BUSINESS REPORT: CAR PRICE PREDICTION

Objective

To build a regression model based on sample data and predict the price in the test data of the used cars.

Training Data

Granularity of Data

The Data in the provided dataset consists of cars In India. Each column provides data of company, car name, location, year, kilometer driven, fuel type, transmission, owner type, mileage, engine, power, seats, price.

Shape of Data

The Training data excel file consists of data with 14 columns and 6020 rows.

The 14 columns contain 14 variables namely: ID, Name, Location, Year, Kilometers_Driven, Fuel Type, Transmission, Owner type, Mileage, Engine, Power, Seats, New-Price and Price.

The data is a mixture of both categorical and numerical data.

Dependent & Independent Variables

The Depending variable in the given sample data is price and all the other variables are independent variables. We have to check which independent variables are affecting price and based on that build a model.

Data Cleaning: Finding Missing Values & missing Percentages

Under this we check for any null values, manual errors, blank cells etc. and try to do corrections by filling Central values (Mean/Median/Mode) of that particular column in blank and null places and remove the errors.

Correlation Matrix

After doing the above process and getting our data ready for analysis, the data should be used to create Correlation Matrix.

After creation of the matrix we should do conditional formatting using colour scales to see highly correlated variables. We consider only one of the variables from the pair having high correlation to proceed further in building regression analysis.

The Correlation Matrix is presented Below:

	Company_name_num	*Location_num*	*Year*	*Kilometers_Driven*	*Diesel*	*Petrol*	*CNG*	*LPG*	*Manual*	*Owner_type_num*	*Mileage_num*	*Engine_num*	*Power_num*	*Seats*
Company_name_num	1													
Location_num	0.148960329	1												
Year	0.022712175	0.250466547	1											
Kilometers_Driven	-0.03267296	-0.121642622	-0.436792041	1										
Diesel	0.308705047	0.036097847	0.126185089	0.223135387	1									
Petrol	-0.297079972	-0.031800209	-0.127489784	-0.222404605	-0.977528063	1								
CNG	-0.054089998	-0.016343094	0.019023243	-0.005820396	-0.103422154	-0.088764411	1							
LPG	-0.021529441	-0.012206319	0.031919482	0.002322725	-0.043536235	-0.037365961	-0.003953331	1						
Manual	-0.652957956	-0.159043272	-0.097058578	0.100392132	-0.141613156	0.128988344	0.061297474	0.025803545	1					
Owner_type_num	0.025855855	-0.077446556	-0.397151596	0.212764418	-0.045358468	0.045286809	0.000242725	-0.000255	0.008796	1				
Mileage_num	-0.396602184	-0.078518387	0.290577469	-0.137251586	0.111267815	-0.144141313	-0.165199474	0.010248525	0.358559656	-0.150215398	1			
Engine_num	0.651933563	0.124815438	-0.051708241	0.157464095	0.426750565	-0.406902192	-0.085709	-0.040411202	-0.490829022	0.05045 2033	-0.639786758	1		
Power_num	0.765601049	0.158716426	0.014063439	0.010474505	0.289012384	-0.268777167	-0.086584	-0.040875	-0.630887 3851	0.025113851	-0.548464006	0.859052128	1	
Seats	-0.024144262	0.027730552	0.015204176	0.204182316	0.308869354	-0.302445961	-0.028975 4	-0.010044 4	0.074837 6	0.012210038	-0.341467246	0.392983191	0.100226955	1

Regression Analysis:

After removing the correlated variables, we created the regression model. Insert Avg_Price variable in y-axis as it is dependent variable and insert all the other independent variable in x-axis and perform regression.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.870846217
R Square	0.758653134
Adjusted R Square	0.757991752
Standard Error	5.502684063
Observations	6019

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	7	571259.7205	81608.53151	2695.171503	0
Residual	6011	182010.2662	30.27953179		
Total	6018	753269.9867			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-1805.518708	51.95318311	-34.75280243	2.6053E-241	-1907.365584	-1703.671833	-1907.365584	-1703.671833
Company_name_num	0.548064599	0.013043648	42.01773928	0	0.522494377	0.573634828	0.522494377	0.573634828
Location_num	0.196488786	0.026629436	7.378631005	1.81584E-13	0.144285538	0.248692034	0.144285538	0.248692034
Year	0.892770238	0.025803754	34.59846382	2.2609E-239	0.842185625	0.943354852	0.842185625	0.943354852
Kilometers_Driven	-1.14542E-05	2.19461E-06	-5.219219364	1.85707E-07	-1.57564E-05	-7.15194E-06	-1.57564E-05	-7.15194E-06
Diesel	0.657648876	0.170456796	3.858155797	0.000115441	0.323492401	0.991805342	0.323492401	0.991805342
Engine_num	0.001294433	0.000256437	5.047756221	4.60187E-07	0.000791724	0.001797142	0.000791724	0.001797142
Power_num	0.076706449	0.003192926	24.02387498	7.4452E-122	0.070447169	0.082965729	0.070447169	0.082965729

- **R-squared is a statistical measure that represents the goodness of fit of a regression model. Ideal value for r-square is 1. The closer the value of r-square to 1, the better is the model fitted.**
- The **adjusted R-squared** is a modified version of R-squared that adjusts for the number of predictors in a regression model.
- **The Adjusted R square value should lie between 0 to 1.**
- The Significance of a variable depends on it's **P-Value**. If the **P-Value is less than 0.05** then the Variable is significant for our analysis, else insignificant.

Inference:

The Regression summary Output gives the better performing regression with R square above 72% probability.

Test Data

Granularity of Data

The Data in the provided dataset consists of cars In India. Each column provides data of company, car name, location, year, kilometer driven, fuel type, transmission, owner type, mileage, engine, power, seats, price.

Shape of Data

The Data contains 13 columns which help in classification of each car and there are a total of 1234 records which tells us there are 1234 cars details in the test data.

Dependent & Independent Variables

The Depending variable in the given sample data is price and all the other variables are independent variables. We have to check which independent variables are affecting price and based on that build a model.

Data Cleaning: Finding Missing Values & missing Percentages

Under this we check for any null values, manual errors, blank cells etc. and try to do corrections by filling Central values (Mean/Median/Mode) of that particular column in blank and null places and remove the errors.

Price Prediction

Regression Equation/ Best Fit Line Equation: $\mathbf{Y=M_1X_1+M_2X_2+C.}$

From the above Model, the Regression Equation can be created. The Equation is:

Price = (-1805.51) + 0.5481*Company_name_num + 0.1964*Location_num + 0.8927*Year + (-0.00001)*Kilometers_Driven + 0.6576*Diesel + 0.0013*Engine_num + 0.0767*Power_num .