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)

Answer: I have plotted the categorical variables with the target variables on boxplot and has inferred following effect on target:

- ✓ Season: 3: fall has highest demand for rental bikes
- ✓ I see that demand for the year 2019 has grown
- ✓ Demand is continuously growing each month till Jul. September month has highest demand. After September, demand is decreasing
- ✓ When there is a holiday, demand has decreased.
- ✓ Weekday is not giving clear picture about demand.
- ✓ The clear weathershit has highest demand

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

Answer: 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.

If we do not drop one of the dummy variables created from a categorical variable then it becomes redundant with dataset as we will have constant variable(intercept) which will create multicollinearity issue.

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

Answer: The feature “temp” has highest correlation. It is very well linearly related with target “cnt”

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

Answer: I have checked the following assumptions:

- ✓ Error terms are normally distributed with mean 0.
- ✓ Error Terms do not follow any pattern.
- ✓ Multicollinearity check using VIF(s).
- ✓ Linearity Check.
- ✓ Ensured the overfitting by looking the R2 value and Adjusted R2.

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

Answer: Features “temp”, “yr” and “weathersit_bad” are highly related with target column, so these are top contributing features in model building.

General Subjective Questions:

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

Answer: Linear Regression Algorithm is a machine learning algorithm based on supervised learning. Linear regression is a part of regression analysis. Regression analysis is a technique of predictive modelling that helps you to find out the relationship between Input and the target variable.
Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

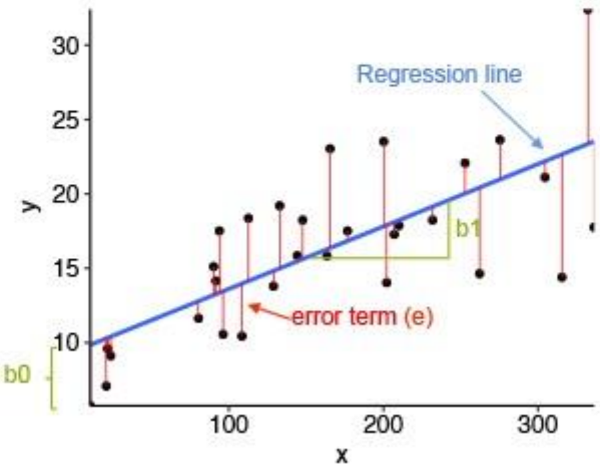
Example for that can be let's say you are running a sales promotion and expecting a certain number of count of customers to be increased now what you can do is you can look the previous promotions and plot if over on the chart when you run it and then try to see whether there is an increment into the number of customers whenever you rate the promotions and with the help of the previous historical data you try to figure it out or you try to estimate what will be the count or what will be the estimated count for my current promotion this will give you an idea to do the planning in a much better way about how many numbers of stalls maybe you need or how many increase number of employees you need to serve the customer. Here the idea is to estimate the future value based on the historical data by learning the behaviour or patterns from the historical data.

In some cases, the value will be linearly upward that means 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.

One example for that could be that the police department is running a campaign to reduce the number of robberies; in this case, the graph will be linearly downward.

Linear regression is used to predict a quantitative response Y from the predictor variable X.

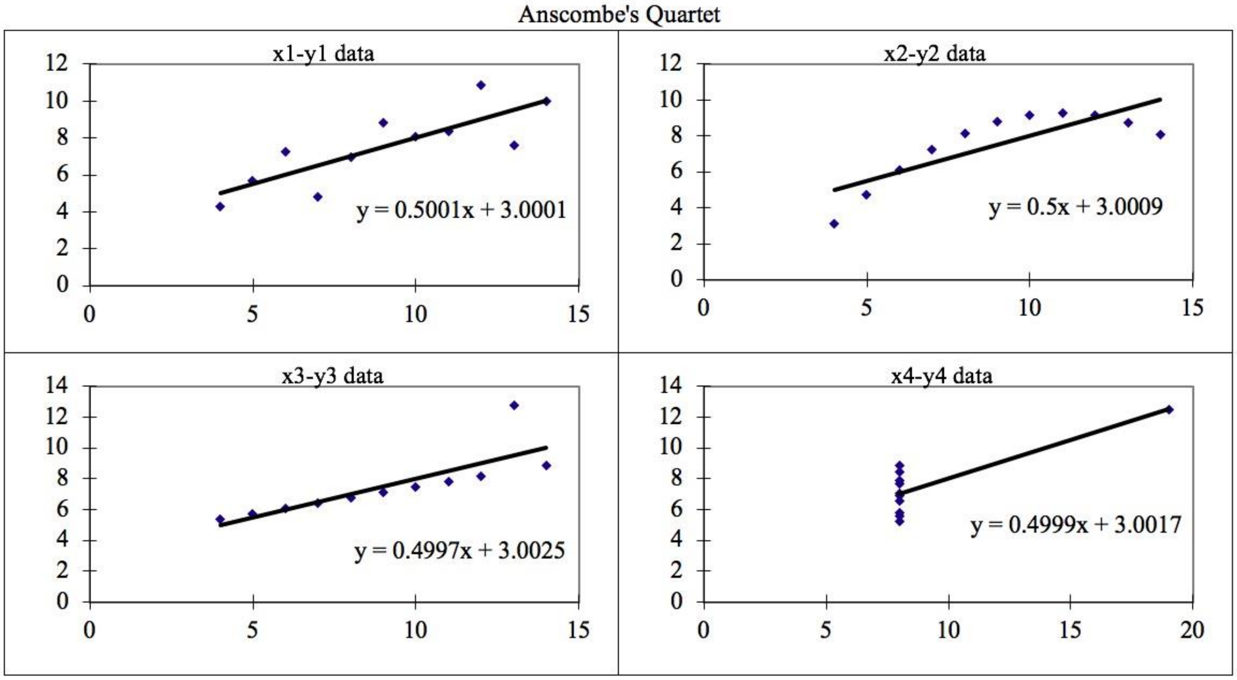
Mathematically, we can write a simple linear regression equation as follow **$y \sim b0 + b1*x$**
Where y is the predicted variable (dependent variable), b1 is slope of the line, x is independent variable, b0 is intercept(constant). It is cost function which helps to find the best possible value for m and c which in turn provide the best fit line for the data points.



Here, x and y are two variables on the regression line.
b1 = Slope of the line. b0 = y-intercept of the line. x
= Independent variable from dataset y = Dependent
variable from dataset

7. Explain the Anscombe's quartet in detail. (3 marks)

Answer: **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. Each dataset consists of eleven (x,y) points.



The four datasets can be described as:

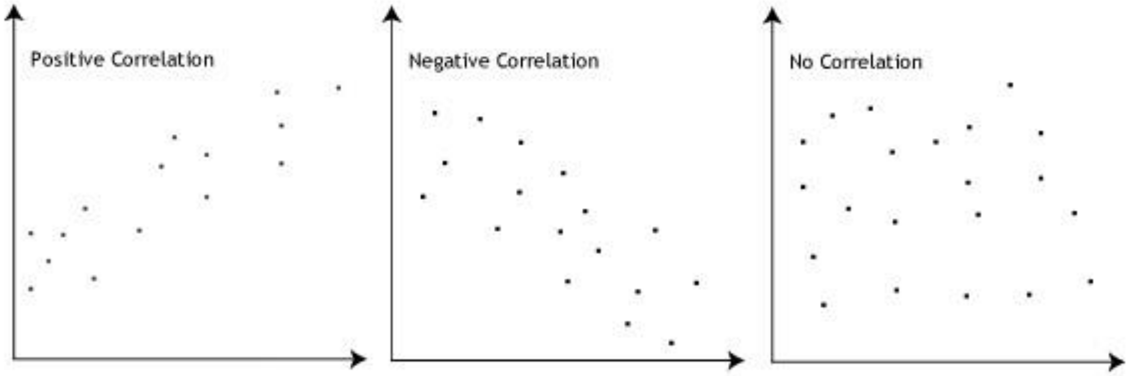
- **Dataset 1:** this **fits** the linear regression model pretty well.
- **Dataset 2:** this **could not fit** linear regression model on the data quite well as the data is non-linear.
- **Dataset 3:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model **Dataset 4:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

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

Answer: **Pearson's r** is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive.
Pearson's r measures the strength of the linear relationship between two variables.

Pearson's r always between -1 and 1.

If data lie on a perfect straight line with negative slope, then r = -1.



Positive correlation indicates the both the variable increase and decrease together. Negative correlation indicates the one the variable increase and the other variable decrease and vice versa.

9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer: Scaling is a method to normalize the range of independent variables. It is performed to bring all the independent variables on a same scale in regression. If Scaling is not done, then regression algorithm will consider greater values as higher and smaller values as lower values.

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.

Example Weight of a device = 500 grams, and weight of another device is 5 kg. In this example machine learning algorithm will consider 500 as greater value which is not the case. And it will do wrong prediction.

Machine Learning algorithm works on numbers not units. So, before regression on a dataset it is a necessary step to perform.

Scaling can be performed in two ways: Normalization: It scale a variable in range 0 and 1. Standardization: It transforms data to have a mean of 0 and standard deviation of 1

10. You might have observed that sometimes the value of VIF is infinite.
Why does this happen? (3 marks)

Answer: When there is a perfect relationship then VIF = Infinity whereas if all the independent variables are orthogonal to each other then VIF = 1.0. Means if a variable is expressed exactly by a linear combination of other variable then it is said that VIF is infinite.
The Variance Inflation Factor (VIF) is a measure used to assess the severity of multicollinearity in a regression analysis. Specifically, it quantifies how much the variance of an estimated regression coefficient increases when your predictors are correlated.
The formula for calculating the VIF for a variable is:
VIF=11− 2VIF=1−R21
where 2R^2 is the coefficient of determination from a regression of the variable of interest against all other independent variables.
In theory, VIF values should be greater than or equal to 1. A VIF of 1 indicates no multicollinearity, and values above 1 suggest increasing levels of multicollinearity. However, a VIF of infinity occurs when the 2R^2 value in the denominator approaches 1.
This situation of 2R^2 approaching 1 (or being exactly 1) typically happens when a variable can be perfectly predicted by a linear combination of the other variables in the model. In other words, the variable is a perfect linear function of the other variables, leading to a situation called perfect multicollinearity.
Perfect multicollinearity can occur for various reasons:

1. **Redundant Variables:** One variable is a constant multiple of another variable or a combination of other variables.
2. **Data Issues:** Data entry errors, duplication of variables, or other issues in the dataset may lead to perfect multicollinearity.
3. **Linear Dependencies:** The presence of linear dependencies among the predictors.

When perfect multicollinearity is present, the matrix of predictor variables becomes singular, making it impossible to compute a unique set of coefficients. As a result, the VIF becomes infinite because of the division by zero (1 - 2R^2 = 0).

To address this issue, it's essential to carefully examine the dataset and the relationships between variables. Identifying and removing redundant variables or addressing data issues can help mitigate problems associated with perfect multicollinearity and infinite VIF values.

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

Answer: **Quantile-Quantile** (**Q-Q**) **plot**, is a graphical tool to help us assess if a set of data plausibly 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

It is used 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. Whether the Distributions is Gaussian, Uniform, Exponential or even Pareto distribution, it can be found out.

Few 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.

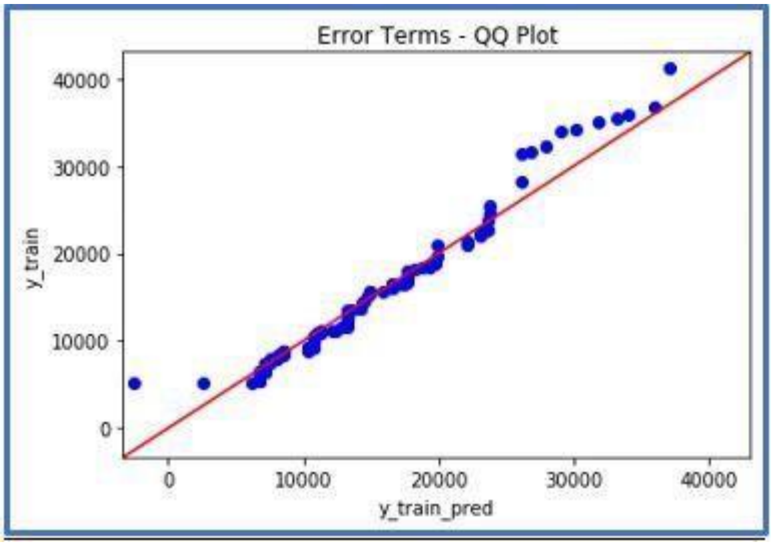
It is used to check following scenarios:

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