In this project, we will analyze data from the most popular used car website in Pakistan, OLX. This project aims to understand the trends in the car market in Pakistan and help potential buyers make accurate and reasonable decisions based on the insights presented with the analysis. We will use Python and Tabluea to make these visualizations and analyses.

```
data_car = pd.read_csv('car_data.csv', encoding='latin-1')
data_car
```

	Brand	Condition	Fuel	KMs Driven	Model	Price	Registered City	Transaction Type	Year
0	Toyota	Used	Diesel	1.0	Prado	2100000	Karachi	Cash	1997.0
1	Suzuki	Used	Petrol	100000.0	Bolan	380000	Karachi	Cash	2006.0
2	Suzuki	Used	CNG	12345.0	Bolan	340000	Karachi	Cash	1998.0
3	Suzuki	Used	Petrol	94000.0	Alto	535000	Karachi	Cash	2010.0
4	Toyota	Used	Petrol	100000.0	Corolla XLI	1430000	Karachi	Cash	2013.0
...	...	...	...	...	...	...	...	...	...
24968	Toyota	Used	CNG	200000.0	Corolla XE	1070000	Lahore	Cash	2001.0
24969	Daihatsu	New	Petrol	10000.0	Cuore	390000	Karachi	Cash	2004.0
24970	Other Brands	Used	CNG	158715.0	Other	180000	NaN	Cash	2000.0
24971	Suzuki	Used	Petrol	1.0	Alto	470000	Rawalpindi	Cash	2003.0
24972	Toyota	Used	Petrol	48500.0	Corolla GLI	2050000	Lahore	Cash	2017.0

24973 rows × 9 columns

Fig 1: Shows the dataset that we have, showing data of 25000 used cars in Pakistan.

To begin the analysis, handling the missing values in this dataset was essential. To handle the nan values, row-wise removal of nan values was done since first, even after the row-wise removal, we had sufficient data to make conclusions. Furthermore, removing column-wise nan did not make any sense. Furthermore, to add missing values, it was very difficult to do them for categorical variables, as a lot of our data has categorical variables.

Starting off with our analysis, it was essential to know which brands of cars have the highest number of cars in the used market, giving the user an idea on how hard or easy would be to find that brand, and also giving some hint about the resale value of the car.

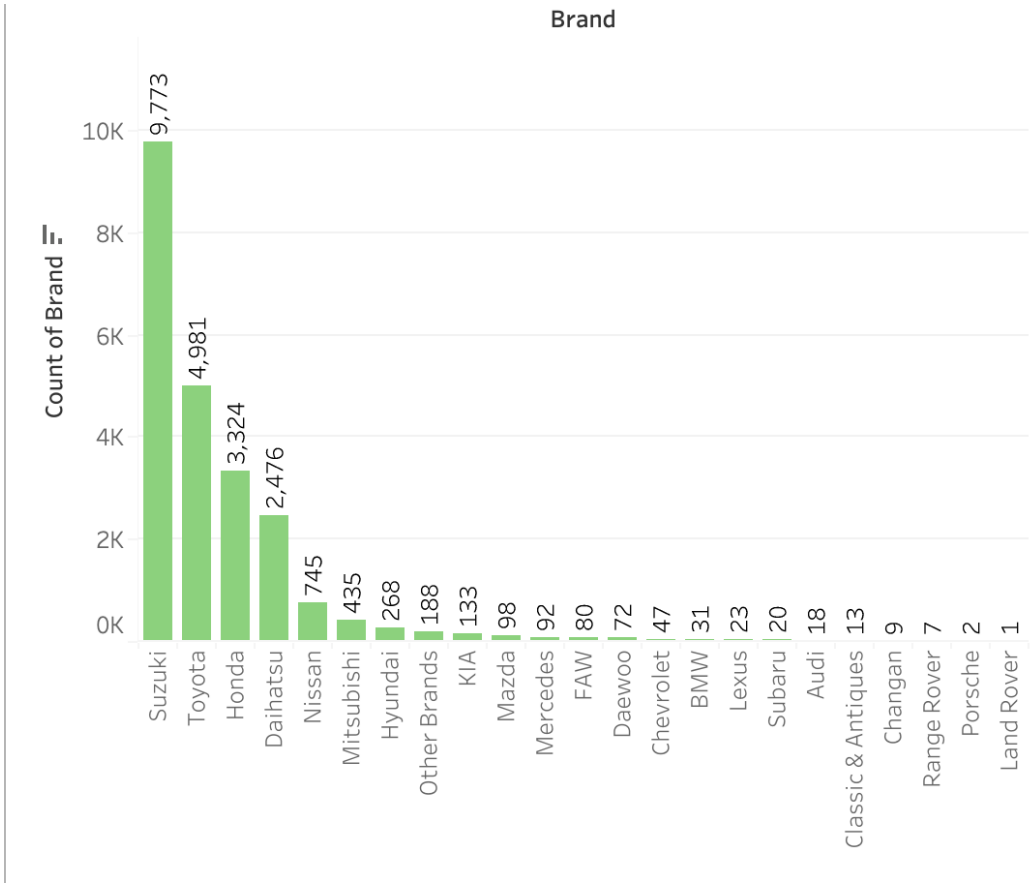

Fig 2: All brands and the number of cars of each brand available in OLX.

Figure 2 shows that Suzuki, Honda, and Toyota are the most popular brands in Pakistan and have the most amount of cars. A primary reason is their pricing is affordability. Finding something readily available on OLX would be challenging for someone interested in luxurious cars like Porsche and land rover. However, for Suzuki cars, different models would be available in sufficient quantity. Thus from this insight, we have an idea that OLX is better for standard and affordable cars and does not provide much variety for luxurious cars.

It is also helpful to find the most common car model in Pakistan. This would give us an insight on which cars have the most resale demand.

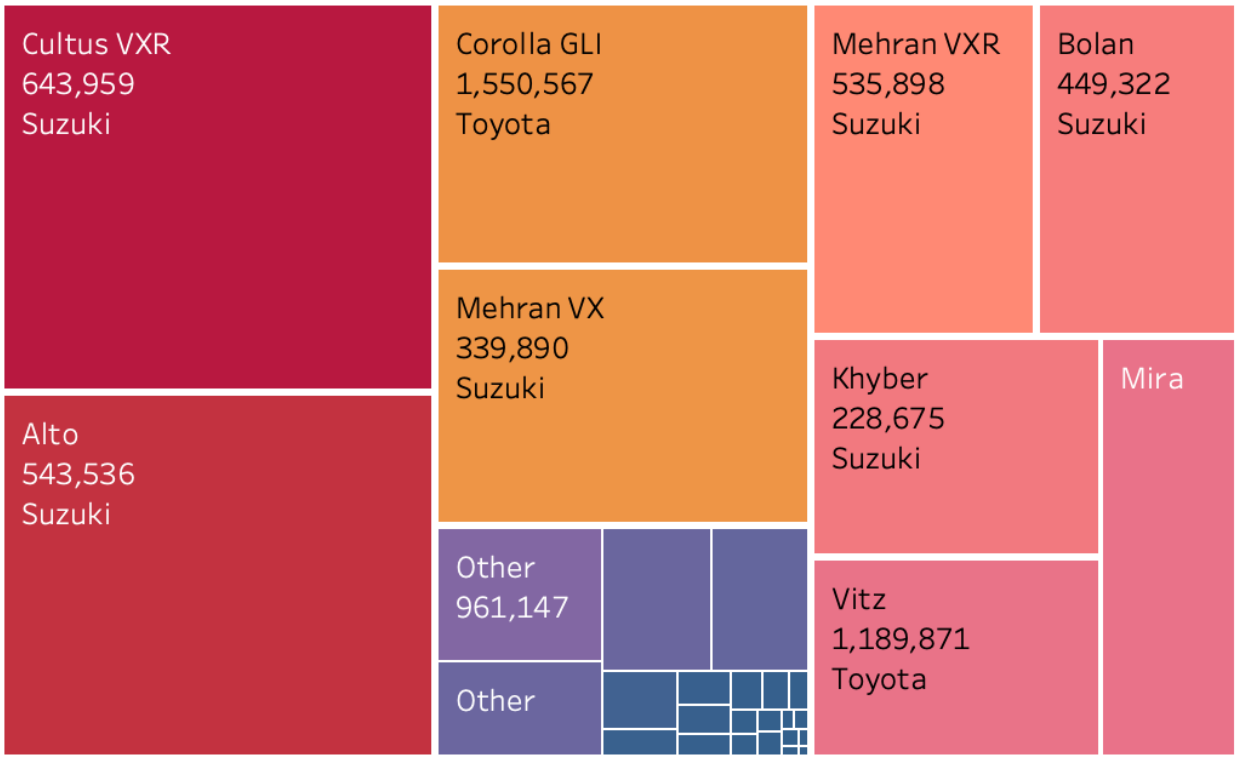

Fig 3: The top 10 most popular car models, their average price and their manufacturing company.

From the above figure, we can see that Cultus VXR is the most in-demand car, and we can also see that Suzuki manufactured it. We can also notice that almost all of the cars that are in high demand fall in the lower bracket of pricing, except for Toyota GLI and VITZ. From the figure, there are almost 6 Suzuki car models in the most demanded cars, so buying Suzuki would be a safe bet in terms of resale demand. If planning to spend more than a million PKR, Toyota Corolla and Vitz would have the highest resale demand.

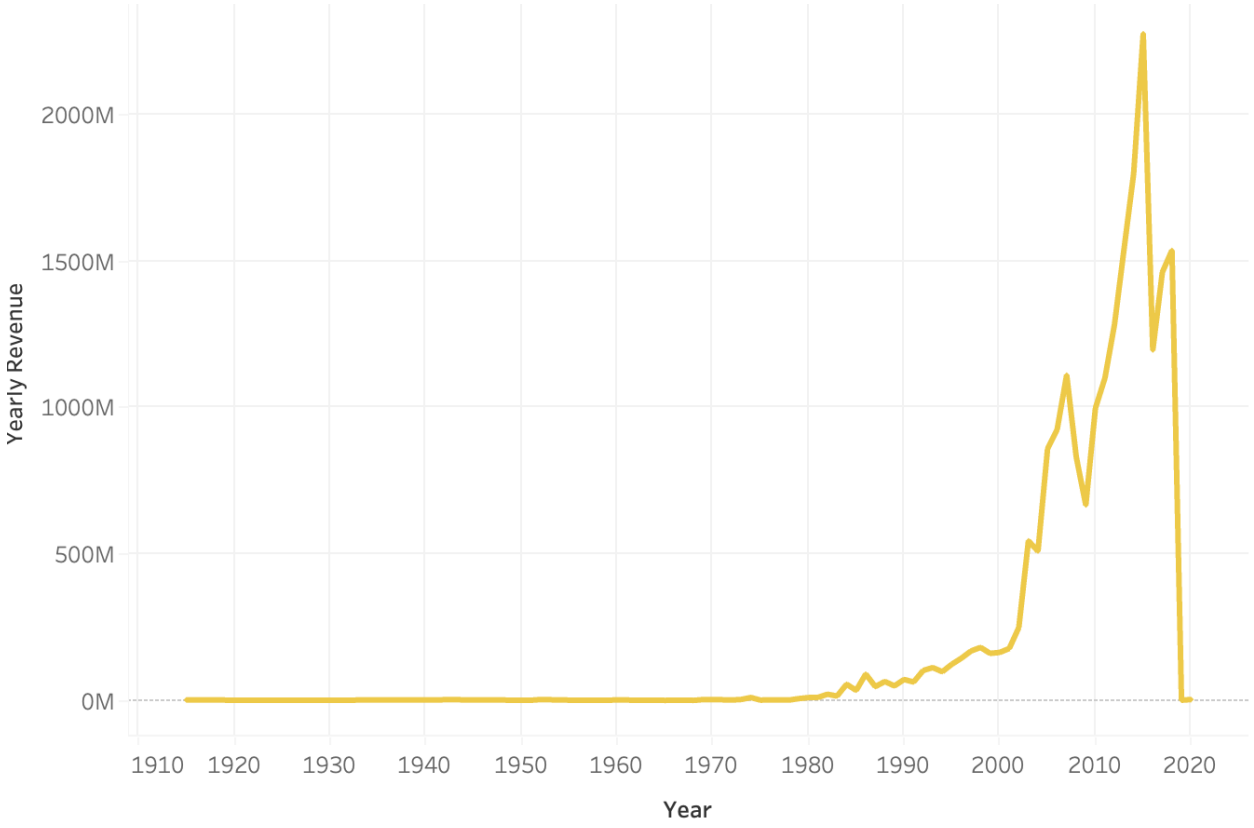

Fig 4: Figure 4 shows the year and the yearly revenue OLX obtained from selling cars for the specific year.

From this figure 4, we can see that Cars that were released in 2014-2016 had the highest demand and highest yearly revenue. The potential reason behind this is that the data is from 2020, which we can show that graph is almost zero, as almost no people are selling their used cars, as this entails losses. Cars before 1990 almost have zero revenue because they might have a lot of problems and hard-to-find spare parts, and that is why buyers do not prefer them. Cras after 2010 tend to have the highest demand because a lot of sellers have had the car when it was released new and are making to make a change, that is why they are in supply as well. Also, users tend to prefer cars 5-6 years older than the current model, as they are still considered reliable but are more affordable for everyone.

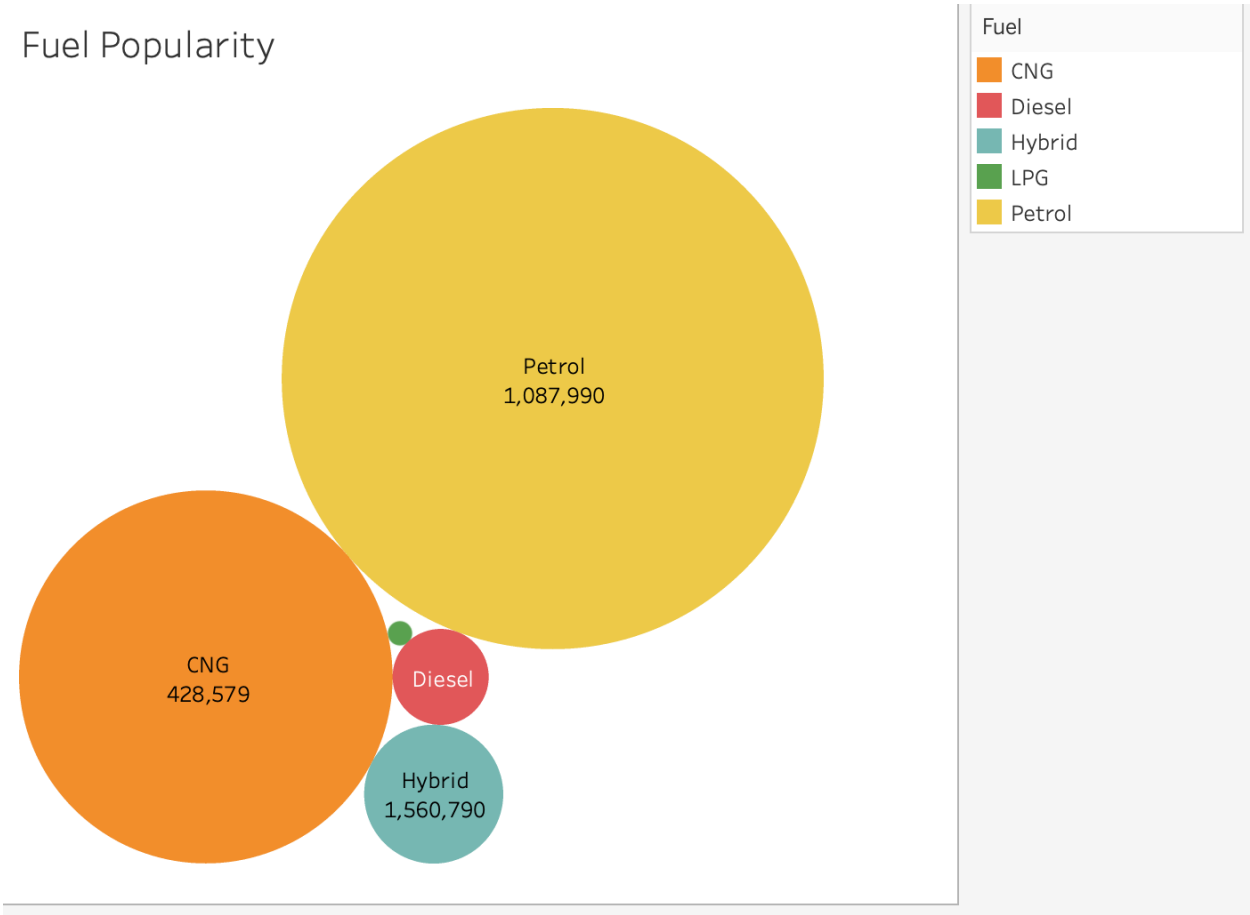

Fig 5: Shows the fuel type in Pakistan and their average price.

We can see that cars that run on petrol are the most common in Pakistan. Although the average price of petrol price is higher after a bit of research, the reason petrol cars are preferred is they are safer, car performance is higher, and the cars are not coming with a CNG gas kit. Hybrid cars tend to be not so famous and cars on diesel had a higher average price, making it fall into the expensive category, leading to lower demand.

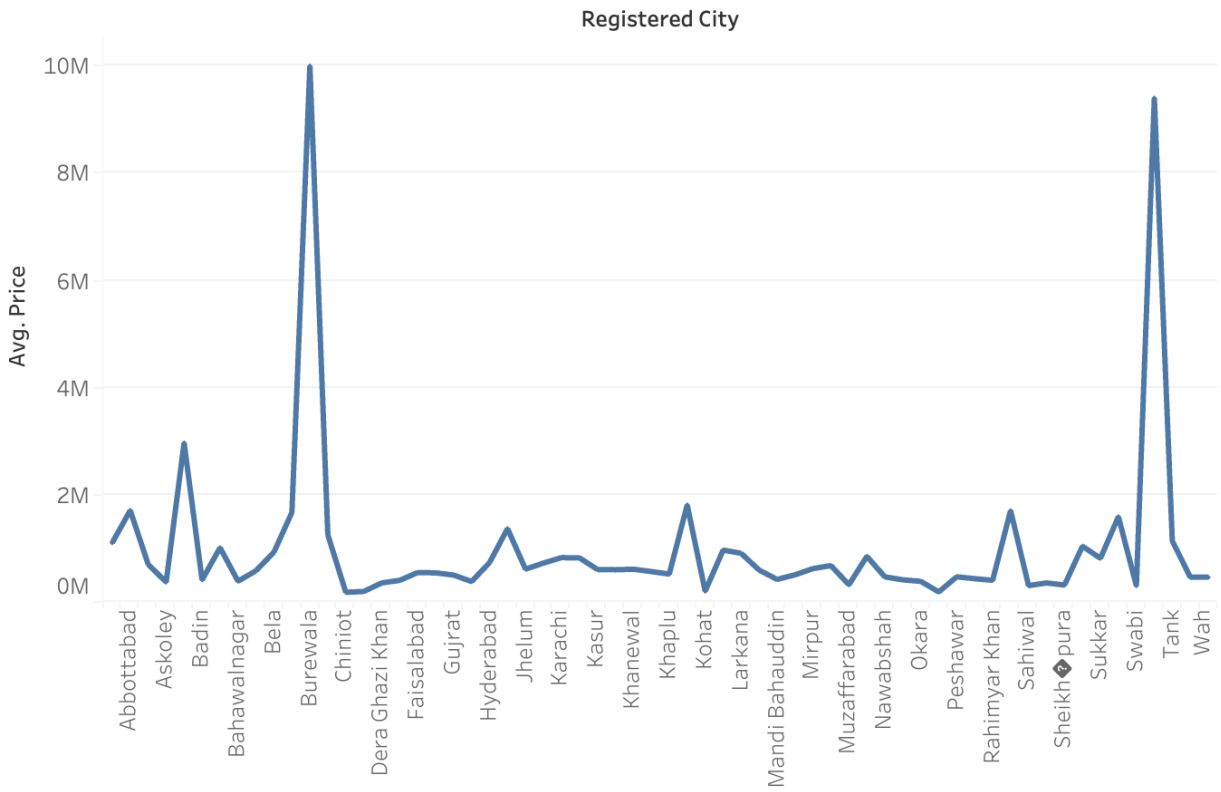

Fig 6: Average price of a car in each city

This plot gives an exciting insight. The average price in smaller cities like Burewala, tends to be very high. Analyzing the land area of Pakistan, these cities tend to have Steepy Hills, and 4 x4 cars are usually preferred on the roads, thus showing that the average car price is higher. Large cities like Karachi tend to have more affordable cars. Kohat is known for having very low-income in residents, and the graph shows that the average car price in Kohat is lower, which means cheaper cars are preferred in that region. This graph shows the user of the type of car and the price bracket he can expect depending on the part of Pakistan he is living in.

It is also important to understand the convention in cars that the higher the mileage of the car, the lower the price applies in this data case as well. Figure 7

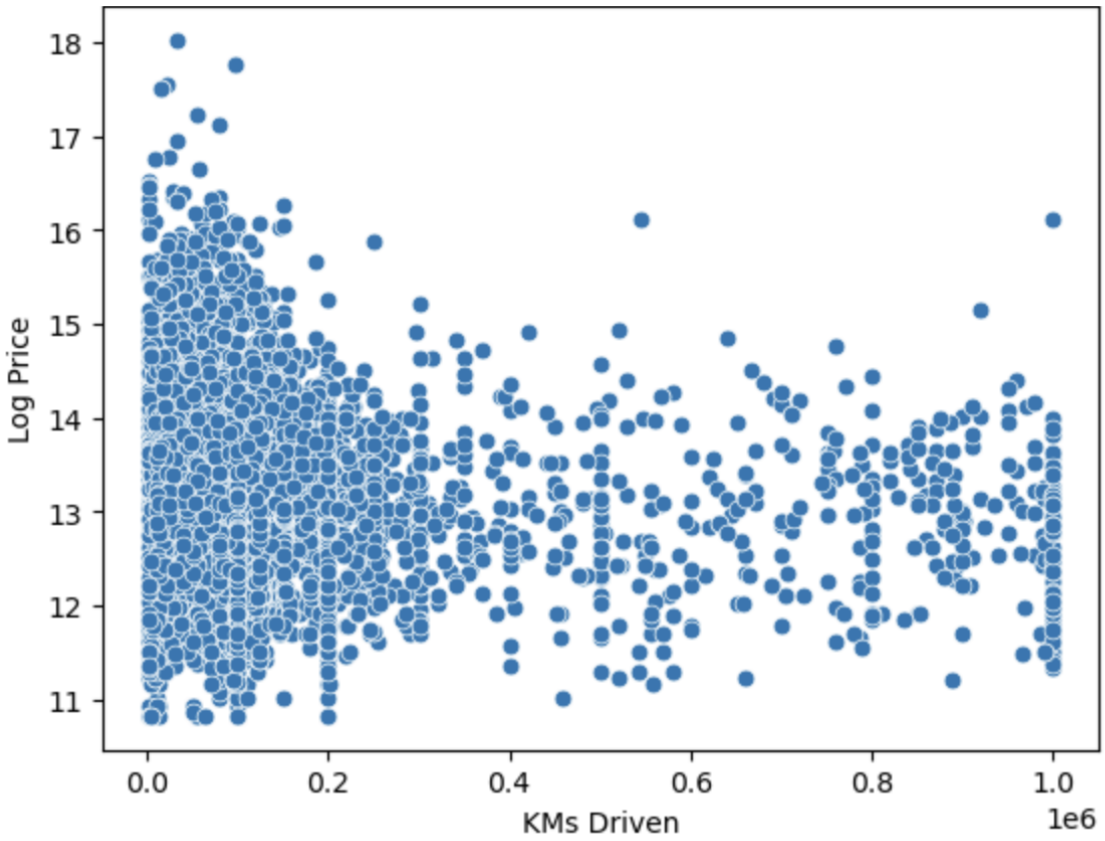

Fig 7: The KMs Driven against the Log Price

From figure 7, we can see that the lesser the mileage, the higher the price of the car, thus, considering the car's mileage before buying or selling the car would be an important factor in deciding the prices of the cars.

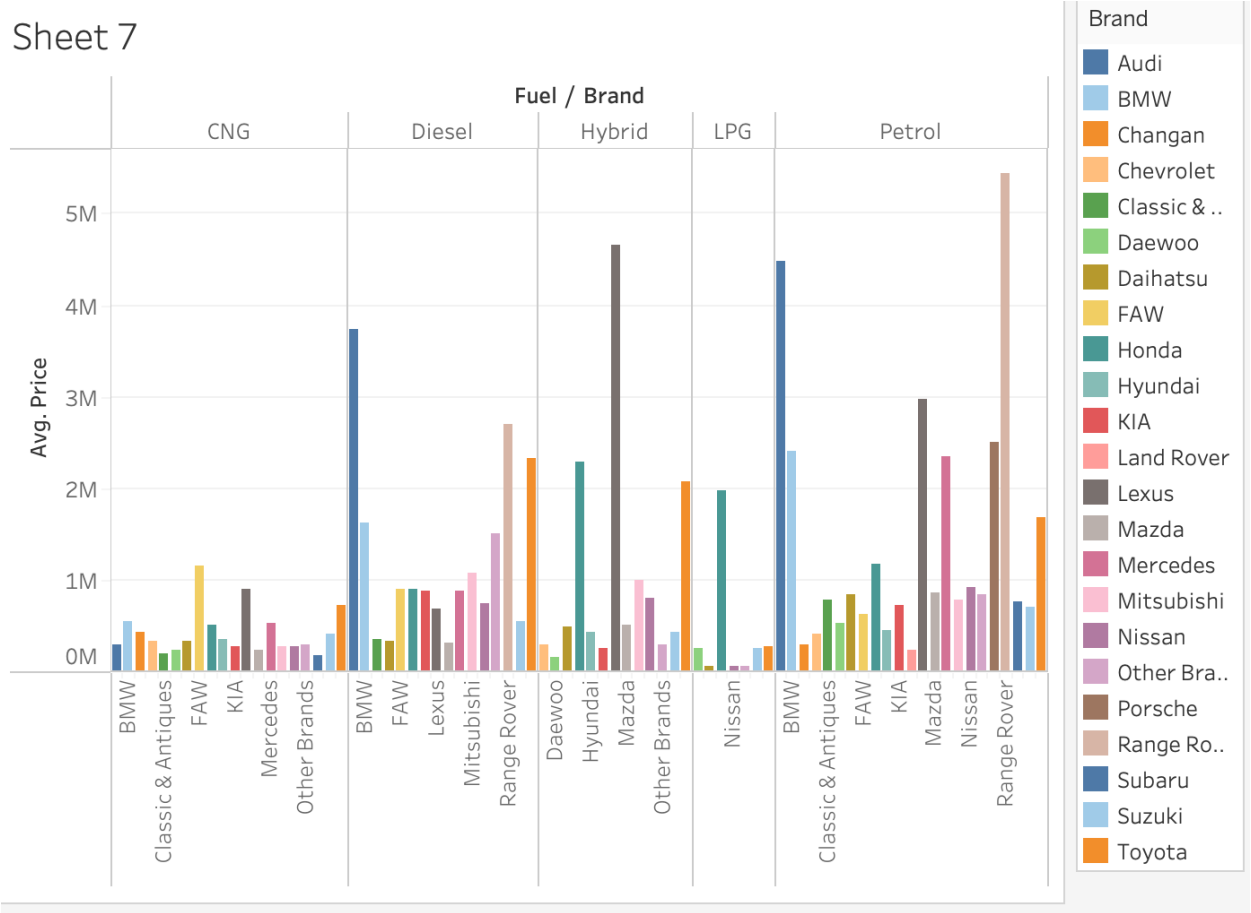

Fig 8: Shows the fuel types of all the brands in Pakistan and the corresponding average price.

This figure provides us with many significant trends, including which brands have cars that run on which kind of fuel. This allows the customer to analyze different options of brands he can buy from if he plans to buy petrol, etc. This also gives us the average price, which allows us to consider the pricing between different brands, provided that their fuel type is similar.

After the analysis, it was also essential to use Machine learning to predict the price of used cars in Pakistan and develop a system where users can estimate their price to protect them from being written off. To find out which factors contributed to the price of the car, we used the ANOVA test to find out which factor will impact the price.

```
ANOVA Results:
Predictor: Condition
F-value: 332.0415541694152
p-value: 1.67364723176846 5e-73
------------------------------------
Predictor: Fuel
F-value: 286.2620029471761
p-value: 8.927187128155894e-239
------------------------------------
Predictor: Brand
F-value: 107.94519842179086
p-value: 0.0
------------------------------------
Predictor: Year
F-value: 28.661547336291402
p-value: 5.775395e-318
------------------------------------
Predictor: Model
F-value: 26.86605122714554
p-value: 0.0
------------------------------------
Predictor: Registered City
F-value: 5.550532412105659
p-value: 5.1336609533510 02e-39
------------------------------------
Predictor: KMs Driven
F-value: 1.499657999669 6946
p-value: 5.488582543907 5855e-45
------------------------------------
```

Fig 9: ANOVA results

By the results and our analysis, we found out that KMs driven, registered city, model, brand, fuel, and condition all will impact the pricing, as all predictors are statistically significant.

A simple linear regression model was trained and tested to train the model and then test out the predictions. The predictions compared to the actual price are depicted in Figure 10.

	Actual Price	Predicted Price	Difference
12958	370000	4.651106e+05	95110.601831
1193	1180000	1.268403e+06	88402.697526
10669	590000	5.794114e+05	10588.588112
12002	1265000	1.255810e+06	9190.455756
201	55000	2.791586e+04	27084.142558
...	...	...	...
6023	400000	5.885938e+05	188593.757799
275	1135000	1.289025e+06	154024.993192
6247	660000	7.750770e+05	115077.031991
9078	1715000	1.578459e+06	136541.350100
23884	250000	-1.103590e+04	261035.897264

5205 rows × 3 columns

Figure 10: The results that were found using the linear regression machine learning model.

Figure 10 shows some of the car price predictions that the linear regression model made compared with the actual price.

Furthermore, I also ran a random forest, to make predictions of the price.

	Actual Price	Predicted Price	Difference
0	370000	3.360333e+05	33966.666667
1	1180000	1.210050e+06	30050.000000
2	590000	6.137500e+05	23750.000000
3	1265000	8.994000e+05	365600.000000
4	55000	1.145050e+05	59505.000000
...	...	...	...
5200	400000	4.794714e+05	79471.428571
5201	1135000	1.103767e+06	31233.333333
5202	660000	7.846500e+05	124650.000000
5203	1715000	1.860250e+06	145250.000000
5204	250000	2.753200e+05	25320.000000

Figure 11: shows the difference in the actual and predicted price, which is trained and tested on a random forest classifier.