# 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?**

Ans - The inference from each categorical variable is as follows:

- **Season** – As compared to other seasons Spring season has the least demand for shared bikes while Fall season has the most
- **Weekday** – All the weekdays have their median demand very close to each other. Thursdays and Fridays have a slightly more demand as compared to other weekdays.
- **weathersit** – There is more demand of shared bikes when the weather is clear as compared to when the weather is cloudy.
  The demand drops drastically when weather is bad I.e., when there is Light Snow/Rain and Thunderstorm and Scattered clouds
- **mnth** – The demand for shared bikes is highest in the months of June, July August and September which starts decreasing from October and is lowest in the month of January. It then starts rising from January till June
- **yr** – The demand has increase in the year 2019 as compared to 2018
- **workingday** – There is not as much of a difference in demands on a working day or on a non-working day
- **holiday** – Although the 75$^{th}$ percentile of holiday or non-holiday demands are very close, the median and 25$^{th}$ percentile demands on a holiday are much lower as compared to a non-holiday. Also, the demands on a non-holiday have a large range as compared to the range of demand on holidays.

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

Ans - drop_first=True is important to use, as it removes the extra column created while dummy variable creation which is not needed as we can describe that column the other dummy variable columns.

Example:

Let's say we have 4 different categories namely 'spring', 'summer', 'fall', 'winter' in categorical column season and we want to create dummy variables for that column.

Now each dummy variable can have values 1 or 0, and when a dummy variable has value 1 in a row all other dummy variables corresponding to that categorical column will have values 0,

So, for 4 dummy variables if 3 have the value 0 for a particular row we know for sure that the 4th dummy variable will have the value 1.

So, as we can represent the nth category by n-1 categories having 0 as their values we can use n-1 dummy variables for n categories without any loss of information.

This is the reason it is important to use drop_first=True during dummy variable creation

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

Ans - Looking at the pair-plot among the numerical variables , The 'temp' variable as well as the 'atemp' variable has the highest correlation

We have temp and atemp variables having equal correlation because they are almost perfectly correlated

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

Ans - The assumptions were validated as follows:

1) **There is a linear relationship between X and y:**

This was validated by plotting true values of y against predicted values of y

2) **Error terms are normally distributed with mean equal to zero:**

This was validated by plotting a distribution plot of residual values

3) **Error terms are independent of each other:**

This was validated using the Durbin-Watson (DW) statistic through which we can determine the autocorrelation between residual values.

For the final model it was 2.043 which indicates that the errors are independent to each other as DW = 2 indicates that there is no autocorrelation

This was also validated by plotting a scatter plot of each residual value across an index value

4) **Error terms have constant variance (homoscedasticity):**

This was validated by plotting residuals against values of y (independent) variable

5) **Multicollinearity:**

This was validated by looking at VIF which was less than 5 for each variable in final model

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

Ans - The top 3 features based on the final model are

1) “temp” variable which had a coefficient of 0.4896, meaning for each unit increase in “temp” the cnt (demand) variable increased by 0.4896

2) “weathersit_Light Snow/Rain + Thunderstorm + Scattered clouds” variable which had a coefficient of -0.2998, meaning for each unit increase in “weathersit_Light Snow/Rain + Thunderstorm + Scattered clouds” the cnt (demand) variable decreased by –0.2998

3) “yr” variable which had a coefficient of 0.2332, meaning for each unit increase in “yr” the cnt (demand) variable increased by 0.2332

# General Subjective Questions

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

Ans - Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task and builds a linear regression model explaining the relationship between a dependent variable (y) and independent variables (X) using a straight line.

It tries achieve a best fit line which explains the linear relationship between dependent variable and independent variables by minimizing the cost function of Linear Regression which is Root Mean Squared Error (RMSE) between predicted values of y (y_pred) and true values of y (y_true).

Cost function (j) for linear regression is

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

where J = Cost function

n = sample size

$pred_i$ = ith predicted value of y

$y_i$ = ith true value of y

For minimizing the cost function, it uses the Gradient Descent algorithm which optimizes the cost function (here minimizing RMSE) to reach optimal solution

While training the model we are given:

x: input training data (one or more variable)

y: Dependent variable (variable to predict using X)

When training the model – it fits the best line to predict the values of y for given values of X. The model gets the best regression fit line by finding the best Intercept and coefficients of each variables in our model.

The equation of regression line looks like:

$y = c + m_1X_1 + m_2X_2 + m_3X_3 + .... + m_nX_n$

Where, y = Independent Variable

$X_i$ = ith independent variable

$m_i$ = coefficient of ith variable

c = Intercept of regression line

Once we find the best Intercept and Coefficients of all variables in our model, we get the best fit line. So, when we are finally using our model for prediction, it will predict the values of y for the input values of X.

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

Ans – Anscombe's quartet comprises of 4 datasets having nearly identical simple descriptive statistics and yet have very different distributions and appear very different from each other when graphed. Each dataset consists of eleven (x,y) point. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties

We can plot 4 graphs having the same linear regression line for all four datasets even when they have very different data distributions which can be seen in these graphs

1) The first graph describes the linear relationship between x and y, where we can see a general correlation between two variables by observing the plotted points.

2) In the second graph where the points show a clear non-linear relationship, we can observe that we are able still fit a linear regression line here even when relationship between x and y is not linear and Pearson's R is not a relevant parameter here

3) In the 3rd graph where there is only one outlier and all other 11 points can lie on a line, we are able to observe that even though x and y have a linear relationship the fitted regression line here is incorrect and is offset by presence of outliers in the data

4) 4th graph where all 11 points line on the same x co-ordinate having different y co-ordinates and a single point lying on extreme (x,y) position, we can observe that even when other data points

show no linear relationship a high leverage point is enough to produce a high correlation coefficient which in reality is misleading.

This tells us about the importance of visualizing 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.

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

Ans - In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, is the product-moment correlation coefficient, or the bivariate correlation, which is a measure of linear correlation between datasets of two variables. It is the covariance of two variables, divided by the product of their standard deviations. So, it is essentially a normalized measurement of the covariance.

It always lies between +1 and –1.

A correlation coefficient of +1 means the two variables are perfectly correlated with a positive slope

A correlation coefficient of –1 means the two variables are perfectly correlated with a negative slope

A correlation coefficient equal to 0 indicate that the two variables have no linear correlation

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

Ans - Scaling is a process of Normalizing or Standardizing the data within a particular range so to ease the process of calculation and speed it up.

Algorithms that compute the distance between the features are biased towards numerically larger values if the data is not scaled, also when we have data with features having huge difference in magnitude the algorithm takes in account magnitude and not the units so we get an incorrect model, this is the reason scaling is performed.

Difference between Normalized and Standardized scaling:

Normalization (Min-Max Scaling):

- It transforms all the data between a range of 0 to 1,
- The new points are calculated using the formula-

$$X_{new} = \frac{(X - X_{min})}{X_{max} - X_{min}}$$

- Normalization is best when there are no outliers in data

Standardization or Z-score Normalization:

- It transforms the data to its corresponding Z-scores
- It brings all the data into a standard normal distribution with mean ($\mu$ ) = 0 and Standard deviation ($\sigma$ ) = 1
- The new points are calculated using the formula-

$$X_{new} = \frac{X - mean(X)}{sd(X)}$$

- Here the data is not restricted to a particular range
- It is much less affected by outliers

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

Ans - VIF can range from 1 to infinity depending on the value of $R^2$ of a particular variable.

The formula of VIF is given by:

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

When we have no correlation of a variable to other variables, we can get a $R^2$ equal to 0.00 and thus, VIF equals 1, but when we have a perfect collinearity between two variables, we can get a $R^2$ value for any of both variables as 1.00.

Thus VIF becomes equal to $\frac{1}{0}$ which is not defined and so in such cases we get an infinite value of VIF.

Such situations can be delt by dropping either of these perfectly related variables.

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

Ans - The Q-Q plot or the Quantile-Quantile plot is a graphical technique for determining if two datasets come from populations having a common distribution.

These are plots of quantiles of 1st dataset against quantiles of 2nd dataset which are used is to find out if two sets of data come from the same distribution

First the intervals of quantiles are chosen and the they are plotted against each other.

A point on the plot corresponds to quantile of 1st distribution plotted against the same quantile of 2nd distribution. This gives us a parametric curve with parameter which is the number of intervals for the quantiles

A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line.

**use and importance of a Q-Q plot in linear regression -**

Q-Q plots is very useful to determine If two populations are of the same distribution

If residuals follow a normal distribution, skewness of distribution, etc.

Q-Q plots helps us in a scenario where we receive training and testing datasets separately as we can confirm if both datasets are from population with same distribution