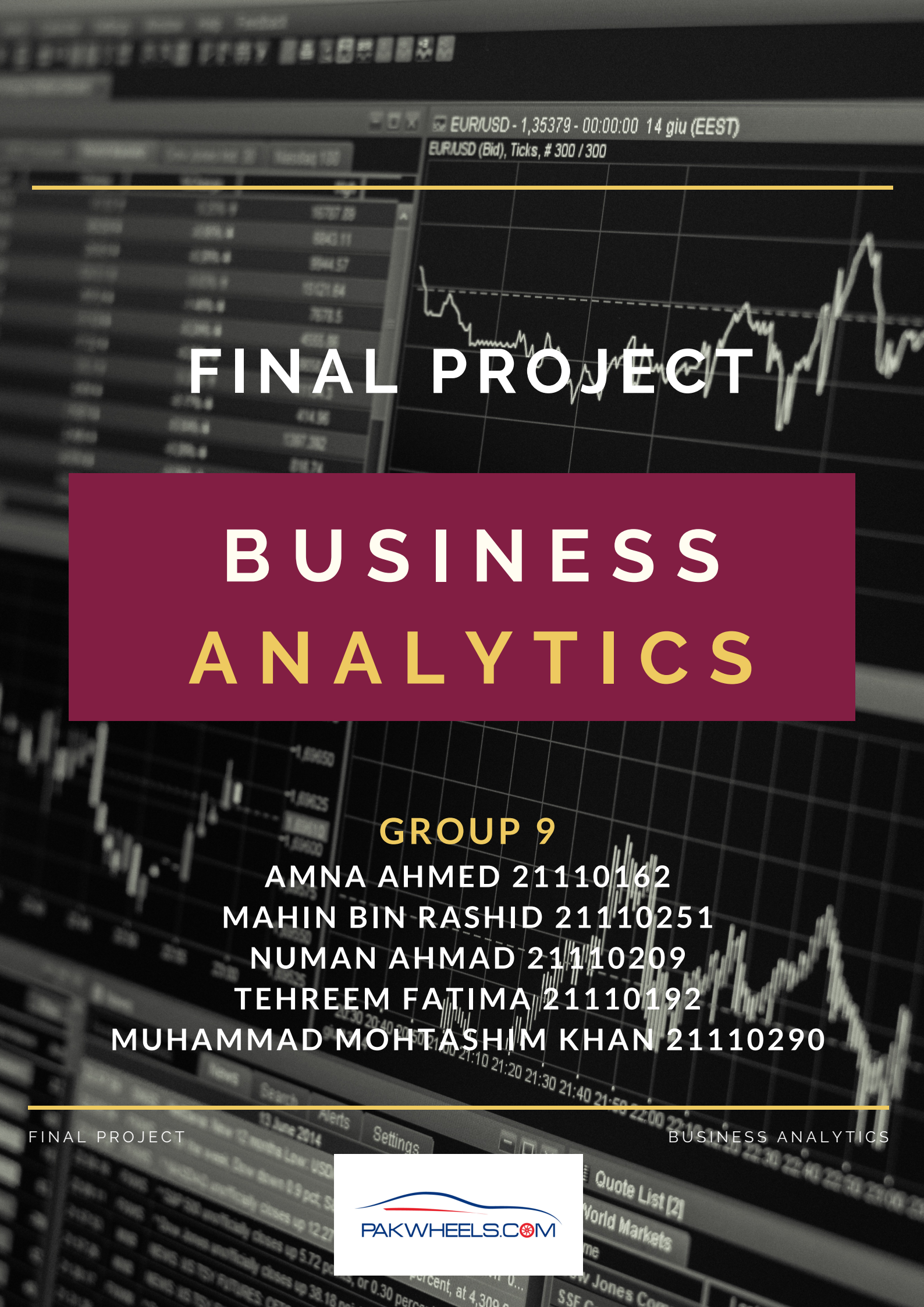


FINAL PROJECT

BUSINESS ANALYTICS

GROUP 9

AMNA AHMED 21110162

MAHIN BIN RASHID 21110251

NUMAN AHMAD 21110209

TEHREEM FATIMA 21110192

MUHAMMAD MOHTASHIM KHAN 21110290

INTRODUCTION

BACKGROUND

Pakwheels.com is Pakistan's leading e-commerce platform for autombiles.They have been in the industry since 2002. PakWheels itself is is an online marketplace for buyers and sellers of cars, taking out the middleman and providing them with a platform to engage in such dealings. PakWheels also provides users with automotive reviews, comparison tools, insurance information, and other car informatory tools.
Being an online platform, every day thousands of gigabytes of data are posted on this platform, but most of it is not processed to generate meaningful insights. In our project we have focused to utilize the data present on their website to recommend PakWheels regarding how they can optimize the usage of data posted by advertisers to enhance the experience of not only sellers but also of buys.

LITERATURE REVIEW

According to Dawn, the Ministry of Commerce Pakistan has stated a massive growth of e-commerce market size; growing over 35 percent in just the first quarter of the fiscal year 2021 (Khan, 2021). According to an online resource, this growth in Pakistan's eCommerce industry reflects an increased number of people buying online. Along with the increased online purchase of other items, vehicle shopping and information gathering online is also rising ("How eCommerce Is Changing The Way We Shop For Cars In Pakistan - PakWheels Blog," n.d.).
When we talk about e-commerce specifically in the automobile industry "PakWheels" is a well-established name that comes to mind. Founded in 2003, PakWheels was acquired by entrepreneurs Raza Saeed and Suneel Munj in 2008. Once being a vertically classified automotive portal, to date, PakWheels continues to revolutionize the traditional channels of buying and selling of (certified used and new) cars online (Arif and Sarfraz, 2017).
According to BusinessWire "Pakistan's automotive market is slowly shifting towards a service-oriented model with new players focusing extensively on customer experience and consumer data" ("2019 Future of Pakistan Automobile Market - Trends, Outlook and Growth Opportunities - ResearchAndMarkets.com," 2019). With each and every online interaction of buyers and sellers, more and more data is being collected. Several useful behavioral patterns, insights, as well as trends, can be gathered from this data that can form the basis of strategy formation. It is often claimed that the Geographical Diversification Model remains the most important strategy for leading car manufacturers and sellers within Pakistan. ("2019 Future of Pakistan Automobile Market - Trends, Outlook and Growth Opportunities - ResearchAndMarkets.com," 2019). This project tends to explore this claim in depth along with looking at relationships between car features and overall prices. Through this project, we also intend to explore customer segmentation based on their preferences towards cars. All of this will be done in an attempt to solve PakWheel's on going problem.

THE PROBLEM

Pakwheels is facing several important challenges which we will particularly address in this project. First of all their management wants to know if either city plays any significant role over other features of a car. Moreover, Currently they are not able to segment the advertisements properly to show them to the right customers, which is costing them waste of advertisement cost and a lot of opportunity cost which can be avoided by showing the right ad to the right person. Along with that they are yet confused about which features impact the price of a resale car, due to which Pakwheels is not able to understand the pattern of pricing yet.

DATA SELECTION

For the given problem, the data was scraped, in a structured format, from the PakWheels website. As stated before, PakWheels is an online marketplace for car buyers and sellers within Pakistan and all this includes new, used, and certified second-hand car, providing other car services such as car insurance tools, repair checkups etc. We have used Python's Beautiful4soup and selenium library to extract this data.

Explanation of Raw Dataset:
In our data set, we have 17 columns and 2675 rows, which were scrapped from the official website of pakwheel.com. In the raw format, we had the following variables in our data:

Name;
Modelyear;
total mileage;
consumption;
geartype;
registration_city;
Assembly;
bodytype;
color;
engine_capacity;
posted_date;
posted_location;
price;
added_via;

Null values were majorly seen only in 2 columns, out of total observation 4.1% values were Null in posted_date columns and 8.9% were Null in the added_via column. All the variables are in the character format initially. There were independent issues in each variable, which were addressed using different pieces of codes. We will explain their problems and their solutions over the course of the report.

PROCESSING OF DATA

1)EXTRACTING CAR'S BRAND NAME FROM NAME COLUMN

By printing the full unique names of the cars posted in ads, we can see that the first word is always the brand name, followed by the model name and then the rest of the details of that car, e.g. engine capacity, model year edition, engine type and etc. By printing the unique values we can see that almost every row has a unique name, which is not useful for our analysis as we cannot classify this information and can get quite long and redundant. Therefore, we tried to separately extract the name of the manufacturer and model.

To do this, we separated these names into three parts by making two additional columns in our data set, using the "separate" function. In the first column we have now got the brand's name, and in second and third column the brand name. Then we have used the "unite" function to combine column# 2 and 3 to get the final model name.

In this way we have divided the original name column into two columns, with the names of "Manufacturer" and "model_name". And we have removed the rest of the information in the name. As that was not required in our analysis due to lack of uniqueness in that information.

2) SEPARATING LOCATION, CITY AND PROVINCE FROM THE ORIGINAL POSTED_LOCATION COLUMN.

In the posted_location we have the complete address of the person posting an ad. This is again unique information in every column. We need to extract useful information to classify this information. After printing the unique names we notice that we can get 3 important pieces of information, i.e., town, city, and province.

In order to implement this, we again used the "separate" function to create two new columns, to portray the above-mentioned information. Hence, from the original "posted_location" column now we have got "Main_Location", "Posting_City" and "Province" columns in our dataset.

In some ads, there was no Main_location instead there was only city and province. And after observing the data, we observed that these two were there in every ad regardless of the Main location. Hence, we used the "coalesce" function to bring this city and province data from "Main_Location" to their respective columns.

3) TREATING WITH JUNK VALUES IN POSTED_LOCATION COLUMN

After using the "coalesce" function we have got empty cells in the "Main_Location" column. Hence, as we cannot predict these values using any means therefore, we have made a separate category for it, with the name of "Not Applicable".

4) TREATING THE SHORT FORM OF NAME WITH THEIR FULL NAMES IN ORIGINAL POSTED_LOCATION COLUMN

After using the separate function, some cities which had spaces in their names gave wrong values in their respective columns, e.g. for the city "Dera Gazi Khan" we got the "Dera" in the City column and "Gazi in the province columns. Hence, this needs to be treated.

As these values were only a few in numbers, therefore, we manually treated them to fill the right values in each cell.

5)TREATING NULL VALUES IN ADDED_VIA COLUMN

In the "Added_via" column we had two type of values, one of them was "Added_via_phone" and the other one was "null". After observing the ad posting options on Pakwheels we observed that the "null" value represents "Added_via_desktop". Hence, we replaced these values in our Data set. We had to employ the "if else" condition inside the "mutate" function to perform this job.

PROCESSING OF DATA

6) CONVERTING MILEAGE TO A NUMERIC VALUE IN ORIGINAL TOTALMIL COLUMN

In the original column that represented total mileage, we had “km” written with every value. Due to which our column was being read as a character by R. We needed to remove the “km” word to convert it into a continuous numeric variable. In order to perform this operation we the “strsplit” function to split the value into different parts, and then unlisted the last value, i.e. “km” from every observation using the “unlist” function.

7) TREATING NULL VALUES IN BODYTYPE COLUMN

In body type we had Null and N/A values. We used two approached to solve this problem.
First of all we found the given model in the “Model_name” column, and replaced the body type of that car with the given observation. Secondly, some cars did not have any body type in the entire data. For them we made a category and names it as “Not Specified”.

8) TREATING OUTLIERS IN ENGINE_CAPACITY COLUMN

In the engine_capacity column we had some outliers which were basically the wrongly entered data values, e.g. a 1000CC car had been inputted as 1 CC. There were few observations with this kind of error. Hence, first of all we figured out all of these error values by printing the unique values. And then we replaced them with their correct values.

9) CONVERTING PRICE TO NUMERIC VARIABLE

In the original column price was mentioned with the units, e.g. lac and crore. These were present in every column. Due to which our column was a character, we need to convert them in the right format to make our column numeric for further analysis. Therefore, once again we used the “strsplit” function to separate string variables, and then “unlist” function to remove the unit figure at the end of the figures. Hence, now we have a numeric column.

10) CONVERTING DESIRED COLUMNS TO FACTORS

Finally we have converted all the remaining character columns to “factor” variables using the as.factor function, assigning it back to their respective variables. And to reassure the numeric type of our column have once again applied the “as.numeric” function for price and mileage columns.

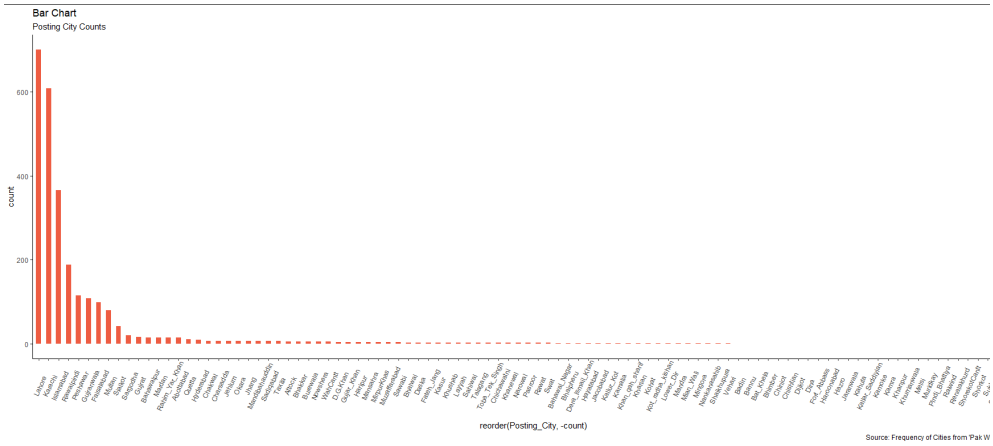
DATA EXPLORATION

For data exploration, we mostly made use of different graphs (using ggplot2) and other specific techniques such as Association Rules. Though we did come up with some hypothesis just by looking at the data, we wanted to explore the data further and try to look at how other different patterns may exist within our dataset as well. Furthermore, we made some hypothesis-specific data exploratory graphs as well.

GRAPHS

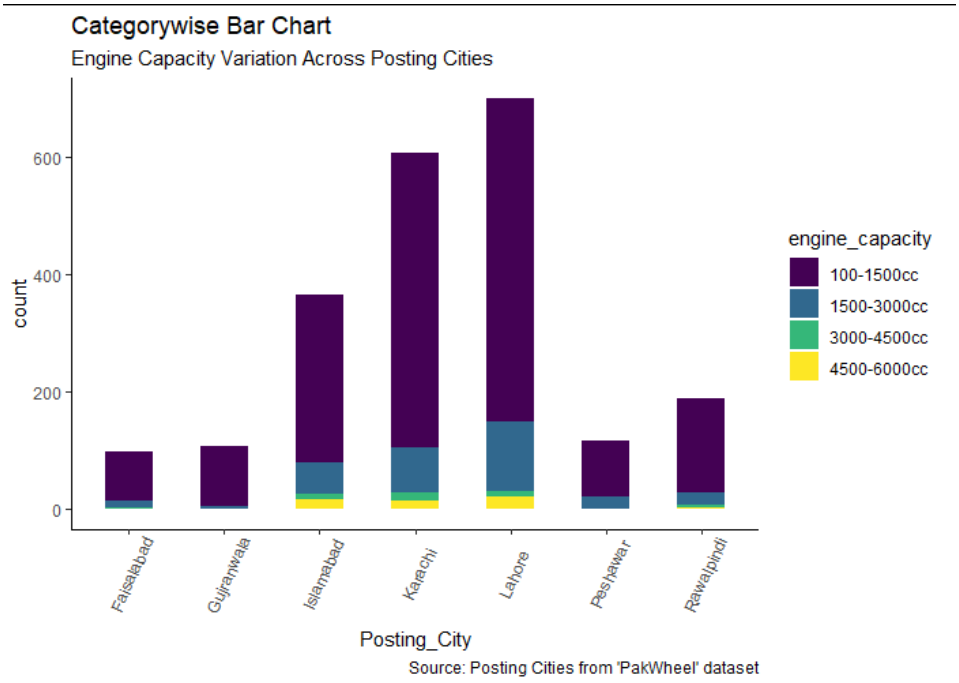
Following are some of the graphs we obtained for exploratory data analysis. Due to the limitation of time & space, we have put some of the graphs in the Appendix as well.

FIGURE 1



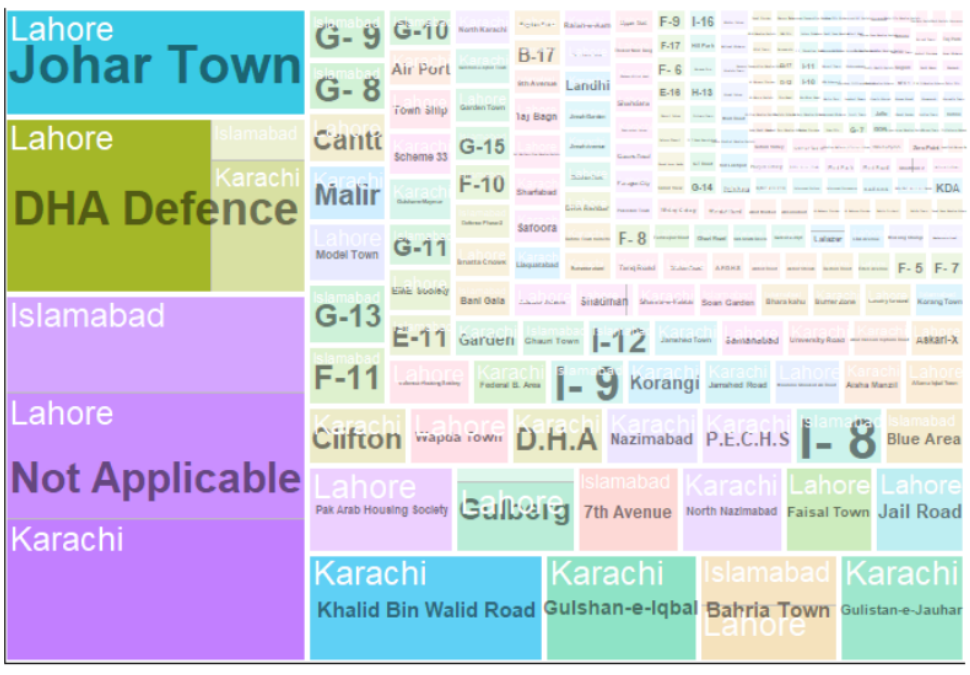
This graph helps to explain the frequency of cars in each city, in the ascending order. According to the graph most of the cars belong to Lahore, Karachi and then Islamabad.

FIGURE 2



We also obtained different bar-charts to better analyze variation in preferences on engine capacity as well as the colour of vehicles across different cities of Pakistan. Our visualisations showed that the majority of people in top 7 cities of Pakistan were using cars with engine capacity between 100 to 1500cc. On the other hand, cities like Karachi, Islamabad, and Lahore also had some cars with engine capacities falling in the range 4500 to 6000cc. Moreover, white colour on cars was predominantly the most popular across all cities. It is also interesting to note that the second consistent colour present in top 7 cities was silver. Cities like Islamabad, Karachi, and Lahore also preferred to have black-coloured cars, perhaps appearing to some as a classic yet posh look.

FIGURE 3



Upon further data exploration, through a treemap, it was also indicated that the majority of ads, in Lahore, were posted from Johor Town locality. Similarly, we also had posted cars from DHA defence of Lahore, Karachi, and Islamabad. Some other hot areas for ads posting included I-8 and G-9 in Islamabad, and Gulshan-e-Iqbal and Gulistan-e-Jauhar in Karachi. “Not Applicable” in our visualisation basically meant that users failed to provide their residential data.

Other exploratory graphs are given in the Appendix (refer to Figure 1a, 2a, and 3a)

ASSOCIATION RULES (ARULES)

Furthermore, for the purpose of exploring the data and different patterns within it, we also ran arules. Though this technique is mostly used for things like recommendations systems etc., we used this to explore general patterns and trends within our dataset, essentially allowing us to also see what linkages were being created within the different variables. Since these were meant to purely explore the data, we included all the variables just to see what sort of results we would get. We first created general rules with no limitations on the values of either the support or confidence and the results can be seen in Figures 30,31 and 32 (Appendix). Then, we experimented with the values of support, in the range of 0.1 to 0.5 and confidence's value, from 0.5 to 1, inspecting each of the top 10 rules and then later sorting it by lift as well. These values for support and confidence were chosen based on not only the hit&trial method but also by looking at the summary of the first general grules where we used no restrictions on either the support or confidence or lift.

If we look at model#4, it gives us 2203 rules (Figure 37). The summary shows that the confidence values range from 0.5 to 1 while the values for lift range from 0.86 to 7.34. The support values ranged from 0.1 to 0.93. The plot (Figure 36) showed that a lot of the rules were clustered for lower values of support (around and before 0.2). It also seemed that those with higher lift ratios were also found where the confidence levels were higher too. However, when we inspect it (Figure 38) though we get rules where the lift is at the maximum level i.e. 7.34 and the confidence is 100%, those rules are not too useful e.g. we can see that one of these rules are that if Posting_city is Islamabad, the province would likely to be Islamabad as well and then the reverse of this rule is present as well. Overall the model did not give us too many useful rules though the values themselves looked quite good. After inspecting all the rules, we believe that model 5 and model 8 gave relatively decent results with more consistent rules. With model 5, the support was kept at 0.1 and the confidence was kept at 0.8. The support values reached a maximum of 0.93, the confidence values reached a maximum of 1 while the lift values reached a maximum value of 7.34. After sorting it by lift, the top 10 rules (Figure 39) give us the same rules as we saw in model#4 where the lift was at a maximum value of 7.34 and the confidence was 100% but the rules themselves were quite how if Posting_city is Islamabad, the province would likely to be Islamabad as well and vice versa. Again, these are not as useful since this pattern itself is quite intuitive. However, there were some other useful rules we obtained where both the confidence was relatively high (higher than the threshold value of 0.5) and the lift values were quite high as well. For instance, we can see that if the car runs on petrol, if it is imported, and belongs to the province of Sindh, its posting city is more likely to be that from Karachi. This rule had pretty decent confidence of 0.986 while its lift was about 4.384.

```
> inspect(grules5[1:10])
     lhs                                                  rhs                      support   confidence coverage  lift     count
[1]  {Posting_City=Islamabad}                          => {Province=Islamabad}     0.1360993 1.0000000  0.1360993 7.347578 351
[2]  {Province=Islamabad}                              => {Posting_City=Islamabad} 0.1360993 1.0000000  0.1360993 7.347578 351
[3]  {consumption=Petrol,Posting_City=Islamabad}       => {Province=Islamabad}     0.1252423 1.0000000  0.1252423 7.347578 323
[4]  {consumption=Petrol,Province=Islamabad}           => {Posting_City=Islamabad} 0.1252423 1.0000000  0.1252423 7.347578 323
[5]  {Assembly=Imported,Province=Sindh}                => {Posting_City=Karachi}   0.1170997 0.9869281  0.1186506 4.388427 302
[6]  {geartype=Automatic,Province=Sindh}               => {Posting_City=Karachi}   0.1388135 0.9862259  0.1407522 4.385304 358
[7]  {consumption=Petrol,Assembly=Imported,Province=Sindh} => {Posting_City=Karachi} 0.1093447 0.9860140  0.1108957 4.384362 282
[8]  {geartype=Automatic,Assembly=Imported,Province=Sindh} => {Posting_City=Karachi} 0.1066305 0.9856631  0.1081815 4.382802 275
[9]  {consumption=Petrol,geartype=Automatic,Province=Sindh} => {Posting_City=Karachi} 0.1306708 0.9853801  0.1326095 4.381544 337
[10] {registration_city=Karachi,Province=Sindh,added_via=Added via Desktop} => {Posting_City=Karachi} 0.1031408 0.9851852  0.1046917 4.380677 266
```

FIGURE 4 - INSPECTION OF MODEL 5 RULES

For model#8, we got some decent rules as well - here the value of support was 0.15 and confidence was 0.8 (Figure 41). A rational to keep the confidence 0.8 (greater than the usual 0.5 threshold value) is that we would want to ideally find those patterns/rules that give the highest level of accuracy. Model 8 gave us around 316 rules - the support values ranged from 0.15 to 0.938, confidence had a maximum value of 1 while the lift ranged from 0.91 till a maximum value of 4.33. We can observe from the graph of grules8 (Figure 43) that out of 316 rules, we have most of the values at the start, around the 0.2 mark on the support side. We can also see that there are quite a few rules clustered around the 100% confidence level. We can also see from the Figure 42 (Appendix) where we sorted the rules by lift, that though the rules may not as high a confidence as those in model 5 or model 4, but we can see that there are relatively more useful rules here with pretty decent confidence and lift values. For instance, rule 9 shows that if the car runs on petrol, has a manual gear type, is locally assembled and if its body type is Sedan, then it is more likely to have an engine capacity of 1300cc.

FIGURE 5 - SCATTERPLOT OF RULES FOR MODEL 5

This rule had a confidence of around 94% and lift ratio of about 3.8 which are both relatively decent values. Similarly rule 8 is that if the car runs on petrol and both its registration city and posting city are Karachi, then it is more likely to belong to the province of Sindh.

However, after inspecting all the different rules of the different models, we believe that going with model 5 would be the best idea since it has quite a lot of rules which will allow us to not miss out on many important rules. Furthermore, the values for confidence and lift ratio for model 5 were higher than that of model 8 and therefore, we believe that going with model 5 would be a better option.

HYPOTHESIS

After exploring the data, we came up with 3 main hypotheses. Each hypothesis, however, may or may not lead to sub-hypothesis to strengthen our claim.

1) H-1: How do features of Posted Cars vary according to the cities?

In order to test the above hypothesis, we have to test following sub-hypothesis as well:

- Which city has the most expensive cars posting?
- Are local cars preferred in a specific city over the imported ones?
- Which Manufacturer is most popular in each of the cities?
- Does Clustering on features maps on the city features?
- Which type of Color is the most popular in each of the cities?
- Is there a difference in preference for total miles across the different cities (New vs Really Old cars being sold)?
- Variation Across Consumption, geartype and Body Type

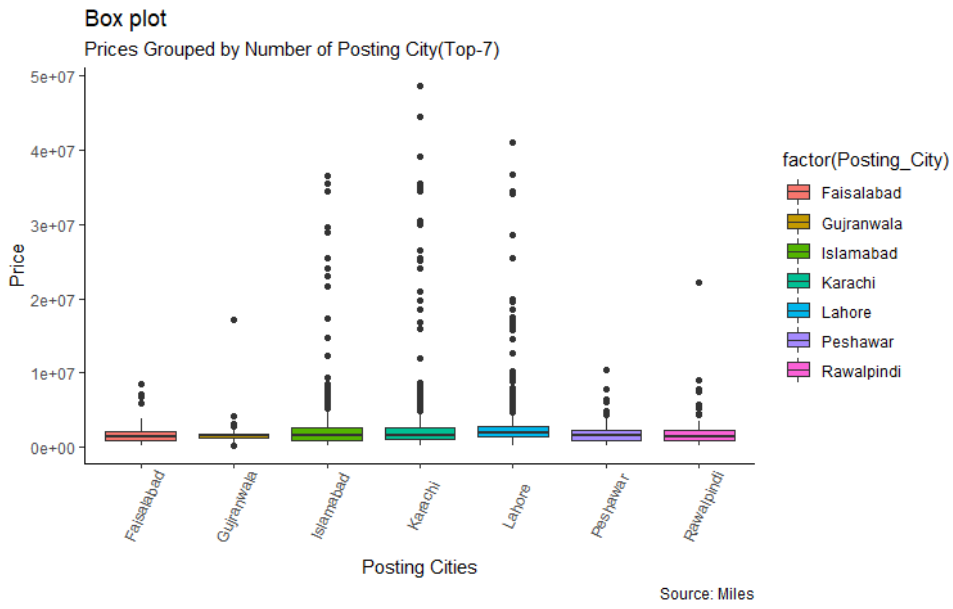
2) H-2: Which features are the most important in determining the price?

3) H-3: What kind of customer Segmentation is possible?

HYPOTHESIS 1: ANALYSIS

How do features of Posted Cars vary according to the cities?

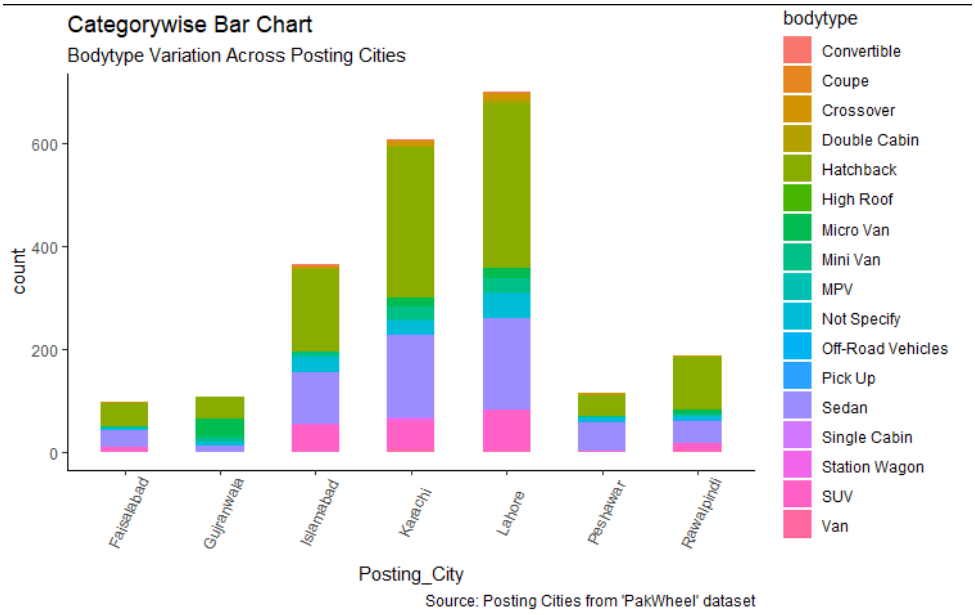
Sub-hypothesis 1: Which city has the most expensive cars posting?



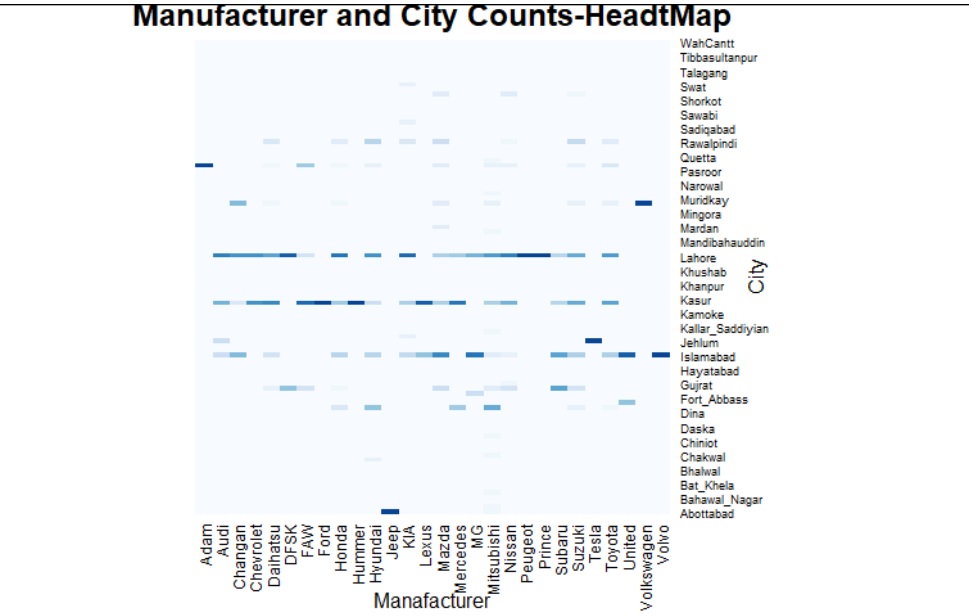
This box-plot, which is plotted against prices and posting city, indicates that most of the posted cars were priced PKR 10 million or less. However, cities like Karachi, Lahore and Islamabad had some cars with prices as high as almost PKR 40 million, thus indicating outliers. This could be because affordability for buying cars is relatively greater in these cities. hence, Pakwheel should ensure that when relatively more expensive cars are available for sale, they are more likely to be sold in these economically stronger cities (advertise here more, introduce here first).

Sub-hypothesis 2: Is body type popularity differing across the major cities?

We were interested in looking at the preference of body type in each city, hence, we made the bar chart below. We can observe that hatchback is the most popular body type of car in all the cities. Moreover, Sedan is the second most popular body type in almost all the cities, except for Gujranwala where Microvan is relatively more popular.

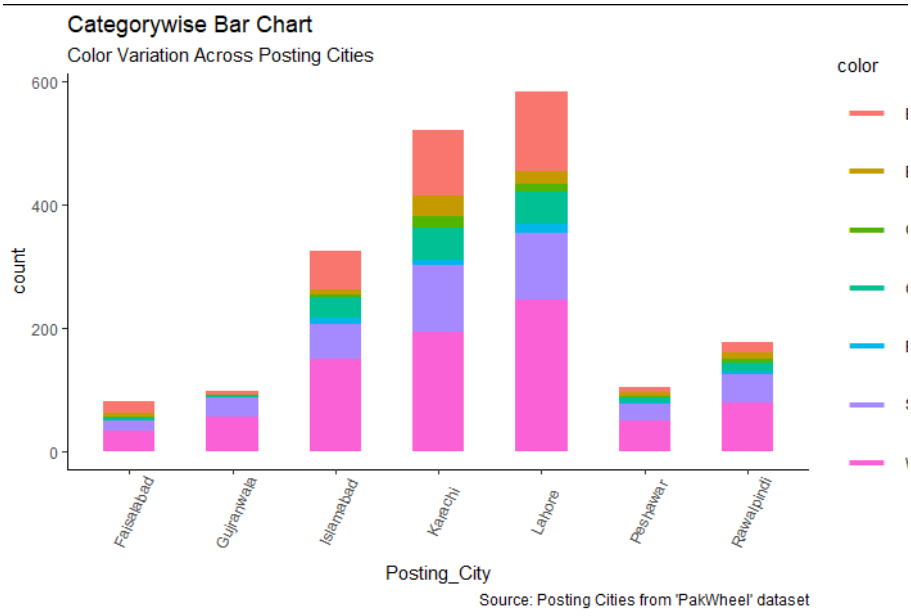


Sub-hypothesis 3: Which Manufacturer is most popular in each of the cities?



This heat map explains the relative frequency of the number of cars of each manufacturer for sale in each city. The darker color represents the high relative frequency. Highest number of Ford and Lexus are in Karachi. And Islamabad has the highest number of Volvo and united. And Abbottabad which is located in a mountain valley has the highest number of jeeps, which is convenient there.

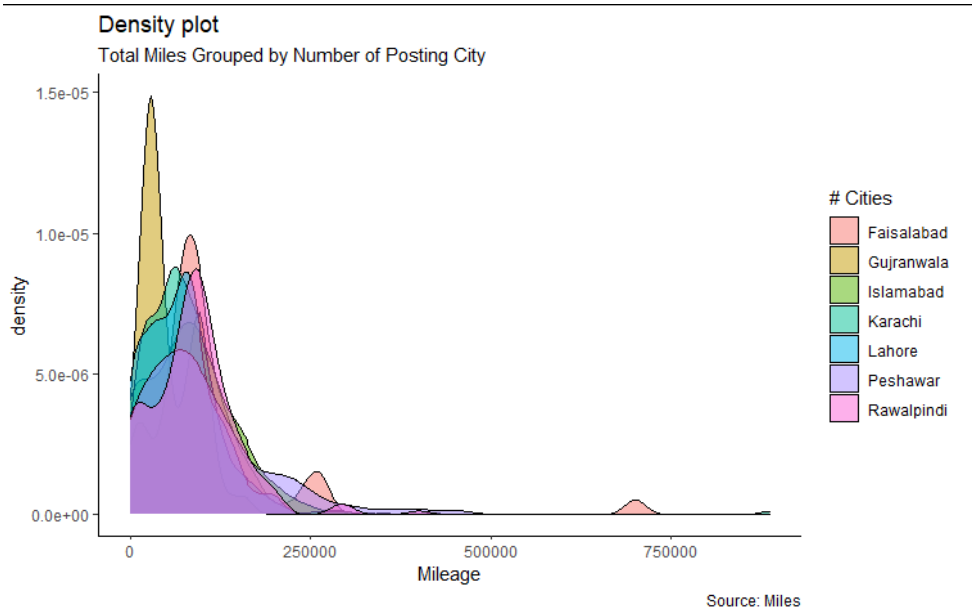
Sub-hypothesis 4: Which type of Color is popular in each of the cities?



According to the bar chart above it is visible that White color is most popular in every city of Pakistan, which is followed by black color and then we have silver color. In Gujranwala, Faisalabad and Rawalpindi Silver color seems to be more in the market as compared to black color. In this graph as well, for the sake of simplicity we have taken the data of top 7 cities.

Sub-hypothesis 35: Is there a difference in preference for total miles across the different cities(New vs Really Old cars being sold)?

The graphs mentioned below represent the city and mileage.



This first graph is the density plot, which shows which city has the highest mileage of cars. And each city is represented by a specific color. In order to keep the visualization readable we have selected the top 7 cities only. We can see that Gujranwala has the highest number of average mileage of each car, followed by Faisalabad. However, almost all these cities have cars that have a mileage of less than 250,000 km with an exception for Faisalabad where some posted cars have a mileage as high as 700,000 km.

In the second graph (Figure 4a in Appendix) we have a box plot, which we have used to represent the deviation in mileage in each city. So we can see that deviation is the same among all the top 7 cities. Highest deviation lies in Peshawar and then Islamabad. Moreover, the visualisation also indicates that Karachi and Faisalabad have outliers as some of the posted cars from these two cities had significantly high mileage - approximately 700,000 km in Faisalabad and more than 99,00,000 km in Karachi.

- ***Sub-hypothesis 5: What is the variation across Consumption, Geartype and Assembly?***

In this graph on the side, we have shown the variation in consumption, gear type and assembly in top 7 cities, frequency wise. And we can observe that there exists different variations in each city. Diesel cars are only available for resale in Lahore, Karachi and Islamabad. Whereas, CNG cars are only available in Karachi. One reason for this could be because cars, by default, come with petrol as means of fuel consumption and, therefore, less people are willing to put in extra effort and install CNG kits in their cars. On the other hand in Lahore and Karachi the automatic gear type cars are more than the manual types.

Whereas in the rest of the cities these variations seem to be the same. Surprisingly the ratio of imported and local cars for resale is the same in every city. Which shows the equal importance of imported and local cars throughout major cities of Pakistan.

- ***Sub-hypothesis 6: Does Clustering on car features map onto the city features?***

To understand whether the features of the cars posting maps on the city , we made use of K-means. This means that if we were to compare the features of segments generated by the K-mean with features of top-k cities, how close these features are. If they were close enough then this means that cities play a major in differentiating the features of the cars. Consequently, we tried out two values of K, the result of which is shown below:

K=3 vs Top-3 Cities

	Posting_City	Manufacturer	totalmil_mean	price_mean	consumption	geartype	Assembly	bodytype	color	engine_capacity
1	Islamabad	Toyota	84053.38	3190521	Petrol	Automatic	Local	Hatchback	White	1300 cc
2	Karachi	Toyota	74821.42	3118967	Petrol	Automatic	Imported	Hatchback	White	1000 cc
3	Lahore	Toyota	69203.15	2946482	Petrol	Automatic	Imported	Hatchback	White	660 cc

Main features of Top-3 Cities

	cluster	Manafacturer	totalmil_mean	price_mean	consumption	geartype	Assembly	bodytype	color	engine_capacity	city
1	1	Suzuki	71353.09	3411162	Petrol	Automatic	Imported	Hatchback	White	660 cc	Lahore
2	2	Toyota	79481.79	3168606	Petrol	Automatic	Imported	Hatchback	Black	1300 cc	Lahore
3	3	Toyota	70952.06	2746899	Petrol	Automatic	Local	Sedan	White	1000 cc	Lahore

Main features of 3 segments generated by K-means

Note that features have been grouped on cities for upper table and grouped on segment for lower table, where mode was used for categorical features and mean for numeric features. As it is apparent that the feature mapping is not that good. Firstly, the K-mean segments are dominated by Lahore. If the cluster were also dominated by three different cities or same city as top-3 cities, then feature mapping would have been significant and city might have played an important role in created segments. Moreover, if we compare the features one by one for both tables, the difference becomes clear e.g., Sedan Body type is present in segments generated by K-means but not in top-3 cities features whereas Toyota is dominated in all 3 cities ,one of the segments is being dominated by Suzuki. Consequently, to further testify the importance of cities in explaining the segment, we ran K-mean again, but this means with K=7.

K=7 vs Top-7 Cities

	Posting_City	Manafacturer	totalmil_mean	price_mean	consumption	geartype	Assembly	bodytype	color	engine_capacity
1	Faisalabad	Suzuki	108858.11	1876465	Petrol	Manual	Local	Hatchback	White	1300 cc
2	Gujranwala	Suzuki	51837.56	1745523	Petrol	Automatic	Imported	Hatchback	White	660 cc
3	Islamabad	Toyota	84053.38	3190521	Petrol	Automatic	Local	Hatchback	White	1300 cc
4	Karachi	Toyota	74821.42	3118967	Petrol	Automatic	Imported	Hatchback	White	1000 cc
5	Lahore	Toyota	69203.15	2946482	Petrol	Automatic	Imported	Hatchback	White	660 cc
6	Peshawar	Toyota	99441.91	1931897	Petrol	Automatic	Local	Sedan	White	1300 cc
7	Rawalpindi	Suzuki	88683.98	2001566	Petrol	Automatic	Local	Hatchback	White	1000 cc

Main features of top 7 Cities

	cluster	Manafacturer	totalmil_mean	price_mean	consumption	geartype	Assembly	bodytype	color	engine_capacity	city
1	1	Toyota	66556.90	2952221	Petrol	Automatic	Local	Hatchback	White	1000 cc	Lahore
2	2	Toyota	86401.72	2314664	Petrol	Manual	Local	Sedan	Black	1300 cc	Karachi
3	3	Daihatsu	68317.49	1684689	Petrol	Automatic	Imported	Hatchback	White	660 cc	Lahore
4	4	Suzuki	73544.31	4328941	Petrol	Automatic	Imported	Hatchback	Black	660 cc	Lahore
5	5	Suzuki	67989.91	3541625	Petrol	Automatic	Imported	Hatchback	White	660 cc	Lahore
6	6	Honda	59657.40	2750029	Petrol	Automatic	Local	Sedan	White	1300 cc	Lahore
7	7	Toyota	115534.05	1607573	Petrol	Manual	Local	Sedan	White	1000 cc	Islamabad

Main features of 7 segments generated by K-means

Again, the tables have quite constraining differences, and it was difficult to map features of top-7 cities on the segment features as per K-mean. Consequently, we concluded cities were not playing that significant of role in segmenting the user posting on the website, which was further testified by the our third main hypothesis.

HYPOTHESIS 2: ANALYSIS

Which feature of the car is the most important in determining the price?

Price is one of the most important factors that makes a car sell on any platform. We have planned to suggest Pakwheels on the features in an advertisement of a car that decides the final price. So that they can ensure the efficiency by the viewpoint of both buyers and sellers. In order to perform this operation we have used the Arules technique. In summary we have fixed the right hand right side of the equation with price and left hand side with the features of a car. Moreover, this price has been classified into categories to better understand the buyer's psychology at each price level. Exploratory graphs for this analysis have been given in the Appendix (Figures 5 till 11). Furthermore, in order to test this hypothesis, we have also carried out logistic regression in order to look at which factors help predict price in a better manner.

Data Preparation for Arules:

First of all, it should be kept in mind that Arules only works on categorical data. Hence, we need to convert our continuous numeric variables, i.e. price and mileage, into categories. For price we have made five categories, which are: Very cheap, cheap, moderate, high, very high. For total mileage we have made three categories, which are low, moderate and high. These were made keeping in mind the ranges and mean values for both the price and total mileage variables.

Findings:

For our analysis for the given hypothesis, we will be keeping the R.H.S fixed with different combinations of price, to see how L.H.S will vary with different combinations of features. And then we will be interpreting features for each category. The Apriori algorithm was used to carry out this technique. (Please note that the accompanying graphs and screenshots of the output of the following are given in the Appendix)

Very cheap cars: After fixing price to be as the "Very Cheap" category, keeping support = 0.07 and confidence =0.7, and then sorting it by confidence, when we inspect the first 10 rules, we get many useful insights. For instance, the first rule which has a value of confidence of about 77% and a lift ratio of 2.988 shows that if the car manufacturer is Suzuki, if the total mileage is moderate, if the car is locally assembled and if the posting has been added via the phone, then it is likely for the car to be priced cheaply. Similarly, we see in the third rule that if the gear type is manual, the engine capacity is 1000cc, and if the posting has been added through a phone it is more likely to be priced cheaply. This rule had a confidence of 77% and a lift ratio of 2.97. The given rules show the combination of factors that may make it more likely for a car to be priced cheaply.

Cheap cars: If we fix price to be cheap on the RHS and keep support=0.05 and confidence=0.7, we get about 141 rules. After sorting it by lift and inspecting the first 10 rules, we can see some meaningful results. If we look at the top rule, it has a decent confidence value of 0.847 and lift ratio of 2.23 - it shows that if total mileage is moderate, the car runs on petrol, if the gear type is automatic, the body type of the car is hatchback and the engine capacity is 660 cc, then the price of the car is likely to be cheap. Similarly, rule 6 shows that if the gear type is automatic, the registration city is Lahore, if the car is imported and the body type is hatchback, then it is likely to be priced cheaply. The confidence for this rule is 0.839 and the lift ratio is 2.21. Other rules can be interpreted in the same manner, showing what combination of features or what features make a car more likely to be priced cheaply.

Moderately priced cars: For moderately priced cars, we kept support to be equal to 0035 and the confidence to be 0.7. After sorting it by lift, inspection of the top 10 rules showed very interesting results. For instance, the first rule which has a 90% confidence and lift ratio of 2.9, it shows that if it is a Toyota car and is a 2017 model, then it is likely to be priced moderately. Similarly, rule 5 (with 86.9% confidence and lift ratio of 2.78) shows that if the total mileage is low, the car runs on petrol, if it is automatic and locally assembled along with being a sedan, then it is more likely to be priced moderately.

Highly priced cars: For highly priced cars, we kept support to be 0.01 and confidence to be equal to 0.2 and then sorted them by lift. An issue with these rules is that their confidence and support values are really low - however their lift values reach as high as 15.1. For instance, the second rule which has a 35% confidence and lift value of 15.1 shows that if the manufacturer is a Toyota, the car runs on petrol, is automatic, imported, and a SUV and the posting is added through a desktop, then the car is likely to be priced higher. Similarly, the other rules can be interpreted.

Very High cars: If we fix price to be very highly priced, keeping support=0.01 and confidence=0.5, then we can see from the top 10 rules (after sorting by lift) that there are some very interesting results. It is important to note that all top 10 rules have a 100% confidence and a very high lift ratio making them very good rules. For instance, if engine capacity is 4600cc it is lowly to be a very highly priced car. This rule has 100% confidence and 37.92. Similarly, if the car is a Land Cruiser and has an engine capacity of 4600cc then it will be priced highly most likely. This rule again has 100% confidence and lift ratio. Similarly, the other rules can be interpreted.

Logistic Regression

Data Preparation

Similar to the data preparation for Arules, for logistic regression, the column of price was first converted into 2 categories i.e. either high or low. These 2 categories were then converted into a binary variable with 1 representing High price and 0 representing Low price

Findings:

We have experimented with 6 models in our R script to find the optimal set of independent variables. In all of these models we have kept dependent variable constant, i.e. Price. After running it against various combination we have concluded with the model#1 (as seen in Figure 27 of the Appendix). In which we have got consumption, Body type, Assembly, gear type, total mileage as dependent variable. The results are mentioned below. The type of relationship has been kept binomial.

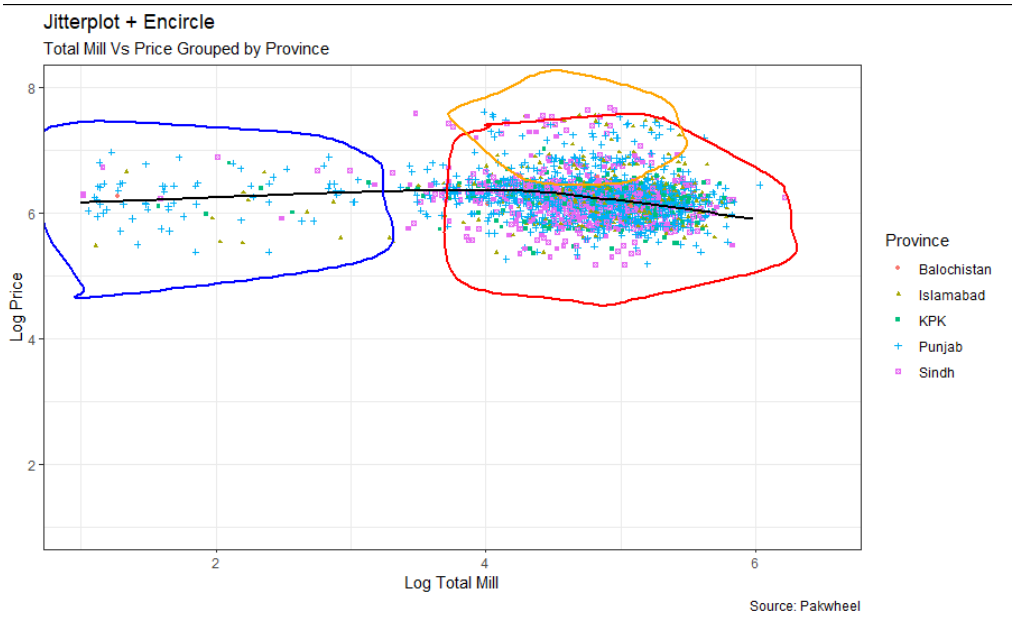
In the summary of the model we can observe that the features which have p value less than 0.05 are portraying significant results, this means that they are actually affecting price. Hence, Manual gear type and low total mileage has the most significant impact to change the category of price. Other significant factors are the bodytype of the car being Crossover, Hatchback or SUV.

HYPOTHESIS 3: ANALYSIS

What kind of customer Segmentation is Possible?

Is Segmentation Possible?

In order to give recommendations to PakWheels regarding the segmentation of the ads on their platform we have performed customer segmentation. First of all we need to ensure whether it is possible to make clusters from the given data set. In order to perform this operation we have made a jitterplot and encircled the potential clusters. We have taken log values for price and total, to test our hypothesis. And then by plotting them on both the axis we have made clusters against provinces, to get the following results.



We can see that from the given experiment we have got 3 clusters, which proves that clustering is possible.

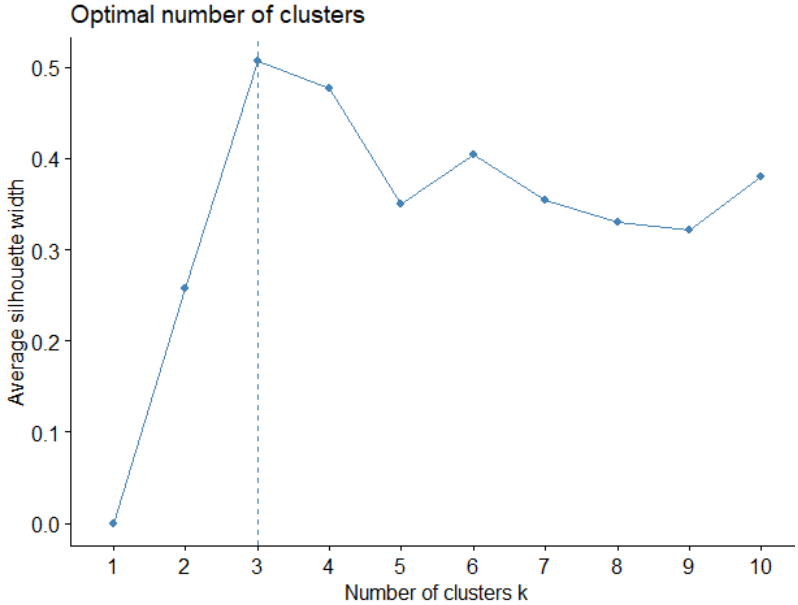
Selecting technique and variables for segmentation

In this part we will make clusters, and find the optimal number of clusters. We have used the k-means clustering technique to perform this operation. We have used the following variables against each observation to distinguish among Ads: Manufacturer's name, total mileage, consumption type, gear type, Assembly, body type, color, model year, engine capacity, Added_via, price and province. We have eliminated Model_name, as this gives us the same information as the manufacturer at the end of the day. And as to classify ads provincial information was enough this we removed other geo-location variables.

Finding optimal number of clusters:

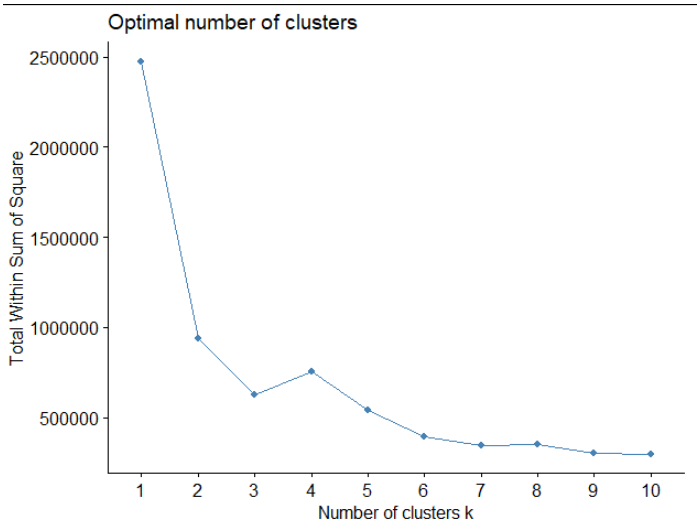
We have used multiple methods to confirm the optimal segments in the training data which can be labelled as follows:

1) Silhouette Width

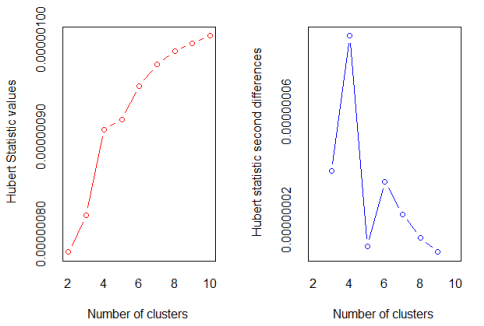


2) Total sum of squares

We can observe from the graph below that after 5 clusters there is no significant reduction in the value. Hence, according to this approach optimal clusters are 5.

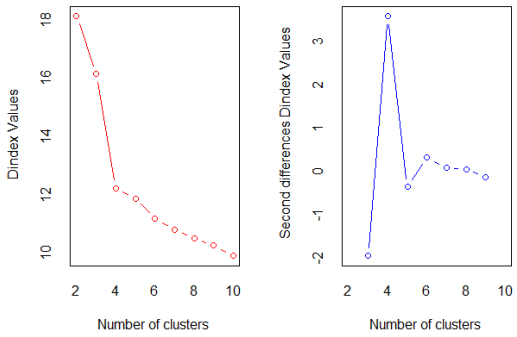


3) Hubert statistics



According to Hubert Statistics the minimum second difference and optimal number of clusters is at 5.

4) Dindex Value



Dindex Value is also suggesting the optimal number of clusters at 5 .

Manual observation on the basis of significant differences.

From all the techniques that we have used above, the optimal number of clusters we are getting is between 3 and 5. Hence, in order to reach a certain conclusion we decided to form segments of these clusters, and then manually analyze which meets our requirement according to the problem at PakWheels.

a) Number of cluster = 3

	cluster	Manafacturer	totalmil_mean	price_mean	consumption	geartype	Assembly	bodytype	added_via	color	engine_capacity
1	1	Suzuki	69609.55	2720243	Petrol	Automatic	Imported	Hatchback	Added via Desktop	White	660 cc
2	2	Toyota	83863.90	2909628	Petrol	Automatic	Local	Hatchback	Added via Phone	Black	1300 cc
3	3	Toyota	81264.66	2357452	Petrol	Manual	Local	Sedan	Added via Phone	White	1300 cc

The above table shows the 3 segments from the data. We can see that, there is no one variable which is totally distinguishing each cluster clearly, there are overlapping values in each column. Hence, 3 clusters are not enrich enough to differentiate.

b) Number of clusters = 4

	cluster	Manafacturer	totalmil_mean	price_mean	consumption	geartype	Assembly	bodytype	added_via	color	engine_capacity
1	1	Suzuki	74016.50	4325110	Petrol	Automatic	Imported	Hatchback	Added via Desktop	Black	660 cc
2	2	Toyota	88149.70	2275662	Petrol	Manual	Local	Sedan	Added via Phone	Black	1300 cc
3	3	Suzuki	69705.10	2721043	Petrol	Automatic	Imported	Hatchback	Added via Desktop	White	660 cc
4	4	Toyota	81264.66	2357452	Petrol	Manual	Local	Sedan	Added via Phone	White	1300 cc

Using 4 clusters we are able to distinguish between the values better than 3 clusters. Columns like Manufacturer, total mileage, price, geartype, body type, color and engine capacity are portraying different means in each segment. Hence, these are enriched enough to differentiate. Still we will go on to 5 clusters in the hope of even better segmentation.

c) Number of clusters = 5

	cluster	Manafacturer	totalmil_mean	price_mean	consumption	geartype	Assembly	bodytype	added_via	color	engine_capacity
1	1	Toyota	63898.95	2883209	Petrol	Automatic	Local	Sedan	Added via Phone	White	1300 cc
2	2	Toyota	88149.70	2275662	Petrol	Manual	Local	Sedan	Added via Phone	Black	1300 cc
3	3	Suzuki	69021.02	2654509	Petrol	Automatic	Imported	Hatchback	Added via Desktop	White	660 cc
4	4	Suzuki	74016.50	4325110	Petrol	Automatic	Imported	Hatchback	Added via Desktop	Black	660 cc
5	5	Toyota	116608.12	1463260	Petrol	Manual	Local	Sedan	Added via Phone	White	1000 cc

We are getting an even better difference in each cluster at this stage (cluster plots for k=3,4,5 given in Figures 44, 45 and 46). Therefore, we will finalize 5 clusters to suggest to the management of Pakwheels. In Segment#1 prefers Pak suzuki, with 74,000 km average mileage. They have the highest price affordability range of 4 million plus. And they seem to prefer imported Hatchback 660cc cars. This category is generally expensive in the car market. For the segment#2, we have Toyota with a tradeoff between price and mileage, they prefer to buy a car that has a high mileage in order to pay less. Hence, this can be the category of cheap but high brand cars. Which is also seen in the gear type which is manual, assembly is local but body type is Sedan, so this category can be middle class families. In Segment#3 We again have cars for families but with a high price range, they are willing to get extra features and less mileage, moreover, they will prefer to get a 1300cc engine Sedan with Automatic gear. Hence, this segment of ads is suitable for upper-middle class families. In Segment#4 we can once again have upper-middle class targeted people as the preference for mileage and price is similar to that of segment#3, but this time it seems more like individual cars instead of family cars. As the body type is a hatchback and 660cc car, which is normally driven by young people. Moreover, this segment of Ads will show Automatic gear type and imported cars which causes the prices to shoot up. In segment#5 We have a Toyota as must available car. In this segment people tend not to care about mileage or price, as it seems like the preference is to get a 1300cc sedan in the given price range. Therefore, these ads have compromised on other features, like manual gear type, local assembly and color to deliver the previous mentioned ones. This seems like an ad for families richer than segment#2 families but not as wealthy as people targeted in segment#3.

LIMITATIONS & CONCLUSION

Assumptions:

- Posts by sellers will remain the same throughout the year.
- Each city, brand, manufacturer and model year has a true weighted representation.
- Price will not get affected by inflation in the short term.
- The price mentioned is the net amount, Pakwheels will not deduct any commision from the price written in the advertisement.

Limitations:

- Dataset was collected between 14th March to 16th March. Hence, it only represents advertisements posted/ present with in 3 days.
- We were limited to only use modules of R in our system.
- Dataset only represents a sample of all advertisements present on Pakwheels' platform.
- Some users added data incorrectly. For example, rather than adding 1000 cc for their car, some users ended up adding 10 cc. Although the engine capacity would have been obvious to the buyers due to pictures attached with ads, however, it could not be adjusted while scrapping data from the website.

Conclusion & Recommendations:

After using various types of k-means model it has been verified that cities do not play any significant role in the attributes of cars posted on the platform. However, through sub hypothesis it can be seen that some features are slightly more prominent than the other cities. In Lahore, Karachi and Islamabad we can have more similar sort of attributes in the advertised cars. this can be owing to the fact that Lahore, Karachi and Islamabad are among the top most developed cities of Pakistan. Here the engagement is highest as well. In order to increase the customer footfall Pakwheels should also focus on cities like Fasilabad and Gujranwala where the demand is still relatively higher. This can give Pakwheels a competitive edge as well. Whereas, more mileage can be observed in Faisalabad and Peshawar. Thus when making marketing strategies, these features should be the most highlighted within the specific cities, the sales force and customer service personnel should be informed and trained to highlight these features among cars in depth. Furthermore, considering that black and white are the most popular colors for all the cities so perhaps PakWheels can show cars in these colors more and promote such cars more given their popularity. Furthermore, PakWheels should focus on automatic cars more since the trend is moving more towards on automatic cars rather than manual cars.

Along with that, through our hypothesis#2 we can see that different price ranges are affected by different car features, some prominent among them are total mileage, Body type and engine capacity. Similarly in our Hypothesis#3, we can see that it is possible for Pakwheels to segment the advertisements on their platform. Ideally there should be 5 clusters in this scenario. And each cluster will target the specific consumer segment, some prominent among them could be middle class families, upper middle class families and cars for people in their early careers. The wordings, mediums of communication should be altered in accordance to these segments. In order to have full customer engagement on the website they can use the first 2-3 pages showing cars that have a mileage of less than 250,000 km.

All in all, we can conclude that there is a significant room for data mining at PakWheels, that they can use to optimize their platform for both sellers and buyers.

APPENDIX A

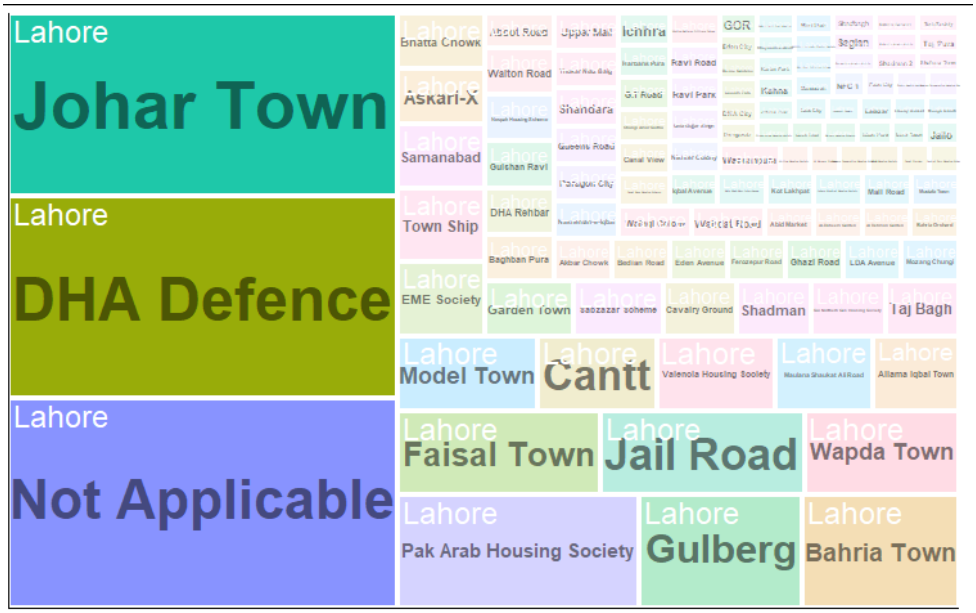


Figure: 1a

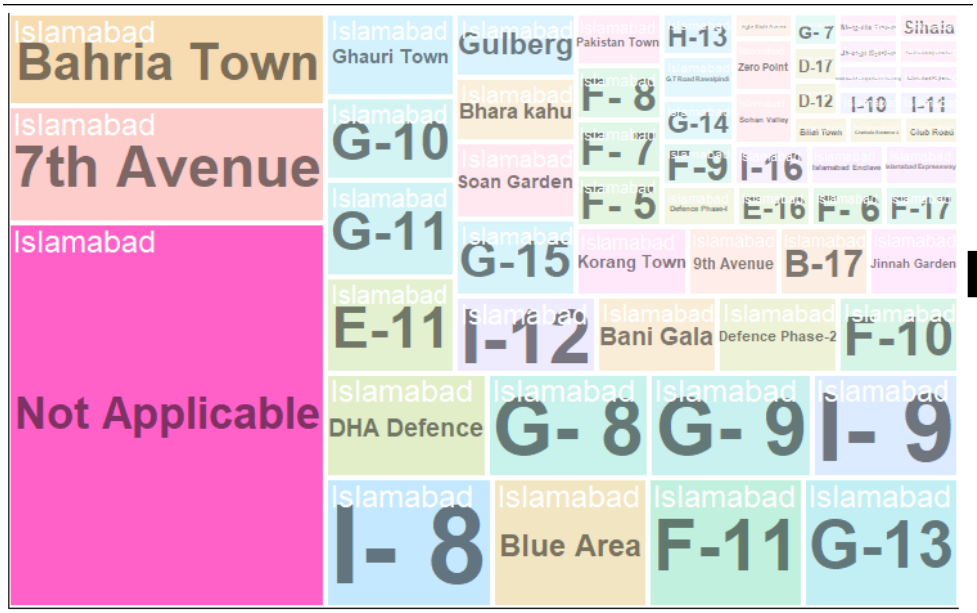


Figure: 2a

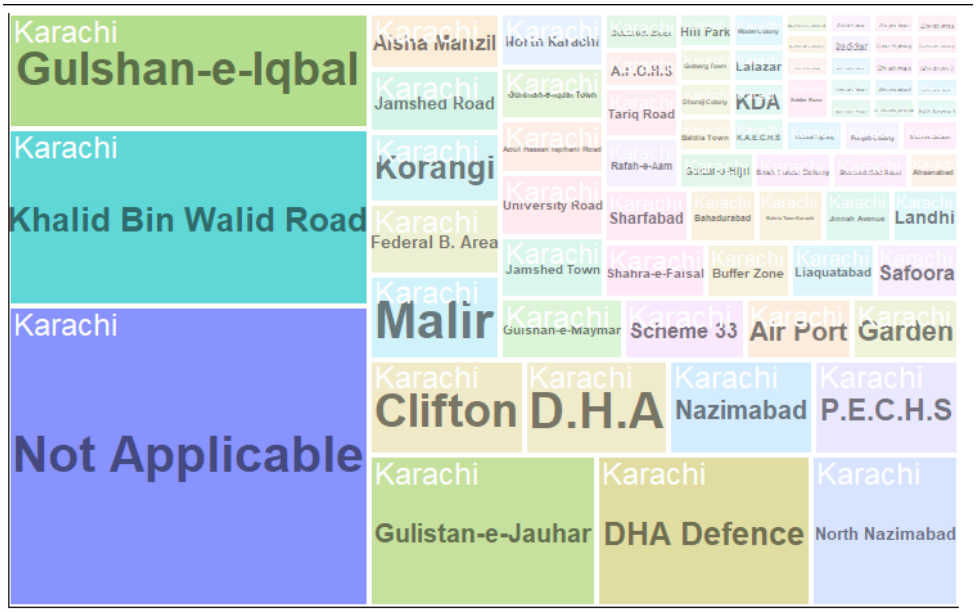


Figure: 3a

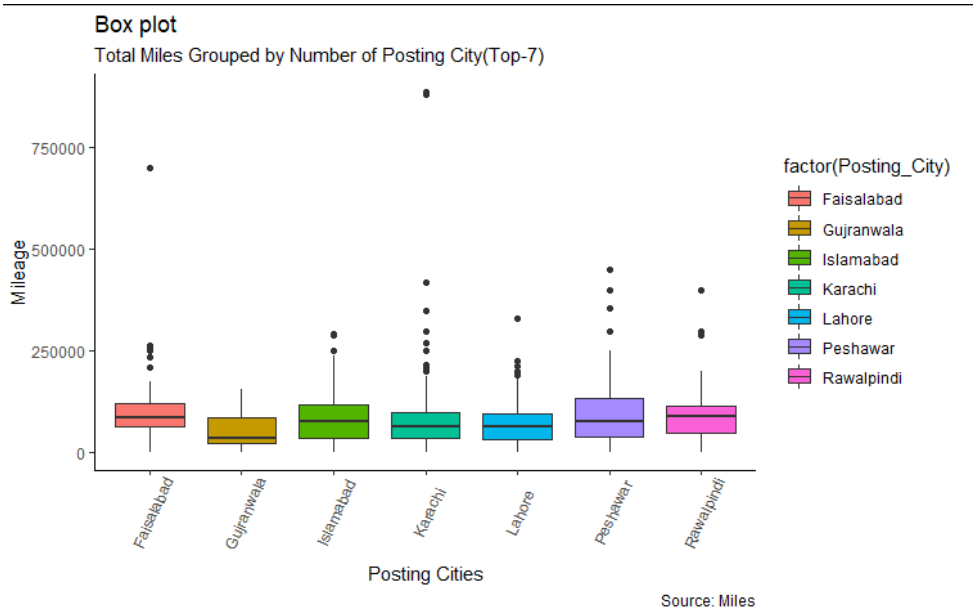


Figure: 4a

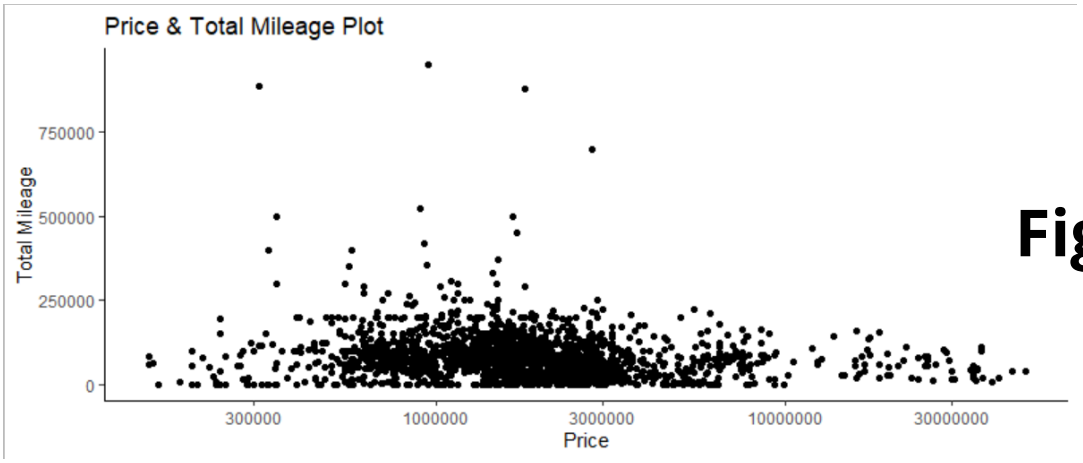


Figure: 5

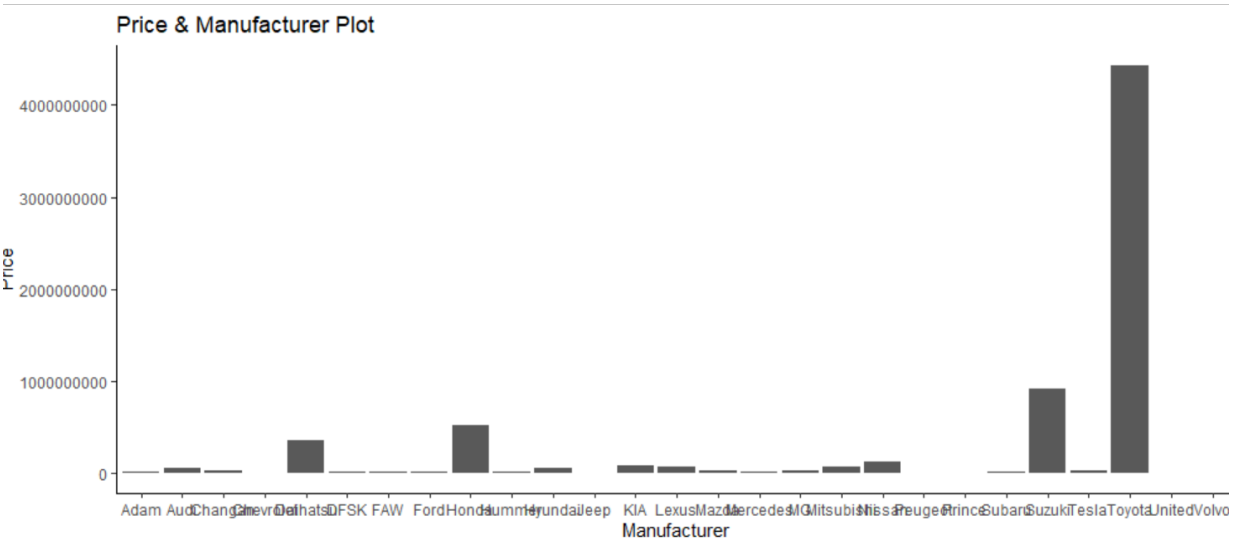


Figure: 6

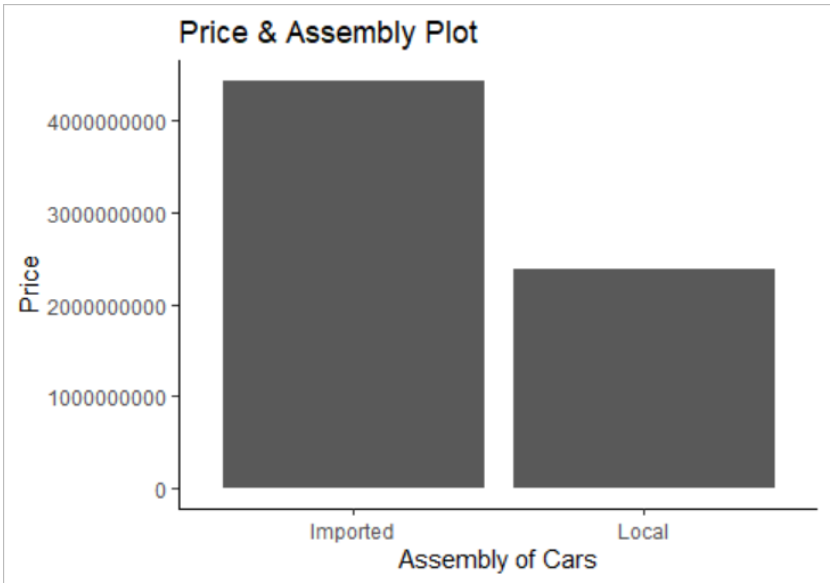


Figure: 7

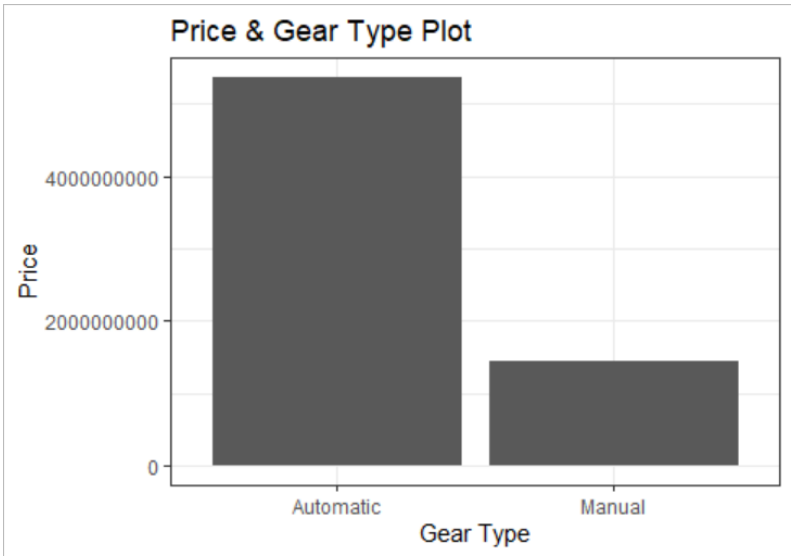


Figure: 8

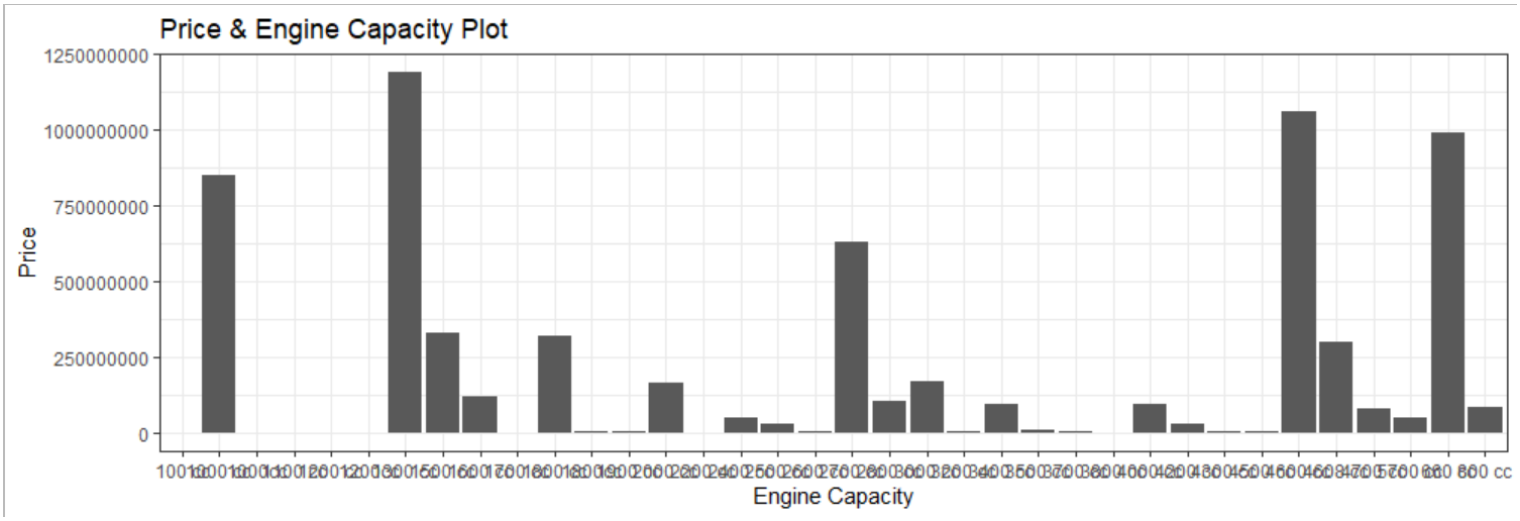


Figure: 9

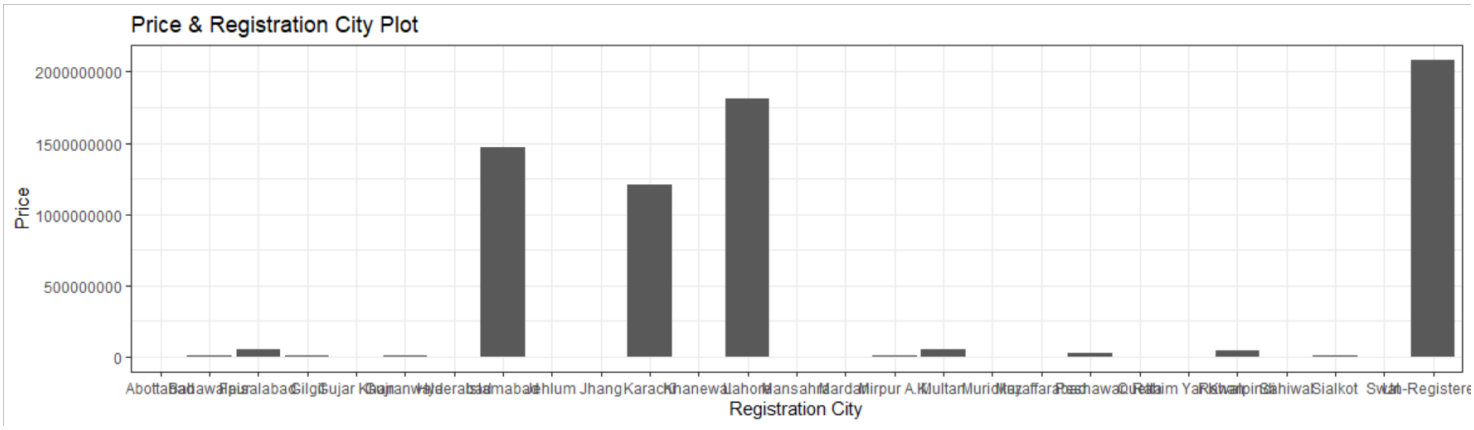


Figure: 10

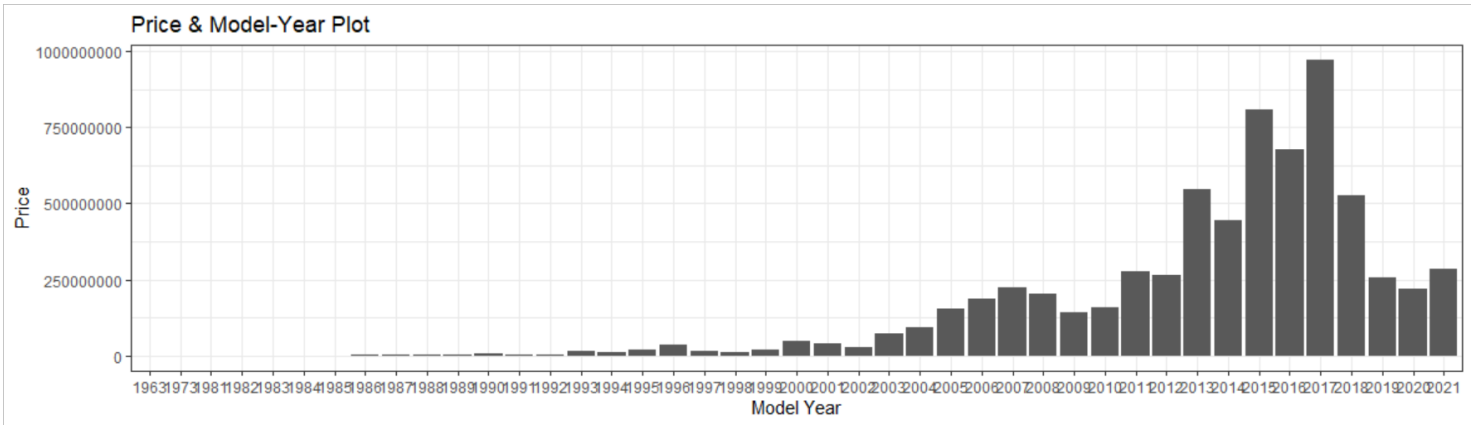


Figure: 11

```
> inspect(cars_vcheap[1:10]) #look at top 10 rules
     lhs                                rhs                  support confidence   coverage     lift count
[1]  {Manafacturer=Suzuki,
      totalmil=Moderate,
      geartype=Manual,
      Assembly=Local,
      added_via=Added via Phone} => {price=Very Cheap} 0.05544785  0.7729730 0.07173323 2.988752   143
[2]  {Manafacturer=Suzuki,
      totalmil=Moderate,
      geartype=Manual,
      added_via=Added via Phone} => {price=Very Cheap} 0.05622334  0.7712766 0.07289647 2.982192   145
[3]  {geartype=Manual,
      engine_capacity=1000 cc,
      added_via=Added via Phone} => {price=Very Cheap} 0.06630477  0.7702703 0.08607988 2.978301   171
[4]  {geartype=Manual,
      Assembly=Local,
      engine_capacity=1000 cc,
      added_via=Added via Phone} => {price=Very Cheap} 0.06359054  0.7699531 0.08259015 2.977075   164
[5]  {Manafacturer=Suzuki,
      totalmil=Moderate,
      Assembly=Local,
      added_via=Added via Phone} => {price=Very Cheap} 0.05583560  0.7659574 0.07289647 2.961626   144
[6]  {totalmil=Moderate,
      geartype=Manual,
      engine_capacity=1000 cc}   => {price=Very Cheap} 0.05273362  0.7640449 0.06901900 2.954231   136
[7]  {totalmil=Moderate,
      geartype=Manual,
      Assembly=Local,
      engine_capacity=1000 cc}   => {price=Very Cheap} 0.05118263  0.7630058 0.06708026 2.950213   132
[8]  {Manafacturer=Suzuki,
      totalmil=Moderate,
      consumption=Petrol,
      geartype=Manual,
      Assembly=Local,
      added_via=Added via Phone} => {price=Very Cheap} 0.05234587  0.7627119 0.06863125 2.949076   135
[9]  {geartype=Manual,
      Assembly=Local,
      bodytype=Hatchback,
      engine_capacity=1000 cc,
      added_via=Added via Phone} => {price=Very Cheap} 0.05971307  0.7623762 0.07832493 2.947779   154
[10] {geartype=Manual,
      bodytype=Hatchback,
      engine_capacity=1000 cc,
      added_via=Added via Phone} => {price=Very Cheap} 0.06087631  0.7621359 0.07987592 2.946849   157
```

Figure: 12

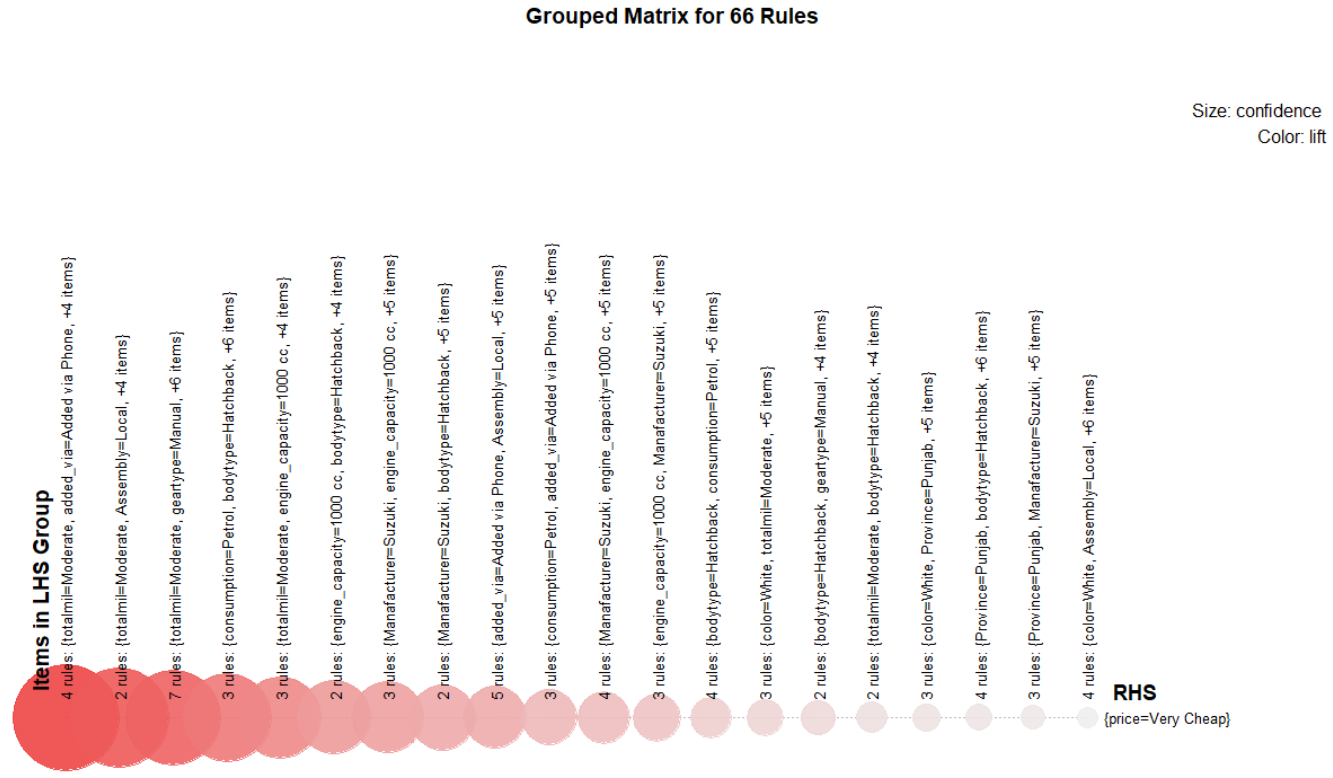


Figure: 13

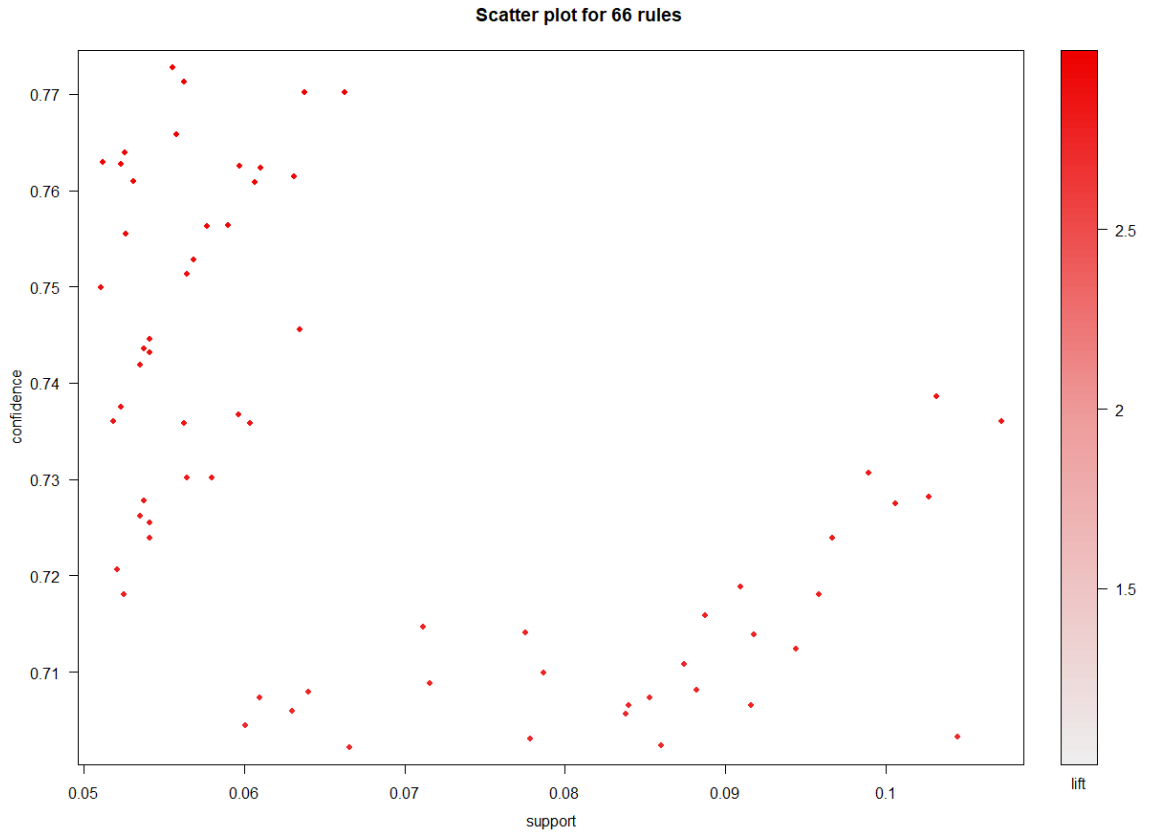


Figure: 14

```
> inspect(cars_cheap[1:10]) #look at top 10 rules
     lhs                                rhs             support confidence   coverage     lift count
[1]  {totalmil=Moderate,
      consumption=Petrol,
      geartype=Automatic,
      bodytype=Hatchback,
      engine_capacity=660 cc}        => {price=Cheap} 0.05389686  0.8475610 0.06359054 2.235030   139
[2]  {totalmil=Moderate,
      consumption=Petrol,
      Assembly=Imported,
      bodytype=Hatchback,
      engine_capacity=660 cc}        => {price=Cheap} 0.05350911  0.8466258 0.06320279 2.232564   138
[3]  {totalmil=Moderate,
      consumption=Petrol,
      geartype=Automatic,
      Assembly=Imported,
      bodytype=Hatchback,
      engine_capacity=660 cc}        => {price=Cheap} 0.05234587  0.8437500 0.06203955 2.224981   135
[4]  {totalmil=Moderate,
      geartype=Automatic,
      bodytype=Hatchback,
      engine_capacity=660 cc}        => {price=Cheap} 0.05389686  0.8424242 0.06397829 2.221485   139
[5]  {totalmil=Moderate,
      Assembly=Imported,
      bodytype=Hatchback,
      engine_capacity=660 cc}        => {price=Cheap} 0.05350911  0.8414634 0.06359054 2.218951   138
[6]  {geartype=Automatic,
      registration_city=Lahore,
      Assembly=Imported,
      bodytype=Hatchback}            => {price=Cheap} 0.05467235  0.8392857 0.06514153 2.213208   141
[7]  {consumption=Petrol,
      geartype=Automatic,
      registration_city=Lahore,
      Assembly=Imported,
      bodytype=Hatchback}            => {price=Cheap} 0.05467235  0.8392857 0.06514153 2.213208   141
[8]  {totalmil=Moderate,
      geartype=Automatic,
      Assembly=Imported,
      bodytype=Hatchback,
      engine_capacity=660 cc}        => {price=Cheap} 0.05234587  0.8385093 0.06242730 2.211161   135
[9]  {consumption=Petrol,
      geartype=Automatic,
      registration_city=Lahore,
      bodytype=Hatchback}            => {price=Cheap} 0.05932532  0.8315217 0.07134548 2.192735   153
[10] {totalmil=Moderate,
      consumption=Petrol,
      bodytype=Hatchback,
      engine_capacity=660 cc}        => {price=Cheap} 0.05506010  0.8304094 0.06630477 2.189801   142
```

Figure: 15

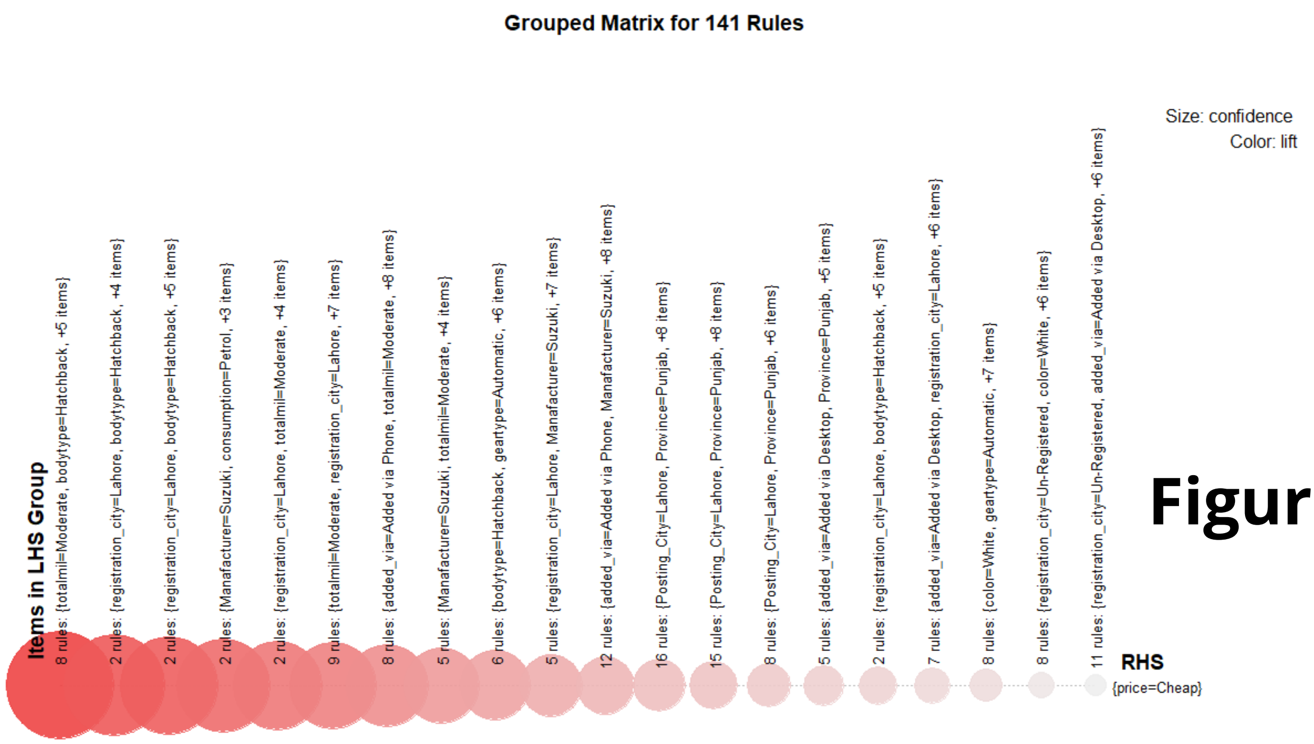


Figure: 16

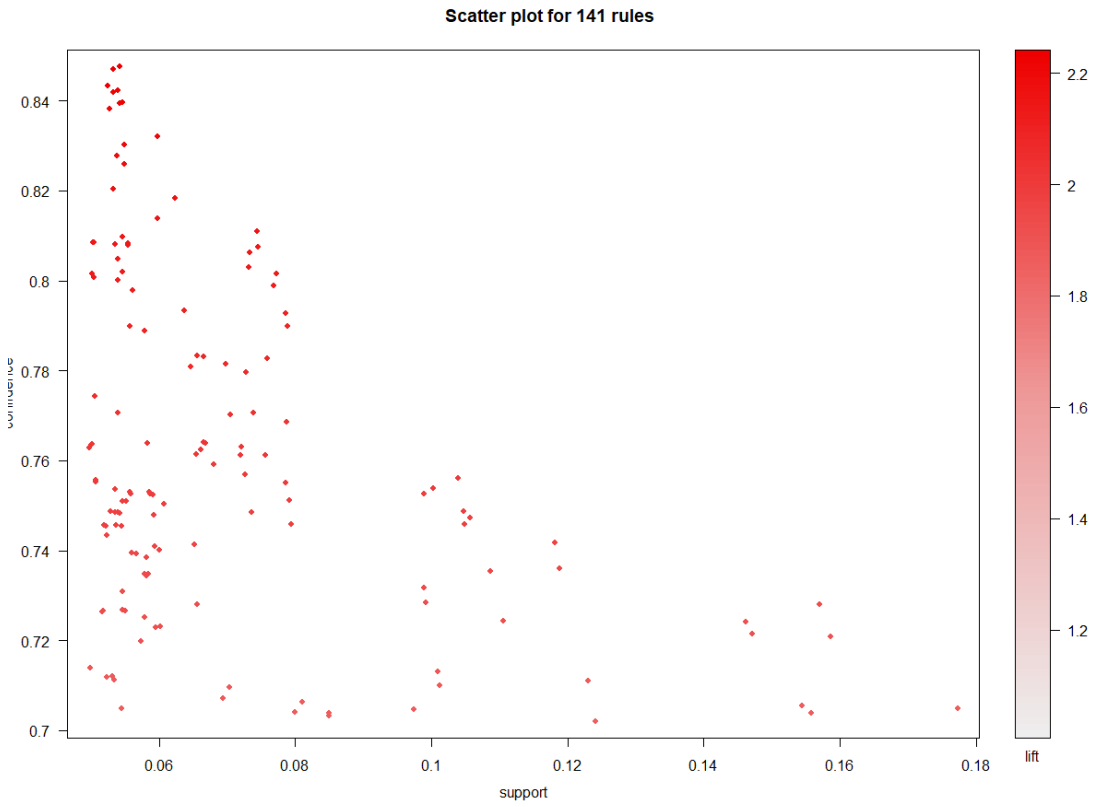


Figure: 17

```
> inspect(cars_moderate[1:10]) #look at top 10 rules
     lhs                                                  rhs                 support confidence   coverage     lift count
[1]  {Manafacturer=Toyota,
      modelyear=2017}                                  => {price=Moderate} 0.04265219  0.9090909 0.04691741 2.908865   110
[2]  {Manafacturer=Toyota,
      modelyear=2017,
      consumption=Petrol}                              => {price=Moderate} 0.03644824  0.8952381 0.04071345 2.864540    94
[3]  {Manafacturer=Toyota,
      consumption=Petrol,
      geartype=Automatic,
      Assembly=Local,
      bodytype=Sedan}                                  => {price=Moderate} 0.04963164  0.8827586 0.05622334 2.824609   128
[4]  {Manafacturer=Toyota,
      geartype=Automatic,
      Assembly=Local,
      bodytype=Sedan}                                  => {price=Moderate} 0.04963164  0.8767123 0.05661109 2.805262   128
[5]  {totalmil=Low,
      consumption=Petrol,
      geartype=Automatic,
      Assembly=Local,
      bodytype=Sedan}                                  => {price=Moderate} 0.03606049  0.8691589 0.04148895 2.781093    93
[6]  {totalmil=Low,
      geartype=Automatic,
      Assembly=Local,
      bodytype=Sedan}                                  => {price=Moderate} 0.03606049  0.8611111 0.04187670 2.755342    93
[7]  {Manafacturer=Toyota,
      geartype=Automatic,
      Assembly=Local,
      added_via=Added via Phone}                       => {price=Moderate} 0.03644824  0.8545455 0.04265219 2.734333    94
[8]  {Manafacturer=Toyota,
      totalmil=Low,
      consumption=Petrol,
      Assembly=Local,
      bodytype=Sedan}                                  => {price=Moderate} 0.04691741  0.8521127 0.05506010 2.726549   121
[9]  {totalmil=Low,
      consumption=Petrol,
      bodytype=Sedan,
      color=White}                                     => {price=Moderate} 0.03528499  0.8504673 0.04148895 2.721284    91
[10] {totalmil=Low,
      consumption=Petrol,
      Assembly=Local,
      bodytype=Sedan,
      added_via=Added via Phone}                       => {price=Moderate} 0.03722373  0.8495575 0.04381543 2.718373    96
```

Figure: 18

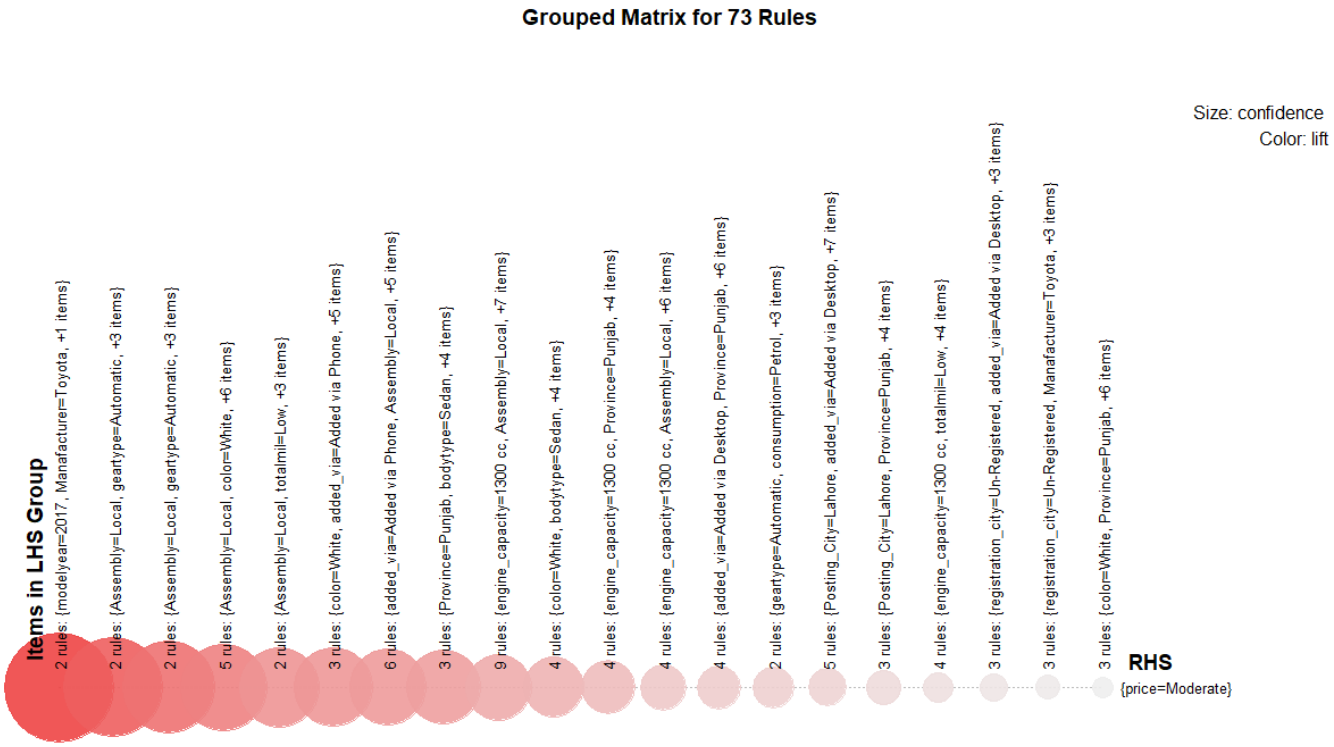


Figure: 19

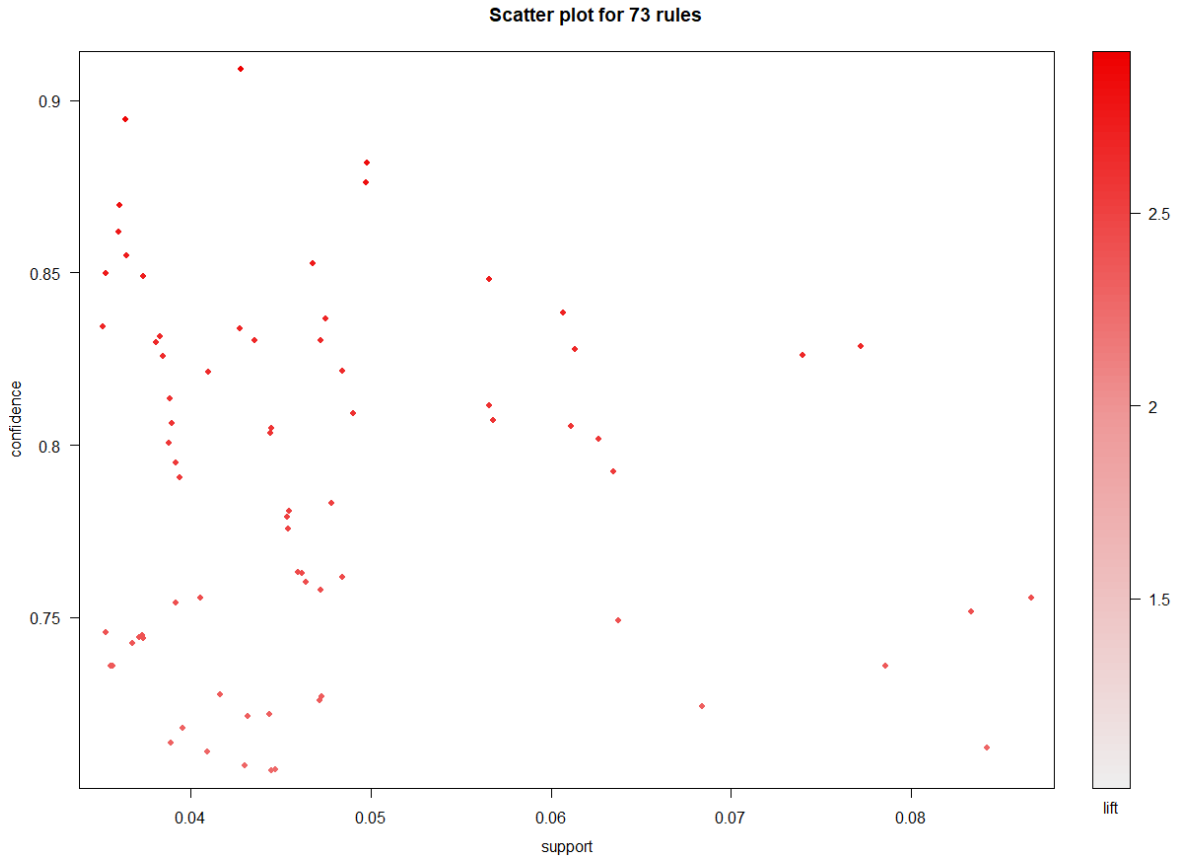


Figure: 20

```
> inspect(cars_high[1:10]) #look at top 10 rules
     lhs                                rhs             support confidence   coverage     lift count
[1]  {Manafacturer=Toyota,
      consumption=Petrol,
      Assembly=Imported,
      bodytype=SUV,
      added_via=Added via Desktop} => {price=High} 0.01008143  0.3513514 0.02869329 15.10225    26
[2]  {Manafacturer=Toyota,
      consumption=Petrol,
      geartype=Automatic,
      Assembly=Imported,
      bodytype=SUV,
      added_via=Added via Desktop} => {price=High} 0.01008143  0.3513514 0.02869329 15.10225    26
[3]  {Manafacturer=Toyota,
      consumption=Petrol,
      bodytype=SUV,
      Province=Punjab}             => {price=High} 0.01046917  0.3417722 0.03063203 14.69051    27
[4]  {Manafacturer=Toyota,
      consumption=Petrol,
      geartype=Automatic,
      bodytype=SUV,
      Province=Punjab}             => {price=High} 0.01046917  0.3417722 0.03063203 14.69051    27
[5]  {Manafacturer=Toyota,
      consumption=Petrol,
      bodytype=SUV,
      added_via=Added via Desktop} => {price=High} 0.01163242  0.3370787 0.03450950 14.48876    30
[6]  {Manafacturer=Toyota,
      consumption=Petrol,
      geartype=Automatic,
      bodytype=SUV,
      added_via=Added via Desktop} => {price=High} 0.01163242  0.3370787 0.03450950 14.48876    30
[7]  {Manafacturer=Toyota,
      geartype=Automatic,
      bodytype=SUV,
      added_via=Added via Desktop} => {price=High} 0.01240791  0.3298969 0.03761148 14.18007    32
[8]  {Manafacturer=Toyota,
      totalmil=Moderate,
      consumption=Petrol,
      bodytype=SUV}                => {price=High} 0.01046917  0.3253012 0.03218302 13.98253    27
[9]  {Manafacturer=Toyota,
      totalmil=Moderate,
      consumption=Petrol,
      geartype=Automatic,
      bodytype=SUV}                => {price=High} 0.01046917  0.3253012 0.03218302 13.98253    27
[10] {Manafacturer=Toyota,
      geartype=Automatic,
      Assembly=Imported,
      bodytype=SUV,
      added_via=Added via Desktop} => {price=High} 0.01008143  0.3250000 0.03101978 13.96958    26
```

Figure: 21

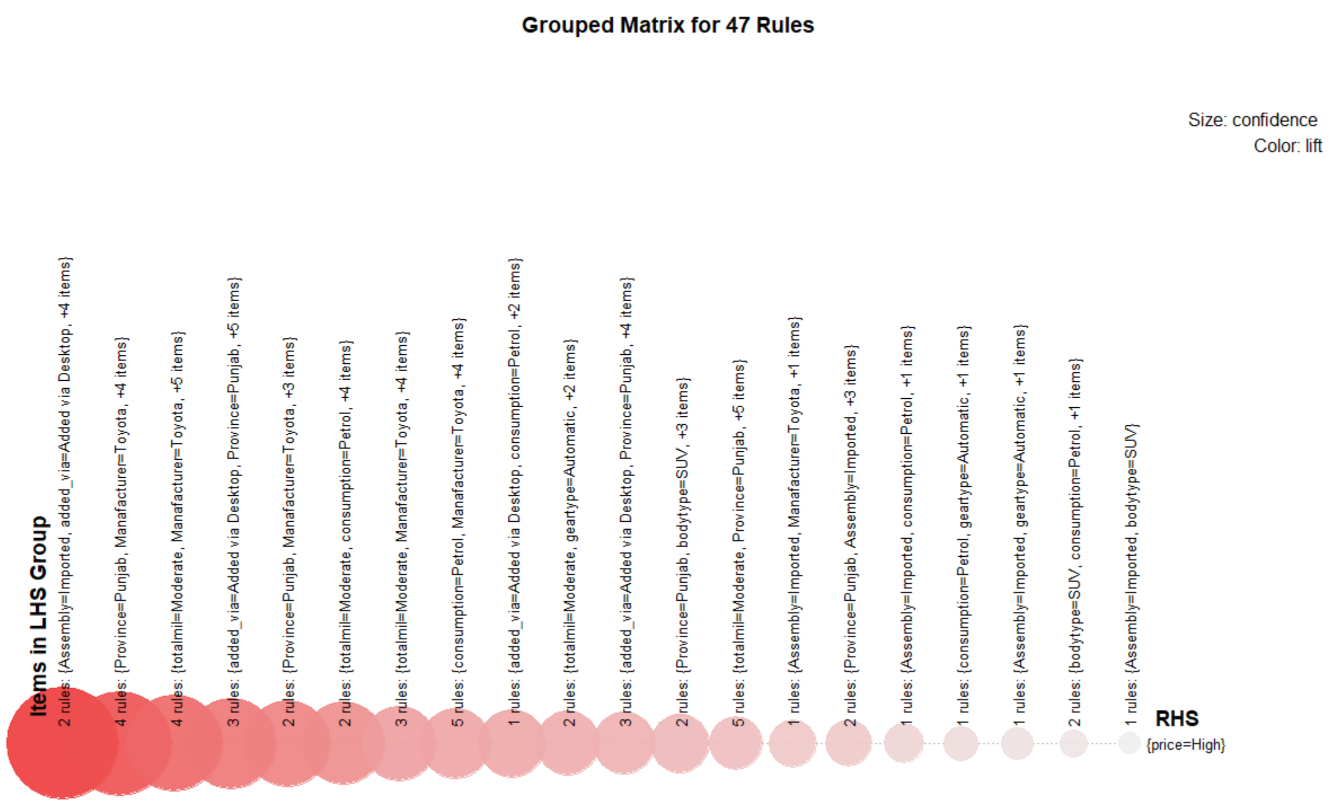


Figure: 22

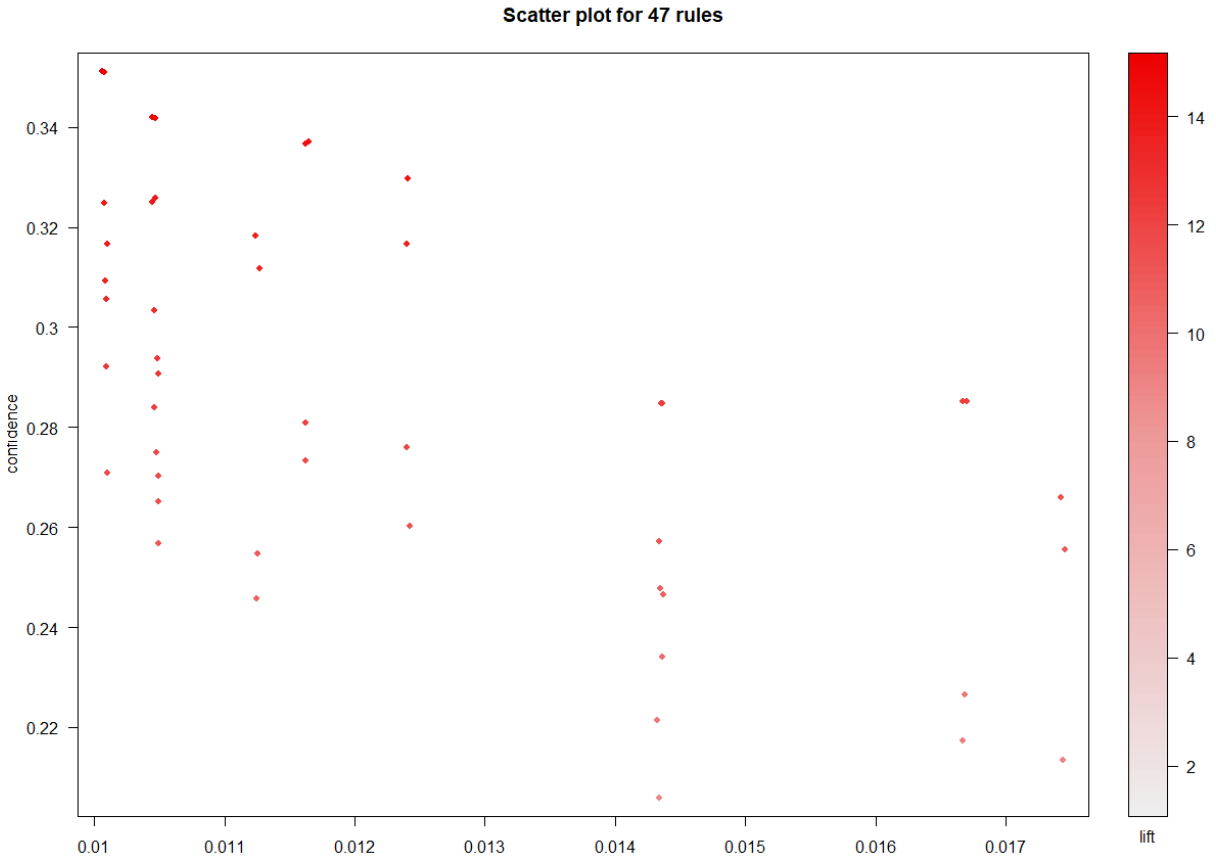


Figure: 23

```
> inspect(cars_vhigh[1:10]) #look at top 10 rules
     lhs                                                                rhs                    support    confidence coverage   lift     count
[1]  {engine_capacity=4600 cc}                                       => {price=Very High} 0.01589763 1          0.01589763 37.92647 41
[2]  {model_name=Land_Cruiser,engine_capacity=4600 cc}               => {price=Very High} 0.01589763 1          0.01589763 37.92647 41
[3]  {bodytype=SUV,engine_capacity=4600 cc}                          => {price=Very High} 0.01589763 1          0.01589763 37.92647 41
[4]  {color=Black,engine_capacity=4600 cc}                           => {price=Very High} 0.01046917 1          0.01046917 37.92647 27
[5]  {Manafacturer=Toyota,engine_capacity=4600 cc}                   => {price=Very High} 0.01589763 1          0.01589763 37.92647 41
[6]  {Assembly=Imported,engine_capacity=4600 cc}                     => {price=Very High} 0.01589763 1          0.01589763 37.92647 41
[7]  {geartype=Automatic,engine_capacity=4600 cc}                    => {price=Very High} 0.01589763 1          0.01589763 37.92647 41
[8]  {consumption=Petrol,engine_capacity=4600 cc}                    => {price=Very High} 0.01589763 1          0.01589763 37.92647 41
[9]  {model_name=Land_Cruiser,color=Black}                           => {price=Very High} 0.01124467 1          0.01124467 37.92647 29
[10] {model_name=Land_Cruiser,bodytype=SUV,engine_capacity=4600 cc} => {price=Very High} 0.01589763 1          0.01589763 37.92647 41
>
```

Figure: 24

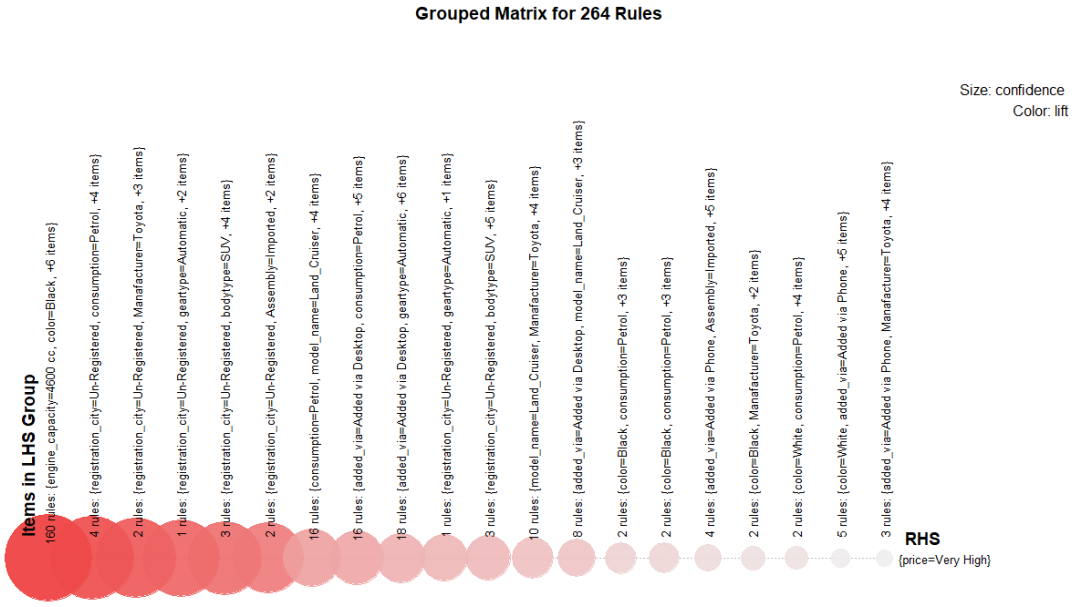


Figure: 25

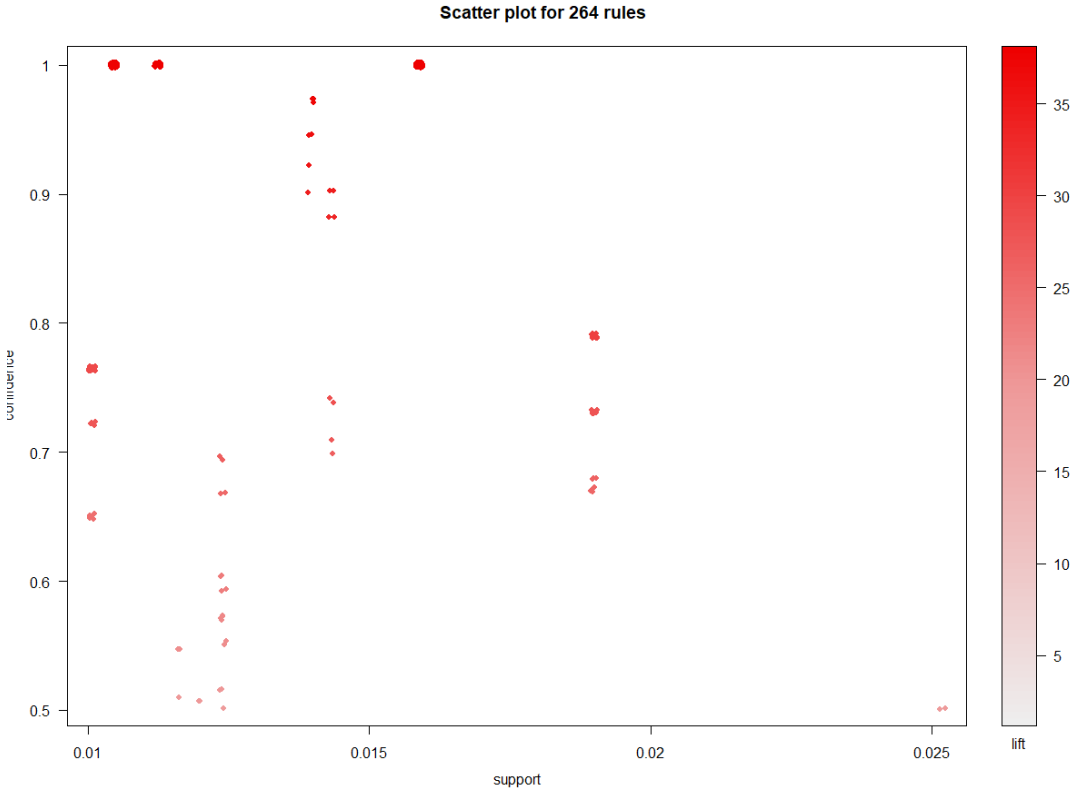


Figure: 26

```
> #Model 1
> LR <- glm(price~consumption+bodytype+Assembly+geartype+totalmil,train,family="binomial")
> summary(LR)

Call:
glm(formula = price ~ consumption + bodytype + Assembly + geartype +
    totalmil, family = "binomial", data = train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.64423  -0.30052  -0.12526  -0.02842   3.01711

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -17.8645  1213.9951  -0.015   0.9883
consumptionDiesel         16.4205  1213.9949   0.014   0.9892
consumptionHybrid         18.3394  1213.9948   0.015   0.9879
consumptionPetrol         16.1173  1213.9947   0.013   0.9894
bodytypeCoupe            -15.7895  6522.6387  -0.002   0.9981
bodytypeCrossover         19.3744  1200.4855   0.016   0.9871
bodytypeDouble Cabin      20.0530  1687.6956   0.012   0.9905
bodytypeHatchback         -2.1438     0.9486  -2.260   0.0238 *
bodytypeMicro Van        -17.3550   822.6971  -0.021   0.9832
bodytypeMini Van          -1.7864     1.3921  -1.283   0.1994
bodytypeMPV               19.3512  2917.0128   0.007   0.9947
bodytypeNot Specify        0.4471     0.9710   0.460   0.6452
bodytypeOff-Road Vehicles 21.4860  6522.6387   0.003   0.9974
bodytypePick Up          -15.2120  2875.7972  -0.005   0.9958
bodytypeSedan              1.5918     0.9595   1.659   0.0971 .
bodytypeStation Wagon      0.8036     1.5770   0.510   0.6103
bodytypeSUV                3.8706     0.9727   3.979 6.91e-05 ***
bodytypeVan                0.9116     1.4561   0.626   0.5313
AssemblyLocal             -0.3796     0.2728  -1.391   0.1641
geartypeManual            -2.3711     0.2666  -8.895  < 2e-16 ***
totalmil.L                -1.7780     0.2126  -8.363  < 2e-16 ***
totalmil.Q                 0.2068     0.1730   1.196   0.2318
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1749.6  on 1811  degrees of freedom
Residual deviance:  738.1  on 1790  degrees of freedom
  (60 observations deleted due to missingness)
AIC: 782.1

Number of Fisher Scoring iterations: 17
```

Figure: 27

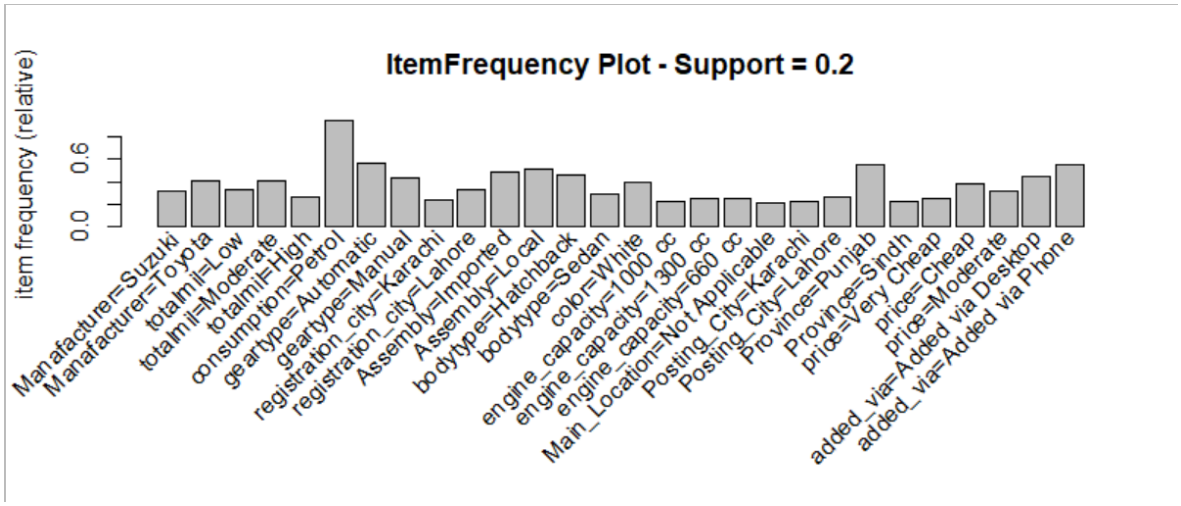


Figure: 28

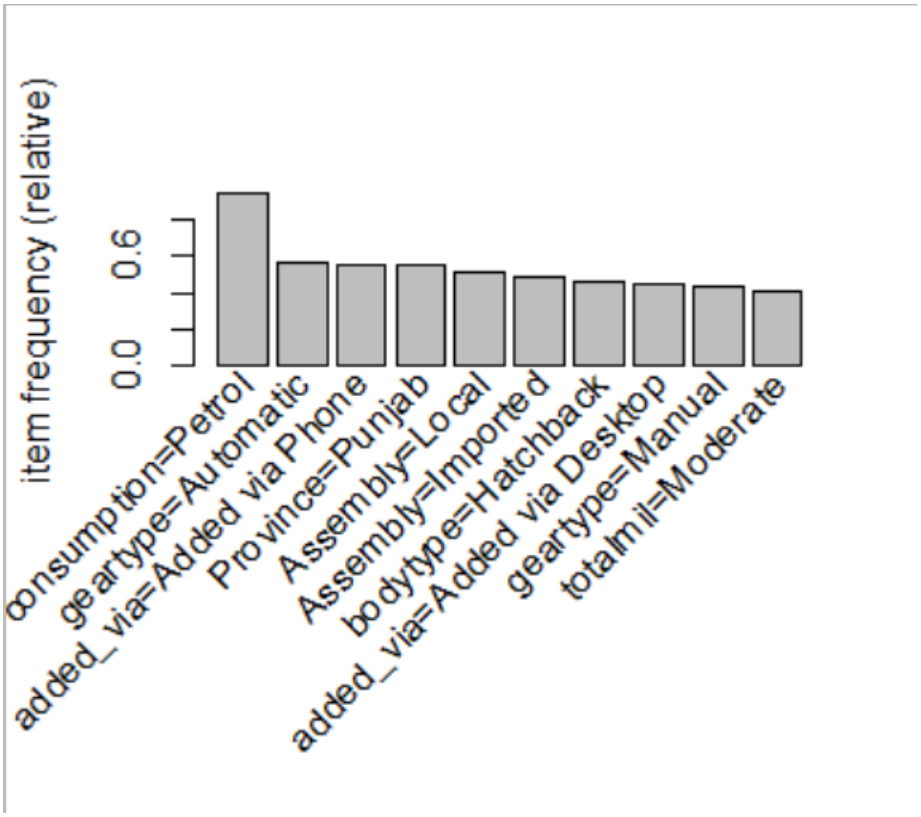


Figure: 29

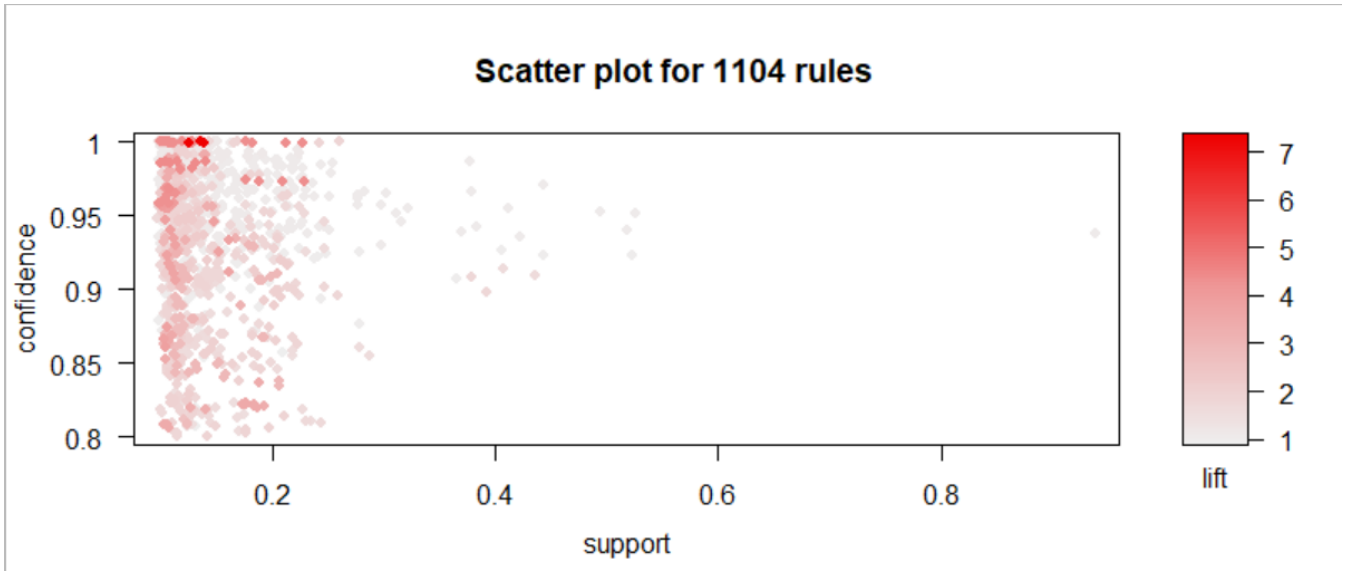


Figure: 30

```
> summary(grules)
set of 1104 rules

rule length distribution (lhs + rhs):sizes
  1   2   3   4   5   6
  1  53 362 456 203  29

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    3.00    4.00    3.81    4.00    6.00

summary of quality measures:
    support         confidence        coverage            lift            count
 Min.   :0.1000   Min.   :0.8000   Min.   :0.1000   Min.   :0.8601   Min.   : 258.0
 1st Qu.:0.1117   1st Qu.:0.9048   1st Qu.:0.1190   1st Qu.:1.0310   1st Qu.: 288.0
 Median :0.1249   Median :0.9447   Median :0.1353   Median :1.6177   Median : 322.0
 Mean   :0.1451   Mean   :0.9332   Mean   :0.1560   Mean   :1.7119   Mean   : 374.1
 3rd Qu.:0.1575   3rd Qu.:0.9719   3rd Qu.:0.1718   3rd Qu.:1.8948   3rd Qu.: 406.2
 Max.   :0.9380   Max.   :1.0000   Max.   :1.0000   Max.   :7.3476   Max.   :2419.0

mining info:
 data ntransactions support confidence
 cars          2579     0.1        0.8
```

Figure: 31

```
> inspect(grules[1:10])
     lhs                                  rhs                          support   confidence coverage
[1]  {}                                => {consumption=Petrol}        0.9379604 0.9379604  1.0000000
[2]  {modelyear=2017}                  => {consumption=Petrol}        0.1085692 0.9090909  0.1194261
[3]  {Posting_City=Islamabad}          => {Province=Islamabad}        0.1360993 1.0000000  0.1360993
[4]  {Province=Islamabad}              => {Posting_City=Islamabad}    0.1360993 1.0000000  0.1360993
[5]  {Posting_City=Islamabad}          => {consumption=Petrol}        0.1252423 0.9202279  0.1360993
[6]  {Province=Islamabad}              => {consumption=Petrol}        0.1252423 0.9202279  0.1360993
[7]  {color=Black}                     => {consumption=Petrol}        0.1368748 0.9265092  0.1477317
[8]  {registration_city=Un-Registered} => {Assembly=Imported}         0.1337728 0.8175355  0.1636293
[9]  {registration_city=Un-Registered} => {geartype=Automatic}        0.1481194 0.9052133  0.1636293
[10] {registration_city=Un-Registered} => {consumption=Petrol}        0.1504459 0.9194313  0.1636293
     lift      count
[1]  1.0000000 2419
[2]  0.9692209  280
[3]  7.3475783  351
[4]  7.3475783  351
[5]  0.9810946  323
[6]  0.9810946  323
[7]  0.9877913  353
[8]  1.6907972  345
[9]  1.5968160  382
[10] 0.9802453  388
>
```

Figure: 32

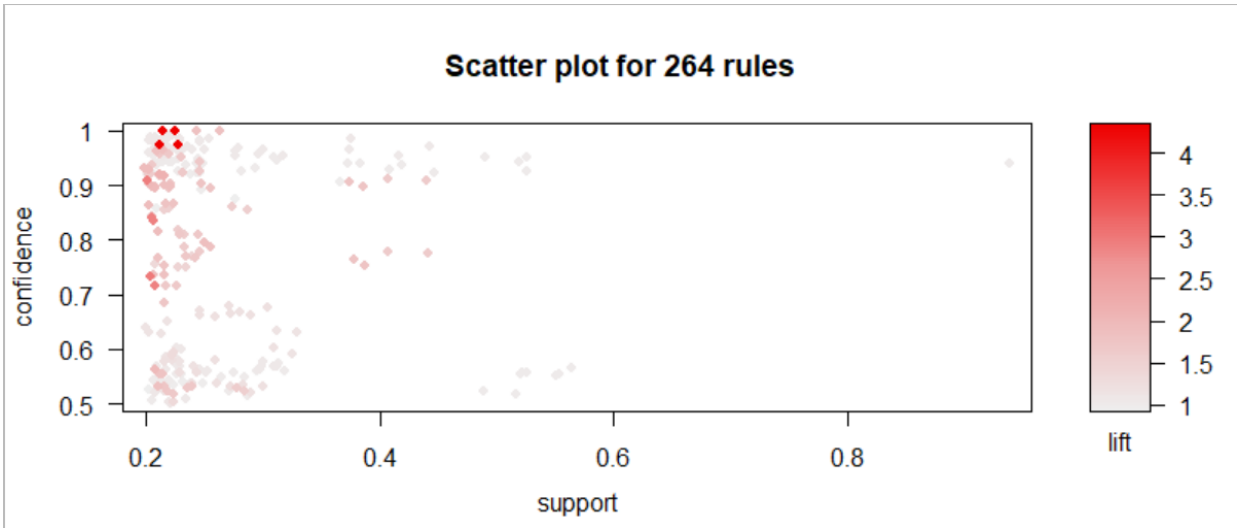


Figure: 33

```
> summary(grules1)
set of 264 rules

rule length distribution (lhs + rhs):sizes
  1   2   3   4
  5  96 133  30

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.000   3.000   2.712   3.000   4.000

summary of quality measures:
    support         confidence        coverage            lift            count
 Min.   :0.2001   Min.   :0.5000   Min.   :0.2036   Min.   :0.9136   Min.   : 516.0
 1st Qu.:0.2140   1st Qu.:0.5670   1st Qu.:0.2408   1st Qu.:1.0150   1st Qu.: 552.0
 Median :0.2288   Median :0.8362   Median :0.3164   Median :1.0568   Median : 590.0
 Mean   :0.2625   Mean   :0.7746   Mean   :0.3611   Mean   :1.3345   Mean   : 676.9
 3rd Qu.:0.2749   3rd Qu.:0.9426   3rd Qu.:0.4393   3rd Qu.:1.6357   3rd Qu.: 709.0
 Max.   :0.9380   Max.   :1.0000   Max.   :1.0000   Max.   :4.3277   Max.   :2419.0

mining info:
 data ntransactions support confidence
 cars          2579     0.2        0.5
```

Figure: 34

```
> inspect(grules1[1:10])
     lhs                                  rhs                                 support   confidence coverage
[1]  {}                                => {Assembly=Local}                   0.5164793 0.5164793  1.0000000
[2]  {}                                => {Province=Punjab}                  0.5498255 0.5498255  1.0000000
[3]  {}                                => {added_via=Added via Phone}        0.5513765 0.5513765  1.0000000
[4]  {}                                => {geartype=Automatic}               0.5668864 0.5668864  1.0000000
[5]  {}                                => {consumption=Petrol}               0.9379604 0.9379604  1.0000000
[6]  {Main_Location=Not Applicable} => {consumption=Petrol}               0.2027918 0.9526412  0.2128732
[7]  {Posting_City=Karachi}         => {Province=Sindh}                   0.2248934 1.0000000  0.2248934
[8]  {Province=Sindh}               => {Posting_City=Karachi}             0.2248934 0.9731544  0.2310973
[9]  {Posting_City=Karachi}         => {consumption=Petrol}               0.2117100 0.9413793  0.2248934
[10] {Province=Sindh}               => {consumption=Petrol}               0.2175262 0.9412752  0.2310973
     lift     count
[1]  1.000000 1332
[2]  1.000000 1418
[3]  1.000000 1422
[4]  1.000000 1462
[5]  1.000000 2419
[6]  1.015652  523
[7]  4.327181  580
[8]  4.327181  580
[9]  1.003645  546
[10] 1.003534  561
>
```

Figure: 35

Figure: 36

```
> summary(grules4)
set of 2203 rules

rule length distribution (lhs + rhs):sizes
  1   2   3   4   5   6
  5 173 808 868 313  36

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   3.000   4.000   3.644   4.000   6.000

summary of quality measures:
    support         confidence        coverage           lift            count
 Min.   :0.1000   Min.   :0.5000   Min.   :0.1000   Min.   :0.8601   Min.   : 258.0
 1st Qu.:0.1113   1st Qu.:0.6035   1st Qu.:0.1344   1st Qu.:1.0467   1st Qu.: 287.0
 Median :0.1256   Median :0.8017   Median :0.1799   Median :1.3289   Median : 324.0
 Mean   :0.1443   Mean   :0.7770   Mean   :0.1960   Mean   :1.5889   Mean   : 372.2
 3rd Qu.:0.1528   3rd Qu.:0.9448   3rd Qu.:0.2237   3rd Qu.:1.8311   3rd Qu.: 394.0
 Max.   :0.9380   Max.   :1.0000   Max.   :1.0000   Max.   :7.3476   Max.   :2419.0

mining info:
 data ntransactions support confidence
 cars          2579     0.1        0.5
```

Figure: 37

```
> inspect(grules4[1:10])
     lhs                          rhs                                support   confidence coverage
[1]  {}                        => {Assembly=Local}                   0.5164793 0.5164793  1.0000000
[2]  {}                        => {Province=Punjab}                  0.5498255 0.5498255  1.0000000
[3]  {}                        => {added_via=Added via Phone}        0.5513765 0.5513765  1.0000000
[4]  {}                        => {geartype=Automatic}               0.5668864 0.5668864  1.0000000
[5]  {}                        => {consumption=Petrol}               0.9379604 0.9379604  1.0000000
[6]  {modelyear=2017}          => {consumption=Petrol}               0.1085692 0.9090909  0.1194261
[7]  {Posting_City=Islamabad} => {Province=Islamabad}                0.1360993 1.0000000  0.1360993
[8]  {Province=Islamabad}      => {Posting_City=Islamabad}           0.1360993 1.0000000  0.1360993
[9]  {Posting_City=Islamabad} => {consumption=Petrol}               0.1252423 0.9202279  0.1360993
[10] {Province=Islamabad}      => {consumption=Petrol}               0.1252423 0.9202279  0.1360993
     lift      count
[1]  1.0000000 1332
[2]  1.0000000 1418
[3]  1.0000000 1422
[4]  1.0000000 1462
[5]  1.0000000 2419
[6]  0.9692209  280
[7]  7.3475783  351
[8]  7.3475783  351
[9]  0.9810946  323
[10] 0.9810946  323
>
```

Figure: 38

```
> summary(grules5)
set of 1104 rules

rule length distribution (lhs + rhs):sizes
  1   2   3   4   5   6
  1  53 362 456 203  29

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    3.00    4.00    3.81    4.00    6.00

summary of quality measures:
    support         confidence        coverage           lift            count
 Min.   :0.1000   Min.   :0.8000   Min.   :0.1000   Min.   :0.8601   Min.   : 258.0
 1st Qu.:0.1117   1st Qu.:0.9048   1st Qu.:0.1190   1st Qu.:1.0310   1st Qu.: 288.0
 Median :0.1249   Median :0.9447   Median :0.1353   Median :1.6177   Median : 322.0
 Mean   :0.1451   Mean   :0.9332   Mean   :0.1560   Mean   :1.7119   Mean   : 374.1
 3rd Qu.:0.1575   3rd Qu.:0.9719   3rd Qu.:0.1718   3rd Qu.:1.8948   3rd Qu.: 406.2
 Max.   :0.9380   Max.   :1.0000   Max.   :1.0000   Max.   :7.3476   Max.   :2419.0

mining info:
 data ntransactions support confidence
 cars          2579     0.1        0.8
```

Figure: 39

```
> inspect(grules5[1:10])
     lhs                                  rhs                          support confidence  coverage     lift count
[1]  {Posting_City=Islamabad}          => {Province=Islamabad}     0.1360993  1.0000000 0.1360993 7.347578   351
[2]  {Province=Islamabad}              => {Posting_City=Islamabad} 0.1360993  1.0000000 0.1360993 7.347578   351
[3]  {consumption=Petrol,
      Posting_City=Islamabad}          => {Province=Islamabad}     0.1252423  1.0000000 0.1252423 7.347578   323
[4]  {consumption=Petrol,
      Province=Islamabad}              => {Posting_City=Islamabad} 0.1252423  1.0000000 0.1252423 7.347578   323
[5]  {Assembly=Imported,
      Province=Sindh}                  => {Posting_City=Karachi}   0.1170997  0.9869281 0.1186506 4.388427   302
[6]  {geartype=Automatic,
      Province=Sindh}                  => {Posting_City=Karachi}   0.1388135  0.9862259 0.1407522 4.385304   358
[7]  {consumption=Petrol,
      Assembly=Imported,
      Province=Sindh}                  => {Posting_City=Karachi}   0.1093447  0.9860140 0.1108957 4.384362   282
[8]  {geartype=Automatic,
      Assembly=Imported,
      Province=Sindh}                  => {Posting_City=Karachi}   0.1066305  0.9856631 0.1081815 4.382802   275
[9]  {consumption=Petrol,
      geartype=Automatic,
      Province=Sindh}                  => {Posting_City=Karachi}   0.1306708  0.9853801 0.1326095 4.381544   337
[10] {registration_city=Karachi,
      Province=Sindh,
      added_via=Added via Desktop}     => {Posting_City=Karachi}   0.1031408  0.9851852 0.1046917 4.380677   266
>
```

Figure: 40

```
> summary(grules8)
set of 316 rules

rule length distribution (lhs + rhs):sizes
  1   2   3   4   5
  1  45 168  93   9

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   3.000   3.000   3.203   4.000   5.000

summary of quality measures:
    support         confidence        coverage            lift            count
 Min.   :0.1501   Min.   :0.8032   Min.   :0.1512   Min.   :0.9136   Min.   : 387.0
 1st Qu.:0.1714   1st Qu.:0.9016   1st Qu.:0.1845   1st Qu.:1.0191   1st Qu.: 442.0
 Median :0.1927   Median :0.9397   Median :0.2117   Median :1.0581   Median : 497.0
 Mean   :0.2106   Mean   :0.9285   Mean   :0.2272   Mean   :1.5675   Mean   : 543.1
 3rd Qu.:0.2198   3rd Qu.:0.9667   3rd Qu.:0.2354   3rd Qu.:1.8188   3rd Qu.: 566.8
 Max.   :0.9380   Max.   :1.0000   Max.   :1.0000   Max.   :4.3308   Max.   :2419.0

mining info:
 data ntransactions support confidence
 cars          2579    0.15        0.8
>
```

Figure: 41

```
> grules8 <- sort(grules8, by = "lift")
> inspect(grules8[1:10])
     lhs                                  rhs                                  support confidence  coverage     lift count
[1]  {consumption=Petrol,
      registration_city=Karachi,
      Province=Sindh}                  => {Posting_City=Karachi}               0.1740985  0.9739696 0.1787515 4.330806   449
[2]  {registration_city=Karachi,
      Province=Sindh}                  => {Posting_City=Karachi}               0.1845677  0.9734151 0.1896084 4.328341   476
[3]  {consumption=Petrol,
      Province=Sindh}                  => {Posting_City=Karachi}               0.2117100  0.9732620 0.2175262 4.327660   546
[4]  {Province=Sindh}                  => {Posting_City=Karachi}               0.2248934  0.9731544 0.2310973 4.327181   580
[5]  {Posting_City=Karachi}            => {Province=Sindh}                     0.2248934  1.0000000 0.2248934 4.327181   580
[6]  {registration_city=Karachi,
      Posting_City=Karachi}            => {Province=Sindh}                     0.1845677  1.0000000 0.1845677 4.327181   476
[7]  {consumption=Petrol,
      Posting_City=Karachi}            => {Province=Sindh}                     0.2117100  1.0000000 0.2117100 4.327181   546
[8]  {consumption=Petrol,
      registration_city=Karachi,
      Posting_City=Karachi}            => {Province=Sindh}                     0.1740985  1.0000000 0.1740985 4.327181   449
[9]  {consumption=Petrol,
      geartype=Manual,
      Assembly=Local,
      bodytype=Sedan}                  => {engine_capacity=1300 cc}            0.1504459  0.9463415 0.1589763 3.813460   388
[10] {consumption=Petrol,
      geartype=Manual,
      bodytype=Sedan}                  => {engine_capacity=1300 cc}            0.1574254  0.9333333 0.1686700 3.761042   406
>
```

Figure: 42

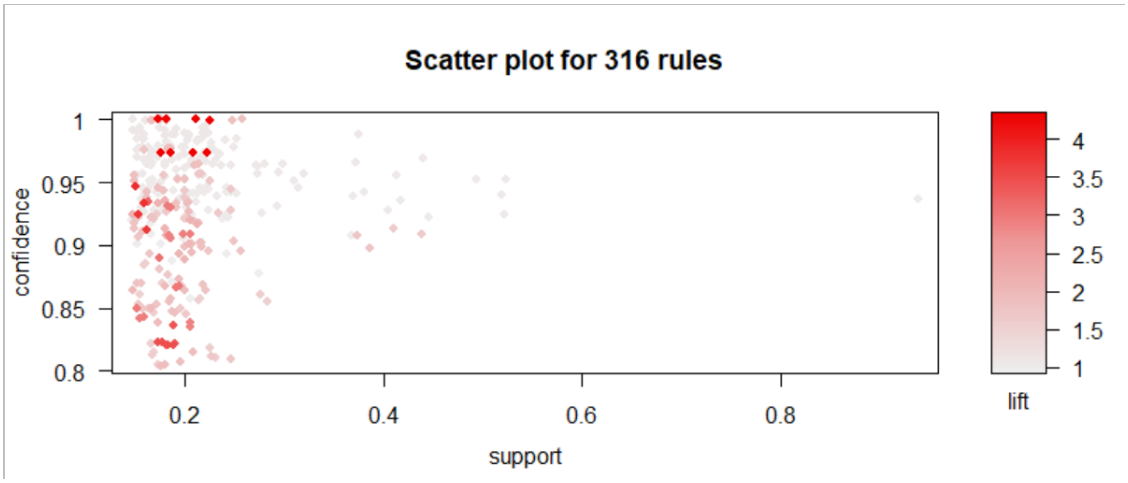


Figure: 43

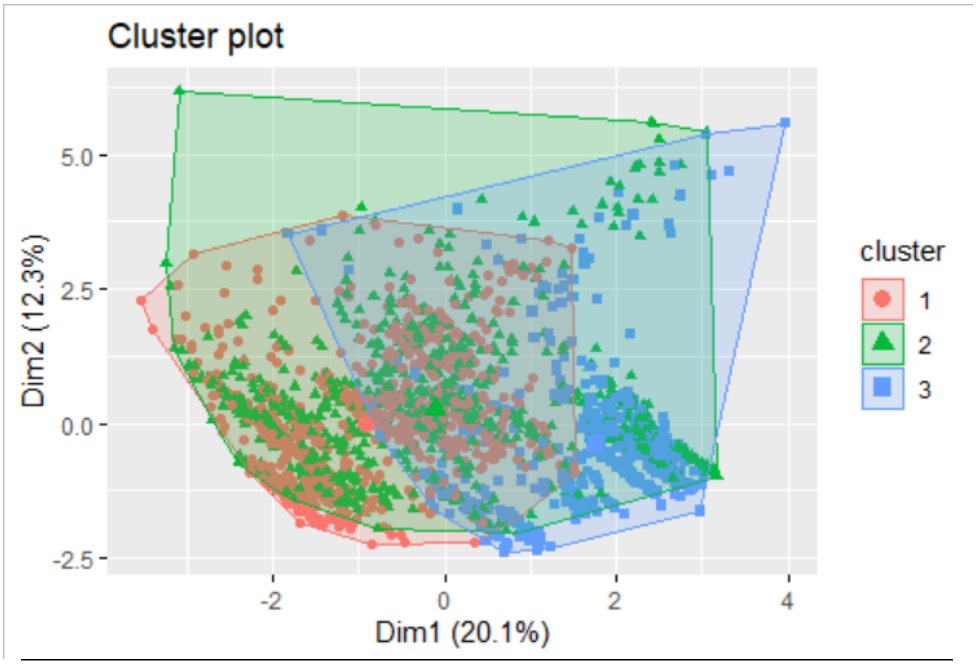


Figure: 44

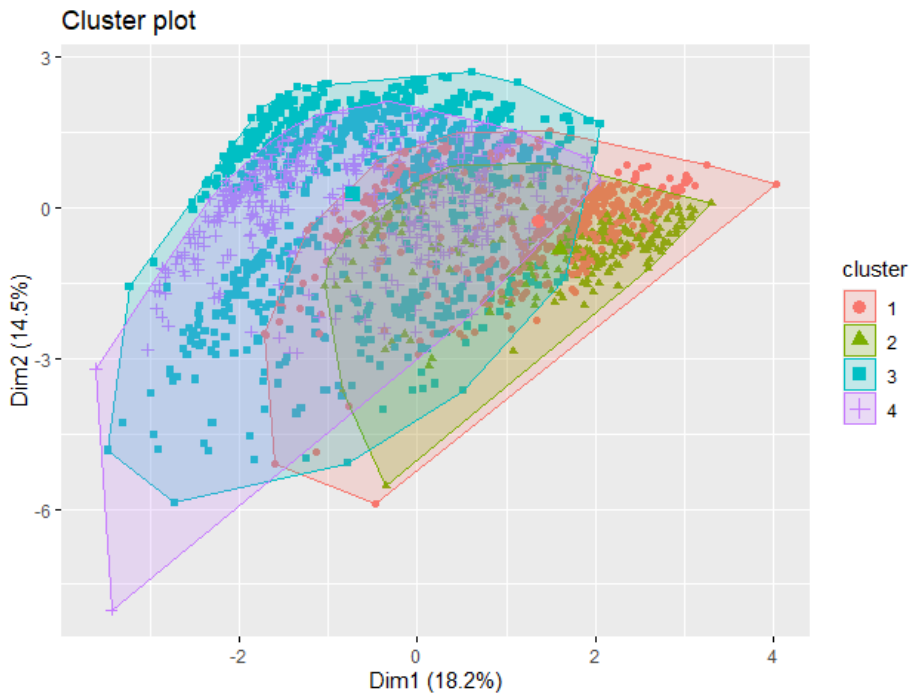


Figure: 45

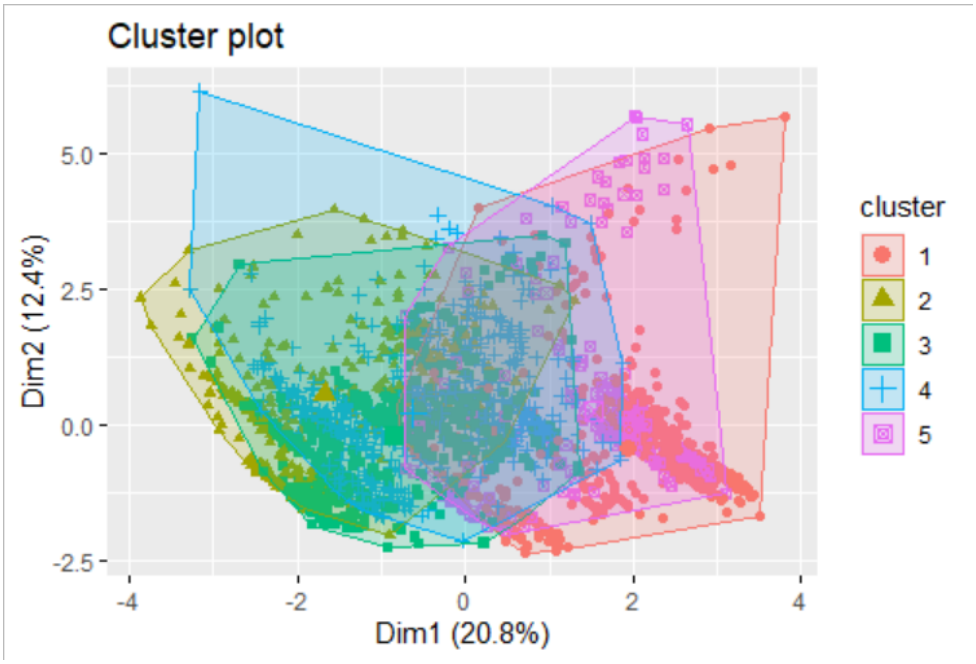

Cluster plot
Dim2 (12.4%)
5.0
2.5
0.0
-2.5
-4
-2
0
2
4
Dim1 (20.8%)
cluster
1
2
3
4
5

Figure: 46

APPENDIX B

CODEBOOK

Variables	Description
Car's Name (name)	This is a character column specifying details about the car for which a post has been made by the user.
Car's Model year (modelyear)	This field contains a character specifying the manufacturing year of the car which has been listed for posting.
Total Miles Travelled (totalmil)	This field represents how many miles the car has travelled ever since its first ownership.
Consumption Type (consumption)	This is a factor specifying the type of fuel the car uses. There are four unique categories in this column: CNG, Diesel, Hybrid and Petrol.
Gear Type (geartype)	This field Specifies whether the car is automatic or manual.
Registration City (registration_city)	This is a factor column specifying the city where the car was originally registered.
Assembly Type (Assembly)	This is a factor column specifying whether the car was assembled locally or imported from other countries.

Body Type (bodytype)	This is a factor column specifying the type of body a car has. There are 18 distinct choices recognized in the data.
Color (color)	This is a factor column that specifies the colour of a car. There are 68 unique colours identified in the dataset.
Engine Capacity (engine_capacity)	A factor column specifying the engine capacity. There are 36 different types of categories in the data.
Date Posted (posted_date)	This field represents date on which the posting was made.
Posted Location (posted_location)	This field represents address of the user who made the post.
Price (price)	This field shows the asking price for the car from the user.
Device Type (added_via)	This field represents through which device was the post made.

BIBLIOGRAPHY

ResearchAndMarkets.com [WWW Document], 2019. URL https://www.businesswire.com/news/home/20190208005417/en/2019-Future-of-Pakistan-Automobile-Market---Trends-Outlook-and-Growth-Opportunities---ResearchAndMarkets.com (accessed 5.3.21).

Arif, F., Sarfraz, S.S., 2017. Pakwheels.com—The Next Challenge! Asian J. Manag. Cases 14, 137–159. https://doi.org/10.1177/0972820117712315

How eCommerce Is Changing The Way We Shop For Cars In Pakistan - PakWheels Blog [WWW Document], n.d. URL https://www.pakwheels.com/blog/how-ecommerce-is-changing-the-way-we-shop-for-cars-in-pakistan/ (accessed 5.3.21).

Khan, M.Z., 2021. Pakistan's e-commerce market growing [WWW Document]. DAWN.COM. URL https://www.dawn.com/news/1606875 (accessed 5.3.21)