Assignment-based Subjective Questions

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

There are the inferences:

Season: Fall season has the highest count of bike sharing.

Year: 2019 has a higher amount of bike sharing.

Month: Months in the middle of the year have the highest bike sharing.

Working days have a higher bike sharing Bike sharing.

clear weather of course has more rental bike sharing.

- **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**

We need to use the drop_first=True option while creating dummy variable to eliminate redundance. If I have n different values in my categorical variable column, I will only need 'n -1' dummy variables to represent the n different values. Without using drop_first=True, the variables will be a good predictor for the variable that supposed to be dropped. So there will be some dependability between some of the variables.

- **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

The highest correlation with the target variable, cnt, is with the independent variable temp.

- How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

The assumptions of Linear Regression are:

- Linearity: There should be a linear relationship between dependent and independent variables. With the help of pairplot, we can check if there were any linear relationships that existed between the independent variable and target variable.

- Normality: Residuals (Errors) must be normally distributed. Using the distplot, we can know if the errors are normally distributed or not.

- Undependability: Error terms must be independent. Scatter plot can show that there is no visible pattern between error terms.

- **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

The top 3 features contributing significantly towards the demand of shared bikes are:

- Temperature (Positive Predictor)
- Year (Positive Predictor)
- weathersit_light_rain_&_snow: (Negative Predictor).

General Subjective Questions:

1. **Explain the linear regression algorithm in detail?**

Linear regression is finding the best linear relationship between the independent and dependent variables. It is mostly done by the Sum of Squared Residuals Method. Linear regression is one of the very basic forms of machine learning algorithms where we can train a model to predict the behaviour of the data based on some variables.

In Linear Regression, whenever X is increasing Y is also increasing or vice versa that means they have a correlation or there will be a linear downward relationship.

Linear regression is used to predict a quantitative response Y from the predictor variable X. Mathematically, we can write a linear regression equation be like below:

$$Y = mx + c$$

Where m and c given by the formulas:

Where m is slope of the line, c is the y intercept of the line.

Assumption of Linear Regression is:

- There is a linear Relationship between dependant and independent variables.
- Error values (ε) are normally distributed for any given value of X that means.
- Error terms are normally distributed around zero.
- Constant variance assumption: It is assumed that the residual terms have the variance, σ2, this assumption is also known as the assumption of homogeneity or homoscedasticity.
- Independent error assumption: residual terms are independent of each other, i.e. their pair-wise covariance is zero.
- The independent variables are linearly independent of each other, i.e. there is no multicollinearity in the data.

2. **Explain the Anscombe's quartet in detail?**

- **Anscombe's Quartet** can be defined as a group of four data sets which are **nearly identical in simple descriptive statistics**, but there are some peculiarities in the dataset that **fools the regression model** if built. They have very different distributions and **appear differently** when plotted on scatter plots.
- There are these four data set plots which have nearly **same statistical observations**, which provides same statistical information that involves **variance**, and **mean** of all x,y points in all four datasets.
- This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted

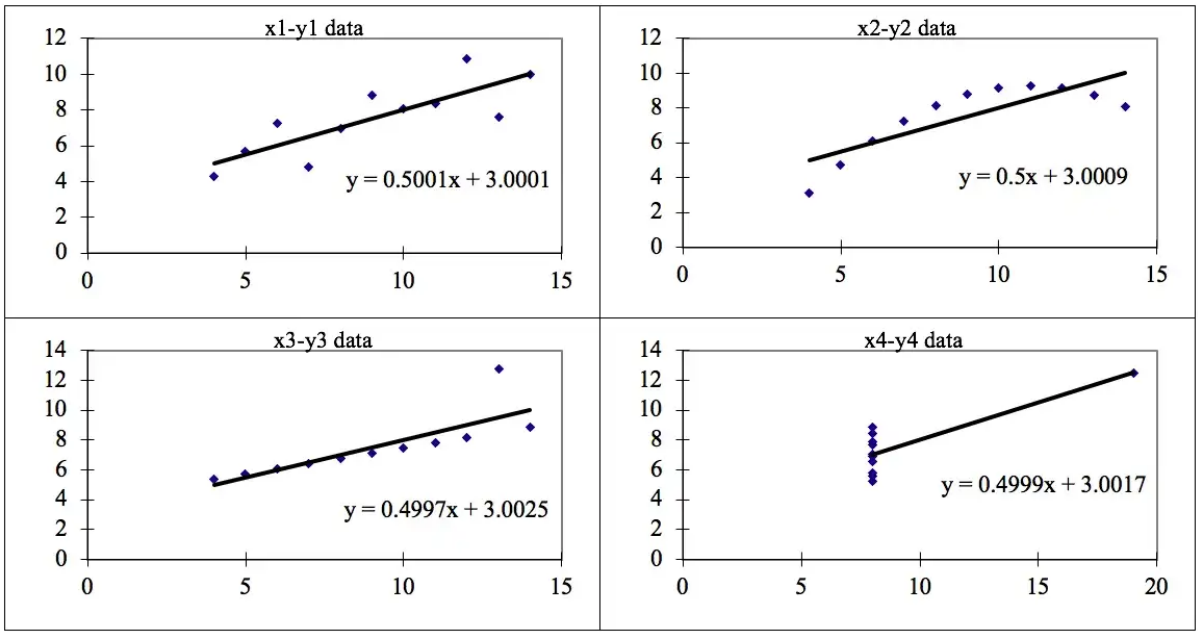
in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the **data with linear relationships** and is incapable of handling any other kind of datasets. These four plots can be defined as follows:

Anscombe's Data										
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89

- The statistical information for all these four datasets are approximately similar and can be computed as follows:

Anscombe's Data										
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

When these models are plotted on a scatter plot, all datasets generates a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows:



The four datasets can be described as:

Dataset 1: this **fits** the linear regression model pretty well.

Dataset 2: this **could not fit** linear regression model on the data quite well as the data is non-linear.

Dataset 3: shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

Dataset 4: shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

3. What is Pearson's R?

Pearson correlation coefficient will calculates the strength of a linear association between two variables that means change in the one variable where the other variable changes and this is denoted by r.

What does the Pearson correlation coefficient test do? It seeks to draw a line through the data of two variables to show their relationship. The relationship can be measured with the calculator of Pearson correlation coefficient. This linear relationship can either be positive or negative.

For example:

Positive linear relationship: The income of a person increases as his/her experience increases.

Negative linear relationship: If the vehicle speed is increases, then time taken to travel decreases, and vice versa. Pearson correlation coefficient formula:

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

Where:

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

Σx2 = the sum of squared x scores

Σy2 = the sum of squared y scores.

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

Scaling:

Scaling is the pre-processing applied to independent variables to normalize the data with in a particular range and also helps in speeding up the calculations.

Why is scaling performed? Often, the data set contains most varying units, magnitudes and range. If we have such data and if we not done scaling then it will only takes the magnitude in account and not units hence the incorrect model will build. To resolve this we will use scaling to bring all the variables to the same level of magnitude and units.

Scaling only affects the coefficients but not the other parameters.

Normalization/Min-Max Scaling:

It is a technique that the values are shifted or rescaled and is ranging between 0 and 1 known as min – max scaling. That means it brings all the data within the range of 0 and 1.

sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Standardization Scaling:

It is another technique in which the values are centred around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant

distribution has a unit standard deviation. 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).

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

sklearn.preprocessing.scale helps to implement standardization in python. Disadvantage of Min-max scaling over standardization: Normalization lose the information of outliers.

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

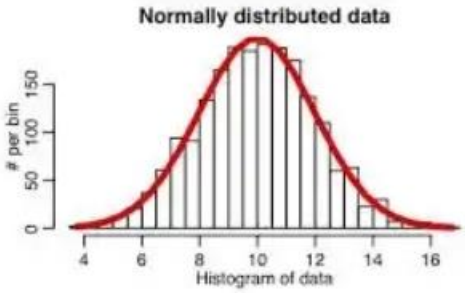
- VIF = infinity is indicating that there is perfect correlation between the two independent variables.
- If such cases are arrived, we get R2 = 1, which leads to 1/(1-R2) = 1/0 = infinity.
- To solve this problem, we need to drop one of the variables which causes the perfect correlation i.e., multicollinearity.
- This infinite value of VIF indicates that the corresponding variable can be expressed by Linear Combination of the other variables.

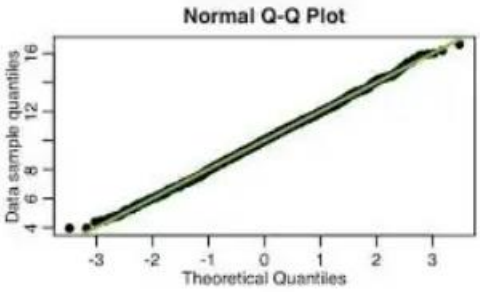
6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression?**

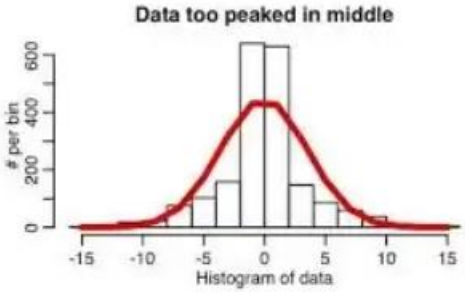
We plot the theoretical quantiles or basically known as the standard normal variate (a normal distribution with mean=0 and standard deviation=1) on the x-axis and the ordered values for the random variable which we want to find whether it is Gaussian distributed or not, on the y-axis. Which gives a very beautiful and a smooth straight line like structure from each point plotted on the graph.

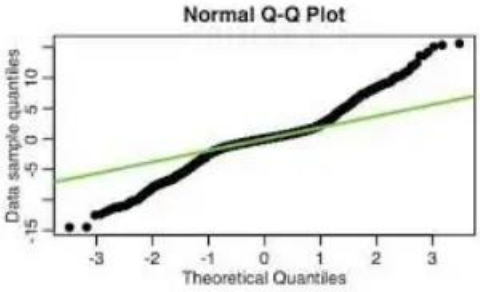
Now we have to focus on the ends of the straight line. If the points at the ends of the curve formed from the points are not falling on a straight line but indeed are scattered significantly from the positions then we cannot conclude a relationship between the x and y axes which clearly signifies that our ordered values which we wanted to calculate are not Normally distributed.

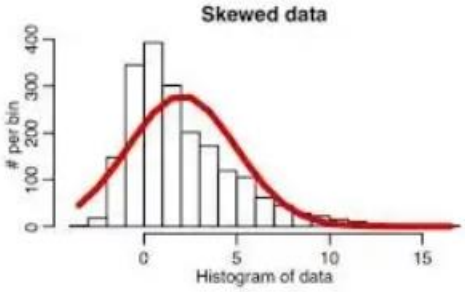
If all the points plotted on the graph perfectly lies on a straight line then we can clearly say that this distribution is Normally distribution because it is evenly aligned with the standard normal variate which is the simple concept of Q-Q plot.

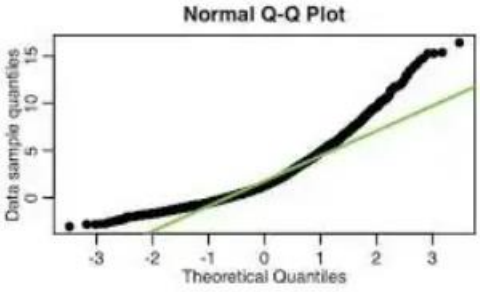

Normally distributed data
per bin
0
50
100
150
4
6
8
10
12
14
16
Histogram of data


Normal Q-Q Plot
Data sample quantiles
4
6
8
12
16
-3
-2
-1
0
1
2
3
Theoretical Quantiles


Data too peaked in middle
per bin
0
200
400
600
-15
-10
-5
0
5
10
15
Histogram of data


Normal Q-Q Plot
Data sample quantiles
-15
-5
0
5
10
-3
-2
-1
0
1
2
3
Theoretical Quantiles


Skewed data
per bin
0
100
200
300
400
0
5
10
15
Histogram of data


Normal Q-Q Plot
Data sample quantiles
0
5
10
15
-3
-2
-1
0
1
2
3
Theoretical Quantiles