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?**

   From the analysis of categorical variables using boxplot, below inferences can be made:

   - From all the seasons, summer and fall seems to attract more customers and hence, there are more bookings.
   - Year 2019 had fetched more customers as compared to year 2018 which shows good progress in terms of business.
   - Most bookings are made in May, June, July, August, September, and October. Booking trend starts to increase from February till the mid of the year and starts decreasing by the end of the year.
   - Most booking are made on Friday, Saturday, and Sunday but the median remains similar for all the days of the week.
   - Approximately similar number of bookings can be observed for both working and non-working day with median remaining the same. But more bookings are made during the holidays, the reason being people wants to spend time with their family or friends.
   - Good weather or clear weather attracted most customers and moderate weather remains second. That seems to be obvious as people wants to ride bikes during warms days.

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

   drop_first = True helps in removing extra column created during dummy variable creation. Thus, it reduces the instance of multicollinearity amongst the dummy variables.

   Syntax: drop_first = bool. The default is always “False” which implies whether to get k-1 dummies out of k categorical levels by removing the first level.

For example, if there are 3 levels, we know that if the value is not in level 1 and level 2, then it automatically is in level 3. So, we only need 2 levels. drop_first = True will drop the first column.

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

   "temp" and "atemp" variables has the highest correlation with the target variable "cnt".

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

   There are 5 assumptions to validate a Linear Regression Model after building:

   - Normality of error terms: Error terms should be normally distributed
   - Multicollinearity check: There should be less than 5 Variance Inflation Factor score amongst the variables. Which means, insignificant multicollinearity amongst the variables.
   - Linear relationship validation: Linearity should be visible amongst variables.
   - Homoscedasticity: There should be no visible patterns in residual values.
   - Residual independency: No auto-correlation

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

   Top 3 features significantly contributing towards the demand of the shared bikes are temperature, season, and year.

General Subjective Questions

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

   Linear regression is a machine learning algorithm to predict numerical values based on input features. It describes linear relationship between the dependent variable and

independent variables. It helps us to understand how independent variable influences the dependent variable and allow us to make predictions based on the learned relationships. Linear regression estimates the best-fit line that represents this relationship, enable to analyze and make predictions on new data points.

The mathematical equation for linear relationship is as below:

Y = mX + C

Here, Y = dependent variable

m = slope of line representing the influence of X on Y

C = constant, known as Y intercept

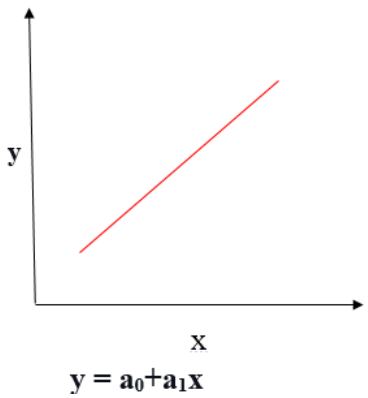
X = independent variable

Linear relationship can be

There are two types of linear regressions. If there is only one input variable, it is called simple linear regression. If there are more than one input variables, it is called multiple linear regression. The linear regression model gives a sloped straight line describing the relationship within the variables. A regression line can be a Positive Linear Relationship or Negative Linear Relationship.

a. Positive Linear Relationship

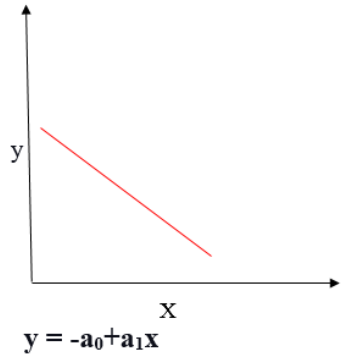
When dependent variable expands on the Y-axis and the independent variable progress on X-axis.



$y = a_0 + a_1 x$

b. Negative Linear Relationship

When dependent variable decreases on Y-axis, independent variable increases on X-axis.



In linear regression, RFE or Mean Squared Error (MSE) or cost function is used, which helps to figure out the best possible values for a0 and a1, which provides the best fit line for the data points.

2. Explain the Anscombe's quartet in detail.

Statistician Francis Anscombe constructed Anscombe's quartet in 1973 which describes a group of four datasets (x,y) that has identical descriptive statistical properties such as mean, median, variance, variance, R-squared, correlations, and linear regression lines but are qualitatively different. When we plot them, they look different. It illustrates the importance of exploratory data analysis rather than just depending on the statistical summary. It also emphasizes the importance of data visualization to spot trends, outliers, and other crucial details that might not be found from statistical summary.

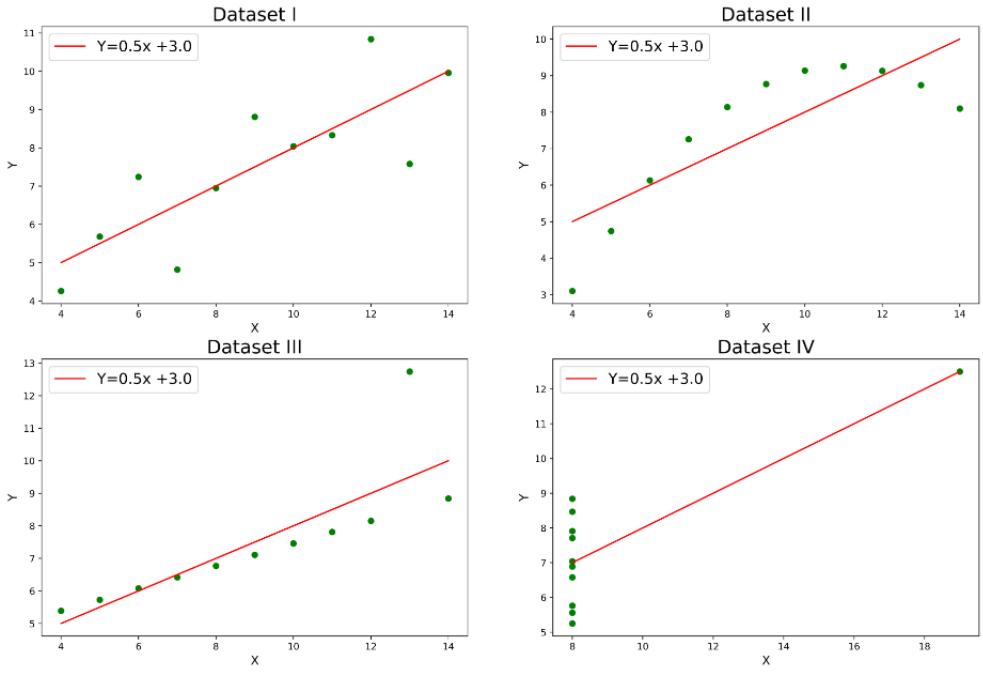
Example:

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Statistical Data:

	I	II	III	IV
Mean_X	9	9	9	9
Variance_X	11	11	11	11
Mean_Y	7.500909	7.500909	7.500909	7.500909
Variance_Y	4.127269	4.127269	4.127269	4.127269
Correlation	0.816421	0.816421	0.816421	0.816421
Linear Regression Slope	0.500091	0.5	0.499727	0.499909
Linear Regression Intercept	3.000091	3.000909	3.002455	3.001727

Data Visualization:



a. First plot shows a linear relationship between x and y.
b. Second plot shows nonlinear relationship between x and y.
c. Third plot shows perfect linear relationship for all data points except one which is far away and can be indicative of an outlier.
d. Fourth plot shows an example when one high leverage point is enough to produce a high correlation coefficient.

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

Pearson's R correlation coefficient is a numerical summary that shows the strength of linear correlation between two variables. Value ranges between -1 and 1, where -1 indicates a negative perfect correlation, +1 indicates a positive perfect correlation, and 0 indicates there is no linear correlation.

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

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

In machine learning, scaling is referred to as putting feature variables into the same range. It is important for the algorithm to measure the distances between data points like in k-nearest neighbour. Collected dataset contains many features varying in different magnitude, units, and range. If we don't perform scaling, the algorithm will consider only high magnitude data points in account and not units. Thus, it will result in incorrect modelling. Scaling is performed to bring all the variables to the same level of magnitude. There are two types of scaling:

A. Normalized/Min-Max scaling:

It brings all the variables in the range of 0 and 1.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

B. Standardized Scaling:

It replaces the value by their Z score. It brings all of the data to a mean (μ) value of 0 and standard deviation (σ) to 1.

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

Difference between Normalized and Standardized Scaling:

i. In Normalized Min and Max values of features are used, whereas in Standardized scaling, mean and standard deviations are used.

ii. Normalized scaling is used when features of different scale, whereas Standardized scaling is used for zero mean and unit standard deviation.

iii. In Normalized scaling, values are fixed in the range of 0 to 1. Whereas in Standardized scaling, values are not bounded to a fixed range.

iv. Normalized scaling is affected by outliers, but Standardized scaling is not affected by them.

v. Normalized scaling is used when we don't know about the distribution, whereas Standardized scaling is used for normal distribution of data.

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

Variance inflation factor (VIF) indicates the relationship between one independent variable with another independent variable or their collinearity. If the independent variables are orthogonal to each other, their VIF is 1. If VIF value is 10, then it is definitely high, and VIF value of 5 should also be inspected carefully. When there is a

perfect correlation, then the VIF value is infinite. It is a case when R2 =1, which gives 1/(1-R2) = infinite. To solve this issue, we drop one of the variables in the dataset that causes perfect multicollinearity.

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

Quantile-Quantile plot is a graphical technique that plot quantiles of sample distribution against quantile of theoretical distribution, which helps us to determine whether dataset follows normal, uniform, or exponential distribution.

Q-Q plot can also be used to determine whether or not two distributions are similar or not. If they are similar, the Q-Q plot will look more linear. The linearity assumptions can be best tested using scatter plots. Moreover, the linear regression analysis requires all variables to be multivariate normal. It can be best tested using histogram or Q-Q plots.

Importance of Q-Q plots:

In linear regression, we create Q-Q on test and train datasets to confirm whether both datasets are from the same population and having same distribution or not.

Advantages:

- It can be used on sample sizes too.
- Many distribution aspects such as shift in location, shift in scale, change in symmetry, and outliers can be detected.

It can be used to check two datasets:

- Coming from same population with common distribution
- Have common location and scale
- Have similar distribution shape.
- Have similar tail behaviour.