## 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:** Based on my analysis of the categorical variables using box plot and bar chart, below are the insights about their influence on dependent variables:

- Whenever there was clear weather there were more bookings observed.
- In a year we could observe that there were more booking done in the month of May,June,July,August,September,October and we see that booking has increased abundantly from 2018 to 2019.Which is good sign of business progress.
- When it is working day, booking will be less as people will be busy in their own work and may not hang out.
- Among the weekdays, we can observe that bookings are more in Thursday,Friday,Saturday and Sunday.

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

**Answer:** Using drop_first = True is very important as it will avoid additional columns created during creation of dummy variables and hence reduces unwanted correlations among multiple dummy variables**.**

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

**Answer:** By looking at the pair plots , 'temp' variable has the highest correlation with the target variables.

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

**Answer:** Below are the validations I have done for the assumptions of Linear Regression

- Normality of error terms - Error terms are normally distributed
- Multicollinearity Check - There is no significant collinearity among variables.
- Linear Relationship Check – Linearity is visible among the variables
- Homoscedasticity – No Visible pattern found in residual values.
- Independence of Residuals - There is no auto correlation found.

**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:** The top 3 features that significantly explains the demand of the shared bikes:

1. 'temp'
2. 'winter'
3. 'sep'

## General Subjective Questions

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

**Answer**: Linear regression is the statistical model that defines the linear relationship that exists between dependent variable with that of the one or more independent variable. This will give more insight on the influence of multiple independent variables on a single/target dependent variable.

Y = mX + c equation defines the mathematical association among the variables.

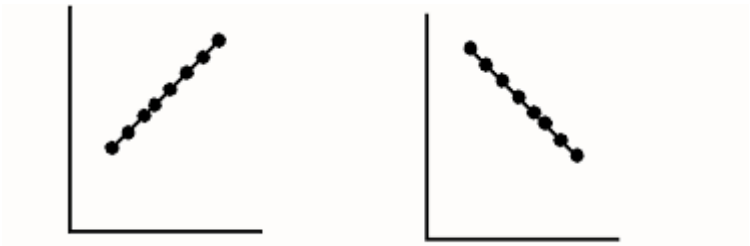
Where Y is the dependent variable

X is the independent variable used for predictions

M is the slope that represents the effect of X on the dependent variable Y.

C is the constant, which is called as Y-intercept .If X=0, Y would be equal to c.

There are two types of linear regression –positive, negative based on the relationship that exists between Y, and X .Below image depicts both the types.



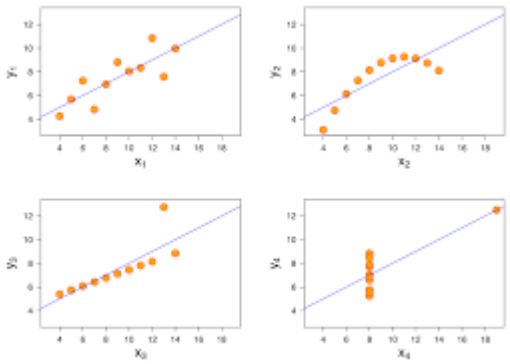
(A) Perfect Positive Correlation ($r = +1$) (B) Perfect Negative Correlation ($r = -1$)

There are certain assumptions exists in Linear Regression:

- Multi-Collinearity
- Auto- Correlation
- Relationship between variables
- Normality of error terms
- Homoscedasticity

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

**Answer:** Anscombe's quartet consists of four datasets having each containing 11 pairs of (x,y). Here descriptive statistics is same,but while they are depicted in graph they will have different insights and observations:



Though the means and variances are same across the groups. The graphical representation says a different observation:

- Dataset 1- It is the clean and well-fitting linear model.
- Dataset-2 is not normally distributed.
- Dataset -3 is linear but outliers are seen.
- Dataset-4 shows outliers are enough to have high correlation coefficients.

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

**Answer:** Pearson's R is the numerical summary of the correlation among the variables. Whether the variables grow as the other variable goes up/increases or variable will decrease as the other variables increase/grow. R can take a value between -1 to +1.Value 0 indicates that there is no association between the two variables. Value greater than 0 indicates positive association and value less than 0 indicates negative association between the variables. Three different scenarios can been seen in the below graph**.**


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

**Answer:** Feature scaling is the technique of standardising independent features present in the data in a fixed range.

It will be usually done in the data pre-processing to manage the data /volume with highly varying units of values.

If Data scaling is not done, then numeric values in the machine learning algorithms will be interpreted just by their value not based on their units so it might depict small values as large without considering the measuring units.

Difference between normalised and standard scaling:

| Normalised Scaling | Standard Scaling |
|---|---|
| Minimum and maximum values are used for scaling | Mean and standard deviation is used for scaling |
| It will be used for the features which are of different scales | It will be used for zero ,mean or unit standard deviation |
| Scales between [0,1] and [-1,1] | It is not bounded by a range |
| It is affected by outliers | It is less affected by outliers |

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

**(3 marks)**

**Answer**: If there is a perfect correlation, then VIF=infinity .Large value of the VIF indicates there is a correlation between the variables. In case of perfect correlation, R-Squared value =1 which lead to 1/(1-R2) as infinity. So from this we can infer that we need to drop the variable from the dataset which is causing the multicollilinearity**.**

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

**(3 marks)**

**Answer:** The quantile-quantile is determining two datasets come from a population with a common distribution.

**Use of Q-Q plot:** It is the quantile of the first dataset against the quantile of second data set. Here 45-degree reference line is plotted .If the two sets come from the population of common distributions, values should fall along this reference line.

**Importance of Q-Q plot:** It can provide more insights into the nature of the differences than the analytical methods such as Chi-Square and 2 sample tests