Cars in the Middle East

Team 6 - Nabeel Bacchus

Data Description

- Data Source Link
  - https://www.kaggle.com/bushnag/cars-in-the-middle-east
- Column Information
  - Engine Capacity - Float, Ratio. Represents the capacity of each cylinder within an engine. Measured in liters
  - Cylinders - Integer, Interval. The amount of cylinders the engine has.
  - Drive Type - Integer, Nominal. States if the vehicle is Front Wheel, Rear Wheel, or All Wheel drive. Front = 0, All = 1 and Rear = 2
  - Fuel Tank Capacity - Float, Ratio. The amount of fuel the vehicle can hold in liters. petrol = 0, diesel = 1 and hybrid = 2
  - Fuel Economy - Float, Ratio. The amount of liters used to travel 100 kilometers
  - Fuel Type - Integer, Nominal. The type of fuel used by vehicle
  - Horsepower - Integer, Ratio. Power of the engine
  - Torque - Float, Ratio. Amount of turning power a car has
  - Transmission - Integer, Nominal. The type of transmission the vehicle has. Automatic = 0, Manual = 1 and CVT = 2
  - Top Speed - Integer, Ratio. The top speed of a vehicle measured in kilometers
  - Seating Capacity - Integer, Interval. The amount of seats the vehicle has.
  - Acceleration - Float, Ratio. How fast the vehicle can go from 0-100 km/h. Measured in seconds
  - Length - Float, Ratio. Length of a vehicle measured in meters
  - Width - Float, Ratio. Width of a vehicle measured in meters
  - Height - Float, Ratio. Height of a vehicle measured in meters
  - Wheelbase - Float, Ratio. Distance in meters between front and rear wheel of a vehicle

- Trunk Capacity - Float, Ratio. Size of the trunk for a vehicle in liters
- Name - String, Nominal. Name of vehicle
- Price - Float, Ratio. Price of vehicle
- Currency - Integer, Nominal. Type of Currency used to show price. SAR = 0, AED = 1, QAR = 2, KWD = 3, OMR = 4, BHD = 5
- Country - Integer, Nominal. Country vehicle was bought in. KSA = 0, UAE = 1, Qatar = 2, Kuwait = 3, Oman = 4, Bahrain = 5

The goal of using this dataset is to use the characteristics of a car that the buyer wants to determine the cost of an imported car within the Middle East.

This dataset contains 5667 rows and 21 columns. It contains information and features about all the imported cars to the Middle East. Since the dataset already has removed any NULL values, there are only 4560 rows remaining. All vehicles in the dataset are the 2021 model. The value of Price is based on the Currency variable in the dataset. Since there are numerous currencies used to represent the price of a car, they will all be converted to USD to have an easier time understanding when comparing the cars. Additionally, there were duplicate rows of cars that were removed to prevent bad training. Some of the rows in Length, Width, and Height were in millimeters instead of meters, these were adjusted to be in meters.

Exploratory Data Analysis

For each one of the variables, we were able to calculate the mean, standard deviation, minimum, first quartile, median, third quartile, and maximum. We also created a heatmap of the correlation between variables.

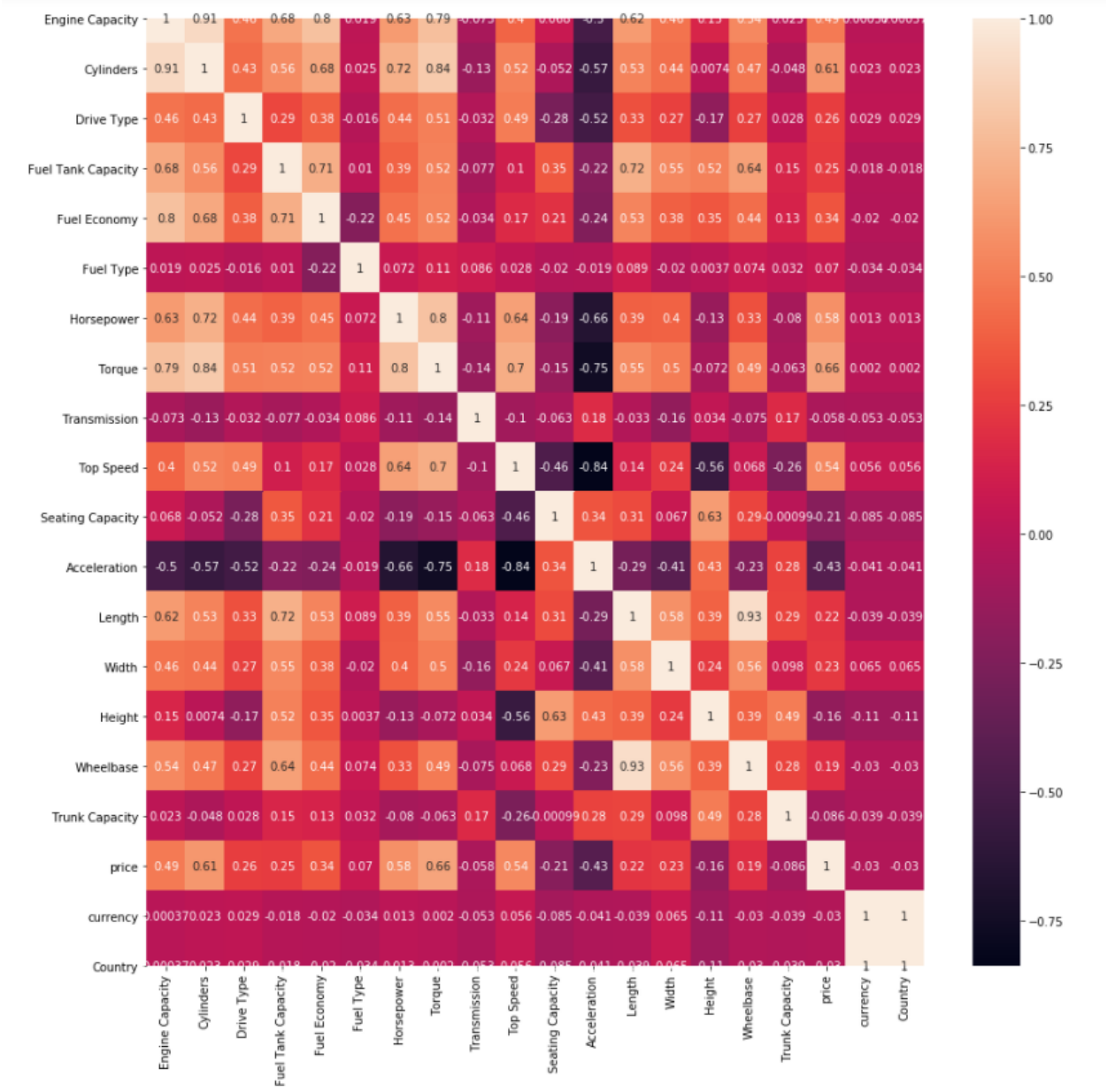
	Engine Capacity	Cylinders	Drive Type	Fuel Tank Capacity
count	4276.000000	4276.000000	4276.000000	4276.000000
mean	3.013471	5.545370	0.994855	70.089687
std	1.384787	1.937274	0.759792	21.587385
min	0.000000	3.000000	0.000000	32.000000
25%	2.000000	4.000000	0.000000	56.000000
50%	2.900000	6.000000	1.000000	66.000000
75%	3.700000	6.000000	2.000000	80.000000
max	8.000000	16.000000	2.000000	159.000000

	Fuel Economy	Fuel Type	Horsepower	Torque	Transmission
count	4276.000000	4276.000000	4276.000000	4276.000000	4276.000000
mean	9.030978	0.061974	295.838167	400.978251	0.087699
std	2.933633	0.330365	180.654371	194.877648	0.352141
min	2.500000	0.000000	67.000000	87.000000	0.000000
25%	6.800000	0.000000	170.000000	246.000000	0.000000
50%	8.600000	0.000000	275.000000	366.000000	0.000000
75%	10.700000	0.000000	380.000000	518.500000	0.000000
max	22.500000	2.000000	5050.000000	1600.000000	2.000000

	Top Speed	Seating Capacity	Acceleration	Length	Width
count	4276.000000	4276.000000	4276.000000	4276.000000	4276.000000
mean	224.523854	4.913237	7.534799	4.748167	1.903897
std	41.007010	1.245606	2.815803	0.410732	0.125461
min	140.000000	2.000000	2.400000	3.500000	1.595000
25%	190.000000	4.000000	5.300000	4.482000	1.822000
50%	220.000000	5.000000	7.100000	4.735000	1.885000
75%	250.000000	5.000000	9.300000	5.005000	1.980000
max	420.000000	9.000000	17.900000	6.758000	2.463000

	Height	Wheelbase	Trunk Capacity	price	currency
count	4276.000000	4276.000000	4276.000000	4.276000e+03	4276.000000
mean	1.560789	2.816144	590.892423	7.666892e+04	2.707905
std	0.190660	0.251186	785.740751	1.144881e+05	1.763315
min	1.136000	2.250000	45.000000	7.813721e+03	0.000000
25%	1.427000	2.660000	370.000000	2.917836e+04	1.000000
50%	1.485000	2.800000	473.000000	5.068153e+04	3.000000
75%	1.693000	2.945250	550.000000	8.992934e+04	4.000000
max	2.796000	4.260000	17000.000000	3.047266e+06	5.000000

	Country
count	4276.000000
mean	2.707905
std	1.763315
min	0.000000
25%	1.000000
50%	3.000000
75%	4.000000
max	5.000000



From the heatmap above we are able to see the correlation between all the variables in the dataset. Some important relations are listed below:

- Engine Capacity, Cylinders, and Torque all have high positive correlation with each other.
- Currency and Country have high correlation with each other because they contain the same information but are in different formats.
- Horsepower and Torque have high positive correlation with each other.

- Acceleration has a high negative correlation with Top Speed and Torque. But Top Speed and Torque have a moderately high positive correlation with each other.
- Length and Wheelbase have high positive correlation because wheelbase is the distance between the two axles. This means that there is significant overlap between the two variables.
- Fuel Economy and Fuel Tank Capacity have moderately high positive correlation.
- Length and Fuel Tank Capacity also has a moderately positive correlation.

Feature Engineering

In order to reduce dimensionality, Length, Width, and Height were multiplied together to form a new predictor variable called “Volume”. The other predictors were dropped from the dataframe. Volume is in meters cubed. After plotting all the numeric variables, we found that most of the graphs were skewed right. In order to normalize the variables, we used log so they may have better correlation and p-value. Additionally, we used the Robust scaler to scale down outliers in the numeric variables within the dataset. Now that the variables are normalized and scaled, we used Pearson Correlation Test to find the correlation and if they have a significant relationship between them.

Variable	Correlation	p-value
Engine Capacity	0.4855696954008153	6.920719997201396e-252
Cylinders	0.6055479585985701	0.0
Drive Type	0.2558418047711355	7.074980464160999e-65
Fuel Tank Capacity	0.2477197301326582	8.250406857246237e-61
Fuel Economy	0.3441918524177089	3.3319051500551337e-119
Fuel Type	0.07026063612934763	4.250069231066791e-06
Horsepower	0.5769138697851373	0.0
Torque	0.6646848540699242	0.0
Transmission	-0.05808898341813288	0.00014434434415605129

Top Speed	0.5445285168868164	0.0
Seating Capacity	-0.2129132064452045	5.019425975259784e-45
Acceleration	-0.4253065787517505	1.8817668959642618e-187
Wheelbase	0.19430669409109835	1.193527602183964e-37
Trunk Capacity	-0.08615090106252046	1.6774456727111715e-08
Volume	0.05422368937808836	0.0003891694246424251

From the test, we find that all the variables have a highly significant relationship with Price, however only a few have a moderate positive or negative correlation. The following variables will be used to create a linear model to predict what price of a car would be in the Middle East:

- Engine Capacity
- Cylinders
- Horsepower
- Torque
- Top Speed
- Acceleration