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?**

Solution:

We have plotted Categorical data on the Box plot against the dependent variable to understand the inference, refer below pointers for the inference.

a. Season column gives a handful insight says that "Fall" and "Summer" season seen a greater booking as compared another season.
b. During the summer period i.e., June, July recorded a greater number of Bike counts.
c. Weather seems to also have a impact on bike rental we can infer on the bar plot.
d. Significant increase in rentals YOY.
e. Holiday seems to be downside as no one interests in bikes as they can opt for a family trip, which can't be accommodated on Bike.

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

Solution:

drop_first = True will helps in reducing the extra column created while dummy variable creation. Hence it reduces the correlations created among dummy variables.

Syntax - drop_first = bool, default False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level.

Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not 1 and 2, then it is obvious 3. So, we do not need 3rd variable to identify the 3.

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

Solution:

'temp' has the highest correlation with the target variable.

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

Solution:

There are four assumptions associated with a linear regression model:

a. Linearity: The relationship between X and the mean of Y is linear.
b. Homoscedasticity: The variance of residual is the same for any value of X.
c. Independence: Observations are independent of each other.
d. Normality: For any fixed value of X, Y is normally distributed.

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

Solution:

Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

a. temp (Feature – 1)
b. winter (Feature – 2)
c. sep (Feature – 3)

General Subjective Questions

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

Solution:

Linear Regression is one of the basic & simple linear models in machine learning which uses the statistical technique of OLS (Ordinary Least squares). It attempts to model the linear relationship between Independent & Dependent variables.

Mathematically the relationship can be represented with the help of following equation –

$Y = mX + c$

Here, Y is the dependent variable we are trying to predict.

X is the independent variable we are using to make predictions.

m is the slope of the regression line which represents the effect X has on Y

c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

Assumptions in Linear Regression

Linear Relationship between the features (X) and target (Y) should Exist.

Residuals are Independent & no correlation between residuals.

Data has a Homoscedasticity feature, which has a constant variance at every level feature.

Data has to be normally distributed.

Best ways to validate the Data before implementing Linear Regression

Visual check on the Independent & Dependent Variables. Initialize the Q-Q plot to understand the assumptions such that the data is normally distributed. This satisfies the normality assumption.

Statistical Methods to understand the normality of the data points.

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

Solution:

Anscombe's Quartet was developed by statistician Francis Anscombe. It contains four data sets that have nearly identical simple descriptive statistics, have very different distributions and appear very different when graphed. Each dataset consists of eleven (x, y) points. It demonstrates both the importance of graphing data when analysing it, and the effect of outliers and other influential observations on statistical properties.

The datasets are as follows. The x values are the same for the first three datasets

Anscombe's quartet

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

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

Solution:

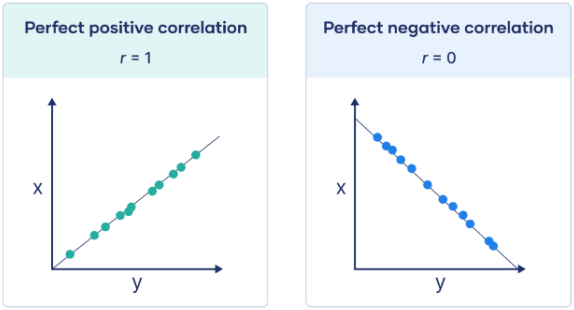
The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Correlation coefficients are used to measure how strong a relationship is between two variables. There are several types of correlation coefficient, but the most popular is Pearson's. Pearson's correlation (also called Pearson's R) is a correlation coefficient commonly used in linear regression.

Below is the table shows the relationship strengths & direction of the R score.

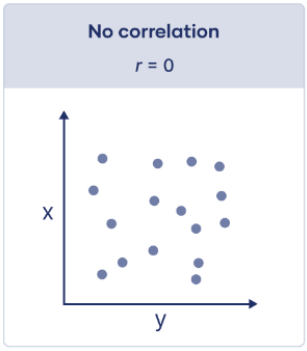
Pearson correlation coefficient (*r*) value	Strength	Direction
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

Some examples for the R Corelation.

Corelation when r is 1 or –1, all the points fall exactly on the line of best fit:



Corelation when r is 0, all the points fall exactly opposite to the best fit line.



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

Solution:

Feature Scaling is a technique to standardize or normalize the independent features present in the data to a fixed range. It is performed during the data pre-processing to handle values or units. If feature scaling is not done, then a machine learning algorithm usually have some weights handled differently, so the exact results will be hampered.

The two most popular techniques for scaling numerical data are normalization and standardization. Normalization scales each input variable separately to the range 0-1, which is the range for floating-point values where we have the most precision. Standardization scales each input variable separately by subtracting the mean (called centring) and dividing by the standard deviation to shift the distribution to have a mean of zero and a standard deviation of one.

Let’s now talk about the differences.

Normalization:

1. Minimum and maximum value of features are used for scaling
2. It is used when features are of different scales.
3. Brings data to [0, 1] or [-1, 1].
4. Outliers as affected.
5. Scikit-Learn provides solution as MinMaxScaler for Normalization.

Standardization:

1. Mean and standard deviation is used for scaling.
2. It is used when we want to ensure zero mean and unit standard deviation.
3. It's bounded in the standard deviation of the data.
4. Outliers are less affected.
5. Scikit-Learn provides solution as StandardScaler for standardization.

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

Solution:

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

When the value of VIF is infinite, it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Solution:

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

Use of Q-Q plot:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted.

Below is a visual representation of the Q-Q Plot which satisfies the linear regression assumptions: