Awarding Body:	Arden University
Programme Name:	Data Analytics and Information System Management
Module Name (and Part if applicable):	Data Design
Assessment Title:	Case Study Report Part 2
Student Number:	STU101620
Tutor Name:	Krithiga Duraisamy
Word Count:	4274

Please refer to the Word Count Policy on your Module Page for guidance

Table of Contents

2.1 Hypothesis Formulation 4

2.1.1 Identification of Independent and Dependent Variables 4

2.1.2 Variables Relationship Diagram 4

2.1.3 Hypothesis 8

2.2 Data Preparation 9

2.2.1 Data Cleaning 9

2.2.2 Data Transformation 12

2.2.3 Data Integration 14

2.3 Data Analysis 16

2.3.1 Identification and Justification of Statistical Test 16

2.3.2 Application of Test 17

2.3.3 Discussion of Test Outcome 22

2.3.4 Graphical Interpretation and Further Test Outcomes 22

2.3.5 Limitations, Assumptions and Enhancement for Better Accuracy 25

2.4 Deployment 26

2.4.1 Risk Analysis and Potential Challenges 26

2.4.2 Ethical Aspects of the Deployment 27

2.4.3 Conclusions and Recommendations 27

Reference 29

Table of Figures

Figure 1 Correlation Heatmap 5
Figure 2 Relationship between Car Make and Price 5
Figure 3 Relationship between Car Make and Price 6
Figure 4 Relationship between Car OfferType and Price 7
Figure 5 Relationship between Fuel Type and Price 8
Figure 6 Correlation Heat Map 8
Figure 7 First Dataset 9
Figure 8 Second Dataset 9
Figure 9 Selecting Germany 10
Figure 10 Dropping Country Column 10
Figure 11 Regex for Fuel Type 10
Figure 12 Dropping PageUrl Column 11
Figure 13 Replacing Symbols 11
Figure 14 Removing SI Unit 11
Figure 15 Replacing Symbols 11
Figure 16 Converting hp to KW 12
Figure 17 Rounding hp 12
Figure 18 Converting Year 13
Figure 19 Selecting Price Average 13
Figure 20 Encoding Categorical Variables 14
Figure 21 Merging Price 14
Figure 22 Rename Column Names 15
Figure 23 Rearrange Column 15
Figure 24 Integrating Datasets 15
Figure 25 Removing Index 16
Figure 26 Kde and QQ Plot 18
Figure 27 Descriptive Analysis 18
Figure 28 Yeo-Johnson Transformation of Continuous Variables 19
Figure 29 Presence of Outliers 20
Figure 30 Removal of Outliers 20
Figure 31 Regression Summary 21
Figure 32 P-values 21
Figure 33 Regression Plot (mileage) 23
Figure 34 Regression Plot (hp) 23
Figure 35 Predicted and Actual Prices 24
Figure 36 Residuals 25

2.1 Hypothesis Formulation

Section 2.1 further explains **Data Understanding** of the CRISP DM Methodology for Autoscout24.

2.1.1 Identification of Independent and Dependent Variables

From the collected data as shown in task 1, the independent variables are the make of the car, the model of the car, the mileage covered, the fuel type, the gear, the offerType, the hp rating and the year it was registered. The dependent variable is the price of the vehicle.

The aim of this analysis is to predict the price of vehicles in the future. The value of the vehicles is dependent on the mileage already covered by the car. If the mileage is high, then the price of the car would be low and vice versa (Autoebid, 2016). The make and model of the car determine the purchase decision which in turn affects the value of the car (Ouet al, 2020). The hp rating of the car determines the acceleration of the car and gives an idea of the overall performance of the car and the price (Threewitt, 2017). A high rated hp is more expensive than a lower one (Threewitt, 2017). The fuel type of the car also determines the price, most electric cars are more expensive than gasoline and diesel cars (Murillo, 2021). Generally automatic cars are more expensive than manual cars and hence is a determining factor for the price of the car (My Car Credit, 2022). The price of the car is also dependent on the offerType. Most definitely new cars would be more expensive than used cars. From the year the car was registered, the price of the car over the years can be monitored.

2.1.2 Variables Relationship Diagram

To show the relationship between the variables, a correlation matrix, bar chart and box plot has been used. The correlation matrix can only show the relationship between numerical variables like the mileage, hp, year and price. While the bar chart shows the relationship between the categorical variables like the make, offerType, Fuel Type and the price.

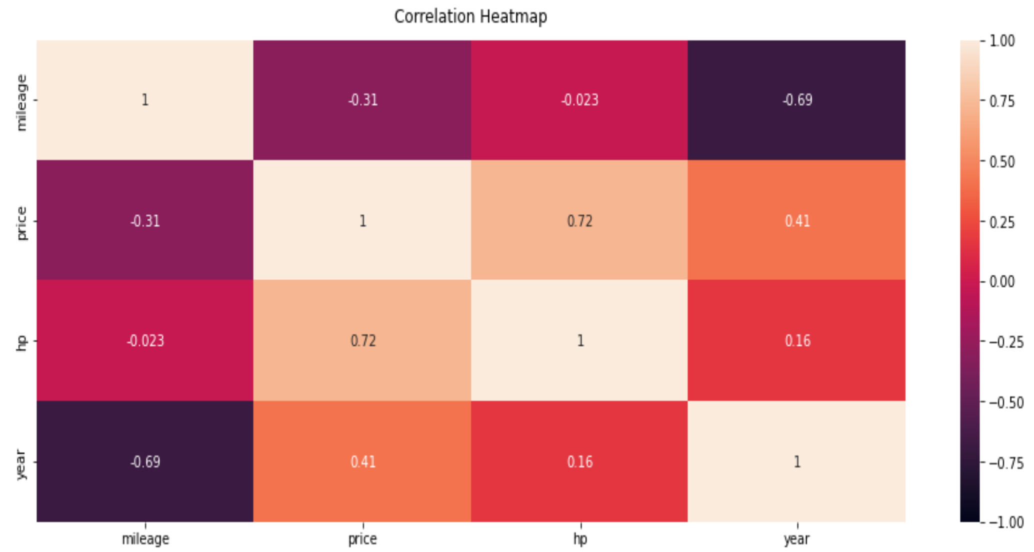


Figure 1 Correlation Heatmap

From the correlation matrix, there is a strong positive correlation of 0.72 between the hp rating of the car and the price. This means that as the hp rating increases, the price of the car increases as well (Preethu, 2022). There is a weak positive correlation of 0.41 between the year and the price. Meaning that the price of vehicles increases in the coming years. There is a weak negative correlation of -0.31 between the mileage and the price. This means that as the mileage increases, there is a decrease in price and vice versa. Between the mileage and the year there is also a strong negative correlation of -0.69. This means that the mileage will reduce as the year increases, but this is not relevant for the analysis.

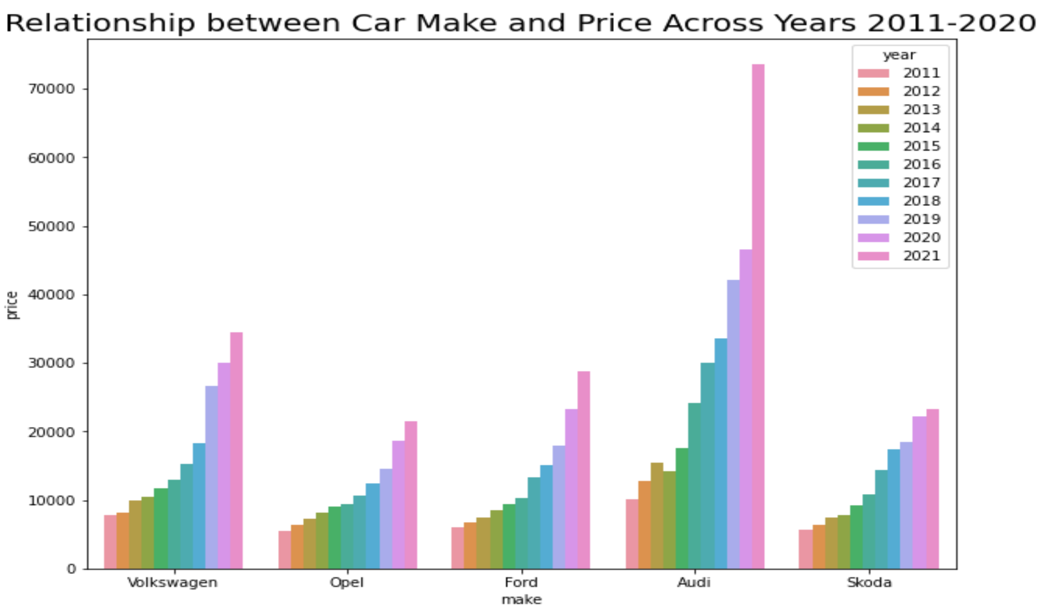


Figure 2 Relationship between Car Make and Price

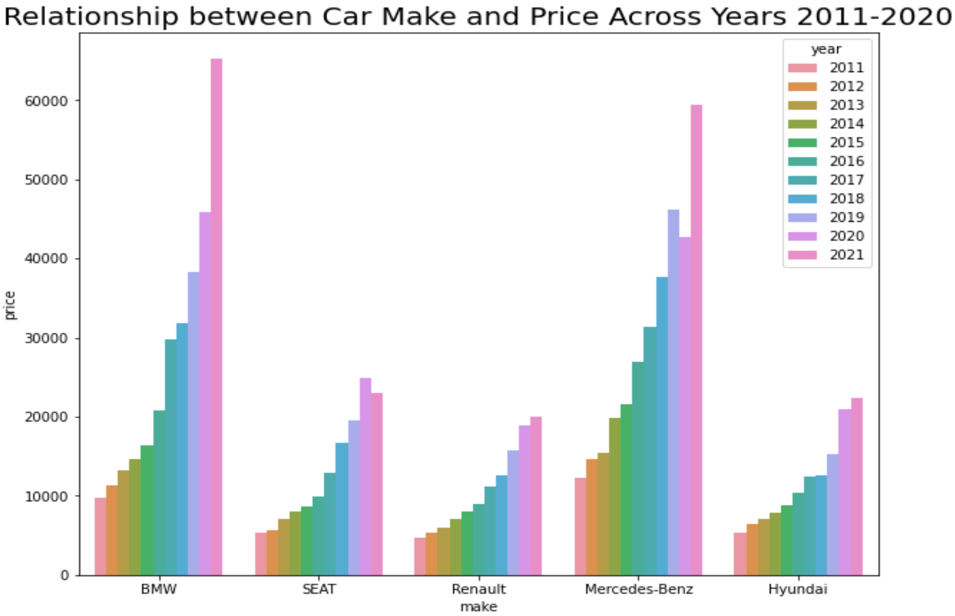


Figure 3 Relationship between Car Make and Price

The grouped bar chart shows the relationship between the make of the car and the prices between the years 2011-2020. The chart indicates that as the year increases, there is a corresponding increase in the price of the cars across all cars makes. However, there are some exceptions like in Audi the price in year 2013 was slightly higher than that of 2014

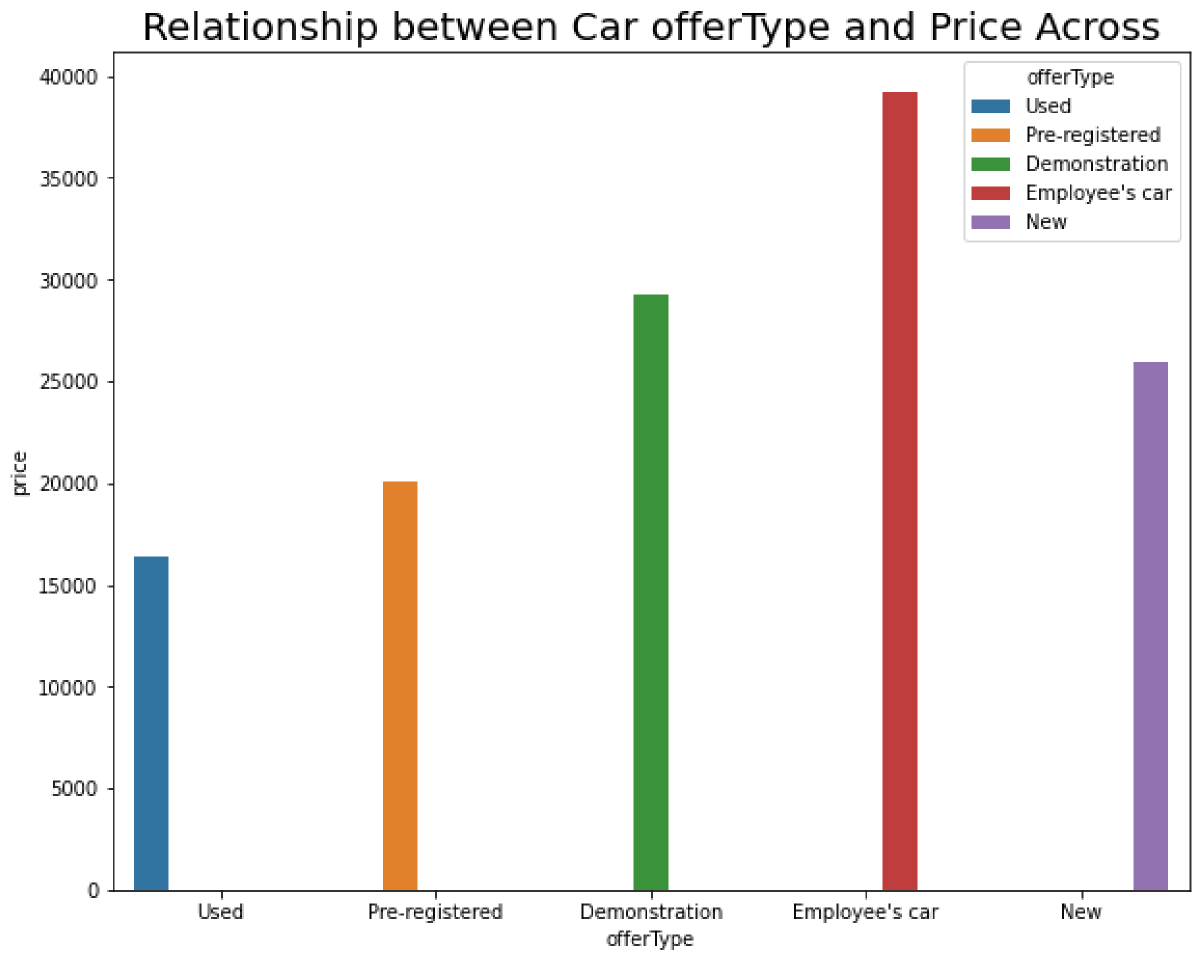
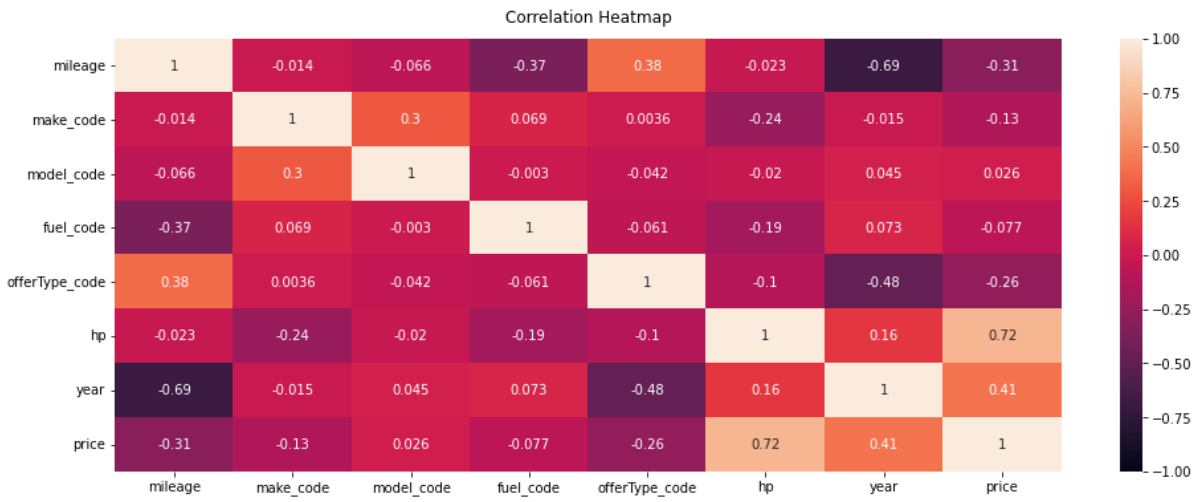


Figure 4 Relationship between Car OfferType and Price

In the bar chart above, the price of used cars is much lower than other offerTypes. Employee's cars and Demonstration cars are even more expensive than new cars. This is important information for the model.

Figure 5 Relationship between Fuel Type and Price

The grouped bar chart above shows that the individual fuel type of the cars increases in price over the years. We can also see gaps where the fuel types are not available in some years.



	mileage	make_code	model_code	fuel_code	offerType_code	hp	year	price
mileage	1	-0.014	-0.066	-0.37	0.38	-0.023	-0.69	-0.31
make_code	-0.014	1	0.3	0.069	0.0036	-0.24	-0.015	-0.13
model_code	-0.066	0.3	1	-0.003	-0.042	-0.02	0.045	0.026
fuel_code	-0.37	0.069	-0.003	1	-0.061	-0.19	0.073	-0.077
offerType_code	0.38	0.0036	-0.042	-0.061	1	-0.1	-0.48	-0.26
hp	-0.023	-0.24	-0.02	-0.19	-0.1	1	0.16	0.72
year	-0.69	-0.015	0.045	0.073	-0.48	0.16	1	0.41
price	-0.31	-0.13	0.026	-0.077	-0.26	0.72	0.41	1

Figure 6 Correlation Heat Map

The correlation matrix above includes the categorical variables after they have been encoded. There is a weak negative correlation of -0.26 and -0.13 between the offType and make respectively and the price.

2.1.3 Hypothesis

From the correlation matrix and bar charts there is a relationship between the explanatory variables (independent variables) and the response variable (dependent variable). Therefore, the null hypothesis would be that there is no significant relationship between the independent variables and the price (dependent variable). While the alternative hypothesis will be that there is a significant relationship between the independent variables and the price. The null hypothesis means the vehicle features and year registered do not determine if there would be an increase or decrease in price in the future. While the alternative hypothesis means that the vehicle features and the year registered determine if there would be an increase or decrease in price of the vehicle in the future.

2.2 Data Preparation

Section 2.3 explains **Data Preparation** as the third stage of the CRISP DM Methodology

For this analysis, data from two different sources will be used as explained in Task 1. The two tables are similar. Due to the completeness of the first Dataset, the second Dataset was prepared to look like the first.

	mileage	make	model	fuel	gear	offerType	price	hp	year
0	235000	BMW	316	Diesel	Manual	Used	6800	116.0	2011
1	92800	Volkswagen	Golf	Gasoline	Manual	Used	6877	122.0	2011
2	149300	SEAT	Exeo	Gasoline	Manual	Used	6900	160.0	2011
3	96200	Renault	Megane	Gasoline	Manual	Used	6950	110.0	2011
4	156000	Peugeot	308	Gasoline	Manual	Used	6950	156.0	2011

Figure 7 First Dataset

	Uniq Id	Crawl Timestamp	Pageurl	Country	Make	Model Name	Mileage	First Registration	Power	Type	Body Type	Description Text
0	c681b07bf81fac6ade9074a2d0ce716b	2020-02-11 01:49:34 +0000	https://www.autoscout24.com/offers/skoda-fabia...	Spain	Skoda	Fabia	193,000 km	12/2006	59 kW	Used	Station wagon	Skoda Fabia 1. 4 TDI 80cv. Motor muy fiable y ...
1	597e7936b4eea5c465e0ecd7a4b9beb7	2020-02-10 00:53:51 +0000	https://www.autoscout24.com/offers/citroen-ds3...	Germany	Citroen	DS3	48,500 km	01/2013	115 kW	Used	Coupe	Polster: Leder Sitzheizung vorn Klimaautomatik...
2	dd359ff8c84b97d75bb15d40a1588223	2020-02-10 07:44:58 +0000	https://www.autoscout24.com/offers/fiat-500x-1...	Italy	Fiat	500X	10 km	12/2019	88 kW	Employee's car	Off-Road/Pick-up	PREZZO CON RITIRO AUTO USATA = € 17.990.00 Mes...
3	b10c30971ed6ead4c6851450105b7138	2020-02-09 19:37:56 +0000	https://www.autoscout24.com/offers/renault-sce...	Europe	Renault	Scenic	95,000 km	04/2006	5 kW	Used	Other	GRANDIAUTO SRLÂ PROPONEÂ RENAULT MEGANE SCENIC...
4	c5568714c40c147d104452aba1aafd3a	2020-02-08 14:28:44 +0000	https://www.autoscout24.com/offers/bmw-118-ser...	Italy	BMW	118	190,000 km	03/2007	90 kW	Used	Sedans	Macchina in perfette condizioni, bollo pagato ...

Figure 8 Second Dataset

2.2.1 Data Cleaning

For the second dataset, firstly, The Uniq Id and Crawl Stamp columns were dropped changing the shape of the dataset from (30,030, 12) to (30,030, 10). The dataset contained observations from different countries in Europe, but for the purpose of this analysis, only observations from Germany were used. This changed the shape of the dataset to (7823, 10).

	Pageurl	Country	Make	Model Name	Mileage	First Registration	Power	Type	Body Type	Description Text
1	https://www.autoscout24.com/offers/citroen-ds3...	Germany	Citroen	DS3	48,500 km	01/2013	115 kW	Used	Coupe	Polster: Leder Sitzheizung vom Klimaautomatik...
12	https://www.autoscout24.com/offers/seat-alhamb...	Germany	SEAT	Alhambra	69,000 km	02/2015	130 kW	Used	Van	Sonderausstattung: Anhängerkupplung (Kugelkopf...
13	https://www.autoscout24.com/offers/opel-mokka-...	Germany	Opel	Mokka	107,000 km	12/2015	100 kW	Used	Off-Road/Pick-up	Airbag-Deaktivierung Beifahrerseite Anhängerzu...
14	https://www.autoscout24.com/offers/dacia-sande...	Germany	Dacia	Sandero	6 km	-/-	74 kW	New	Compact	Modell: Facelift-Modell; 3 Jahre Werksgarantie...
18	https://www.autoscout24.com/offers/opel-corsa-...	Germany	Opel	Corsa	16,892 km	02/2019	66 kW	Used	Compact	Getriebe: Schaltgetriebe Technik: Bordcomputer...

Figure 9 Selecting Germany

Afterwards, the Country column was dropped because it is no longer needed for analysis changing the shape to (7823, 9).

	Pageurl	Make	Model Name	Mileage	First Registration	Power	Type	Body Type	Description Text
0	https://www.autoscout24.com/offers/citroen-ds3...	Citroen	DS3	48,500 km	01/2013	115 kW	Used	Coupe	Polster: Leder Sitzheizung vom Klimaautomatik...
1	https://www.autoscout24.com/offers/seat-alhamb...	SEAT	Alhambra	69,000 km	02/2015	130 kW	Used	Van	Sonderausstattung: Anhängerkupplung (Kugelkopf...
2	https://www.autoscout24.com/offers/opel-mokka-...	Opel	Mokka	107,000 km	12/2015	100 kW	Used	Off-Road/Pick-up	Airbag-Deaktivierung Beifahrerseite Anhängerzu...
3	https://www.autoscout24.com/offers/dacia-sande...	Dacia	Sandero	6 km	-/-	74 kW	New	Compact	Modell: Facelift-Modell; 3 Jahre Werksgarantie...
4	https://www.autoscout24.com/offers/opel-corsa-...	Opel	Corsa	16,892 km	02/2019	66 kW	Used	Compact	Getriebe: Schaltgetriebe Technik: Bordcomputer...

Figure 10 Dropping Country Column

Next, the Page URL contains the fuel type of the car. Therefore, using regular expressions, the fuel type was extracted and placed in a new column called 'Fuel'

index	Pageurl
0	https://www.autoscout24.com/offers/citroen-ds3-thp155-sportchic-hifi-leder-pdc-gasoline-yellow-7df0773d-434b-4011-aad7-171f225cfb06
1	https://www.autoscout24.com/offers/seat-alhambra-style-710-diesel-black-f958b0ad-d02b-4281-88a2-3f8fb6f82683

Fuel
gasoline
diesel

Figure 11 Regex for Fuel Type

The ‘Pageurl’ column was dropped afterwards by selecting only the needed columns in the order of the first dataset: ('Mileage', 'Make', 'Model Name', 'Fuel', 'Type', 'Power', 'First Registration')

	Mileage	Make	Model Name	Fuel	Type	Power	First Registration
0	48,500 km	Citroen	DS3	gasoline	Used	115 kW	01/2013
1	69,000 km	SEAT	Alhambra	diesel	Used	130 kW	02/2015
2	107,000 km	Opel	Mokka	diesel	Used	100 kW	12/2015
3	6 km	Dacia	Sandero	NaN	New	74 kW	-/-
4	16,892 km	Opel	Corsa	gasoline	Used	66 kW	02/2019

Figure 12 Dropping PageUrl Column

Next, in the ‘First Registration’ column, the ‘-/-’ represents a new car. Therefore, using pandas string replace, it was changed to 2020 because that was the crawl time. So, it is assumed the new car was registered in the crawl time.

3	6 km	Dacia	Sandero	NaN	New	74 kW	-/-

3	6 km	Dacia	Sandero	NaN	New	74 kW	2020

Figure 13 Replacing Symbols

The km and KW units found in the ‘Mileage’ and ‘Power’ columns were replaced with an empty string so that it looks like the first dataset.

	Mileage	Make	Model Name	Fuel	Type	Power	First Registration
0	48,500	Citroen	DS3	gasoline	Used	115	2013
1	69,000	SEAT	Alhambra	diesel	Used	130	2015
2	107,000	Opel	Mokka	diesel	Used	100	2015
3	6	Dacia	Sandero	NaN	New	74	2020
4	16,892	Opel	Corsa	gasoline	Used	66	2019

Figure 14 Removing SI Unit

The ‘-’ symbol in the ‘Mileage’ column represents a new car with no mileage, therefore it was replaced to ‘0’ using pandas replace string.

7821	-	Trucks-Lkw	Renault	gasoline	New	103 kW	2020

7821	0	Trucks-Lkw	Renault	gasoline	New	103 kW	2020

Figure 15 Replacing Symbols

The observations with years between 2011 - 2020 were selected from the dataset because the first dataset only contains observations from 2011 - 2021 making the shape now (6500, 7). Lastly, the rows with the missing values were dropped using the dropna() method in pandas. The gear column was supposed to be derived from the 'Description Text' which contained keywords implying that the car was either automatic or manual. However, not all rows contained the information, therefore, it was dropped to have more observations.

2.2.2 Data Transformation

The power in the first dataset is in hp while that of the second dataset is in KW. Therefore, the KW was converted to hp by multiplying the 'Power' column by '1.34102'. Firstly, the data type of the 'Power' column had to be transformed from 'object' to 'float' to be able to multiply with the float number.

	Mileage	Make	Model Name	Fuel	Type	Power	First Registration
0	48,500	Citroen	DS3	gasoline	Used	154.21730	2013
1	69,000	SEAT	Alhambra	diesel	Used	174.33260	2015
2	107,000	Opel	Mokka	diesel	Used	134.10200	2015
3	6	Dacia	Sandero	NaN	New	99.23548	2020
4	16,892	Opel	Corsa	gasoline	Used	88.50732	2019

Figure 16 Converting hp to KW

The 'Power' column was rounded to a whole number just like the first dataset and the type of the column was converted to an integer.

	Mileage	Make	Model Name	Fuel	Type	Power	First Registration
0	48,500	Citroen	DS3	gasoline	Used	154	2013
1	69,000	SEAT	Alhambra	diesel	Used	174	2015
2	107,000	Opel	Mokka	diesel	Used	134	2015
3	6	Dacia	Sandero	NaN	New	99	2020
4	16,892	Opel	Corsa	gasoline	Used	89	2019

Figure 17 Rounding hp

The 'First Registration' column was converted from 'object' to 'datetime'. To resemble the first dataset, only the year was extracted.

First Registration
01/2013
02/2015
12/2015
2020
02/2019

First Registration
2013-01-01
2015-02-01
2015-12-01
2020-01-01
2019-02-01

First Registration
2013
2015
2015
2020
2019

Figure 18 Converting Year

The second dataset does not contain a price column. To solve this, the average car price for each car 'make' in their different years in the first dataset was used as the price for the second column. The groupby pandas function was used. The dataset was grouped by the car 'Make' and 'year' columns and the average of the price was calculated across these two columns.

		price
make	year	
9ff	2018	7000.000000
Abarth	2011	5990.000000
	2012	8299.000000
	2013	12990.000000
	2014	12593.333333
	2015	11850.000000
	2016	14440.000000
	2017	15970.000000
	2018	19490.000000
	2019	20301.428571
	2020	23572.300000
	2021	25690.083333
Aixam	2014	8500.000000
	2021	19775.000000
Alfa	2011	6513.625000
	2012	6002.458333
	2013	8214.666667
	2014	8705.571429
	2015	11860.000000
	2016	14450.000000

Figure 19 Selecting Price Average

After the integration was completed as explained in the next chapter, the categorical variables were converted to numerical values using Label Encoder in pandas.

	mileage	make_code	model_code	fuel_code	offerType_code	hp	year	price
0	235000.0	7.0	38.0	1.0	4.0	116.0	2011	6800.000000
1	92800.0	70.0	415.0	7.0	4.0	122.0	2011	6877.000000
2	149300.0	61.0	340.0	7.0	4.0	160.0	2011	6900.000000
3	96200.0	59.0	529.0	7.0	4.0	110.0	2011	6950.000000
4	156000.0	54.0	37.0	7.0	4.0	156.0	2011	6950.000000
...	...	...	...	...	...	...	...	...
52784	0.0	70.0	726.0	7.0	2.0	296.0	2020	29998.160888
52785	110000.0	70.0	415.0	1.0	4.0	148.0	2014	10519.720466
52786	57000.0	51.0	600.0	1.0	4.0	109.0	2015	9764.921569
52787	22500.0	52.0	538.0	7.0	4.0	114.0	2016	9468.565141
52788	10.0	0.0	81.0	7.0	2.0	158.0	2020	23572.300000

52620 rows × 8 columns

Figure 20 Encoding Categorical Variables

2.2.3 Data Integration

From the dataframe with calculated average price in the first column, the 'make' and 'year' columns were used as a unique key to merge the price column to the second dataset.

	Mileage	Make	Model Name	Fuel	Type	Power	First Registration	price
0	48,500	Citroen	DS3	gasoline	Used	154	2013	6846.804124
1	69,000	SEAT	Alhambra	diesel	Used	174	2015	8695.680365
2	107,000	Opel	Mokka	diesel	Used	134	2015	8994.743542
3	16,892	Opel	Corsa	gasoline	Used	89	2019	14559.588652
4	56,811	Audi	A3	diesel	Used	138	2012	12762.071730

Figure 21 Merging Price

The two dataframes were merged using pandas merge function, joined using left join. The 2nd dataset which was the 'left_on' was merged on the 'Make' and 'First Registration' column and the new dataframe was joined on the 'make' and 'year' column. The price column was created and populated with the mean price in the new dataframe.

After the price was included, the column names were renamed to match the first dataset.

	mileage	make	model	fuel	offerType	hp	year	price
0	48,500	Citroen	DS3	gasoline	Used	154	2013	6846.804124
1	69,000	SEAT	Alhambra	diesel	Used	174	2015	8695.680365
2	107,000	Opel	Mokka	diesel	Used	134	2015	8994.743542
3	16,892	Opel	Corsa	gasoline	Used	89	2019	14559.588652
4	56,811	Audi	A3	diesel	Used	138	2012	12762.071730

Figure 22 Rename Column Names

The columns were also rearranged so that is simple to merge with the first dataset

	mileage	make	model	fuel	offerType	price	hp	year
0	48,500	Citroen	DS3	gasoline	Used	6846.804124	154	2013
1	69,000	SEAT	Alhambra	diesel	Used	8695.680365	174	2015
2	107,000	Opel	Mokka	diesel	Used	8994.743542	134	2015
3	16,892	Opel	Corsa	gasoline	Used	14559.588652	89	2019
4	56,811	Audi	A3	diesel	Used	12762.071730	138	2012

Figure 23 Rearrange Column

The final shape of the transformed 2nd dataset is now (6384, 8). The shape of the first dataset is (46,405, 9). It is not possible to integrate two datasets with different column sizes. For a seamless integration, the 'gear' column was dropped so that it matches the second dataset making the first dataset shape now (46,405, 8).

Using pandas concat function, the two datasets were merged, and the index was reset increasing the observations to 52789.

	index	mileage	make	model	fuel	offerType	price	hp	year
0	0	235000	BMW	316	Diesel	Used	6800.000000	116.0	2011
1	1	92800	Volkswagen	Golf	Gasoline	Used	6877.000000	122.0	2011
2	2	149300	SEAT	Exeo	Gasoline	Used	6900.000000	160.0	2011
3	3	96200	Renault	Megane	Gasoline	Used	6950.000000	110.0	2011
4	4	156000	Peugeot	308	Gasoline	Used	6950.000000	156.0	2011
...	...	...	...	...	...	...	...	...	...
52784	6494	0	Volkswagen	T-Roc	gasoline	New	29998.160888	296.0	2020
52785	6495	110,000	Volkswagen	Golf	diesel	Used	10519.720466	148.0	2014
52786	6496	57,000	Nissan	Pulsar	diesel	Used	9764.921569	109.0	2015
52787	6497	22,500	Opel	Mokka	gasoline	Used	9468.565141	114.0	2016
52788	6498	10	Abarth	595C	gasoline	New	23572.300000	158.0	2020

52789 rows × 9 columns

Figure 24 Integrating Datasets

The extra 'index' column was dropped as well. After removing all NaN values, the final shape of the integrated dataset became (52620, 8)

	mileage	make	model	fuel	offerType	price	hp	year
0	235000	BMW	316	Diesel	Used	6800.000000	116.0	2011
1	92800	Volkswagen	Golf	Gasoline	Used	6877.000000	122.0	2011
2	149300	SEAT	Exeo	Gasoline	Used	6900.000000	160.0	2011
3	96200	Renault	Megane	Gasoline	Used	6950.000000	110.0	2011
4	156000	Peugeot	308	Gasoline	Used	6950.000000	156.0	2011
...	...	...	...	...	...	...	...	...
52784	0	Volkswagen	T-Roc	gasoline	New	29998.160888	296.0	2020
52785	110,000	Volkswagen	Golf	diesel	Used	10519.720466	148.0	2014
52786	57,000	Nissan	Pulsar	diesel	Used	9764.921569	109.0	2015
52787	22,500	Opel	Mokka	gasoline	Used	9468.565141	114.0	2016
52788	10	Abarth	595C	gasoline	New	23572.300000	158.0	2020

52620 rows × 8 columns

Figure 25 Removing Index

2.3 Data Analysis

Section 2.3.1 and 2.3.2 talk about the fourth stage, **Data Modelling** of CRISP DM Methodology.

2.3.1 Identification and Justification of Statistical Test

For this analysis, Multiple linear regression will be used as the statistical tool to justify the initially stated hypothesis. Multiple linear regression is used when there is more than one response variable (dependent variables) and one explanatory variable (independent variable) (Zach, 2021). In this case study, there are 7 dependent variables (make, mileage, hp, offerType, year, model, fuel) and 1 independent variable which is the price. The multiple linear regression is used to establish if there is a relationship amongst the stated variables. To estimate the relationship, the following formula is used (Zach, 2021).

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

Where:

$\hat{y}$ is the estimated response value (estimated dependent variable)

β_0 is the average value of when all the independent variables are equal to zero

β_1 is the average change in y with a unit increase in x_1

x_1 is the value of the independent variable x_1

The multiple linear regression makes use of the following null and alternative hypothesis

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

$$H_A: \beta_1 = \beta_2 = \ldots = \beta_k \neq 0$$

The null hypothesis ($\boldsymbol{H_0}$) states that every coefficient of the independent variable is equal to zero (Zach, 2021). This indicates that the independent variables do not have a significant relationship with the dependent variables. The the alternative hypothesis ($\boldsymbol{H_A}$) states that not all the coefficients of the independent variables are equal to zero (Zach, 2021).

In multiple linear regression, the t-test is used to validate the linearity of the linear relationship. The one sample t-test is used to test that the null hypothesis is equal to zero.

2.3.2 Application of Test

One of the assumptions of multiple linear regression is that the residuals are normally distributed. This applies to only continuous variables. In the integrated dataset there are 3 variables that are continuous: mileage, hp and price, 3 are categorical variables: make, model, offerType and finally, the year is a Discrete Variable.

Normality Test for Continuous Variables

To test for normality in the continuous variables, a histogram with a kde plot was used and a QQ plot. This is used to show the skew in the dataset.

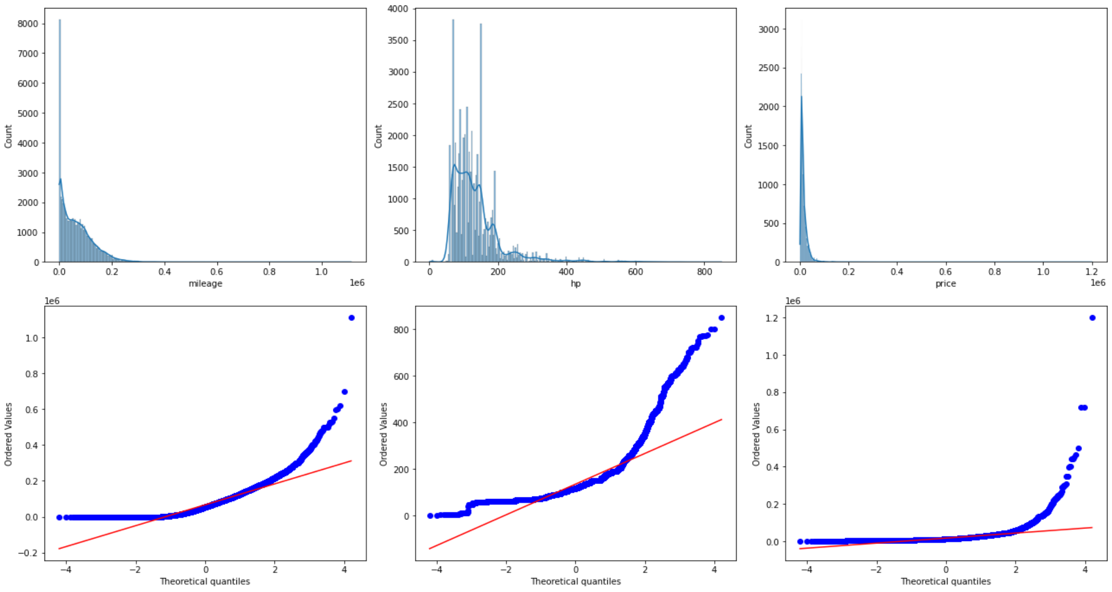


Figure 26 Kde and QQ Plot

From the histogram with kde plot, the data is not normally distributed. All 3 continuous variables are skewed to the right. The skew test result shown below also confirms this as all the values are greater than zero.

```
price      12.194043
hp          2.752529
mileage     1.416851
dtype: float64
```

The descriptive analysis of these three variables is also shown below:

	mileage	hp	price
count	5.262000e+04	52620.000000	5.262000e+04
mean	6.650635e+04	135.260243	1.701375e+04
std	6.184253e+04	75.561881	1.884377e+04
min	0.000000e+00	1.000000	1.100000e+03
25%	1.522600e+04	87.000000	7.897500e+03
50%	5.392650e+04	118.000000	1.195000e+04
75%	9.999900e+04	150.000000	2.045000e+04
max	1.111111e+06	850.000000	1.199900e+06

Figure 27 Descriptive Analysis

To remove the skew, the yeo-johnson technique was applied because it deals with the right skewed data. The value of the skew must be close to zero to be normally distributed.

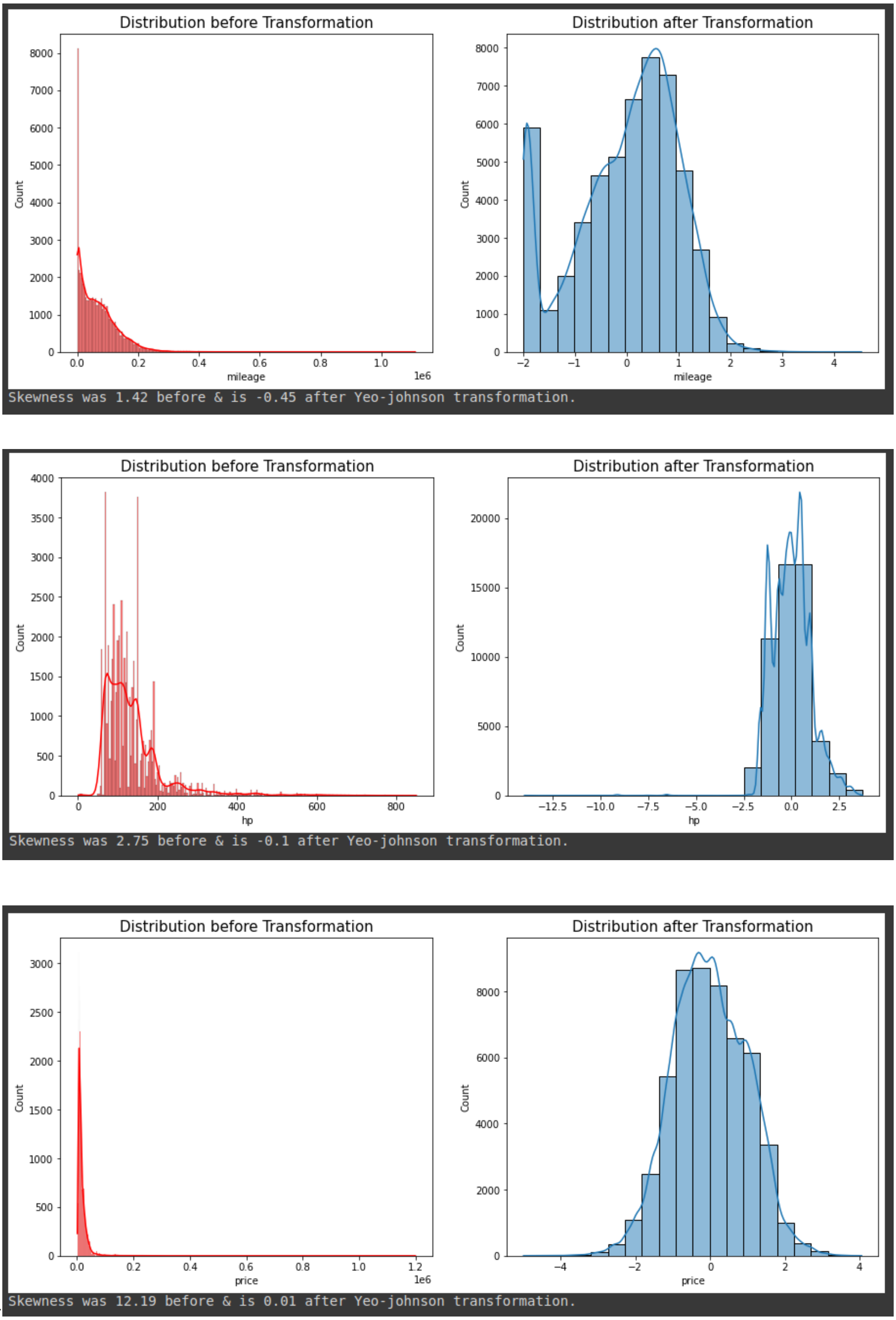


Figure 28 Yeo-Johnson Transformation of Continuous Variables

Removing Outliers

From the graphs above, the skew is now -0.45, -0.1 and 0.01 for mileage, hp and price respectively which is closer to 0 and now normally distributed. But from the graph there is still the presence of outliers which need to be removed. The box plot below shows this.

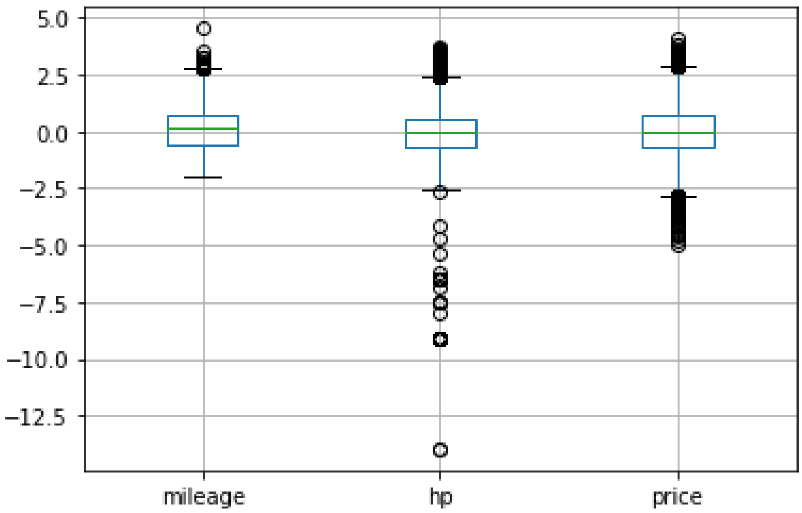


Figure 29 Presence of Outliers

Using the Interquartile range (IQR) technique, the measure of spread from the middle half of the data was noted, and the data points which are further away from the mean were removed (Frost, n.d). The box plot below shows this.

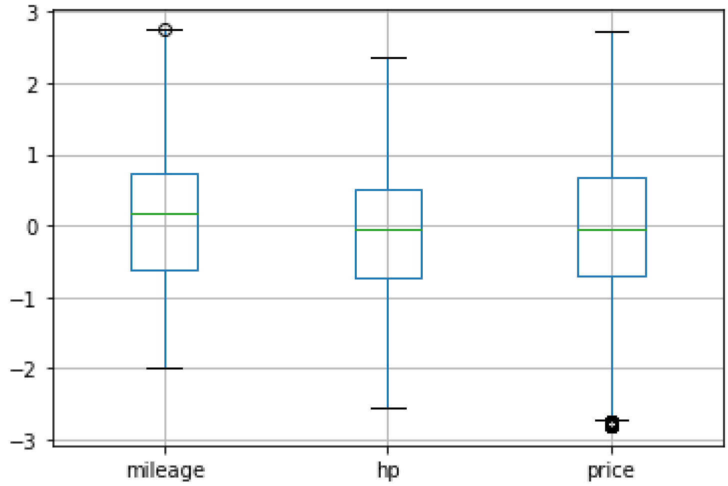


Figure 30 Removal of Outliers

Adding A Constant

A constant value was also included in the dataset to center the residuals. Using the Ordinary Least Square Regression method (OLS) from the statsmodel library in python. The results below show the t-test results, p-value, and the R-squared values.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.828
Model:                            OLS   Adj. R-squared:                  0.828
Method:                 Least Squares   F-statistic:                 3.256e+04
Date:                Wed, 12 Oct 2022   Prob (F-statistic):               0.00
Time:                        20:26:47   Log-Likelihood:                -22571.
No. Observations:               47421   AIC:                         4.516e+04
Df Residuals:                   47413   BIC:                         4.523e+04
Df Model:                           7
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const           -288.3300      1.959   -147.152      0.000    -292.170    -284.490
mileage           -0.2472      0.003    -72.728      0.000      -0.254      -0.241
make_code          0.0004   8.96e-05      4.467      0.000       0.000       0.001
model_code     -8.035e-06   7.19e-06     -1.117      0.264   -2.21e-05    6.07e-06
fuel_code         -0.0213      0.001    -29.285      0.000      -0.023      -0.020
offerType_code     0.0304      0.002     13.211      0.000       0.026       0.035
hp                 0.6173      0.002    270.947      0.000       0.613       0.622
year               0.1430      0.001    147.228      0.000       0.141       0.145
==============================================================================
Omnibus:                     1516.170   Durbin-Watson:                   1.311
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3546.207
Skew:                          -0.166   Prob(JB):                         0.00
Kurtosis:                       4.298   Cond. No.                     2.26e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.26e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Figure 31 Regression Summary

The p-values are truncated as seen in the result above. Below, the p-values are printed out to see more decimal places (Zach, 2022).

```
const            0.0
mileage          0.0
make_code        7.940788589404953e-06
model_code       0.2640637244236653
fuel_code        7.531616691210901e-187
offerType_code   8.967817848171448e-40
hp               0.0
year             0.0
```

Figure 32 P-values

2.3.3 Discussion of Test Outcome

Sections 2.3.3 to 2.3.5 talk about the fifth stage, **Evaluation of the Model** like in CRISP DM Methodology.

From the test result shown in 2.3.2, the derived model is:

$$\hat{y} = -288.330 - 0.2472x_1 + 0.0004x_2 - 8.035e^{-6}x_3 - 0.0213x_3 - 0.0213x_4 + 0.0304x_5 + 0.6173x_6 + 0.1430x_7$$

The coefficients of the x variables indicate the corresponding change of the y value (Rsundery, 2022). For example, the coefficient of x_1 which is -0.2472 means that for a unit change in mileage, the price will reduce by -0.2472, likewise for x_7, the price will increase by 0.1430 for a unit increase in a year. All coefficients are not zero, this gives a hint that we can accept the alternative hypothesis for all variables. It is also noted that the model_code is very close to 0.

Looking further into the variables, the t - test shows by how much the variable deviates from the null hypothesis which is zero (Gillespie, 2018). The calculation is shown below:

$$t = \frac{\beta_0 - null\ hypothesis}{Standard\ error}$$

For example, the t-test for mileage means that the mileage is -72.728 less than the null hypothesis (0). This also means that there is a negative correlation between the mileage and the price.

To check the significance of each vehicle feature, the p-value is used. To reject the null hypothesis, the p-value must be less than 0.05 which is the significance level (Data Courses, 2021). From the p-value results all features except model_code is significant because it is greater than 0.05. Therefore, the null hypothesis will be rejected for all vehicle features except model_code. The alternate hypothesis will be accepted for all vehicle features except the model_code.

The R^2 value is the coefficient of determination, and it explains in percentage how much the price can be determined by the vehicle features. From the result, the price can be explained by the vehicle features by 82.8%.

In conclusion, the vehicle features that are significant for determining the price are mileage, make, fuel, offerType, hp and year.

2.3.4 Graphical Interpretation and Further Test Outcomes

The residual versus mileage plot shows that there is no presence of heteroscedasticity. Meaning that the residuals are equally spread across the range of measured/fitted values and thereby there is a constant variance. The Y and Fitted Vs. X plot shows a decreasing linear relationship between the price and mileage. This means that the assumption for a linear relationship in regression analysis is met for the mileage variable.

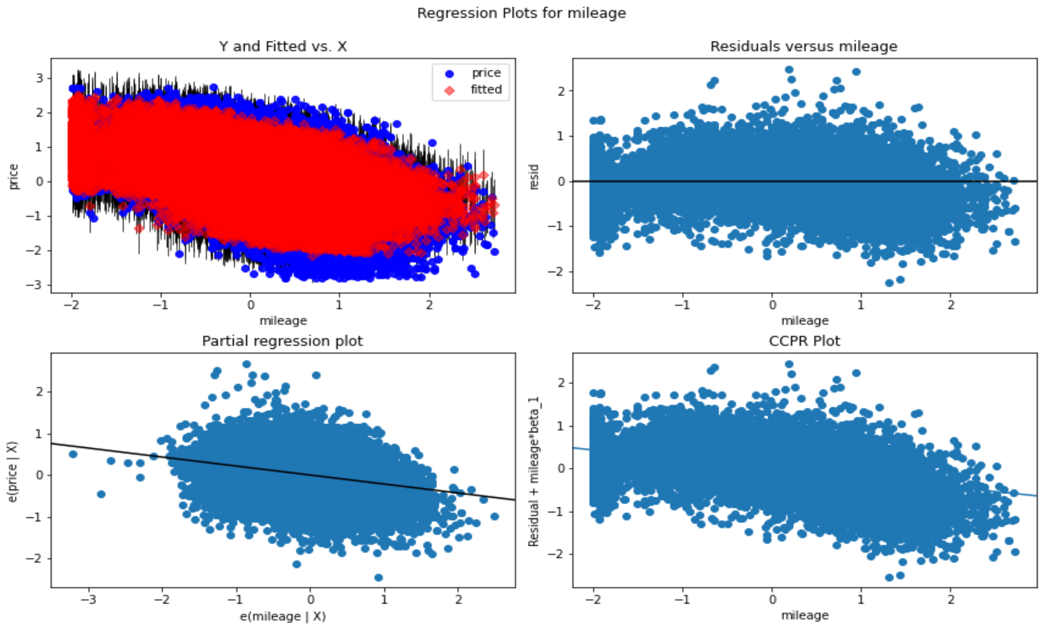


Figure 33 Regression Plot (mileage)

From the Partial regression plot for hp, it is seen that the fitted line is not horizontal. The price of the vehicle increases with an increase in hp.

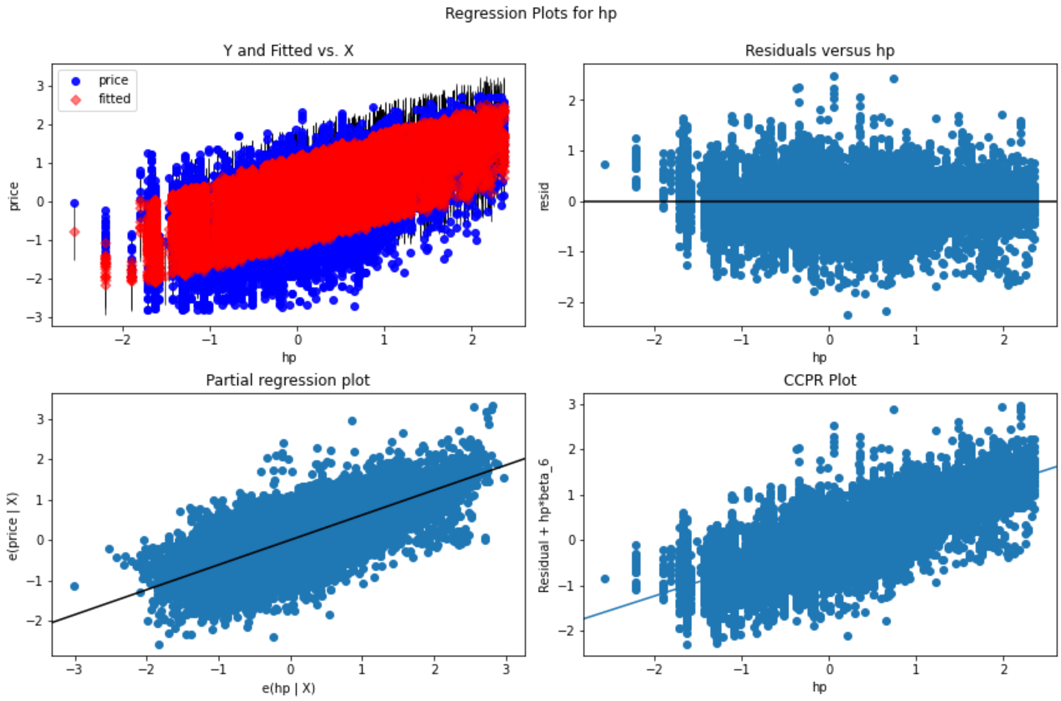


Figure 34 Regression Plot (hp)

Predicted Prices

The dataset was split into train and test sets. The model was trained with the vehicle features from the year 2011 to 2020. While the model was tested with the vehicle features in the year 2021. To evaluate the performance of the model, the model was used to predict the prices in the coming year 2021. The table below shows the model's predicted prices against the real prices in 2021. The predicted prices are not too far from the real prices and hence can provide a good estimate of what the future vehicle prices will be.

	real_price	predicted_price
0	10600.0	14791.169801
1	10680.0	15778.489573
2	10823.0	15982.606713
3	10980.0	15778.489573
4	10980.0	15778.489573
5	10997.0	16059.860896
6	11490.0	16069.838137
7	11490.0	16069.838137
8	11490.0	16069.838137
9	11499.0	14324.252265
10	11790.0	16059.860896
11	11950.0	18768.640855
12	11990.0	16059.860896
13	11990.0	16059.860896
14	11990.0	16059.860896
15	11990.0	16059.860896
16	11990.0	16059.860896
17	11990.0	16059.860896
18	11990.0	16059.860896
19	11990.0	16059.860896

	real_price	predicted_price
3988	10140.0	16282.717876
3989	10140.0	16282.717876
3990	10140.0	16282.717876
3991	10140.0	16282.717876
3992	10140.0	16282.717876
3993	10540.0	16282.717876
3994	10790.0	13596.005044
3995	11805.0	14110.624922
3996	11990.0	15161.313676
3997	12340.0	15815.501153
3998	12340.0	15815.501153
3999	12490.0	16006.376444
4000	12805.0	14076.117552
4001	12805.0	13870.975878
4002	12980.0	11401.828378
4003	12990.0	16006.376444
4004	12990.0	16006.376444
4005	12990.0	16006.376444
4006	12990.0	16006.376444
4007	12990.0	16006.376444

Figure 35 Predicted and Actual Prices

The model works based on the derived formula shown in 2.3.3. With this model, Autoscout24 would be able to predict the future vehicle prices in 2023.

The bar chart below shows the errors between the actual and the predicted prices for 100 observations in the test data. The negative bars mean that the price was overestimated, and the positive bars means that the price was underestimated. The closer the bars are to zero, the more accurate the prediction of the model is.

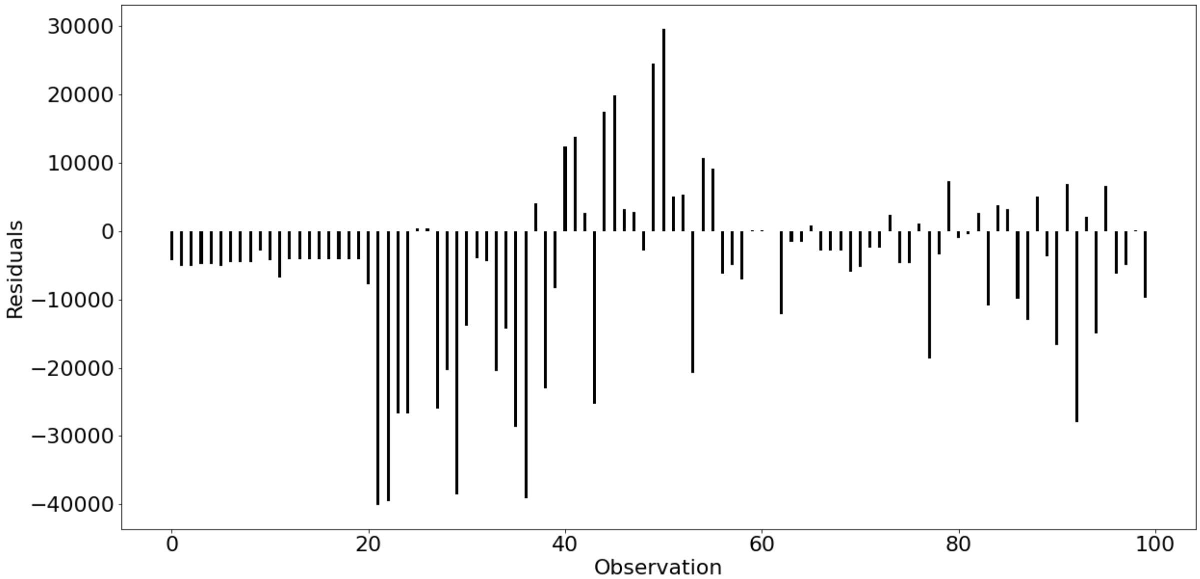


Figure 36 Residuals

2.3.5 Limitations, Assumptions and Enhancement for Better Accuracy

The derived regression model assumes that the cause-and-effect relationship between the vehicle features and the price remains unchanged (Homework1, n.d). Therefore, if there is a change in the relationship in the future, the model would not be able to capture it giving rise to misleading results. Because of the limited data available, it is possible that the regression model was not able to capture the relationship between the vehicle features and the price as it would have when given a larger dataset. The relationship might change with more input. To improve the accuracy of the model, more data is needed.

Although data transformation is important to fulfill the assumptions of multiple linear regression, it is also possible that the transformations done to the dataset may cause misinterpretation of the coefficients of the model (Bryan, 2020).

The regression model assumes that the relationship between the vehicle features and the price is linear, meaning that for a unit change in vehicle feature is a corresponding unit change in the price but in reality, this relationship might not be so, making the assumption of linearity to be restrictive (Hope, 2020). However, to accommodate the non-linearity, the data was transformed.

The problem of multicollinearity rises with the increase in the predictor variables in multiple linear regression models (Hope, 2020). Multicollinearity occurs when two predictor variables which have the same effect on the response variable are present in the model. As such, the correlations between these two variables will be high (Hope, 2020). This is a problem because the goal of multiple linear regression models is to estimate the relationship between the response variable and the predictor variable. Therefore, with highly correlated variables it becomes difficult for the model to estimate (Frost, 2017). The results of the presented model indicate a strong presence of multicollinearity in the model. To improve the accuracy of the model, the correlated variables must be removed.

2.4 Deployment

Section 2.4 talks about the Last stage of the CRISP DM Methodology which is **Model Deployment**.

2.4.1 Risk Analysis and Potential Challenges

For any software development project, security is a great concern. This also applies to machine learning models (Patel, n.d). Attackers can release malicious parties to steal the data, gain authorizations and possibly gain access to the results of the training data (Patel, n.d). To prevent this scenario, the issue of security must be kept in mind and the best practices should be followed from the beginning to the end of the deployment (Patel, n.d).

One potential challenge in model deployment is monitoring both the input and output of the model. It is important that the correct input is fed into the model. This requires high quality data to be fed continuously to the model from various sources. With incorrect data, the model will not be able to perform as expected (Patel, n.d). A post analysis of the model is also important to do. Doing this gives an insight into the performance of the model to spot potential challenges which need to be fixed to prevent even bigger problems in the company's system and processes (Patel, n.d).

Model drift is another common challenge. Sudden changes in the external environment can cause a change in the relationship between the vehicle features and the price (Patel, n.d). This can lead to model degradation. It is important to constantly retrain the model to accommodate the changes in the external environment. Other causes of model drift are faulty pipelines, technological constraints, low data quality or even a change in the data distribution (Patel, n.d).

Rather than using Notebooks, Python files are preferred to improve the maintainability and reusability of the code (Patel, n.d). For the data preparation and training stages, reusing a code makes the workflow more reliable and scalable (Patel, n.d). To create a repeatable pipeline, the machine learning environment should be treated like a code. This way a key event can trigger the end-to-end pipeline (Patel, n.d).

Another consideration is the collaboration between teams like the data scientist, the sales, and the software engineer (Patel, n.d). The team members should be well acquainted with data science processes to contribute positively to the model's performance (Patel, n.d).

2.4.2 Ethical Aspects of the Deployment

There are six general ethical principles that any machine learning model must adhere to as documented in the Charter of Fundamental Rights of the European Union (EU Charter). These principles include Respect for Human Agency, Privacy and Data Governance, Fairness, Individual, Social and Environmental Well-being, Transparency and Accountability and Oversight (European Commission, 2021). Autoscout24 is a European company, therefore they must abide by these laws. The "*Ethics by Design*" states specific tasks to comply with the ethical guidelines for any development methodology like CRISP-DM as used in this case study.

The deployment and implementation aspect are the point where the model is released to the public to be used as well as planning and implementation changes in the company. The way in which the model is deployed can change the ethical characteristics of the model. Therefore, in implementation, the model must meet the ethical requirements always (European Commission, 2021). In order not to violate these principles, the ethical aspects of the deployment of this model should be followed as described in "*Ethics by Design*".

If the presented model contains personal information of the users, they should be deleted except there are justified reasons why the data cannot be deleted or could seriously affect the desired results of the model (European Commission, 2021).

Plans and policies that support operational compliance with the ethical requirements as stated above should be created and implemented in the system (European Commission, 2021).

Regular updates of the data, access, security and risk management procedures and policies should be done for the model to account for the ethical requirements (European Commission, 2021),

The newly created ethics politics should be added to the system when training for the operation of the system. Also, when launching the system, the ethical aspects should be properly implemented (European Commission, 2021).

Throughout the implementation phase of the system, the ethics guidelines should be properly monitored, potential risks and challenges should be identified and mitigations against these risks should be applied (European Commission, 2021).

2.4.3 Conclusions and Recommendations

In summary, the model was built using multiple linear regression. Null and alternative hypotheses were formulated and tested using t-test and p-value. The results of the tests were that all vehicle features except the car's model were significant for determining the price of the car. Therefore, removing the model of the car from the model did not change the performance of the model. The model's R-squared value is 0.828 meaning that only 82.8% of the price is explained by the vehicle features. The model was trained with the data between 2011-2020 and was tested with the data in 2021. This way the model was tested to predict the future prices of the car. A bar graph was used to visualize the errors of the model's prediction, the closer the bars are to zero the more accurate the model is.

It is recommended that more complex machine learning algorithms like K-Nearest Neighbors (KNN), Support Vector Machine and Deep Network are used for price prediction as the errors are substantially smaller (Micro AI, n.d). The model will perform better if more features that have a direct impact on the price are used, however as explained earlier, multicollinearity is an issue that must be looked out for when collecting features. External factors such as economic conditions and exchange rates in the supply chain are also important to get more useful predictions. Also, more observations will enable the model to perform better in training as it has more data to learn from.

With the presented regression model, the future prices of German cars sold by Autoscout24 can be estimated. The values are not exact, but it gives a rough estimate of what the prices can be. With the model, Autoscout24 can determine what car features have the most effect in determining the price of the car and dynamic pricing can be achieved. This prediction tool will also help Autoscout24 to manage their dealership pricing strategy better.

Reference

1. Autoebid, 2016. *Why mileage matters: The importance of mileage on a used car* [Online]. Available from: https://www.autoebid.com/blog/used-cars/importance-of-mileage-on-a-car [Accessed 29 September 2022]
2. Ou, S., Li, W., Li, J., Lin, Z., He, X., Bouchard, J. and Przesmitzki, S., 2020. Relationships between vehicle pricing and features: data driven analysis of the Chinese vehicle market. *Energies*, *13*(12), p.3088.
3. Murillo, A. L., 2021. *The Push for Electric Vehicles Should Determine How Much Your Next Car Costs* [Online]. Available from: https://money.com/electric-car-vs-gas-car-costs-biden/ [Accessed 29 September 2022]
4. Threewitt, C., 2017. *Why does horsepower matter?* [Online]. Available from: https://cars.usnews.com/cars-trucks/advice/why-does-horsepower-matter [Accessed 29 September 2022]
5. My Car Credit, 2022. *Are Automatic Cars More Expensive?* [Online]. Available from: https://www.mycarcredit.co.uk/are-automatic-cars-more-expensive/ [Accessed 29 September 2022]

6. Preethu, 2022. *Positive and Negative Correlation: Definition, Examples, Graphs and Comparison* [Online]. Available from: https://www.embibe.com/exams/positive-and-negative-correlation/#:~:text=If%20the%20value%20of%20one,said%20to%20be%20negatively%20correlated. [Accessed 5 October 2022]

7. Zach, 2021. *Understanding the Null Hypothesis for Linear Regression* [Online]. Available from: https://www.statology.org/null-hypothesis-for-linear-regression/#:~:text=The%20null%20hypothesis%20states%20that%20all%20coefficients%20in%20the%20model,What%20is%20this%3F&text=The%20alternative%20hypothesis%20states%20that%20not,is%20simultaneously%20equal%20to%20zero. [Accessed 5 October 2022]

8. Frost, J., n.d. *Interquartile Range (IQR): How to Find and Use It* [Online]. Available from: https://statisticsbyjim.com/basics/interquartile-range/ [Accessed 11 October 2022]

9. Data Courses, 2021. *P-Value: What is It, How to Calculate It and Why it matters* [Online]. Available from: https://www.datacourses.com/find-p-value-significance-in-scikit-learn-3810/ [Accessed 11 October 2022]

10. Zach, 2022. *How to Extract P-values from Linear Regression in Statsmodel* [Online]. Available from: https://www.statology.org/statsmodels-linear-regression-p-value/ [Accessed 11 October 2022]

11. Gillespie, C., 2018. *What does a Negative T-Value Mean?* [Online]. Available from: https://sciencing.com/negative-tvalue-mean-6921215.html [Accessed 11 October 2022]

12. Rsundery, 2022. *Interpreting the Results of Linear Regression using OLS Summary* [Online]. Available from: https://www.geeksforgeeks.org/interpreting-the-results-of-linear-regression-using-ols-summary/ [Accessed 11 October 2022]

13. Bryan, S., 2020. *Linear Regression in Pricing Analysis, Essential Things to Know* [Online]. Available from: https://www.bryanshalloway.com/2020/08/17/pricing-insights-from-historical-data-part-1/#inference-and-challenges [Accessed 14 October, 2022]

14. Hope, T.M., 2020. Linear regression. In *Machine Learning* (pp. 67-81). Academic Press.

15. Frost, J., 2017. *Multicollinearity in Regression Analysis: Problems, Detection, and Solutions* [Online]. Available from: https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/ [Accessed 14 October 2022]

16. Homework1, n.d. *Limitations of Linear Regression Analysis* [Online]. Available from: https://homework1.com/statistics-homework-help/limitations-of-regression-analysis/ [Accessed 14 October 2022]

17. European Commission, 2021. *Ethics By Design and Ethics of Use Approaches for Artificial Intelligence* [Online]. Available from: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-by-design-and-ethics-of-use-approaches-for-artificial-intelligence_he_en.pdf [Accessed 14 October, 2022]

18. Micro AI, n.d. *Using Machine Learning Algorithms to Predict Pricing Trends* [Online]. Available from: https://www.micro.ai/blog/using-machine-learning-algorithms-to-predict-pricing-trends [Accessed 14 October 2022]

19. Patel, H., n.d. *Challenges in Deploying Machine Learning Models* [Online]. Available from: https://censius.ai/blogs/challenges-in-deploying-machine-learning-models [Accessed 17 October 2022]