it suitable for the model. This included a lengthy process of converting string data into binary data so it can be interpreted by the analysis tool and we can generate the statistics of the data set.

In order to be entirely objective about cleaning the data in the dataset, the number of unique values in each column were counted.

With Price being the dependent variable all other values in the table other than Model were considered possible independent variables. The Model and Fuel_Type columns are strings are separated into unique variables in the next section Creating Value Attributes.

For independent variables with values that are not nominal or ordinal values:

The Scatter plot & box plot graphs were utilized to reach the following:

- The 1st Quartile (Q1), 3rd Quartile (Q3) and Inter Quartile (IQ) were calculated;
- Using a lower boundary of Q1 – 3*IQ and an upper boundary of Q3 + 3*IQ the outliers were identified (SEMATECH, 2019);
- If the outlier was obviously due to incorrectly measured or entered data, the row was removed from the dataset;
- A regression model was created with the independent variable and the dependent variable; and,
- If the outlier did not affect the results of the regression graph, then the outlier rows were not removed from the dataset – if the outlier did affect the results of the regression graph then the rows with the outlier values were removed from the dataset (Grace-Martin, 2019).

The following section describes each variable in the dataset and its significance:

Age_08_04 - 77 unique values:

- This variable contained the age of the vehicle in number of months;
- No obvious outliers on a scatter plot or box plot;
- No values were outside of the lower or upper boundary; and,
- No rows removed from dataset.

Mfg_Month - 12 unique values:

- This variable is the month the vehicle was manufactured;
- There are 12 months in the year and this variable only contained the numbers between 1 and 12; and,
- No rows removed from the dataset.

Mfg_Year - 7 unique values:

- This variable contained the age of the vehicle in number of months;
- No obvious outliers on a scatter plot or box plot;
- No values were outside of the lower or upper boundary; and,
- No rows removed from dataset.

KM - 1263 unique values:

- This variable contained he odometer reading of the vehicle;
- No obvious outliers on a scatter plot but box plot indicated potential outliers;
- 2 rows outside of upper boundary of Q3 + 3*IQ = 219083.0, with 2 unique values: 243000 & 232940;
- Outliers do not affect assumptions and do not make any significant difference to the result; and,
- No rows removed from dataset.

HP - 12 unique values:

- This variable represents the horsepower of the vehicle;
- Both scatter plot and box plot indicated potential outliers;
- 11 rows outside of upper boundary of Q3 + 3*IQ = 170, all with the same value of 192;
- Outliers do not significantly affect results of linear regression;
- Research determined that a HP value of 192 is realistic for some models of Toyota Corolla (CARFOLIO, 2015); and,
- No rows removed from dataset.

CC - 13 unique values:

- This variable contains the cubic centimetres of the vehicle's engine;
- Both scatter plot and box plot indicated an obvious outlier;
- 1 row significantly outside of upper boundary of Q3 = 3*IQ = 2200 with a value of 16000;
- The value of 16000 is not a valid value for cc for a Toyota Corolla and must have been entered into the dataset incorrectly; and,
- The row containing this value was removed from the dataset.

Quarterly_Tax - 13 unique values

- Both scatter plot and box plot indicated potential outliers;
- 72 rows outside of below boundary of Q1 - 3*IQ = 21 all with a value of 19;
- 151 rows outside of upper boundary of Q3 + 3*IQ = 133 with 6 unique values;
- Outliers do not make any significant difference to the result;
- Unknown if the values are reasonable for this variable; and,
- No rows removed from dataset.

Weight - 59 unique values

- This variable represents the weight of the vehicle;
- Both scatter plot and box plot indicated potential outliers;
- 30 rows outside of upper boundary of Q3 + 3*IQ = 1220 with 10 unique values;
- Outliers do not significantly affect results of linear regression;
- Research determined that a weight value as high as the max value of 1615 is realistic for some models of Toyota Corolla (Carfolio, 2019b); and,
- No rows removed from dataset.

The remainder of the columns had only a few nominal or ordinal value with most having only 2 values: 0 & 1.

Creating Value Attributes

The Fuel_Type column contained 3 distinct values, 3 columns were added to the dataset to identify which of these values were applicable to the row. The column IF_Diesel was added and set to 1 when Fuel_Type was 'Diesel', the column IF_petrol was added and set to 1 when Fuel_Type was 'Petrol' and IF_cng was added and set to 1 when Fuel_Type was 'CNG'.

Analysis of the Model column in the dataset appeared to contain some distinct strings repeated on multiple rows that were not represented by other columns in the dataset.

New columns were added to the data set for 'D4D', 'TERRA', 'SOL', 'VVT', 'SPORT', 'VVTLI', 'COMF', 'LINEA', 'LUNA', 'WAGON'/'WAGEN', 'LIFT', 'HATCH', 'SEDAN' and 'VERSO', the new columns isD4D, isTERRA, isSOL, isVVT, isSPORT, isVVTLI, isCOMFORT, isLINEA, isLUNA, isWAGON, isLIFT, isHATCH, isSEDAN and isVERSA were defaulted with a value of 0 and populated with a 1 if the associated string was contained in the Model column.

Most values of Model contained a decimal number – either 2.0, 1.8, 1.6, etc. but it was clear that this value was represented by the 'cc' column so a new column was not created for this.

Data Visualization

Now that we have a nice clean dataset, the next step is exploratory data analysis to discover patterns, relationships and anomalies to inform our subsequent analysis. Using pair plots on various features of the dataset we discover a clear negative relationship between 'Price' and

'Age_08_04' and another negative relationship between 'Price' and 'KM'. We also see a relationship between 'Age_08_04' and 'KM'. Calculating the correlation coefficient between 'Age_08_04' and 'KM' shows only a moderate relationship between the two.[1]

Modeling

As a first step to explore the data and determine which parameters to model, we visualized selected pairwise relationships using seaborn pair plot and determine the correlation (slope and intercept of linear regression), conceptually illustrated in the table below. We then estimated association measures for each pair to inform our selection of variables for the fair pricing model. Results are included in the Tables and Figures section.

Table 1: Illustrative table showing pairwise correlations

	Price	**Age**	**KMs**	**Engine (cc)**	**Fuel type**	**Colour**
Price	1					
Age		1				
KMs			1			
Engine (cc)				1		
Fuel type					1	
Colour						1

Evaluation

Once one or more models have been built that appear to have high quality based on whichever loss functions have been selected, these need to be tested to ensure they generalize against unseen data and that all key business issues have been sufficiently considered. The end result is the selection of the champion model(s).

Deployment:

Generally, this will mean deploying a code representation of the model into an operating system to score or categorize new unseen data as it arises and to create a mechanism for the use of that new information in the solution of the original business problem. Importantly, the code representation must also include all the data prep steps leading up to modeling so that the model will treat new raw data in the same manner as during model development.

[1] Refer to table # titled 'Price, Age_08_04 & KM' for a visual representation.

You may well observe that there is nothing special here and that's largely true. From today's data science perspective this seems like common sense. This is exactly the point. The common process is so logical that it has become embedded into all our education, training, and practice.

Cost/Benefit Analysis

Consumer Reports, which offers an independent perspective to buyers, recommends that buyers use this type of tool to understand used cars' 'book value' prior to entering into negotiations regarding a specific purchase. Their website advises that prices can be affected by mileage, condition, trim level, optional equipment, and location of sale - but stops short of providing an independent valuation tool (Consumer Reports CR, 2014). The site also notes that there is both a retail price (the higher price you would expect to pay at a dealership) and wholesale price (the trade-in value to a dealer). The difference between these two prices is what makes a used car dealership profitable -- and the reason that providers of valuation tools who also sell and purchase used vehicles may not be wholly independent on price.

Our approach will complement existing services and tools that are publicly available by offering an unbiased perspective on fair pricing. We are a disinterested party, as we are neither sellers nor purchasers of Toyota Corollas. Moreover, we will make transparent the correlation of price with specific parameters. This increased transparency will enable buyers to consider trade-offs between specific features and price in making their purchasing decisions.

The modelling approach we are trialling here would be easily scalable in future to other vehicle types and in different markets to provide a more complete perspective on fair pricing for used vehicles. Potential future applications could be for a site like Consumer Reports; or alternately a used car sales aggregator website that wants to benchmark available deals to fair pricing expectations.

Conclusion

In conclusion, we have developed a strong base for the continuation of our research. By accurately differentiating our analysis from other price prediction companies, we will use specific variables such as the car's model year, the number of kilometers, engine capacity, and total weight of the car to determine if there is a correlation that is strong enough to predict a fair price for any used Toyota Corolla vehicle. The association measures have been prepared for each variable for the fair pricing model. The results of which will be further analyzed for our final paper.

Roles & Responsibilities

Name	Working on which section?	Date Started	Date Completed (Suggested: Thursday Night or Friday Morning)	Current Status
Angie	Data Prep	2/7/2019	14 Feb 2019	
Ayca	Proofreading-Final Table of contents Abstract Conclusion Work Cited	2/15/2019	Proposal completed and sent out to group for review by: **Friday** afternoon.	
Kari	Competitive Analysis, Cost/Benefit Analysis	9-Feb-2019	10-Feb-2019	First Draft completed for team review
Oye	Proofreading-Initial before Friday Introduction	2/14/2019	2/15/2019	
Nasir	CRISP-DM approach	12/2/2019	14/2/2019	
Vaibhav	Introduction Data Modelling and Analysis	2/7/2019	2/15/2019	

Tables and Charts

Table 2: Toyota Corolla data set - all columns, first 100 rows.

Id	Model	Price	Age_08_04	Mfg_Month	Mfg_Year	KM	Fuel_Type	HP	Met_Color	Auto	cc	Doors	Cylinders	Gears	Quarterly_	Weight	Mfr_Guar	BOVAG_G	Guarantee	ABS	Airbag_1	Airbag_2	Airco	Auto	Boardcom	CD_Player	Central_L	Powered_	Power_St	Radio	Mistlamps	Sport_Mo	Backseat_	Metallic_R	Radio_cas	Tow_Bar
1	TOYOTA Corolla	13500	23	10	2002	46986	Diesel	90	1	0	2000	3	4	5	210	1165	0	1	3	1	1	1	0	0	1	0	1	1	1	0	0	0	1	0	0	0
2	TOYOTA Corolla	13750	23	10	2002	72937	Diesel	90	1	0	2000	3	4	5	210	1165	0	1	3	1	1	1	1	0	1	1	1	0	1	0	0	0	1	0	0	0
3	?TOYOTA Corolla	13950	24	9	2002	41711	Diesel	90	1	0	2000	3	4	5	210	1165	1	1	3	1	1	1	0	0	1	0	0	0	1	0	0	0	1	0	0	0
4	TOYOTA Corolla	14950	26	7	2002	48000	Diesel	90	0	0	2000	3	4	5	210	1165	1	1	3	1	1	1	0	0	1	0	0	0	1	0	0	0	1	0	0	0
5	TOYOTA Corolla	13750	30	3	2002	38500	Diesel	90	0	0	2000	3	4	5	210	1170	1	1	3	1	1	1	1	0	1	0	1	1	1	0	1	0	1	0	0	0
6	TOYOTA Corolla	12950	32	1	2002	61000	Diesel	90	0	0	2000	3	4	5	210	1170	0	1	3	1	1	1	1	0	1	0	1	1	1	0	1	0	1	0	0	0
7	?TOYOTA Corolla	16900	27	6	2002	94612	Diesel	90	1	0	2000	3	4	5	210	1245	0	1	3	1	1	1	1	0	1	0	1	1	1	0	0	1	1	0	0	0
8	TOYOTA Corolla	18600	30	3	2002	75889	Diesel	90	1	0	2000	3	4	5	210	1245	1	1	3	1	1	1	1	0	1	1	1	1	1	0	0	0	1	0	0	0
9	?TOYOTA Corolla	21500	27	6	2002	19700	Petrol	192	0	0	1800	3	4	5	100	1185	0	1	3	1	1	0	1	0	0	0	1	1	1	1	0	0	0	1	1	0
10	?TOYOTA Corolla	12950	23	10	2002	71138	Diesel	69	0	0	1900	3	4	5	185	1105	0	1	3	1	1	1	1	0	1	0	0	0	1	0	0	0	1	0	0	0
11	TOYOTA Corolla	20950	25	8	2002	31461	Petrol	192	0	0	1800	3	4	6	100	1185	1	1	12	1	1	1	1	1	0	1	1	1	1	0	0	0	0	1	0	0
12	TOYOTA Corolla	19950	22	11	2002	43610	Petrol	192	0	0	1800	3	4	6	100	1185	1	1	3	1	1	1	1	1	1	0	1	1	1	0	1	1	1	1	0	0
13	TOYOTA Corolla	19600	25	8	2002	32189	Petrol	192	0	0	1800	3	4	6	100	1185	1	1	3	1	1	1	1	1	1	0	1	1	1	0	1	1	1	1	0	0
14	TOYOTA Corolla	21500	31	2	2002	23000	Petrol	192	1	0	1800	3	4	6	100	1185	1	1	3	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	0
15	TOYOTA Corolla	22500	32	1	2002	34131	Petrol	192	1	0	1800	3	4	6	100	1185	1	1	3	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	0
16	TOYOTA Corolla	22000	28	5	2002	18739	Petrol	192	0	0	1800	3	4	6	100	1185	0	1	3	1	1	1	1	1	1	0	1	1	1	0	1	1	1	1	0	0
17	?TOYOTA Corolla	22750	30	3	2002	34000	Petrol	192	1	0	1800	3	4	5	100	1185	0	1	3	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	0	0
18	?TOYOTA Corolla	17950	24	9	2002	21716	Petrol	110	1	0	1600	3	4	5	85	1105	0	0	18	1	1	0	1	0	0	0	1	1	1	1	0	0	0	0	1	1
19	TOYOTA Corolla	16750	24	9	2002	25563	Petrol	110	0	0	1600	3	4	5	19	1065	0	0	3	1	1	1	1	1	1	1	1	1	1	0	1	0	0	0	0	0
20	TOYOTA Corolla	16950	30	3	2002	64359	Petrol	110	1	0	1600	3	4	5	85	1105	1	1	3	1	1	1	1	0	1	1	1	1	1	0	0	1	1	0	0	0
21	?TOYOTA Corolla	15950	30	3	2002	67660	Petrol	110	1	0	1600	3	4	5	85	1105	1	1	3	1	1	1	1	0	1	1	1	1	1	0	0	1	1	0	0	1
22	TOYOTA Corolla	16950	29	4	2002	43905	Petrol	110	0	1	1600	3	4	5	100	1170	0	1	3	1	1	1	1	1	1	0	1	1	1	0	1	1	1	0	0	0
23	TOYOTA Corolla	15950	28	5	2002	56349	Petrol	110	1	0	1600	3	4	5	85	1120	0	1	3	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1
24	?TOYOTA Corolla	16950	28	5	2002	32220	Petrol	110	1	0	1600	3	4	5	85	1120	1	1	3	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0
25	TOYOTA Corolla	16250	29	4	2002	25813	Petrol	110	1	0	1600	3	4	5	85	1120	1	1	3	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0
26	TOYOTA Corolla	15950	25	8	2002	28450	Petrol	110	1	0	1600	3	4	5	85	1120	0	1	3	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1
27	?TOYOTA Corolla	17495	27	6	2002	34545	Petrol	110	1	0	1600	3	4	5	85	1120	1	1	3	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	0	0
28	TOYOTA Corolla	15750	29	4	2002	41415	Petrol	110	1	0	1600	3	4	5	85	1120	1	1	3	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	0
29	TOYOTA Corolla	16950	28	5	2002	44142	Petrol	110	0	0	1600	3	4	5	85	1120	1	1	3	1	1	1	1	1	1	0	1	1	1	0	1	1	1	0	0	0
30	TOYOTA Corolla	17950	30	3	2002	11090	Petrol	110	1	0	1600	3	4	5	85	1120	1	1	3	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0
31	?TOYOTA Corolla	12950	29	4	2002	9750	Petrol	97	1	0	1400	3	4	5	69	1100	1	1	12	1	1	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0
32	TOYOTA Corolla	15750	22	11	2002	25199	Petrol	97	1	0	1400	3	4	5	85	1100	1	1	3	1	1	1	1	0	1	1	1	1	1	0	0	1	1	1	0	1
33	TOYOTA Corolla	15950	27	6	2002	29510	Petrol	97	1	0	1400	3	4	5	85	1100	1	1	3	1	1	1	1	0	1	0	1	1	1	0	0	1	1	0	0	0
34	?TOYOTA Corolla	14950	26	7	2002	32692	Petrol	97	1	0	1400	3	4	5	85	1100	0	1	3	1	1	1	1	0	1	0	1	1	1	0	0	1	1	0	0	0
35	?TOYOTA Corolla	15500	22	11	2002	41000	Petrol	97	1	0	1400	3	4	5	85	1100	0	0	3	1	1	1	1	0	1	1	1	1	1	0	0	1	1	0	0	0
36	TOYOTA Corolla	15750	26	7	2002	48000	Petrol	97	0	0	1400	3	4	5	85	1100	1	1	3	1	1	1	1	0	1	1	1	1	1	0	0	1	1	0	0	0
37	TOYOTA Corolla	15950	25	8	2002	30000	Petrol	97	0	0	1400	3	4	5	85	1100	0	1	3	1	1	1	1	0	1	1	1	1	1	0	0	1	1	0	0	0
38	TOYOTA Corolla	14950	23	10	2002	30000	Petrol	97	1	0	1400	3	4	5	85	1100	1	1	3	1	1	1	0	0	1	0	1	1	1	0	0	1	1	0	0	0
39	TOYOTA Corolla	15750	32	1	2002	25329	Petrol	97	1	0	1400	3	4	5	85	1100	1	1	3	1	1	1	1	0	1	1	1	1	1	0	0	1	1	0	0	0
40	?TOYOTA Corolla	14750	27	6	2002	37500	Petrol	97	0	0	1400	3	4	5	85	1100	1	1	3	1	1	1	0	0	1	0	0	0	1	0	0	1	1	0	0	0
41	?TOYOTA Corolla	13950	22	11	2002	49059	Petrol	97	0	0	1400	3	4	5	85	1100	0	1	3	1	1	1	0	0	1	0	0	0	1	0	0	1	1	0	0	0
42	TOYOTA Corolla	16750	27	6	2002	44068	Petrol	97	1	0	1400	3	4	5	85	1100	1	1	3	1	1	1	1	0	1	1	1	1	1	0	0	1	1	0	0	0
43	TOYOTA Corolla	13950	22	11	2002	46961	Petrol	97	0	0	1400	3	4	5	85	1100	1	1	3	1	1	1	0	0	1	0	0	0	1	0	0	1	1	0	0	0
44	TOYOTA Corolla	16950	27	6	2002	110404	Diesel	90	0	0	2000	5	4	5	234	1255	1	1	3	1	1	1	1	0	1	0	1	1	1	0	0	1	1	0	0	0
45	TOYOTA Corolla	16950	22	11	2002	100250	Diesel	90	0	0	2000	5	4	5	234	1255	1	1	3	1	1	1	1	0	1	0	1	1	1	0	0	1	1	0	0	0
46	?TOYOTA Corolla	19000	23	10	2002	84000	Diesel	90	0	0	2000	5	4	5	234	1270	0	1	3	1	1	1	1	1	1	0	1	1	1	0	0	0	1	0	0	0
47	TOYOTA Corolla	17950	27	6	2002	79375	Diesel	90	1	0	2000	5	4	5	234	1255	0	1	3	1	1	1	1	0	1	1	1	1	1	0	0	1	1	0	0	0
48	TOYOTA Corolla	15800	22	11	2002	75048	Petrol	97	1	0	1400	5	4	5	85	1110	0	1	3	1	1	1	1	0	1	1	1	1	1	0	0	1	1	0	0	1
49	?TOYOTA Corolla	17950	22	11	2002	72215	Diesel	90	1	0	2000	5	4	5	234	1255	1	1	3	1	1	1	1	0	1	0	1	1	1	0	0	1	1	0	0	0
50	TOYOTA Corolla	21950	31	2	2002	69982	Petrol	192	1	0	1800	5	4	6	100	1195	1	1	3	1	1	1	1	1	1	0	1	1	1	0	1	1	1	1	0	0
51	TOYOTA Corolla	17950	22	11	2002	65636	Diesel	90	1	0	2000	5	4	5	234	1255	1	1	3	1	1	1	1	0	1	1	1	1	1	0	0	1	1	0	0	0
52	?TOYOTA Corolla	15750	30	3	2002	57086	Petrol	97	1	0	1400	5	4	5	85	1110	0	1	3	1	1	1	1	0	1	1	1	1	1	0	0	0	1	1	0	0
53	TOYOTA Corolla	20500	26	7	2002	58000	Petrol	110	1	1	1600	5	4	5	100	1180	1	1	3	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	0
54	TOYOTA Corolla	21950	27	6	2002	49866	Petrol	192	1	0	1800	5	4	5	100	1195	0	1	12	1	1	0	1	0	0	0	1	1	1	1	0	0	0	1	1	0
55	?TOYOTA Corolla	15500	25	8	2002	49162	Petrol	110	0	1	1600	5	4	5	100	1165	0	1	3	1	1	1	0	0	1	1	0	0	1	0	0	1	1	0	0	0
56	?TOYOTA Corolla	12250	32	1	2002	45725	Petrol	110	1	0	1600	5	4	5	85	1075	1	1	3	1	1	1	1	0	1	1	1	1	1	0	0	0	1	0	0	1
57	TOYOTA Corolla	15250	28	5	2002	48210	Petrol	97	0	0	1400	5	4	5	85	1110	0	1	3	1	1	1	1	0	1	0	1	1	1	0	0	1	1	0	0	0
58	TOYOTA Corolla	15250	26	7	2002	48000	Petrol	97	0	0	1400	5	4	5	85	1110	1	1	3	1	1	1	1	0	1	0	1	1	1	0	0	1	1	0	0	0
59	?TOYOTA Corolla	18950	23	10	2002	28704	Petrol	110	1	1	1600	5	4	5	100	1180	1	1	3	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0
60	?TOYOTA Corolla	15999	30	3	2002	28950	Petrol	110	1	0	1600	5	4	5	85	1120	0	1	3	1	1	1	1	1	1	0	1	1	1	0	1	1	1	0	0	0
61	TOYOTA Corolla	14950	22	11	2002	37400	Petrol	97	1	0	1400	5	4	5	85	1110	0	1	3	1	1	1	1	0	1	1	1	1	1	0	0	1	1	0	0	0
62	?TOYOTA Corolla	16500	27	6	2002	37177	Petrol	110	0	0	1600	5	4	5	85	1120	0	1	3	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0
63	TOYOTA Corolla	16750	31	2	2002	30544	Petrol	110	1	0	1600	5	4	5	85	1120	1	1	3	1	1	1	1	1	1	0	1	1	1	0	1	1	1	0	0	1
64	TOYOTA Corolla	17950	30	3	2002	28511	Petrol	110	1	0	1600	5	4	5	85	1120	1	1	3	1	1	1	1	1	1	0	1	1	1	0	1	1	1	0	0	0
65	TOYOTA Corolla	17950	27	6	2002	28809	Petrol	97	1	0	1400	5	4	5	85	1110	0	1	3	1	1	1	1	0	1	1	1	1	1	0	0	1	1	0	0	1
66	TOYOTA Corolla	16950	26	7	2002	32181	Petrol	110	1	0	1600	5	4	5	69	1075	0	1	12	1	1	0	1	1	1	1	1	1	1	0	1	0	0	0	0	0
67	TOYOTA Corolla	18950	28	5	2002	30993	Petrol	110	1	0	1600	5	4	5	85	1120	1	1	3	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0
68	TOYOTA Corolla	14950	22	11	2002	30400	Petrol	97	1	0	1400	5	4	5	85	1110	0	1	3	1	1	1	1	0	1	1	1	1	1	0	0	1	1	0	0	0
69	TOYOTA Corolla	22250	22	11	2002	30000	Diesel	110	1	0	2000	5	4	5	234	1275	1	1	3	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	0	0
70	TOYOTA Corolla	15950	25	8	2002	29719	Petrol	97	1	0	1400	5	4	5	85	1110	1	1	3	1	1	1	0	0	1	1	0	1	1	0	0	1	1	0	0	0
71	?TOYOTA Corolla	15950	28	5	2002	29206	Petrol	97	1	0	1400	5	4	5	85	1110	1	1	3	1	1	1	1	0	1	1	1	1	1	0	0	1	1	0	0	0
72	TOYOTA Corolla	12995	32	1	2002	29198	Petrol	97	1	0	1400	5	4	5	85	1060	1	1	3	1	1	1	1	0	1	1	0	1	1	0	0	0	1	1	0	0
73	?TOYOTA Corolla	18950	28	5	2002	28817	Petrol	110	1	0	1598	5	4	5	85	1120	1	1	3	1	1	1	1	1	0	1	1	1	1	0	1	0	1	0	0	0
74	?TOYOTA Corolla	15750	23	10	2002	28227	Petrol	97	1	0	1400	5	4	5	85	1110	0	1	3	1	1	1	0	0	1	1	0	0	1	0	0	1	1	0	0	0
75	?TOYOTA Corolla	18950	28	5	2002	28000	Petrol	110	0	0	1600	5	4	5	85	1120	1	0	3	1	1	1	1	1	0	1	1	1	1	0	1	0	1	1	0	0
76	TOYOTA Corolla	16950	23	10	2002	28000	Petrol	110	1	0	1600	5	4	5	85	1115	1	1	3	1	1	1	1	0	1	1	1	1	1	0	0	1	1	0	0	0
77	TOYOTA Corolla	18750	31	2	2002	25266	Petrol	110	1	0	1600	5	4	5	85	1120	1	1	3	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0
78	TOYOTA Corolla	18450	27	6	2002	23489	Petrol	110	0	0	1600	5	4	5	85	1115	0	1	3	1	1	1	0	0	1	0	1	1	1	0	0	1	1	0	0	0
79	TOYOTA Corolla	16895	29	4	2002	22575	Petrol	110	1	0	1600	5	4	5	85	1115	1	1	3	1	1	1	1	0	1	1	1	1	1	0	0	0	0	0	0	0
80	TOYOTA Corolla	14900	30	3	2002	22000	Petrol	97	1	0	1400	5	4	5	85	1110	1	1	3	1	1	1	0	0	1	1	1	1	1	0	0	1	1	0	0	0
81	TOYOTA Corolla	18950	25	8	2002	20019	Petrol	110	1	1	#####	5	4	5	100	1180	0	0	3	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0
82	TOYOTA Corolla	17250	29	4	2002	20000	Petrol	110	1	0	1600	5	4	5	85	1115	1	1	3	1	1	1	1	0	1	1	1	1	1	0	0	1	1	0	0	0
83	?TOYOTA Corolla	15450	25	8	2002	17003	Petrol	97	1	0	1400	5	4	5	85	1110	0	1	3	1	1	1	0	0	1	1	0	0	1	0	0	1	1	0	0	0
84	TOYOTA Corolla	17950	31	2	2002	16238	Petrol	110	1	1	1600	5	4	5	100	1180	1	1	3	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0
85	?TOYOTA Corolla	16850	25	8	2002	15414	Petrol	97	1	0	1400	5	4	5	85	1110	1	1	3	1	1	1	1	0	1	1	1	1	1	0	0	1	1	0	0	0
86	TOYOTA Corolla	17450	28	5	2002	8537	Petrol	110	1	0	1600	5	4	5	85	1120	0	1	3	1	1	1	1	1	1	0	1	1	1	0	1	1	1	0	0	0
87	?TOYOTA Corolla	14900	30	3	2002	7000	Petrol	97	1	0	1400	5	4	5	85	1100	0	0	12	1	1	0	1	0	0	0	1	0	1	1	0	0	0	0	1	0
88	TOYOTA Corolla	17950	20	1	2003	65956	Diesel	90	1	0	2000	3	4	5	210	1245	0	1	3	1	1	1	1	0	1	1	1	1	1	0	0	1	1	0	0	0
89	TOYOTA Corolla	15950	19	2	2003	51884	Petrol	97	1	0	1400	3	4	5	85	1100	1	1	3	1	1	1	1	0	1	1	1	1	1	0	0	1	1	0	0	0
90	TOYOTA Corolla	23950	19	2	2003	50005	Diesel	110	1	0	2000	3	4	5	234	1265	1	1	3	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	0	0
91	TOYOTA Corolla	16950	20	1	2003	48110	Petrol	97	1	0	1400	3	4	5	85	1100	1	1	3	1	1	1	1	0	1	1	1	1	1	0	0	1	1	0	0	0
92	?TOYOTA Corolla	22250	20	1	2003	37500	Diesel	90	1	0	2000	3	4	5	234	1260	0	1	3	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	0	0
93	TOYOTA Corolla	18950	16	5	2003	34472	Diesel	90	1	0	1995	3	4	5	234	1260	0	0	3	1	1	1	1	1	0	1	1	1	1	0	0	0	0	1	0	0
94	TOYOTA Corolla	15950	20	1	2003	28229	Petrol	97	1	0	1400	3	4	5	85	1100	0	1	3	1	1	1	1	0	1	1	1	1	1	0	0	1	1	0	0	0
95	TOYOTA Corolla	18900	20	1	2003	21850	Petrol	110	0	0	1600	3	4	5	85	1120	0	1	20	1	1	0	1	0	0	0	1	1	1	1	0	0	0	0	1	0
96	TOYOTA Corolla	18950	17	4	2003	20251	Diesel	90	1	0	1995	3	4	5	234	1260	0	0	3	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0
97	?TOYOTA Corolla	15950	19	2	2003	29435	Petrol	97	1	0	1400	3	4	5	85	1100	0	1	3	1	1	1	1	0	1	1	1	1	1	0	0	1	1	0	0	0
98	TOYOTA Corolla	15950	19	2	2003	25948	Petrol	97	1	0	1400	3	4	5	85	1100	1	1	3	1	1	1	1	0	1	1	1	1	1	0	0	1	1	0	0	0
99	TOYOTA Corolla	18750	11	10	2003	24500	Petrol	110	1	0	1600	3	4	5	85	1120	0	0	3	1	1	1	1	1	1	0	1	1	1	0	1	1	1	1	0	0

Figure 5: Visualization of pairwise relationships to explore data set.

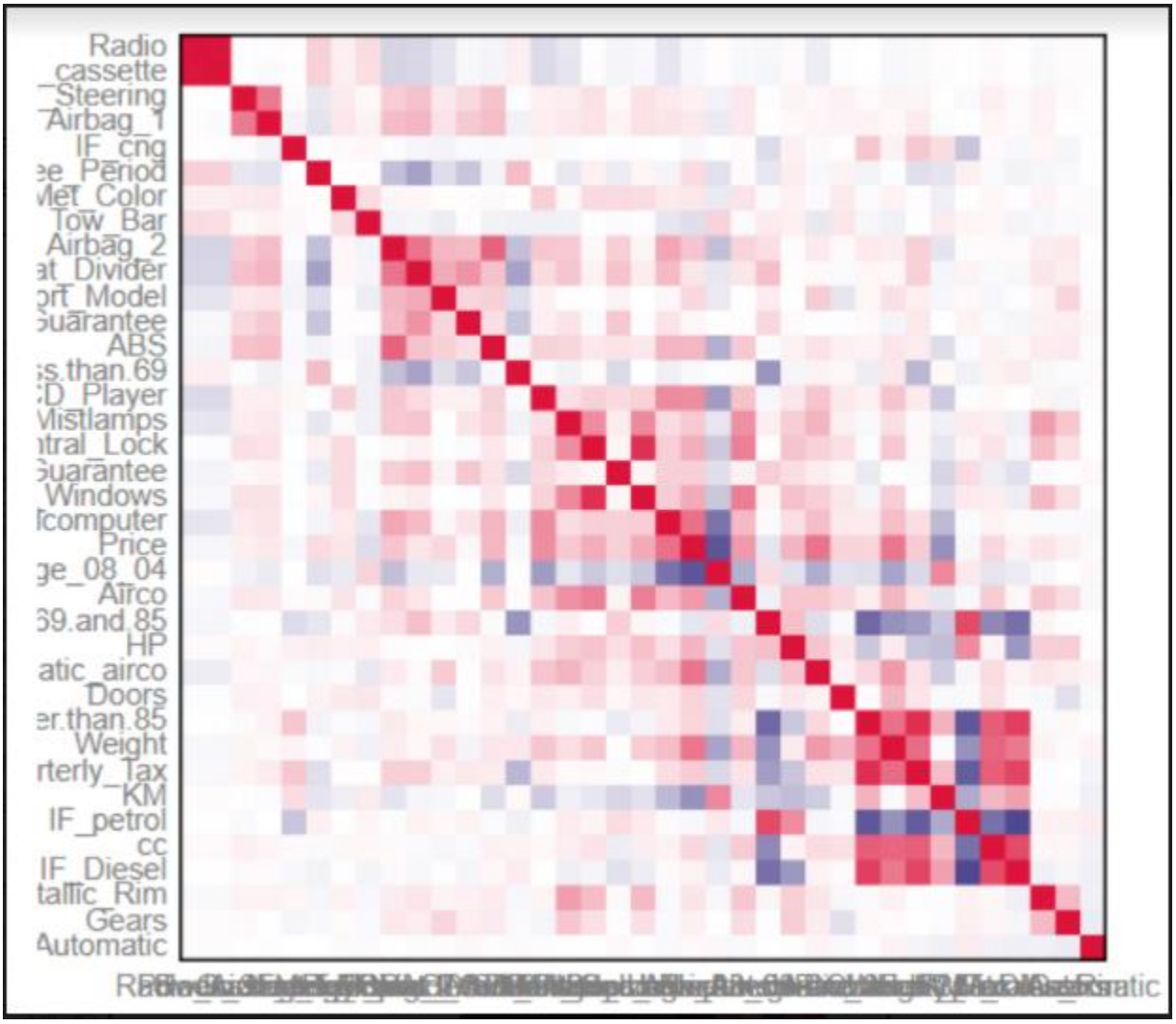


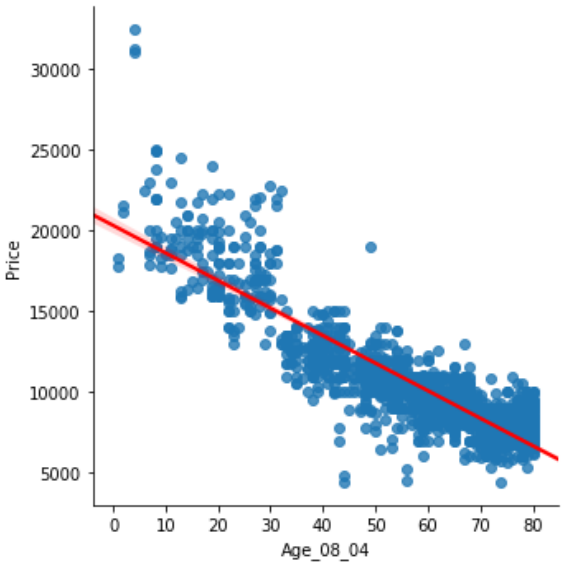
Table 3: Analysis of association measures to inform selection of variables to include in model.

Focused Analysis on Field Price

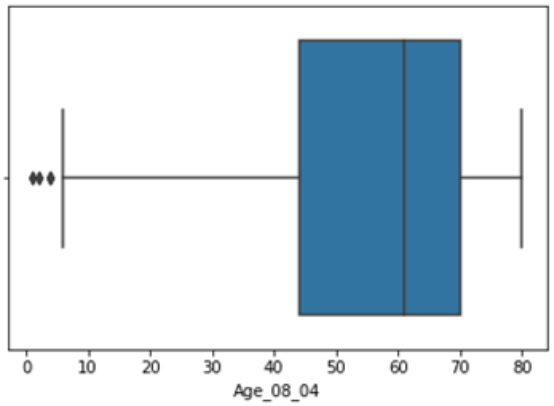
	Association Measure	p-value	
Age_08_04	-0.876377	0.00E+00	***
Boardcomputer	0.603482	0.00E+00	***
Automatic_airco	0.586273	0.00E+00	***
Weight	0.579851	0.00E+00	***
KM	-0.569268	0.00E+00	***
CD_Player	0.479845	0.00E+00	***
Airco	0.428618	0.00E+00	***
Powered_Windows	0.355858	0.00E+00	***
Central_Lock	0.342814	0.00E+00	***
HP	0.314693	0.00E+00	***
ABS	0.305954	0.00E+00	***
Airbag_2	0.248474	0.00E+00	***
Mistlamps	0.223439	0.00E+00	***
Quarterly_Tax	0.219102	0.00E+00	***
Mfr_Guarantee	0.199523	2.38E-14	***
Doors	0.184118	2.08E-12	***
Tow_Bar	-0.171719	5.83E-11	***
Sport_Model	0.165477	2.85E-10	***
cc	0.165085	3.15E-10	***
Guarantee_Period	0.147322	2.06E-08	***
Metallic_Rim	0.109572	3.19E-05	***
Met_Color	0.10802	4.12E-05	***
Backseat_Divider	0.105774	5.95E-05	***
Airbag_1	0.093482	3.91E-04	***
Power_Steering	0.064152	1.51E-02	*
Gears	0.06344	1.62E-02	*
IF_Diesel	0.054734	3.82E-02	*
Radio_cassette	-0.042606	1.07E-01	
Radio	-0.04131	1.18E-01	
IF_cng	-0.039434	1.35E-01	
IF_petrol	-0.039171	1.38E-01	
BOVAG_Guarantee	0.032916	2.13E-01	
Automatic	0.026783	3.11E-01	

Data prep charts

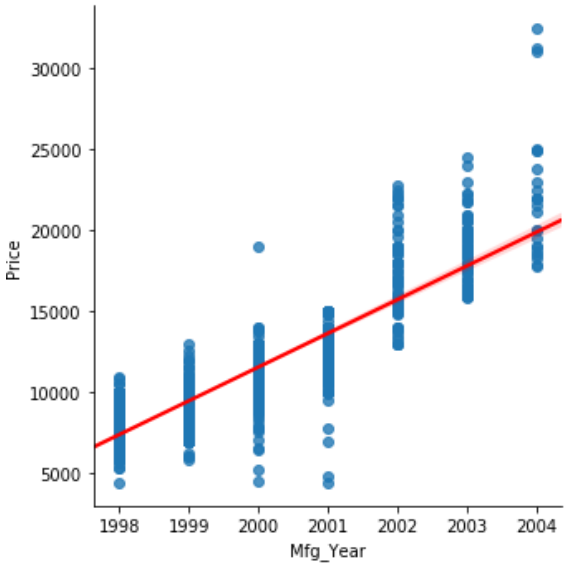
Scatter plot of Age_08_04 x Price



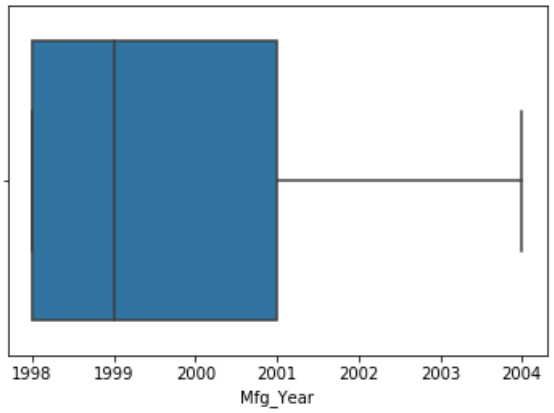
Box plot of Age_08_04



Scatter plot of Mfg_Year x Price

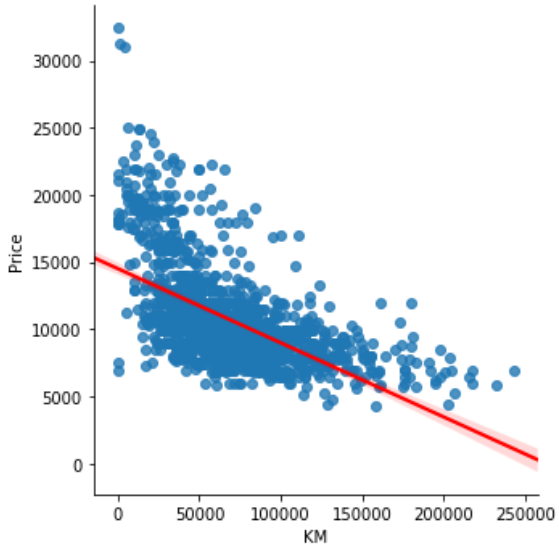


Box plot of Mfg_Year

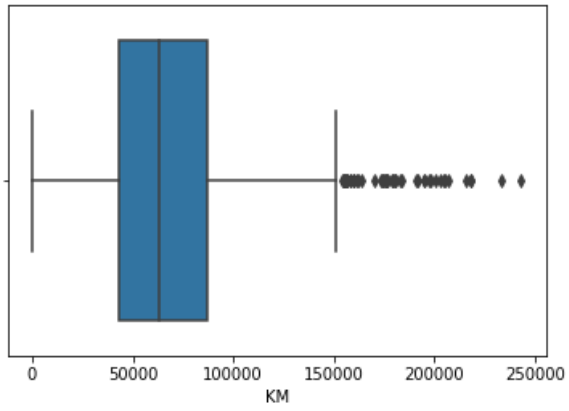


Data prep charts (Continued)

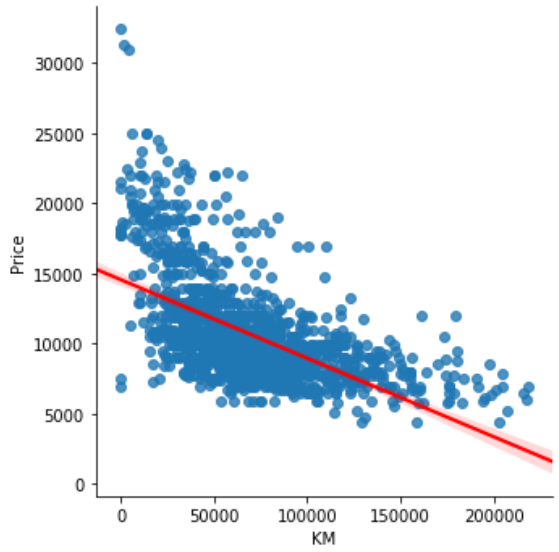
Scatter plot of KM x Price



Box plot of KM

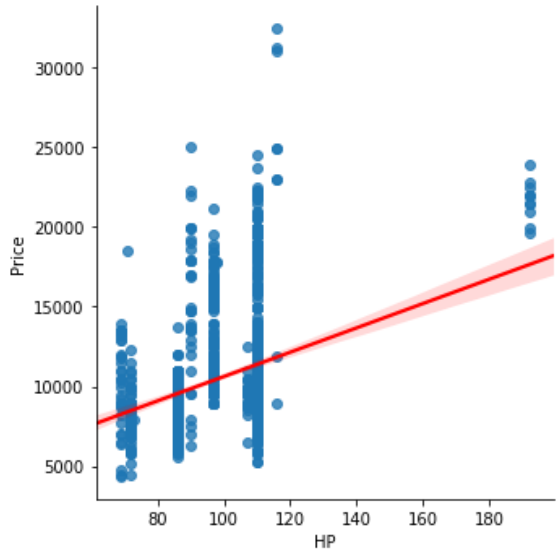


Scatter plot of KM x Price with outliers removed

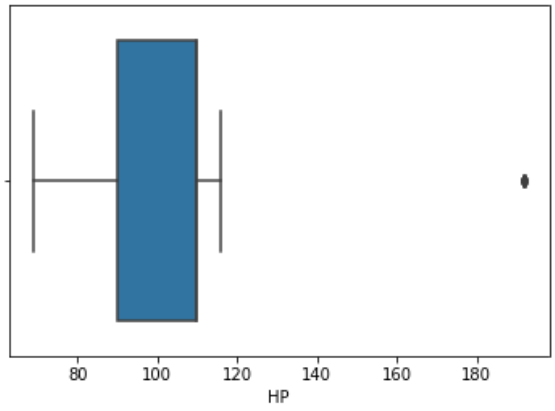


Data prep charts (Continued)

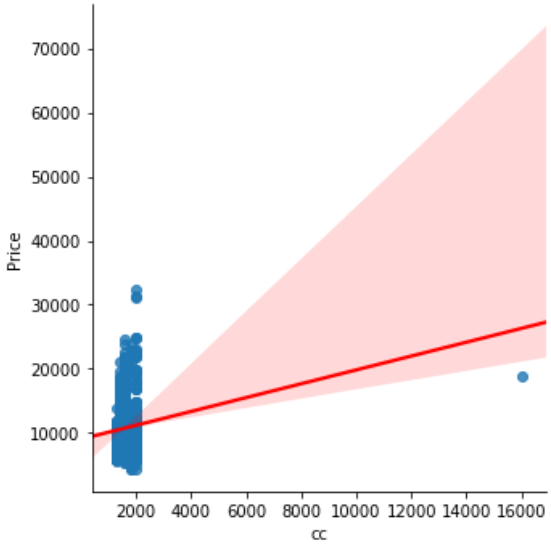
Scatter lot of HP x Price



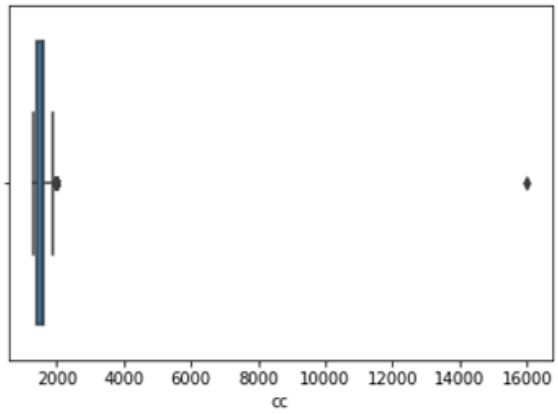
Box plot of HP



Scatter plot of cc x Price

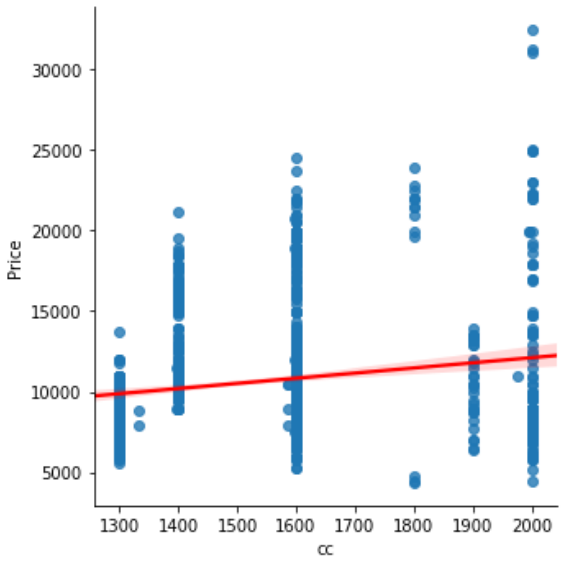


Box plot of cc

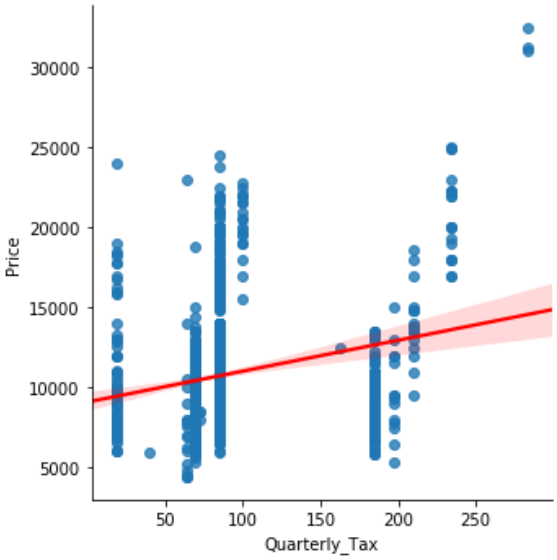


Data prep charts (Continued)

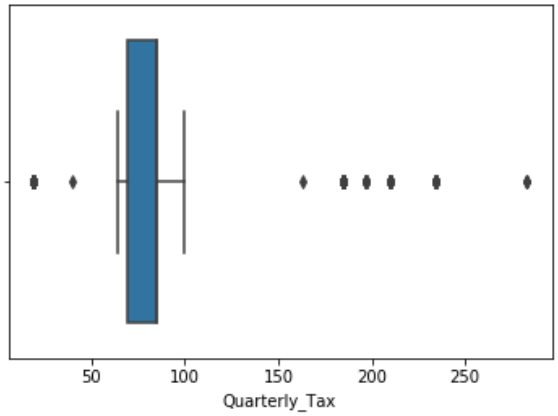
Scatter plot of cc x Price with outlier removed



Scatter plot of Quarterly_Tax x Price

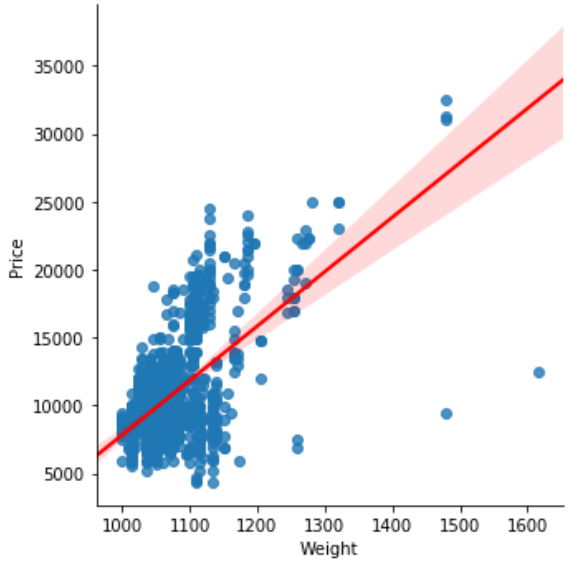


Box plot of Quarterly_Tax

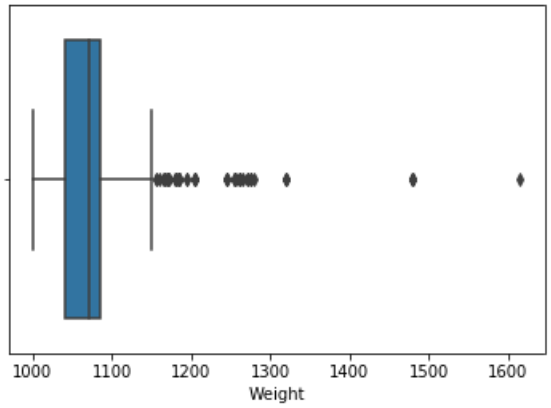


Data prep charts (Continued)

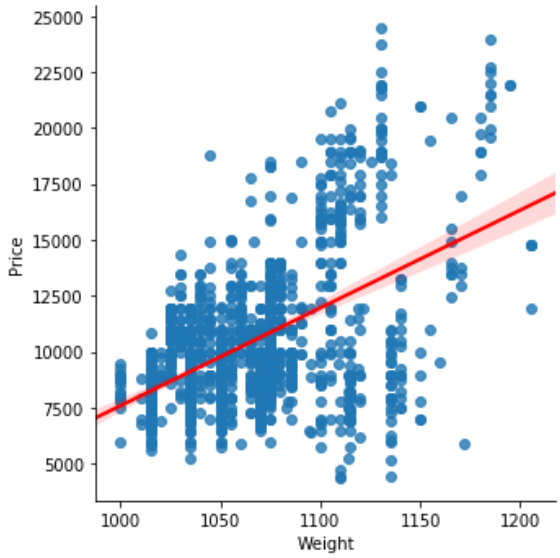
Scatter plot of Weight x Price



Box plot of Weight

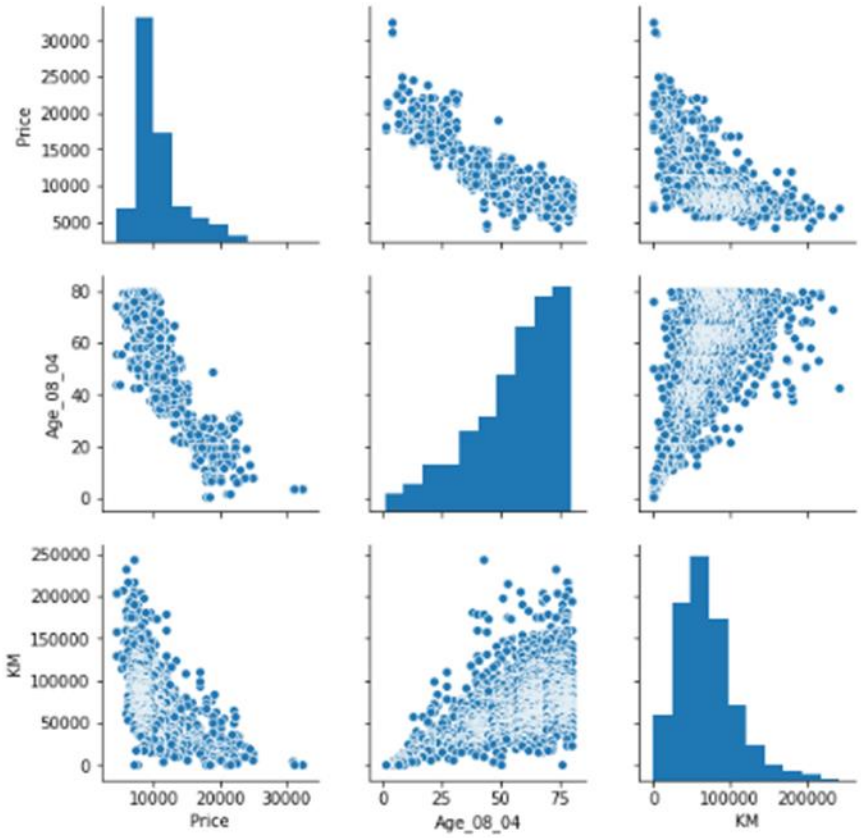


Scatter plot of Weight with outliers removed



Data prep charts (Continued)

'Price, Age_08_04 & KM'



Works Cited

AutoTrade. (2019, 02 15). *Price a car*. Retrieved from AutoTrader: https://www.autotrader.ca/valuations/

Book, B. (2019, 02 15). *Toyota Corolla - Average Asking Price*. Retrieved from Black Book: https://www.canadianblackbook.com/Toyota/Corolla/average-asking-price#4

CARFAX CANADA. (2019, 02 15). *2009 TOYOTA COROLLA CE - CARFAX CANADA TRUE VALUE RANGE*. Retrieved from CARFAX CANADA: https://www.carfax.ca/car-valuation/report//165861

CARFOLIO. (2015, 02 15). *2002 Toyota Corolla 1.8 VVTL-i T Sport*. Retrieved from https://www.carfolio.com/specifications/models/car/?car=534622

Carfolio. (2019, 02 15). *2010 Toyota Corolla 2.0 D-4D*. Retrieved from https://www.carfolio.com/specifications/models/car/?car=218317

Consumer Reports CR. (2014, 05 01). *How much is the used car really worth?* Retrieved from https://www.consumerreports.org/cro/2012/12/how-much-is-the-used-car-really-worth/index.htm

Grace-Martin, K. (2019, 02 15). *Outliers: To Drop or Not to Drop*. Retrieved from THE ANALYSIS FACTOR: https://www.theanalysisfactor.com/outliers-to-drop-or-not-to-drop/

SEMATECH, N. (2019, 02 15). *7.1.6. What are outliers in the data?* Retrieved from Engineering Statistics Handbook: https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm