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: To analyze the effect of categorical variables on the dependent variable, we plotted a boxplot and made the following inferences

i) . The number of bikes 'cnt' have **increased with increase in year**, that is bike usage is more in 2019

ii) We note that the **demand for bikes go down with change in weather to light rain**

iii) The number of bikes **'cnt' remains almost the same for both working days and non-working days**

iv) The number of **bikes booked is more in “summer’ and “fall” seasons** compared to “winter” and “spring”, with “spring” taking the least spot.

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

When we create dummy variables for a categorical variable with n categories , without using drop_first=True, we end up having n number of dummy variables, which much result in correlation between the dummy variables, because any one dummy variable is perfectly collinear with remaining set of dummies. This needs to be avoided and hence **we need only n-1 dummy variables**. This can be achieved through drop_first=True.

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

Plotting numerical variables in pairplots, we noticed that **‘temp’ and ‘atemp**’ variables have 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?

After building the model on the training set, we did ‘**Residual analysis**’ to validate the assumptions of linear regression.

We predicted the y value for the training dataset using the training model built on the x values(independent variables). We then calculated the residual values (y_train-y_train_pred) . We then went on and visualized these residuals in a histogram and made sure that the residuals are **normally distributed** and are **centred around zero.**

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

Top 3 predictors influencing the demand of bikes are

1. With increase in temperature – ‘**temp**’, more people are likely to book bikes

2. With increase in light rain (case of light rain) – ‘**weathersit_light rain**’ , bike count has decreased.

3. The bikes usage has increased with increase in year – ‘**yr**’, that is in the later year - 2019

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Machine learning algorithms are classified into **Supervised** learning and **Unsupervised** learning methods.

Supervised learning refers to that method where **past data with labels** are used for building the model. Under supervised learning, we use **regression models** to predict output variables that are **continuous variables**.

A linear regression model attempts to find the **linear relationship between a dependent variable (target variable) and one or more independent variables (predictor variables).** This can also be explained as linear regression models try to fit the relationship between dependent and independent variable using a straight line.

If the algorithm is between one target variable and one predictor variable, it's a **Simple Linear regression** model, and if it is between one target and multiple predictors then it is **Multiple Linear Regression**.

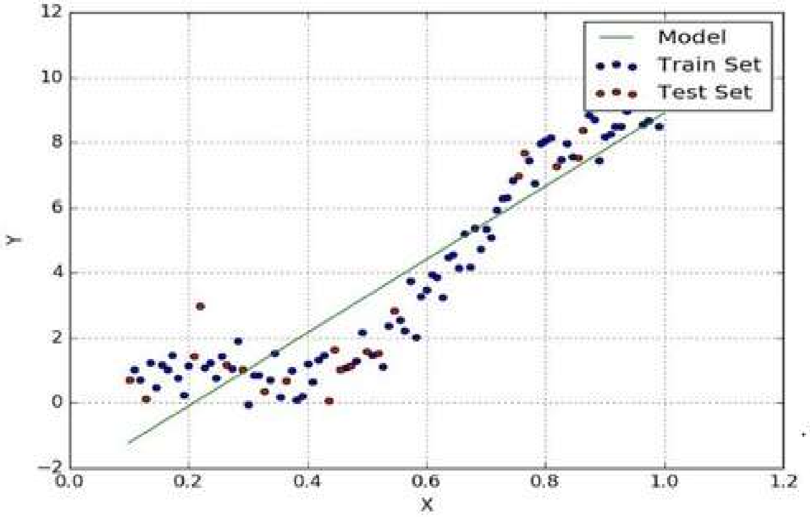
Linear regression algorithms learns and builds a model from a set of **training samples** and predicts the same model on the set of **testing samples**.

The equation for the best-fit line in a linear regression model is defined as

$$Y = \beta_0 + \beta_1 X$$

There are four assumptions associated with a linear regression model:

1. **Linearity**: The relationship between independent variables and the mean of the dependent variable is linear.
2. **Homoscedasticity**: The variance of residuals should be equal.
3. **Independence**: Observations are independent of each other.
4. **Normality**: For any fixed value of an independent variable, the dependent variable is normally distributed.



2. Explain the Anscombe’s quartet in detail.

Anscombe’s quartet comprises **four data sets in the form of x-y pairs** and each data has **eleven points**. All these four datasets have **highly similar summary statistics**, yet have **different distributions and represent altogether different graphs.**

Anscombe’s quartet tells us about the i**mportance 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?

Correlation is a method for checking the relationship between two quantitative, continuous variables. There are several types of correlation coefficient, but the most popular is Pearson’s.

Pearson's correlation coefficient (r) is a **measure of the strength of the association between the two variables**. It is also known as the “product moment correlation coefficient” (PMCC) or bivariate correlation. It is the **method of linear correlation between two sets of data.**

Pearson's correlation coefficient (r) for continuous data ranges **from -1 to +1**, where 1 indicates a strong positive relationship, -1 indicates a strong negative relationship and a result of zero indicates no relationship at all.

The most commonly used formula of Pearson’s correlation coefficient:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

4. i) What is scaling?

Scaling also known as feature scaling is a **technique to standardize the independent features/variable**s in the data to a fixed scale.

ii) Why is scaling performed?

The range of values of different features in a dataset can vary widely. If this is the case then, machine learning algorithms will not yield the desired predictions without scaling. For example, **if a feature in the dataset is big in scale compared to others then in algorithms, this big scaled feature becomes dominating over other features** and needs to be normalized.

iii)What is the difference between normalized scaling and standardized scaling?

Standardized scaling– here the features will be rescaled so that the **resulting values will have the properties of a standard normal distribution** with $\mu = 0$ and $\sigma = 1$

$$z = (x - \mu) / \sigma$$

Normalized scaling – The min-max normalized scaling is the simpler way of **rescaling the feature values to distribution values between 0 and 1**. It is similar to z-score normalization in that it will replace every value in a column with a new value using a formula. In this case, that formula is:

$$z = (x - x_{min}) / (x_{max} - x_{min})$$

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

VIF (Variance Inflation Factor) is used to detect the presence of multicollinearity between variables in a multiple regression model. VIF explains the extent of correlation between one predictor variable and other predictors in the model.

When VIF is infinity it means that the corresponding variable has an exact linear relationship with other predictor variables(whose VIF is also infinity).

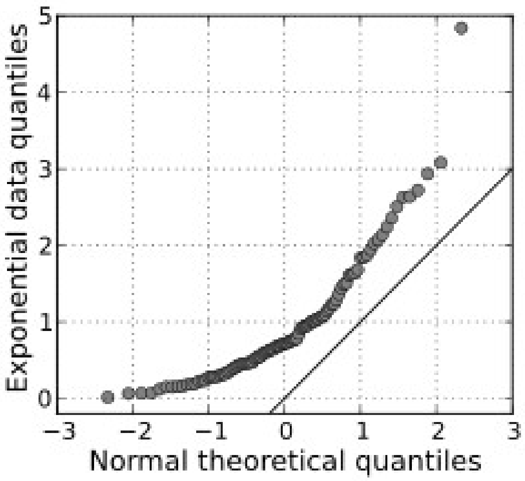
Also, VIF is also defined by VIF = 1 / (1-R2).

So if R2 reaches 1, VIF reaches infinity. So we can conclude that when a linear regression model is overfit, and **If there is perfect correlation between two predictor variables, then VIF = infinity.**

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

A Quantile-Quantile (Q-Q) plot is a **scatter plot created by plotting two sets of quantiles against one another, usually the quantiles of the first dataset against quantiles of second dataset**. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it.

The purpose of Q Q plots is **to find out if two sets of data come from the same distribution**. It lets you compare how close two distributions are, and is often used to assess normality in linear regression. That is, in a scenario of linear regression when we have training *and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.*



It can be interpreted as :

1. ***Similar distribution**: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x –axis*
2. ***Different distribution:** If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis*