upGrad

iit-b
ज्ञानमुत्तमम

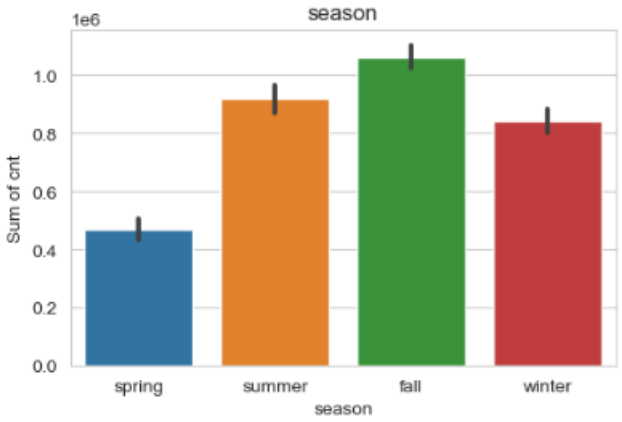
Boom Bikes

Linear Regression Case Study

Assignment-based Subjective Questions

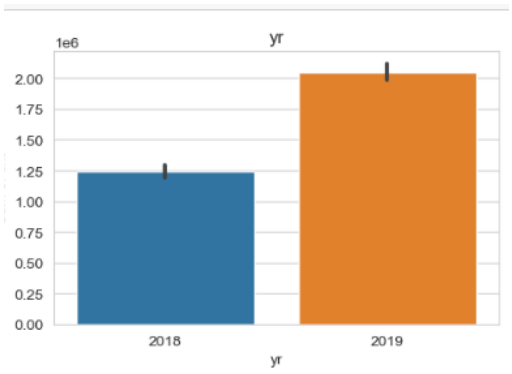
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Here is the inference about the effect of categorical columns on the dependent variable :



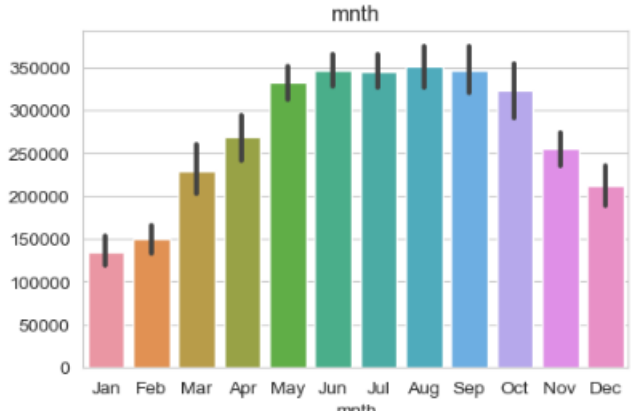
Max Rides are from the fall Season

season		
fall	5644	31.40
spring	2608	14.51
summer	4992	27.78
winter	4728	26.31



Rides have increased year on year

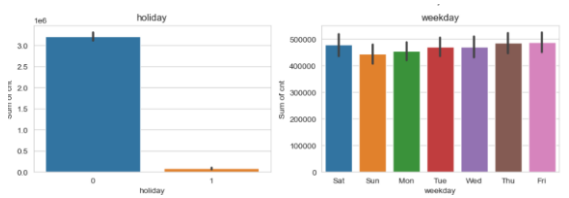
yr		
2018	3406	37.77
2019	5610	62.23



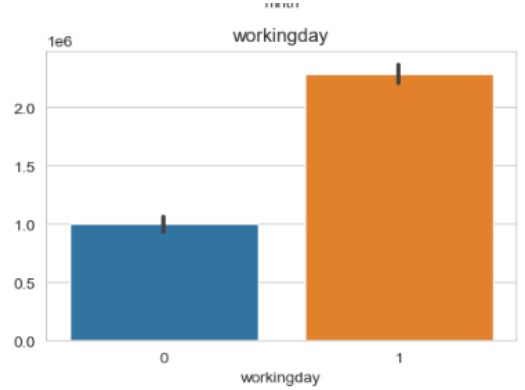
Rides growth visible till June and after that it started declining, Peak are between May to Sep.

Apr	4485	8.31
Aug	5664	10.49
Dec	3404	6.30
Feb	2670	4.95
Jan	2176	4.03
Jul	5564	10.30
Jun	5772	10.69
Mar	3692	6.84
May	5350	9.91
Nov	4247	7.87
Oct	5199	9.63
Sep	5767	10.68

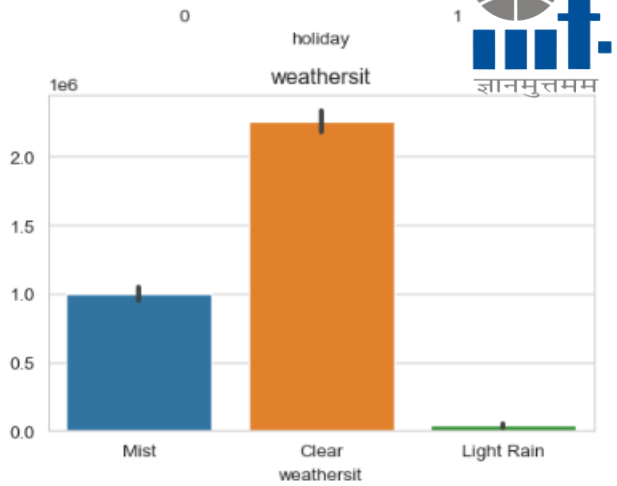
Assignment-based Subjective Questions



No Major Impact on holiday and weekday



Working days have higher ride counts



Clear Weather Situation have higher Rides

Pivot table for column 'holiday':

holiday	cnt	Percentage
0	4531	54.81
1	3735	45.19

Pivot table for column 'weekday':

weekday	cnt	Percentage
Fri	4690	14.86
Mon	4338	13.75
Sat	4551	14.42
Sun	4229	13.40
Thu	4667	14.79
Tue	4511	14.29
Wed	4575	14.50

workingday		
0	4330	48.54
1	4590	51.46

weathersit		
Clear	4877	45.47
Light Rain	1803	16.81
Mist	4045	37.71

Assignment-based Subjective Questions

2. Why is it important to use drop_first=True during dummy variable creation?

When we apply dummy variable technique to a variable , it basically convert all the characters to True/ False (Abbreviated as 0,1). If value of a row is true that means the other variable can be considered as false itself. Having that column will add an extra space which can be very well interpreted even without its presence , so it drops the first column.

For Example : Lets say we have a column as Weather condition which has three distinct values : clear , light rain , ,mist. The same will be represented as below :

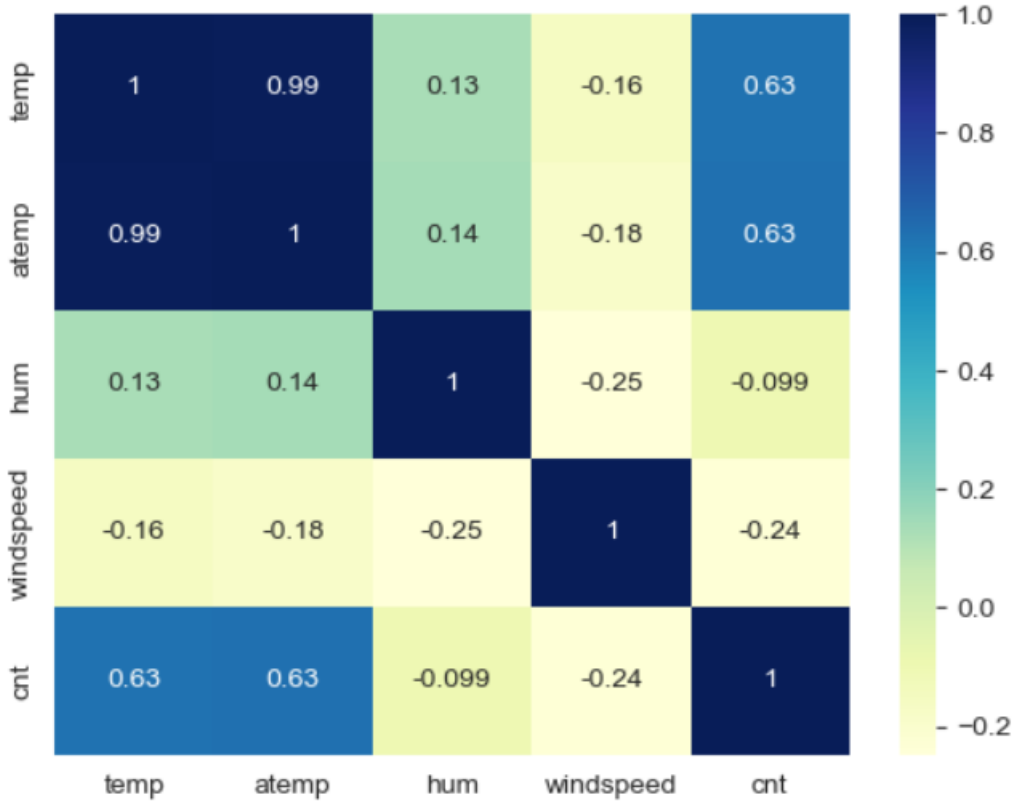
Clear	Light Rain	Mist	Represents
0	1	0	Light Rain
0	0	1	Mist
1	0	0	Clear

Same can be Achieved With Just 2 Columns

Light Rain	Mist	Represents
1	0	Light Rain
0	1	Mist
0	0	Clear

Assignment-based Subjective Questions

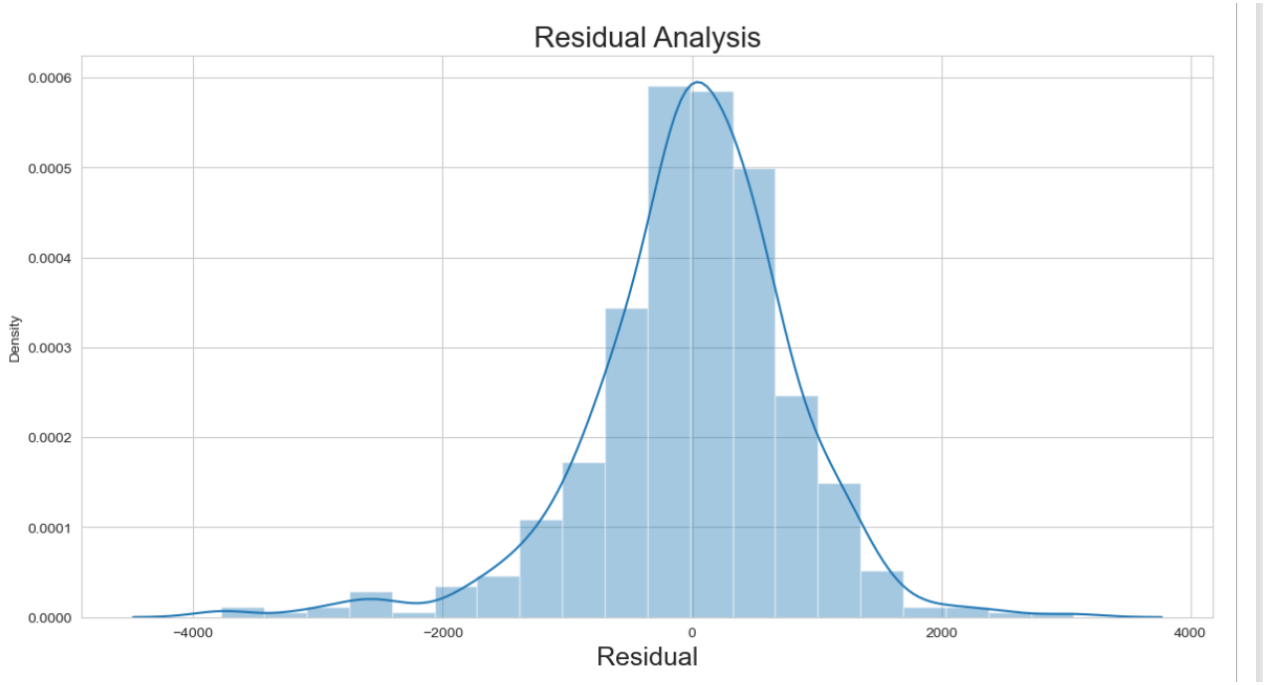
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?



Temp and **Atemp** variable have the Highest correlation with the target variable

Assignment-based Subjective Questions

4. How did you validate the assumptions of Linear Regression after building the model on the training set?



In linear regression, the assumption of normally distributed residuals is crucial for valid hypothesis testing, confidence intervals, and model inference. In our case , we calculated the residual error , By plotting it , we found it to be normally distributed

Assignment-based Subjective Questions

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

atemp	4035.135845
windspeed	-1208.200603
season_spring	-901.326303
season_summer	224.988355
season_winter	492.005336
yr_2019	2058.407693
weekday_Mon	-263.063060
weekday_Sun	-466.628145
weathersit_Light Rain	-2412.902855
weathersit_Mist	-665.155583

The top 3 Features that contribute significantly are :

1) Atemp – High Ride as Feeling Temperature rises
2) Yr_2019 – High Count of Rides As the year increase
3) Weathersit_Light Rain – Low Rides With Light Rain

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a supervised machine learning algorithm used to predict a continuous numerical outcome based on one or more input features. It assumes that there is a straight-line relationship between the input variables and the target variable. The main objective of linear regression is to find the best line that fits the data points and minimizes the difference between the predicted and actual values of the target variable.

Linear Regression is only possible when there is a linear relation found between different variables (also referred as features) , if there is no linear relation is visible on a plotted scatter chart of different variables , the model will not work as expected and the accuracy of model will get down drastically. Linear regression assumes linearity, independence of errors, constant variance (homoscedasticity), normality of errors, and absence of multicollinearity between independent variables. These assumptions should be verified and validated.

Linear regression equation is referred as Y = β0 + β1X1 + β2 x X2 + β3 x X3 ... βn x βn where n is the number of independent variable. This equation is basically the best fit line equation that has the least distance between actual and predicted values of the target variable. While establishing , this equation and during the process of Feature selection , it is very important to make sure that the selected features for the model creation should not have multicollinearity issue between them because if they have multicollinearity , the weightage of similar feature will be taken into account making our model accuracy not as expected.

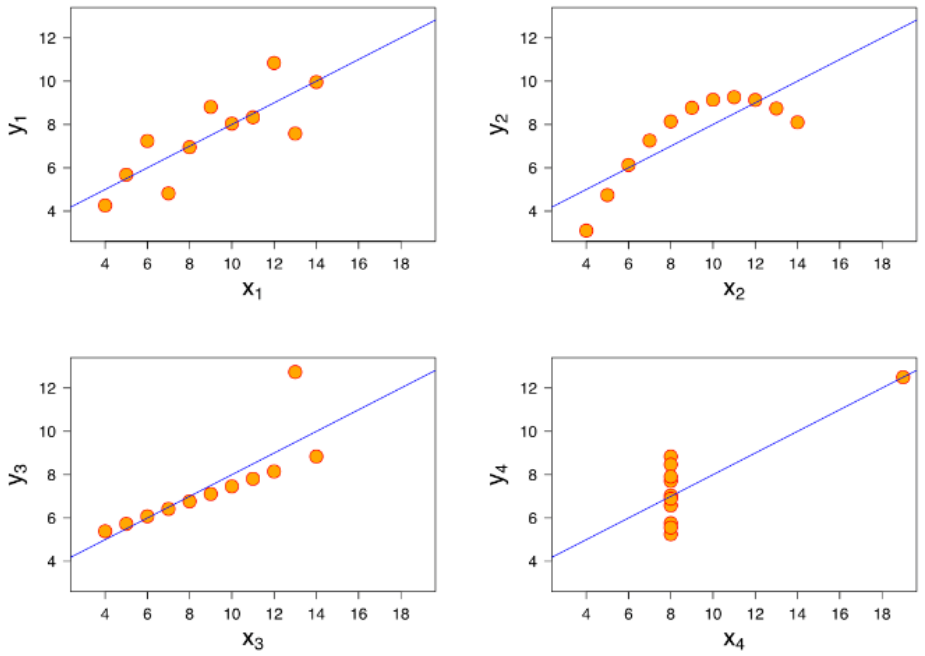
In simple linear regression, we have only one input feature (X), while in multiple linear regression, we can have multiple input features (X1, X2, X3, ...).
The linear regression model assumes that the relationship between the input features and the target variable can be represented by an equation.
For example, in simple linear regression, the equation is y = b0 + b1 * X, where b0 is the starting point (y-intercept), b1 is the slope, and X represents the input feature.
For multiple linear regression, the equation becomes y = b0 + b1 * X1 + b2 * X2 + ... + bn * Xn, where each X represents a different input feature.

During model training, we try to find the best values for the coefficients (b0, b1, b2, ..., bn) in the equation. These coefficients determine the shape and position of the line. We use an optimization algorithm, such as Ordinary Least Squares (OLS), to minimize the difference between the predicted and actual values of the target variable.

After training the model, we evaluate its performance. We use evaluation metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared to measure how well the model predicts the target variable. These metrics help us understand the accuracy of the predictions and how well the model fits the data.

General Subjective Questions

2. Explain the Anscombe’s quartet in detail.



In the figure above, lets explore what actually data holds and why these charts are drawn in different ways for a similar dataset :
X1 : Dataset have linear relationship between x and y
X2 : Dataset do not have linear relationship.
X3 : Dataset have linear relation but few points significantly deviates from the pattern including outliers.
X4 : Dataset have similar values for x and y but an outlier is present that impact the best fit line.

Anscombe’s quartet explain the importance of data visualization.

In 1973, a paper was written by a statistician named Anscombe, emphasizing the importance of graphing data rather than depending solely on statistical analysis. Four sets of XY data pairs were created, Each dataset consists of 11 points with two variables: x and y , which are nearly identical in simple descriptive statistics. It was demonstrated that graphical representation plays a crucial role in comprehending and interpreting data effectively, as shown by Anscombe's work.

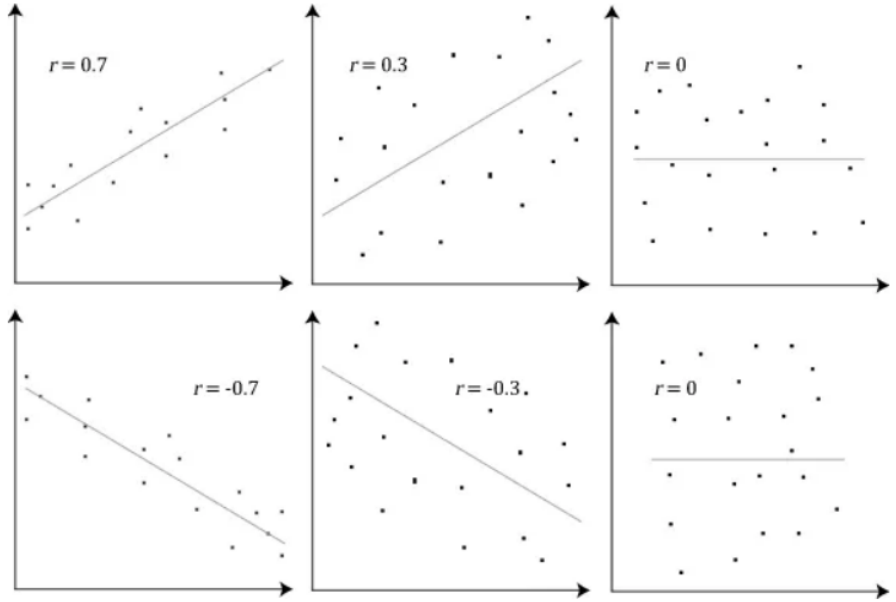
While choosing linear regression model , one needs to take special attention in identifying outliers and before selection , its extremely important to plot the data and get insights of it as to how data is being drawn and whether it is actually showing linearity or not. Like, in the above graphs plotted , if we look at it , the data plotted has identical summary statistics (such as mean, variance, correlation) but the datasets have different distributions and relationships between the variables.

In Summary Anscombe’s quartet emphasizes to do exploratory data analysis before finalizing the model selection. It also makes us understand how outliers can impact the entire modelling exercise. One need to take special steps to understand such data points and take necessary corrective actions to fix them. I would summarize this with below lines to keep this in my mind always :

The combination of Statistical Analysis and Visual Analysis is necessary to understand patterns effectively. These two pillars form the foundation of a robust Exploratory Data Analysis (EDA) process. Utilizing them in the appropriate combination is crucial for achieving the desired outcomes.

General Subjective Questions

3. What is Pearson's R?



An example of how the Pearson correlation coefficient (r) varies with the **strength and the direction of the relationship** between the two variables. Note that when no linear relationship could be established (refer to third (top right) graph), the Pearson coefficient yields a value of zero.

Pearson's R, which is also referred to as Pearson's correlation coefficient or simply correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two variables. It was developed by Karl Pearson and finds extensive usage in statistics and data analysis.

The range of Pearson's R is -1 to +1, with the sign indicating the relationship's direction: positive values denote a positive or direct relationship, while negative values indicate a negative or inverse relationship. A value of 0 signifies the absence of a linear relationship between the variables.

The magnitude of Pearson's R signifies the strength of the relationship, with values near +1 or -1 representing a strong linear relationship, and values near 0 indicating a weak relationship.

Pearson's R is sensitive to outliers and assumes linearity in the relationship between the variables. In cases where the relationship is nonlinear, Pearson's R may not accurately represent the true association between the variables.

When building a linear regression model, Pearson's R can assist in identifying relevant predictor variables. Variables with a high correlation (positive or negative) with the response variable are more likely to have a significant impact on the model's predictive power.

General Subjective Questions

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

While preparing data for model building , a lot of times we see that there are variables with different units for example there could be a column that denotes temperature and an another unit that denotes windspeed. For machine , both these column need to be understood under a common scale or else it would give higher weightage to column that have higher value units. To avoid this problem , we use a technique called scaling , there are two types of scaling techniques widely used in the field of machine learning :

Normalized scaling (or feature scaling): Normalization transforms the values of variables to a common range, typically between 0 and 1. It is achieved by subtracting the minimum value of the variable and dividing by the range (maximum value minus minimum value). Normalization preserves the shape of the original distribution and is suitable when the distribution of the variable is skewed or does not follow a normal distribution.

Standardized scaling (or z-score normalization): Standardization transforms the values of variables to have a mean of 0 and a standard deviation of 1. It is achieved by subtracting the mean value of the variable and dividing by the standard deviation. Standardization centers the variable distribution around zero and rescales it using the standard deviation. Standardization is useful when the variable distribution is approximately normal and when the algorithm or analysis assumes variables to be normally distributed.

Lets understand this with example ,
Assume you have below dataset for temperature values & Windspeed :

Temperature: [25.6, 22.3, 28.9, 26.5, 20.1, 24.7, 21.8, 23.2, 27.4, 19.8]

Windspeed: [8.3, 6.5, 7.1, 9.2, 5.7, 6.9, 7.8, 6.2, 9.7, 5.1]

Here is the Result after Normalized Scaling

Temperature : [0.59, 0.34, 0.86, 0.66, 0.0, 0.52, 0.21, 0.41, 0.76, 0.03]

Windspeed (Normalized): [0.71, 0.29, 0.41, 0.88, 0.06, 0.35, 0.59, 0.24, 1.0, 0.0]

Notice that all numbers lie between 0 and 1

Here is the Result after Standardized Scaling

Temperature (Standardized : [0.54, -0.59, 1.66, 0.84, -1.34, 0.23, -0.76, -0.28, 1.15, -1.44]

Windspeed (Standardized): [0.74, -0.53, -0.11, 1.38, -1.09, -0.25, 0.39, -0.74, 1.73, -1.52]

Notice that mean and standard deviation is 0 and 1 respectively

General Subjective Questions

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

VIF stands for Variance Inflation Factor. It is a statistical measure used to assess the severity of multicollinearity in regression analysis. Multicollinearity refers to the presence of high correlation or interdependence among predictor variables in a regression model. In simple words, it calculates correlation of the variable with all other available variables that are being considered for the model creation and give result as the numerical value for the given variable, the same process is repeated for the rest of the variables. Statistically , The VIF value for a particular predictor variable is calculated as the ratio of the variance of the estimated coefficient for that variable to the variance of the coefficient when that variable is not related to the other predictor variables.

Before doing the calculation for VIF , we should do a correlation study during the Exploratory Data Analysis (EDA) process by plotting a correlation plot , where we can see the correlation values of each variable to corresponding other variables. Based on these values and business understanding , we should identify the variable that are highly correlated to avoid multicollinearity between the features. If not done correctly , the VIF value will than return the values that can further guide you the features that are having issues of multicollinearity.

But when two or more feature are highly correlated to each other which can also be referred as perfect multicollinearity , than the VIF value will be returning infinite as result. It could be possible to have such result for features that are derived from an existing feature apart from the linear dependency between the variable.

Suppose we have a regression model with two predictor variables: X1, X2. If X1 and X2 are perfectly correlated, meaning they have a perfect linear relationship (e.g., X2 = 2*X1), then we have perfect multicollinearity. In this case, the correlation coefficient (R) between X1 and X2 is 1.

Now, when calculating the VIF for X1, the formula is:
VIF(X1) = 1 / (1 - R(X1)^2)
Substituting the value of R(X1) as 1, we get:
VIF(X1) = 1 / (1 - 1^2) VIF(X1) = 1 / (1 - 1) VIF(X1) = 1 / 0
Since division by zero is undefined, the VIF for X1 becomes infinite.

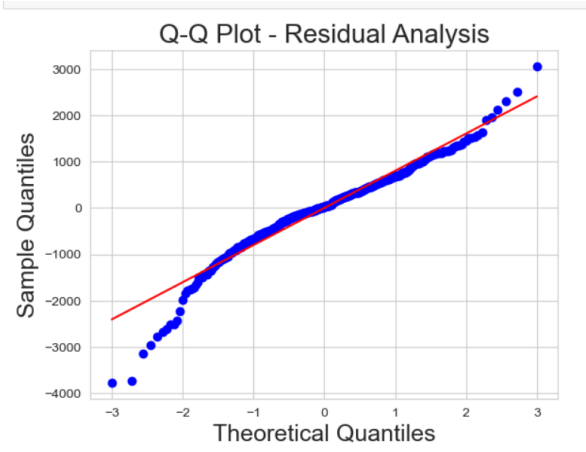
Thus , VIF will be infinite for cases where there is perfect multicollinearity and to avoid this we need to always look into correlation metrics during EDA process and eliminates such duplicate features and keep one out of them to have an efficient model.

General Subjective Questions

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

When we do analysis, we normally look at the distribution of the complete dataset to understand data while doing data exploration. But a lot of times , subset of the data and complete dataset do not show same distribution. To understand this , Q-Q plot comes into play . Q-Q plot as the name suggest is a graphical tool which is used to assess the distributional similarity between a sample of data and a theoretical distribution, such as the normal distribution. It is commonly used in statistics to visually examine whether the data follows a particular distribution or to check the assumption of normality in linear regression.

The Q-Q plot compares the quantiles of the observed data against the quantiles expected from a specific theoretical distribution. In the case of linear regression, the Q-Q plot is used to assess the assumption of normality for the residuals (the differences between the observed and predicted values).



A chart from the submitted case study

With this qqplot, we can make the observation as follows :

- The data point tend to follow a relatively linear pattern on the plot indicating that there is a normal distribution.
- Most of the datapoint fall along the expected diagonal line suggesting that they are consistent with normal distribution.
- There are also few points that deviates from the diagonal line particularly in the upper and lower tails of distribution. These points represent potential outlier or data points that do not conform to a normal distribution.

A Q-Q plot is a valuable tool in linear regression for assessing the assumption of normality in the residuals. By comparing the observed residuals to the expected quantiles of a normal distribution, the plot helps identify departures from normality, such as skewness or outliers. Deviations from the straight diagonal line indicate non-normality, which can affect the validity of statistical inferences. The Q-Q plot is used for model assessment and diagnostics, verifying the assumption of normally distributed residuals, and ensuring reliable estimation of regression coefficients. It aids in decision-making by providing evidence for the suitability of linear regression and supporting inference and prediction based on the model.

upGrad

iit-b
ज्ञानमुत्तमम

Thank you