Student Name: Gaurav Kalra
IIITB Roll Number: EML22010008
Assignment Name: Linear Regression Assignment
Date of Submission: April 13, 2022

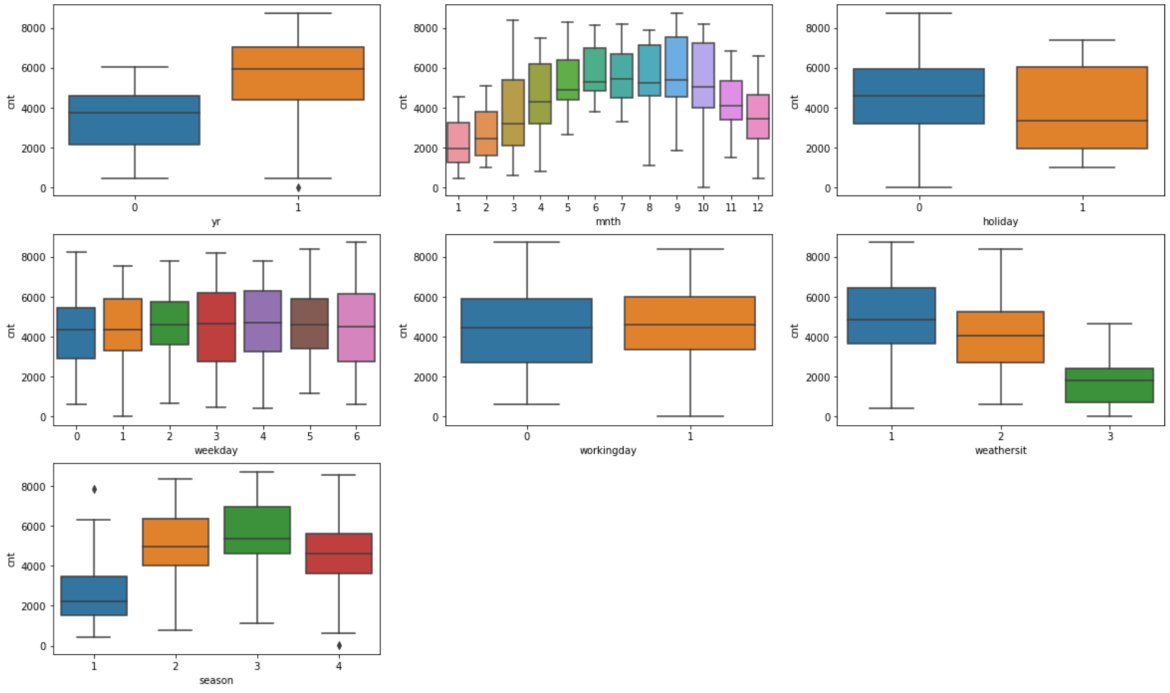
Answer to

Assignment-based Subjective Questions

Question-1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer: The impact of categorical varialbes on target varialbe cnt can be analyzed with the help of box plots.
In the following image, we can clearly see that some values of the categorical variables have higher impact on count than others.

Year: 2019 has signifincantly higher value of cnt than 2018.
Months: June throuh October the cnt is signifincantly higher.
Weather Situation: Clear Weather has significantly higher impact on cnt.
Season: Summer and Fall have higher people riding the bike.



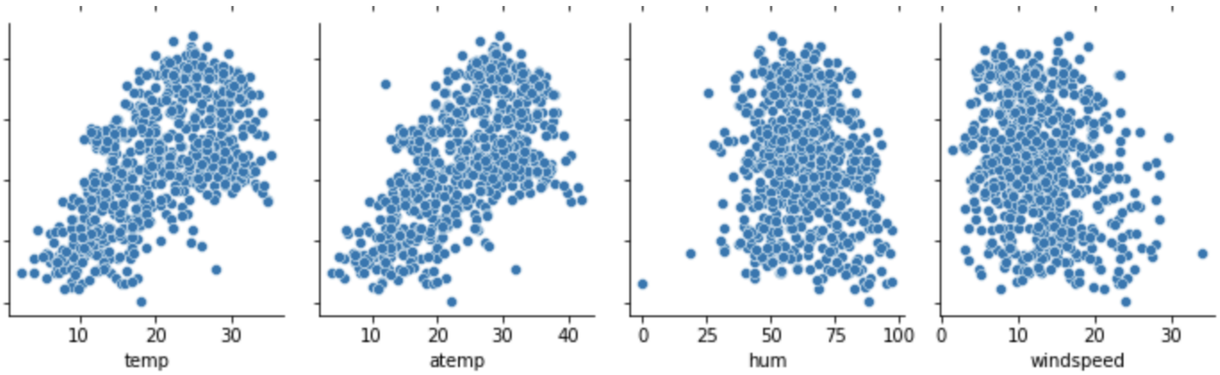
Since Categorical variables show impact on target variable, we should create dummy variables from it to see the impact on the values of categorical variables on target.

Question-2: Why is it important to use drop_first=True during dummy variable creation?

Answer: It is important to use drop_first=True as dropping one dummy variable helps in reducing the extra column created during dummy variable creation.
Therefore, it reduces the correlations created among dummy variables.

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

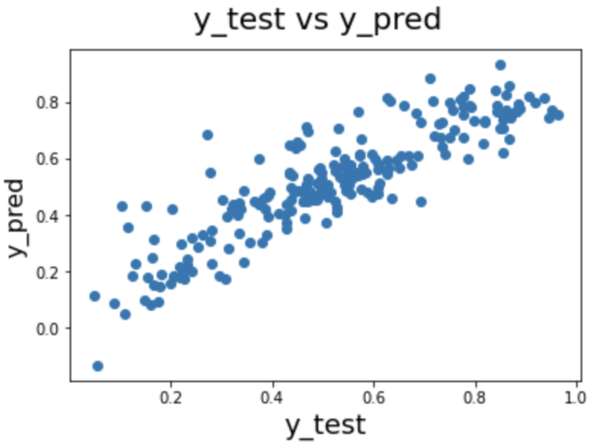
Answer: Out of all the numeric variables, **temp** and **atemp** seem to have highest correlation with cnt. Since **temp** and **atemp** represent similar information with few degrees difference, we can drop **atemp** and just use the **temp**.



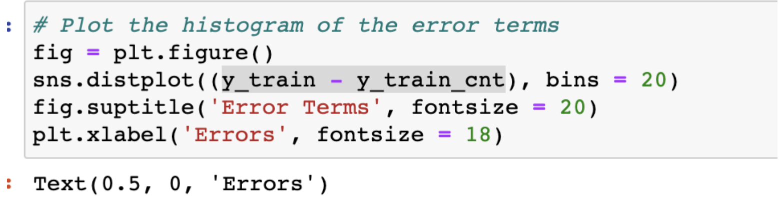
Question-4: How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer: Following are the Assumptions of Linear Regression with details of verification:

a. **Linearity**: There should be linear relationship between dependent and independent variables. To check linearity plot was created between observed vs. predicted values or residuals vs. predicted values. The desired outcome is that points are symmetrically distributed around a diagonal line in the plot.

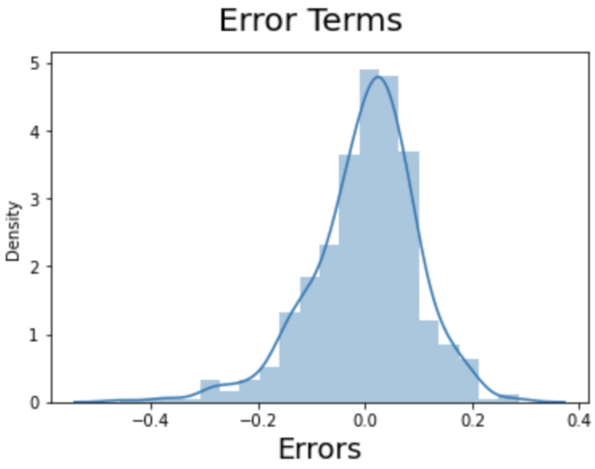


b. **Error Terms are Normally Distributed with mean 0.**
To ensure that error terms are normally distributed, histogram was drawn based on the residuals (y_train - y_train_cnt).



```
# Plot the histogram of the error terms
fig = plt.figure()
sns.distplot((y_train - y_train_cnt), bins = 20)
fig.suptitle('Error Terms', fontsize = 20)
plt.xlabel('Errors', fontsize = 18)
```

```
Text(0.5, 0, 'Errors')
```

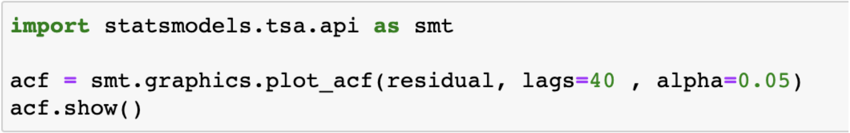


Mean of Residuals

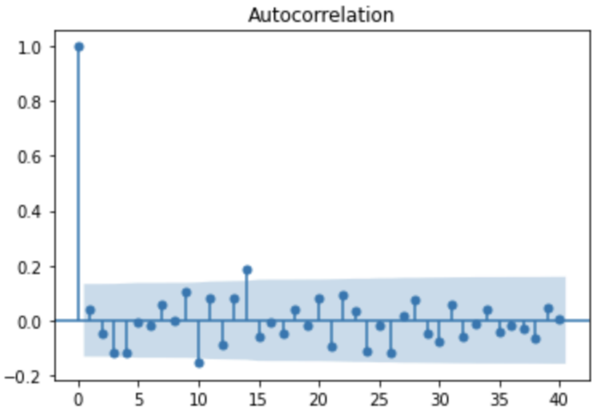
The rounded value of mean of Residuals was calculated to be 0. It is also evident fom the graph that mean is centering towards zero.

c. Error terms are independent of each other - No Autocorrelation

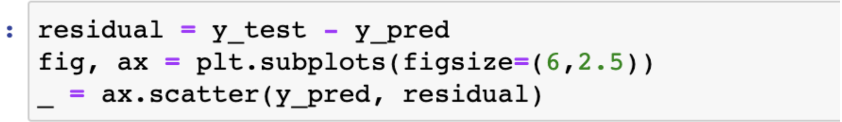
This was verified by plotting residual.



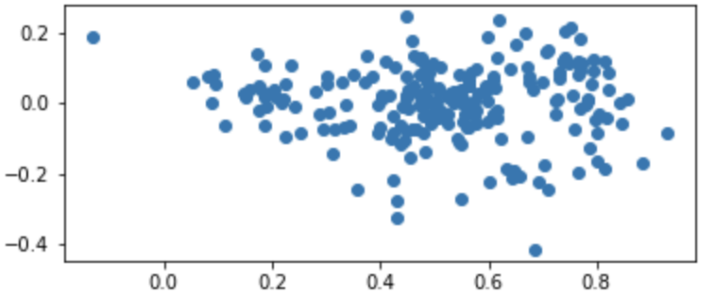
```
import statsmodels.tsa.api as smt

acf = smt.graphics.plot_acf(residual, lags=40 , alpha=0.05)
acf.show()
```



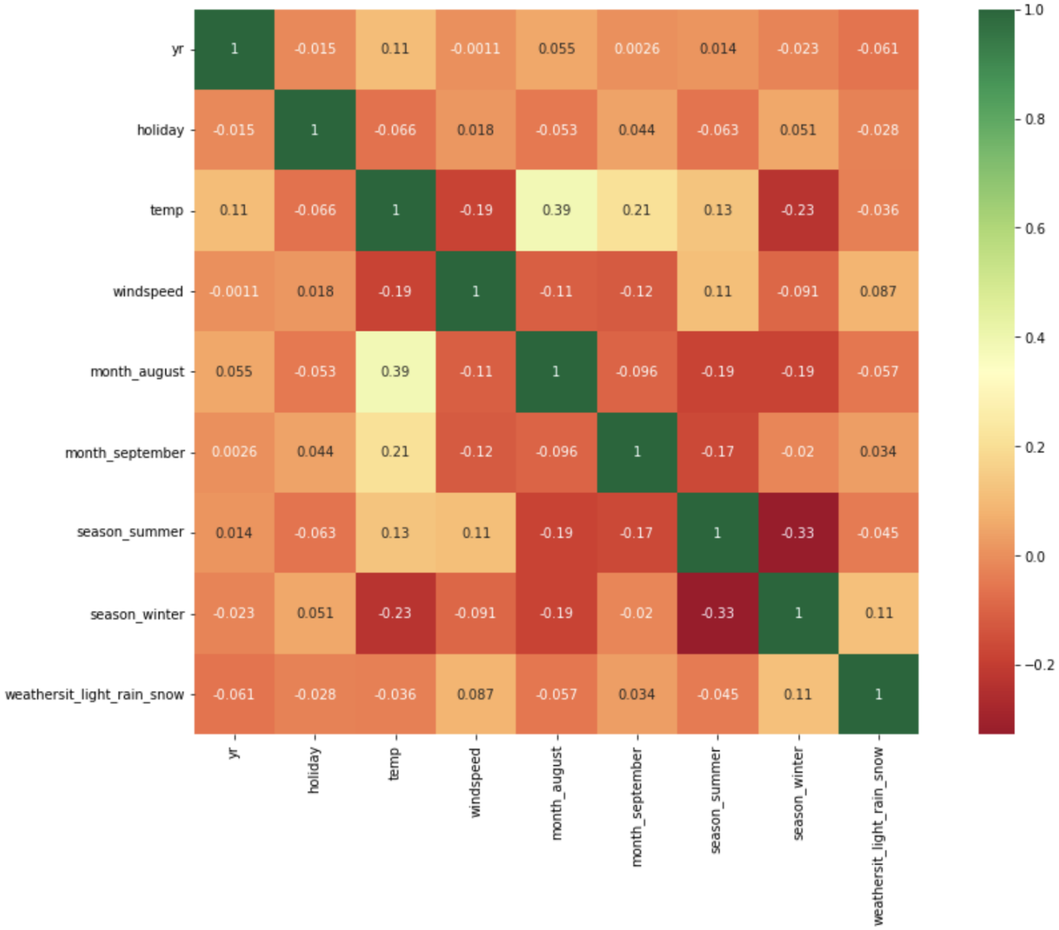
d. Error terms have constant variance



```
residual = y_test - y_pred
fig, ax = plt.subplots(figsize=(6,2.5))
_ = ax.scatter(y_pred, residual)
```

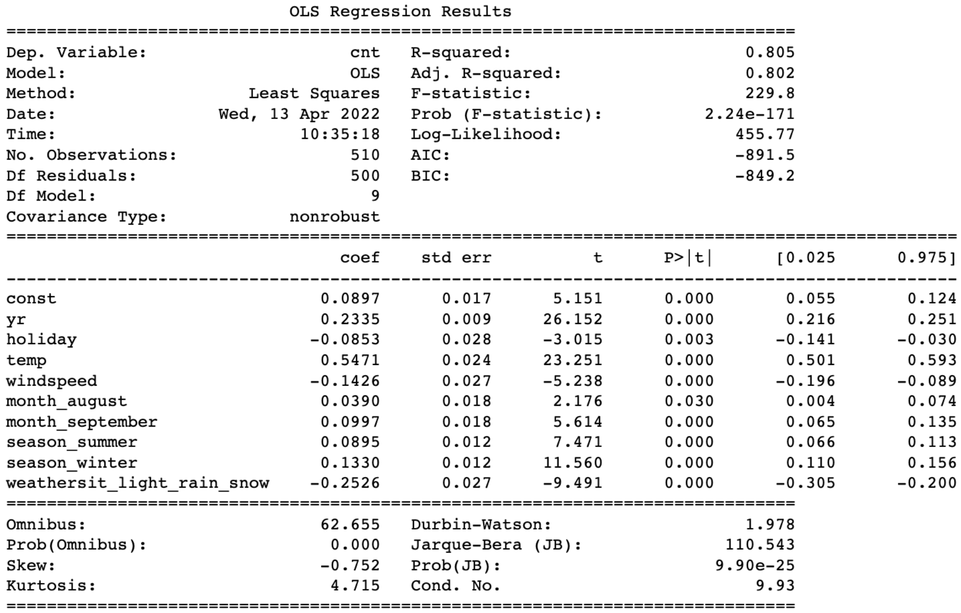


e. **No perfect multicollinearity** **-** The independent variables are not related to each other. This is also evident from following heat graph made from final list of indpendent variables.



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

Answer: **temp**, **weathersit**, and **yr** are the top most features that significantly impact the demand of shared bikes.
For quick reference attached is the summary of OLS Regression Results:



```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.805
Model:                            OLS   Adj. R-squared:                  0.802
Method:                 Least Squares   F-statistic:                     229.8
Date:                Wed, 13 Apr 2022   Prob (F-statistic):          2.24e-171
Time:                        10:35:18   Log-Likelihood:                 455.77
No. Observations:                 510   AIC:                            -891.5
Df Residuals:                     500   BIC:                            -849.2
Df Model:                           9                                         
Covariance Type:            nonrobust                                         
==============================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                          0.0897      0.017      5.151      0.000       0.055       0.124
yr                             0.2335      0.009     26.152      0.000       0.216       0.251
holiday                       -0.0853      0.028     -3.015      0.003      -0.141      -0.030
temp                           0.5471      0.024     23.251      0.000       0.501       0.593
windspeed                     -0.1426      0.027     -5.238      0.000      -0.196      -0.089
month_august                   0.0390      0.018      2.176      0.030       0.004       0.074
month_september                0.0997      0.018      5.614      0.000       0.065       0.135
season_summer                  0.0895      0.012      7.471      0.000       0.066       0.113
season_winter                  0.1330      0.012     11.560      0.000       0.110       0.156
weathersit_light_rain_snow    -0.2526      0.027     -9.491      0.000      -0.305      -0.200
==============================================================================
Omnibus:                       62.655   Durbin-Watson:                   1.978
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              110.543
Skew:                          -0.752   Prob(JB):                     9.90e-25
Kurtosis:                       4.715   Cond. No.                         9.93
==============================================================================
```

General Subjective Questions

Question-1 Explain the linear regression algorithm in detail.
Answer: In Modeling, machines learn from data. A machine learning model is a file that has been trained to recognize certain types of patterns.
Regression is a supervised learning method and in this model, the target variable is a continuous variable.
Linear Regression is a form of predictive modelling technique which explains the relationship between the dependent (target variable) and independent variables (predictor variables).. There are two types of linear regressions

a. Simple linear regression - Model with only one independent Vaiable
b. Multiple linear regression - Model with multile independent Vaiables

Simple Linear Regression
This most fundamental type of regression model is the simple linear regression which explains the relationship between a dependent variable and the target variable using a straight line. A simple linear regression model attempts to explain the relationship between a dependent and an independent variable using a straight line.

The independent variable is also known as the predictor variable. And the dependent variables are also known as the output variables.

The standard equation of the regression line is: $Y = \beta_0 + \beta_1X$
β_0 is intercept and β_1 is slope.

Regression Line:
The Gradient Calculations are done using OLS Method (Ordinary Least Square).
There can be many regression lines but we need the optimized line to predict the values. In regression, there is a notion of a best-fit line — the line which fits the given scatter-plot in the best way. This is done using some mathematical techniques.

Strength of Simple Linear Regression
The best-fit line of Linear Regression is found by minimising the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot.

Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable:

The strength of the linear regression model can be assessed using 2 metrics:
1. R^2 or Coefficient of Determination
2. Residual Standard Error (RSE)

R^2 is a number which explains what portion of the given data variation is explained by the regression model. It always takes a value between 0 & 1.

Mathematically, it is represented as: $R^2 = 1 - (RSS / TSS)$

RSS (Residual Sum of Squares): In statistics, it is defined as the total sum of error across the whole sample. It is the measure of the difference between the expected and the actual output. A small RSS indicates a tight fit of the model to the data.

TSS (Total sum of squares): It is the sum of errors of the data points from mean of response variable.

Physical Significance of R^2

R^2 =1 is the best fit and as R^2 approaches 0 we have more scattered Plot.

Assumptions of Linear Regression:

1. Linear Relation Between X and Y
2. Error Terms are normally distributed
3. Error Terms are independent of each other
4. Error terms have constant variance.

Multiple Linear Regression

It represents the relationship between two or more independent input variables and a response variable. Multiple linear regression is needed when one variable might not be sufficient to create a good model and make accurate predictions.

Impact on R-square

When you add more variables in the regression model,
The R-squared will always either increase or remain the same when you add more variables. Because you already have the predictive power of the previous variable so the R-squared value can definitely not go down. And a new variable, no matter how insignificant it might be, cannot decrease the value of R-squared.

Formulation

Most of the concepts in multiple linear regression are quite similar to those in simple linear regression. The formulation for predicting the response variable now becomes:

$Y=\beta0+\beta1X1+\beta2X2+...+\beta pXp+\epsilon$.

Comparison

Apart from the formulation, there are some other aspects that still remain the same:

a. The model now fits a hyperplane instead of a line

b. Coefficients are still obtained by minimising the sum of squared errors, the least squares criteria.
c. For inference, the assumptions from simple linear regression still hold - zero-mean, independent and normally distributed error terms with constant variance.

Moving from SLR to MLR: New Considerations

1. **Overfitting**: Overfitting is the condition wherein the model is so complex that it ends up memorising almost all the data points on the train set. Hence, this condition is more probable if the number of data points is less since the model passing through almost every point becomes easier

2. **Multicollinearity**: A phenomenon where adding multiple independent variables to the model can sometimes bring about dependency within themselves or cause redundence. In simple terms, in a model which has been built using several independent variables, some of these variables might be interrelated, due to which the presence of that variable in the model is redundant. You drop some of these related independent variables as a way of dealing with multicollinearity.
   Multicollinearity can be detected by Checking the Variance Inflation Factor (**VIF**). **VIF** calculates how well one independent variable is explained by all the other independent variables combined excluding the trget variable.

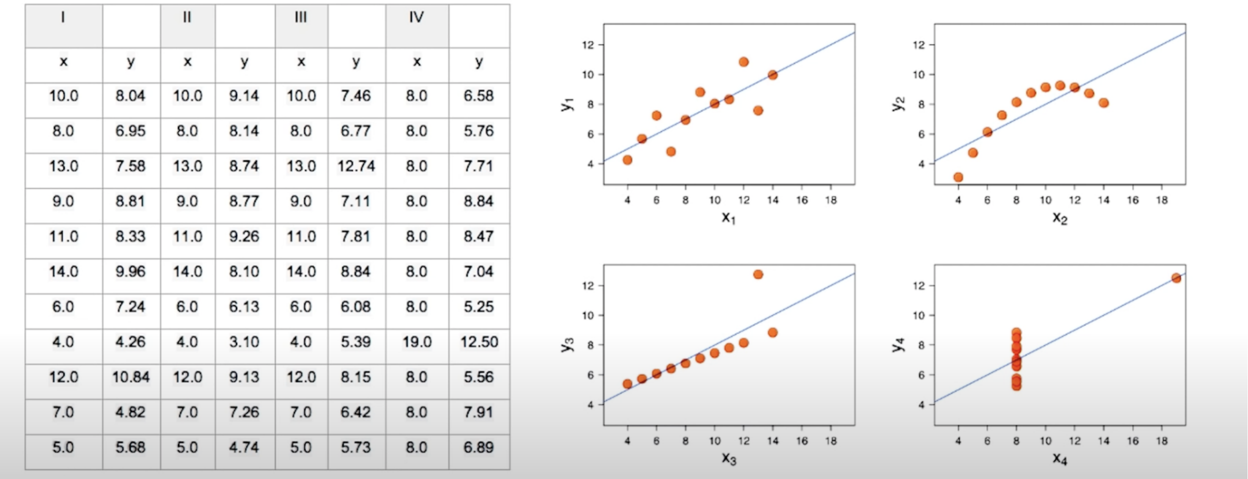
3. **Feature Selection:** Selecting the optimal set from a pool of given features, many of which might be redundant becomes an important task. Feature Selection is done using Recursive Feature Elimination RFE Technique. Recursive feature elimination is based on the idea of repeatedly constructing a model (for example, an SVM or a regression model) and choosing either the best or worst performing feature (for example, based on coefficients), setting the feature aside and then repeating the process with the rest of the features. This process is applied until all the features in the dataset are exhausted. Features are then ranked according to when they were eliminated. As such, it is a greedy optimisation for finding the best performing subset of features.

Dealing with Categorical Variables - Categorical variables are analyzed by creating dummy variables. For n values of a categorical variable n-1 columns are created.

Scaling: Helps with interpretation and for Faster convergence of gradient descent. It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Model Assessment and Comparison in MLR
Adjusted R-squared - The adjusted R-squared value increases only if the new term improves the model more than would be expected by chance.

Question-2 Explain the Anscombe’s quartet in detail.
Answer: Anscombe’s quartet highlights the importance of plotting data instead of just relying on the major statistical summaries.
Example consists of 4 datasets that have alomost similar simple statistical properties, still appear very different when graphed.
Every dataset consists of eleven (x,y) points. These were constructed by the statistician Francis Anscombe in 1973 to showcase that both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties.



I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

All above data points have almost identical

1. mean of x
2. sample variance of x
3. mean of y
4. sample variance of y
5. Correlation between x and y
6. Linear regression line
7. Coefficient of determination of the linear regression

Yet when they are plotted above have totally different spread.

Question-3 What is Pearson’s R?
Answer:
Correlation is a measure of strength of association between two variables as well as the direction.

One significant type correlation measurement is **Pearson's correlation coefficient**. This type of correlation measurement type is used to measure the relationship between two continuous variables.
Pearson's Correlation Coefficient is also referred to as **Pearson's r**, the **Pearson product-moment correlation coefficient** (PPMCC), or **bivariate correlation.**

Note: Pearson’s r cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

How it is calculated:
Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.

There are certain requirements for Pearson's Correlation Coefficient:

1. Scale of measurement should be interval or ratio
2. Variables should be approximately normally distributed
3. The association should be linear
4. There should be no outliers in the data

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

Steps:

1. Make a Pearson correlation coefficient table. Make a data chart using the two variables and name them as X and Y. Add three additional columns for the values of XY, X^2, and Y^2.
2. Use basic multiplications to complete the table
3. Add up all the columns from bottom to top.
4. Use these values in the formula to obtain the value of r.

Strength of the Pearson product-moment correlation coefficient
The more inclined the value of the Pearson correlation coefficient to -1 and 1, the stronger the association between the two variables.

Question-4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Answer:
Feature Scaling is a technique to standardize the independent variables present in the data. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a ML algorithm tends to consider greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

When we have a lot of independent variables in any model, few of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. Also, in the machine learning algorithms if the values of the features are closer to each other there are chances for the algorithm to get trained well and faster instead of the data set where the data points or features values have high differences with each other will take more time to understand the data and the accuracy will be lower

We need to scale features because of two reasons:

1. Ease of interpretation
2. Faster convergence for gradient descent methods

We can scale the features using two very popular method:

Standardizing: The variables should be scaled in a way that the mean is 0 and the standard deviation is 1. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

Normalized Scaling: The variables are scaled in a way that all the values fall between 0 and 1 using the maximum and the minimum values in the data.

Differences:
Normalized scaling is really affected by outliers. Whereas the standardizing scaling is less affected.
Normalized scaling is useful when we don’t know about the distribution. Standardizing scaling is useful when the feature distribution is Normal or Gaussian.

Question-5: You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Answer: An infinite VIF value indicates that the variable under consideration may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well). So If there is perfect correlation, then VIF = infinity.

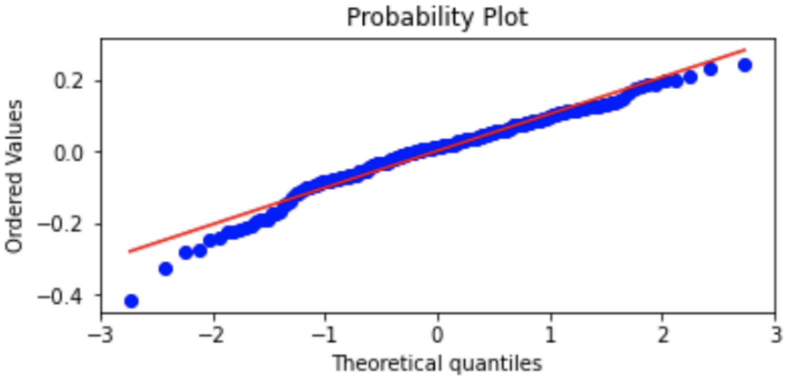
Mathematical Explaination
In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.

To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

Question-6 What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Answer: The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

Q-Q Plot for Linear Regression:
Q-Q Plot helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Example of QQ Plot:



While building machine learning model, we need to check the distribution of the error terms or prediction error using a Q-Q plot. If there is a significant deviation from the mean, we might have to check the distribution of the feature variables and consider transforming them into a normal shape.