Assignment-based Subjective Questions

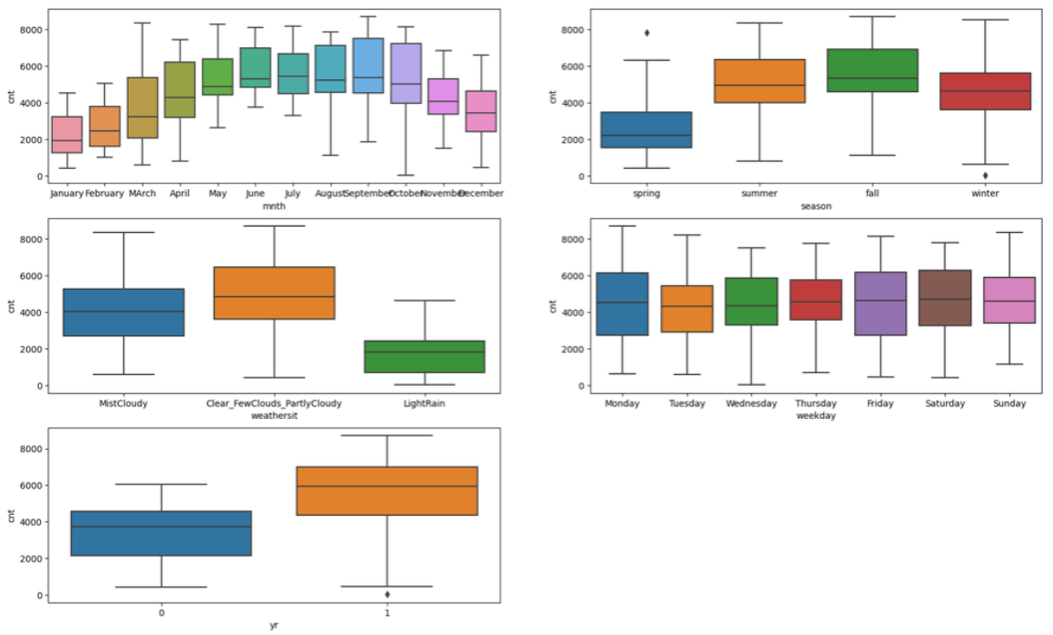
Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Ans: From the categorical variables analysis, we infer following things:

(i) Demand of bikes increased as compared to last year i.e. 2018.

(ii) In rainy weather, demand of bikes is lesser and in cloudy weather, demand of bikes is higher.

(iii) Demand of bikes is higher in mid of the year especially in time of fall.



Q2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Ans: drop_first=True is useful while creating dummy variables because it reduces columns means if one column can explain variables clearly then why we use two e.g. in weekdays if we write for 6 columns then seventh will explain automatically.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans: **A**s per pair plot, “temp” variable is highly correlated with target variable.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

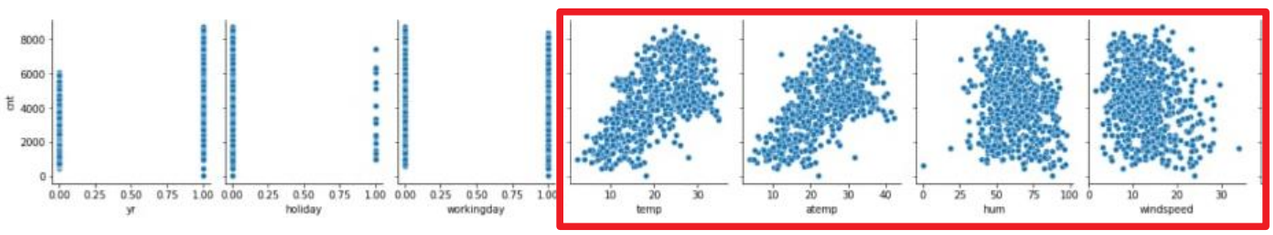
Ans:

Assumption 1: **No Multicollinearity.**

Validation: By checking VIF value, we can validate our assumption. All the variables have a VIF value below 5.

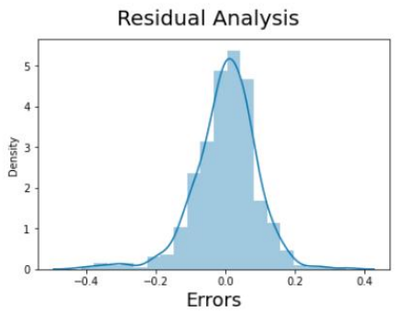
Assumption 2: **Linear relationship in dependent variable and independent variables**

Validation: From the pair plot, we can say that there is a linear relationship in target variable and independent variables.



Assumption 3: **Residuals must be normally distributed**

Validations: From the graph it is clear that residuals are distributed normally.



Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans: (i) aTemp
(ii) Weather situation
(iii) Working day

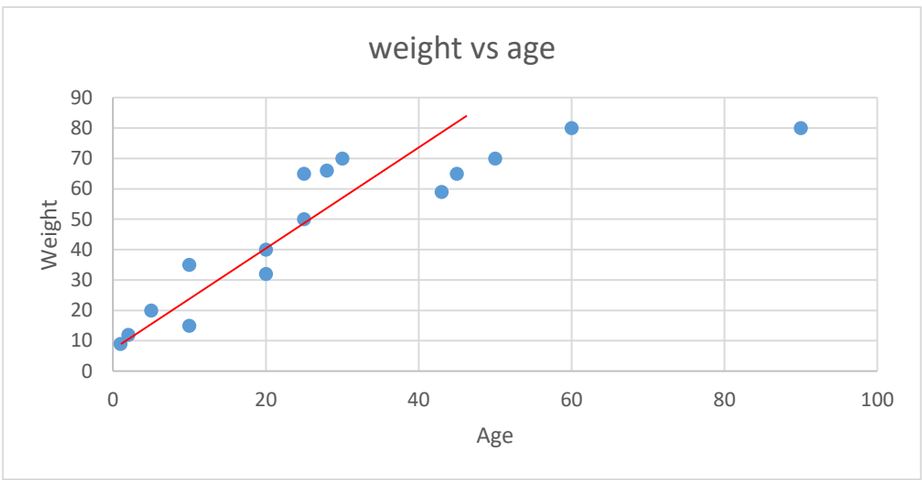
General Subjective Questions

Q1. Explain the linear regression algorithm in detail. (4 marks)

Ans: Linear regression is a Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous variables such as price, age etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (x) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the target variable is changing according to the value of the independent variable.

The linear regression model provides a straight liner representing the relationship between the independent and dependent variables. E.g.



Mathematically, we can represent a linear regression as: y = mx + c

Here,

Y=Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)
c= intercept of the line
m = Linear regression coefficient

<u>Types of Linear Regression</u>

- **Simple Linear Regressi**on

  When a single independent variable is used to predict the value of a numerical dependent variable, called Simple Linear Regression.

- **Multiple Linear regression:**

  When more than one independent variable is used to predict the value of a numerical dependent variable, called Multiple Linear Regression.

Linear Regression Line

- A linear line showing the relationship between the dependent and independent variables is called a **regression line**. A regression line can show two types of relationship:

- **Positive Linear Relationship:**
  If the dependent variable increases on the Y-axis and independent variable increases on X-axis, then such a relationship is termed as a Positive linear relationship.

- **Negative Linear Relationship:**
  If the dependent variable decreases on the Y-axis and independent variable increases on the X-axis, then such a relationship is called a negative linear relationship.

Residuals: The distance between the actual value and predicted values is called residual.

Model Performance:

The process of finding the best model out of various models is called **optimization**. It can be achieved by below method:

1. R-squared method:

- R-squared is a statistical method that determines how fit our model is.
- It measures the strength of the relationship between the dependent and independent variables on a scale of 0-100%.
- The high value of R-square determines the less difference between the predicted values and actual values and hence represents a good model.
- It is also called a **coefficient of determination,** or **coefficient of multiple determination** for multiple regression.

Assumptions of Linear Regression

1. Linear relationship between the features and target variables.
2. No multicollinearity between the features.
3. Normal distribution of residuals.
5. No correlation in between residuals.

Q2. Explain the Anscombe's quartet in detail. (3 Marks)

Ans: Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics like mean, standard deviation etc, 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.

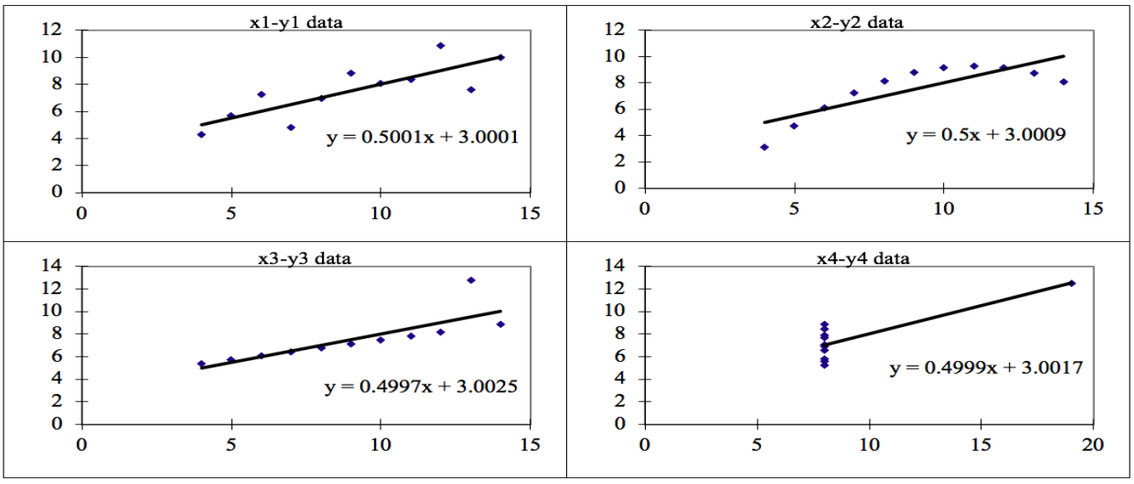
It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analysing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly same statistical observations, which 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

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

Q3. What is Pearson's R? (3 marks)

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:

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

- If R > 8 means there is a strong association

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

Here,
r = correlation coefficient

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

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
Ans: Scaling: It is a process of data Pre-Processing which is applied to independent variables to standardize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Why is scaling performed: Dataset received from client may be in various formats. Numerical values in that dataset may have different ranges. So if we build model on that dataset without scaling, the coefficient of independent variables have different values. So based on that value, it is quite difficult to predict the significance of that variable.

Secondly, after scaling values all variables lies within a range, so computation becomes easy and fast.

Difference between normalized scaling and standardized scaling:

1. **Normalized Scaling :** It is also called Min-Max Scaling. All the values lies in between 0 and 1 after normalized scaling. In python, sklearn library is commanly used for utilizing this funtion as:
   sklearn.preprocessing.MinMaxScaler
   Formula used for achieving this scaling is:

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

2. **Standardized 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**). In python, **sklearn.preprocessing.scale** is used to implement this type of scaling.
   Formula Used:

$$x = \frac{x - \text{mean}(x)}{\text{sd}(x)}$$

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

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
Ans: We know that formula of VIF is :

$$VIF = \frac{1}{1 - R^2}$$

- If value of R will be 1 then VIF will be infinity.
- We know that value of R will be 1 only when there is a perfect correlation in variables.
- This means this variable is fully explained by some another variable.

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

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Ans: Quantile-Quantile plot also known as Q-Q plot, is a graphical tool to help us assess if a set of data possibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This 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.

Advantages:

a) It can be used with sample sizes also.

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:

If two data sets —

i. come from populations with a common distribution
ii. have common location and scale
iii. have similar distributional shapes
iv. have similar tail behaviour

Interpretation:

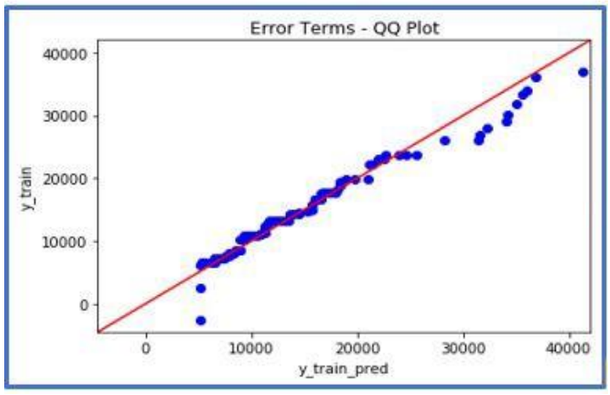
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set

.

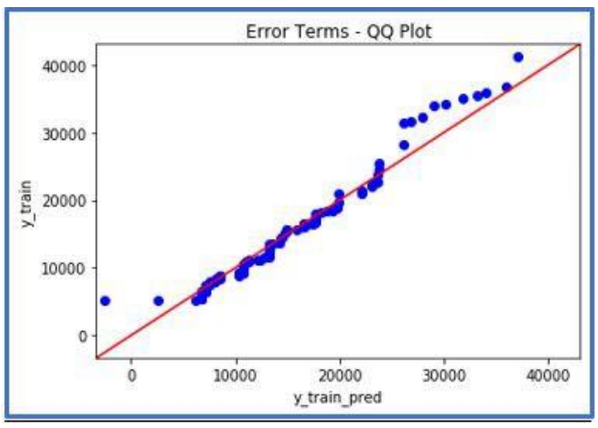
Below are the possible interpretations for two data sets.

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x –axis.

b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.



d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis