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?

Ans: In snowy climate target variable is negatively correlated
In September month target variable is positively correlated.

	cnt	yr	holiday	workingday	temp	windspeed	Misty	Snowy	summer	winter	saturday	september
cnt	1.000000	0.573177	-0.118934	0.109079	0.648869	-0.247968	-0.153314	-0.216487	0.146895	0.065160	-0.009970	0.207293

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

Ans: drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

- Let's say we have 4 types of values in Categorical column and we want to create dummy variable for that column. If one variable is Misty ,Snowy and Sunny, then It is obvious Rainy. So we do not need 4th variable to identify the Rainy.

  Example:

	Misty	Snowy	Sunny
0	0	0	0
1	0	0	1
2	0	1	0
3	1	0	0

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Ans: atemp is highly correlated to cnt(Target variable)
After that temp,yr,spring,sunny

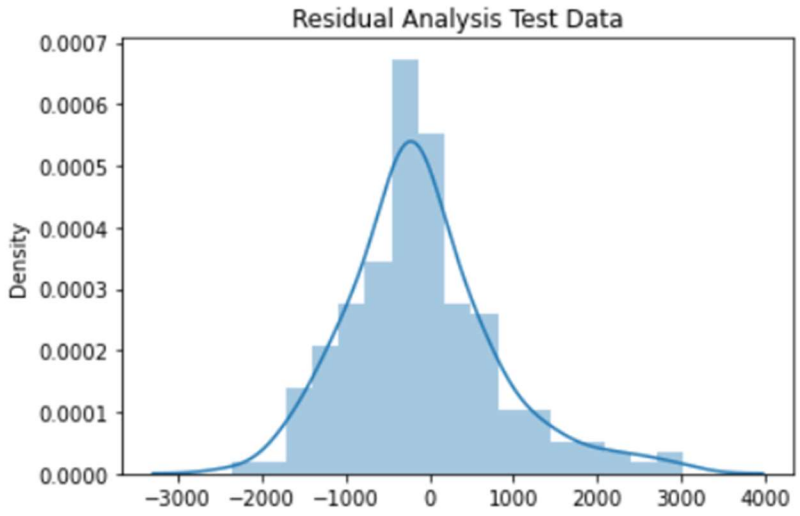
4. How did u Validate the assumptions of Linear regression after building the model on the training set?

Ans:

Assumptions of Linear regression:

1. Linear relation ship between X and Y
2. Error term are normally distributed with mean "0"

```
plt.title("Residual Analysis Test Data")
sns.distplot(y_test_rfe_pred-y_test_rfe)
plt.show()
```

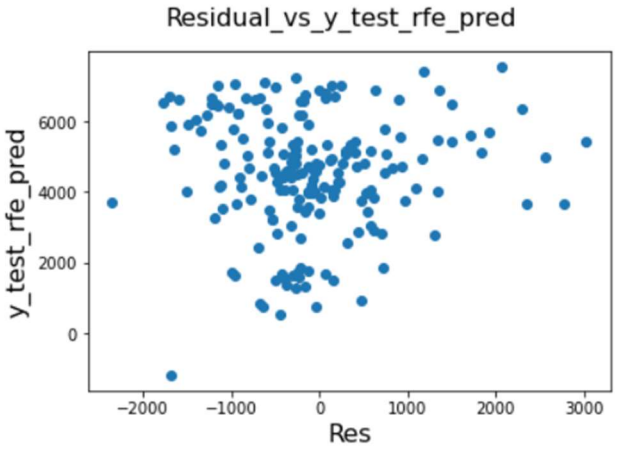


- error term shows normal distribution with mean "0"

3. Error terms are independent of each other (error plot should not follow any patterns)

```
# understanding the spread target variable
plt.scatter(y_test_rfe_pred-y_test_rfe,y_test_rfe_pred)
plt.suptitle('Residual_vs_y_test_rfe_pred', fontsize =16)
plt.xlabel('Res', fontsize=16)              # X-label
plt.ylabel('y_test_rfe_pred', fontsize=16)  # y-label
```

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

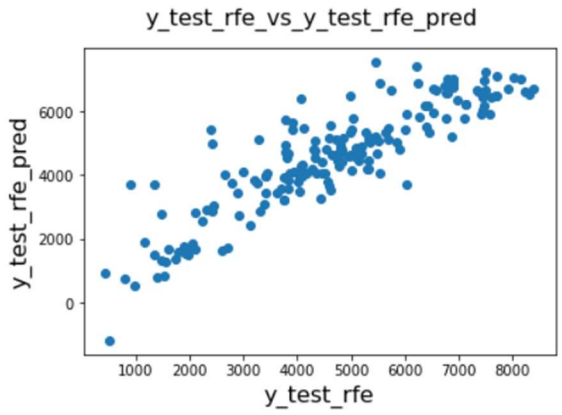


- error term is not following any specific patterns and spreaded

4. Error term have constant variance

```
# understanding the model fit
plt.scatter(y_test_rfe,y_test_rfe_pred)
plt.suptitle('y_test_rfe_vs_y_test_rfe_pred', fontsize =16)
plt.xlabel('y_test_rfe', fontsize=16)          # X-label
plt.ylabel('y_test_rfe_pred', fontsize=16)     # y-label
```

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



- this is to confirm that the model fit is not by chance it predicted decently and having conatnat variance

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

Ans:

Year , temp, windspeed variables contribute more in bike sharing

	cnt	yr	holiday	workingday	temp	windspeed	Misty	Snowy	summer	winter	saturday	september
cnt	1.000000	0.573177	-0.118934	0.109079	0.648869	-0.247968	-0.153314	-0.216487	0.146895	0.065160	-0.009970	0.207293

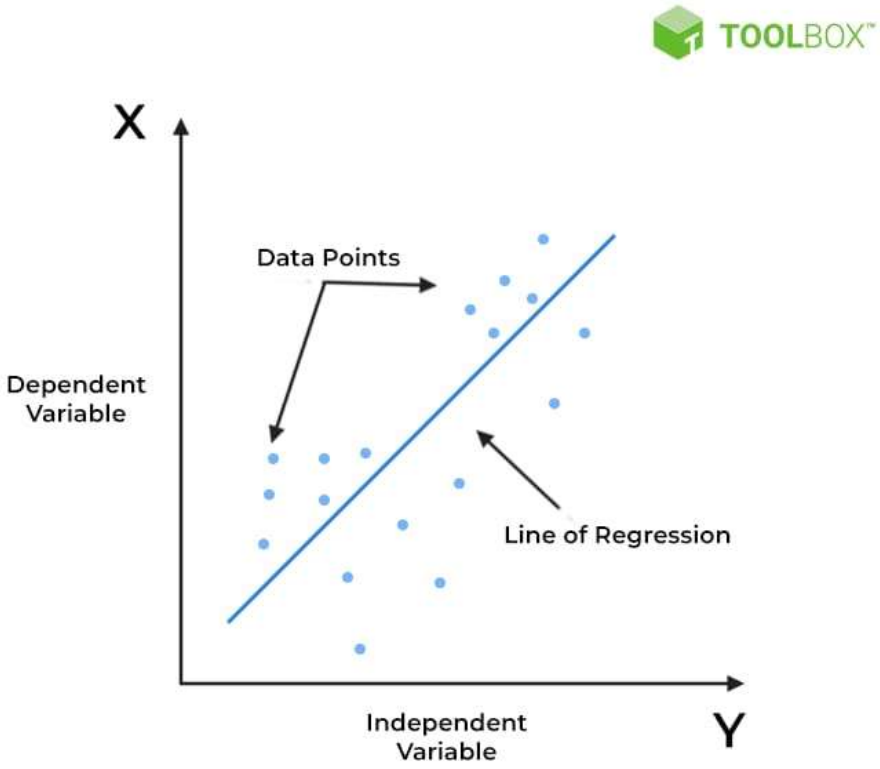
1.Explain the Linear Regression algorithm in detail?

Ans: Linear regression is an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. It is a statistical method used in data science and machine learning for predictive analysis.

The independent variable is also the predictor or explanatory variable that remains unchanged due to the change in other variables. However, the dependent variable changes with fluctuations in the independent variable. The regression model predicts the value of the dependent variable, which is the response or outcome variable being analyzed or studied.

Thus, linear regression is a supervised learning algorithm that simulates a mathematical relationship between variables and makes predictions for continuous or numeric variables such as sales, salary, age, product price, etc.

This analysis method is advantageous when at least two variables are available in the data, as observed in stock market forecasting, portfolio management



Best Fit Line for a Linear Regression Model

In the above figure,

X-axis = Independent variable

Y-axis = Output / dependent variable

Line of regression = Best fit line for a model

Here, a line is plotted for the given data points that suitably fit all the issues. Hence, it is called the 'best fit line.' The goal of the linear regression algorithm is to find this best fit line seen in the above figure.

Least Square Method – Finding the best fit line

Least squares is a statistical method used to determine the best fit line or the regression line by minimizing the sum of squares created by a mathematical function. The "square" here refers to squaring the distance between a data point and the regression line. The line with the minimum value of the sum of square is the best-fit regression line.

Regression Line**, y = mx+c** where,

y = Dependent Variable

x= Independent Variable ; c = y-Intercept

m= Slope of the fitted line (one of the co -efficeint)

c =intercept

2.Explain the Anscombe's quartet in detail?

Ans: Anscombe's Quartet – the importance of graphs

Anscombe's Quartet was devised by the statistician Francis Anscombe to illustrate how important it was to not just rely on statistical measures when analyzing data. To do this he created 4 data sets which would produce nearly identical statistical measures.

Statistical measures

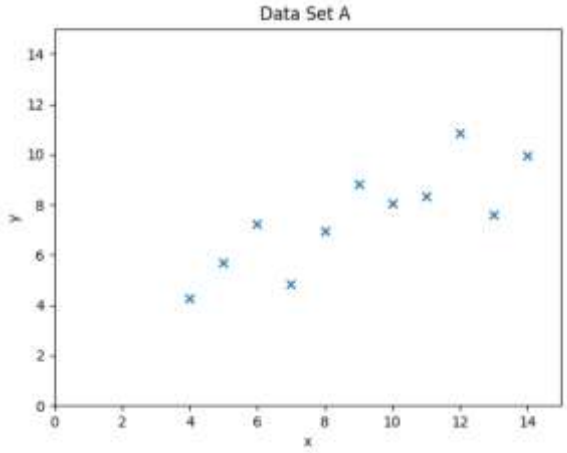
1) Mean of x values in each data set = 9.00
2) Standard deviation of x values in each data set = 3.32
3) Mean of y values in each data set = 7.50
4) Standard deviation of x values in each data set = 2.03
5) Pearson's Correlation coefficient for each paired data set = 0.82
6) Linear regression line for each paired data set: y = 0.500x + 3.00

When looking at this data we would be forgiven for concluding that these data sets must be very similar – but really they are quite different.

Data Set A:

x = [10,8,13,9,11,14,6,4,12,7,5]

y = [8.04, 6.95,7.58,8.81,8.33, 9.96,7.24,4.26,10.84,4.82,5.68]

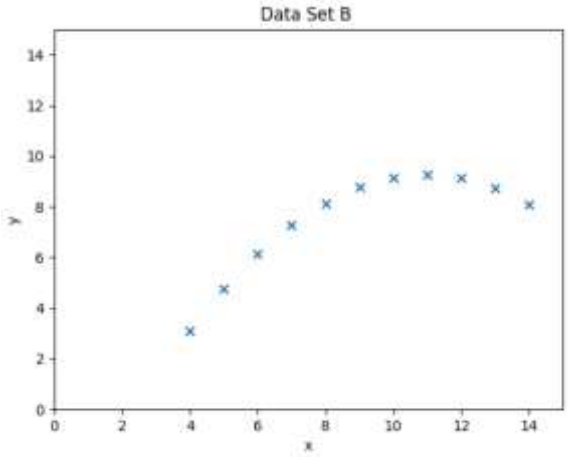


Data Set A does indeed fit a linear regression – and so this would be appropriate to use the line of best fit for predictive purposes.

Data Set B:

x = [10,8,13,9,11,14,6,4,12,7,5]

y = [9.14,8.14,8.74,8.77,9.26,8.1,6.13,3.1,9.13,7.26,4.74]

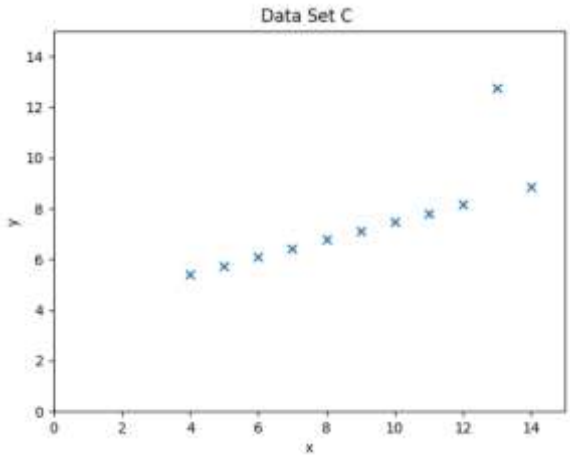


You could fit a linear regression to Data Set B – but this is clearly not the most appropriate regression line for this data. Some quadratic or higher power polynomial would be better for predicting data here.

Data Set C:

x = [10,8,13,9,11,14,6,4,12,7,5]

y = [7.46,6.77,12.74,7.11,7.81,8.84,6.08,5.39,8.15,6.42,5.73]

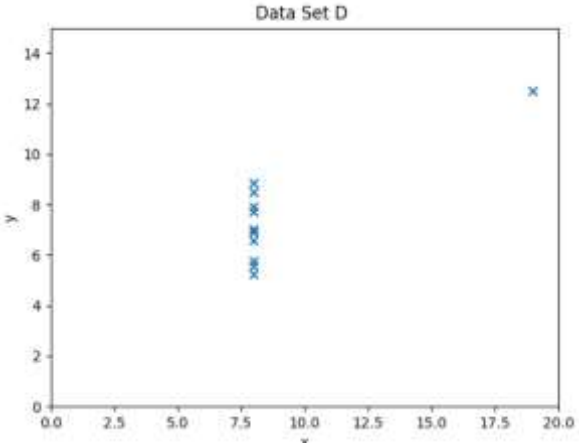


In Data set C we can see the effect of a single outlier – we have 11 points in pretty much a perfect linear correlation, and then a single outlier. For predictive purposes we would be best investigating this outlier (checking that it does conform to the mathematical definition of an outlier), and then potentially doing our regression with this removed.

Data Set D:

x = [8,8,8,8,8,8,8,19,8,8,8]

y = [6.58,5.76,7.71,8.84,8.47,7.04,5.25,12.50,5.56,7.91,6.89]



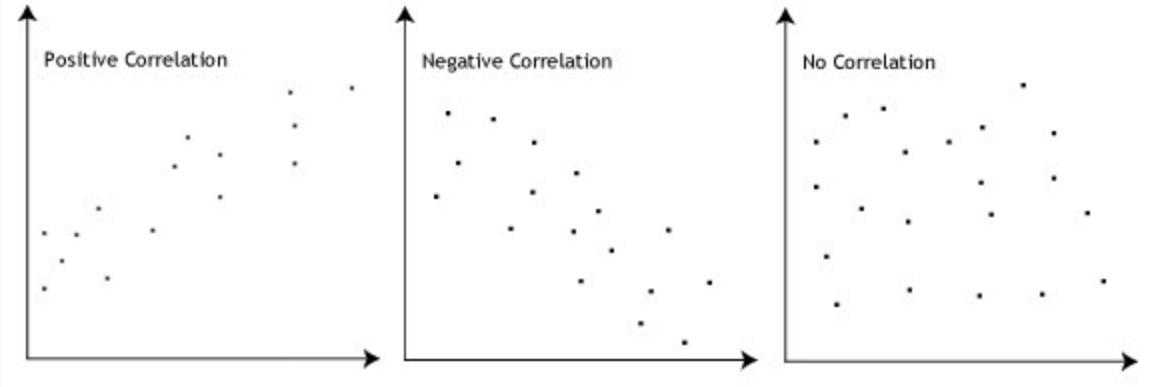
In Data set D we can also see the effect of a single outlier – we have 11 points in a vertical line, and then a single outlier. Clearly here again drawing a line of best fit for this data is not appropriate – unless we remove this outlier first.

3.**What is Pearson's R?**

Ans: In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association



Pearson r Formula

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

Here,

- r=correlation coefficient
- =values of the x-variable in a sample
- =mean of the values of the x-variable
- =values of the y-variable in a sample
- =mean of the values of the y-variable

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

Ans:

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

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.

Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and
  1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

  MinMaxScaling : X = x -min(x) /(max(x) -min(x))

Standardization Scaling:

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

  Standardisation : X =x-mean(x)/ Sd(x)

- sklearn.preprocessing.scale helps to implement standardization in python.

- One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

Ans: If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. 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.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Ans: Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.