ASSIGNMENT BASED SUBJECTIVE QUESTIONS.

1.)

There are 6 categorical variables in the dataset.

We used Box plot to study their effect on the dependent variable (‘cnt’).

The inference that we can derive are:

1.) **Season:** Almost 32% of the bike booking were happening in season3 with a median of over 5000 booking (for the period of 2 years). This was followed by season2 & season4 with 27% & 25% of total booking. This indicates, season can be a good predictor for the dependent variable.

2.) **mnth** : Almost 10% of the bike booking were happening in the months 5,6,7,8 & 9 with a median of over 4000 booking per month. This indicates, **mnth** has some trend for bookings and can be a good predictor for the dependent variable.

3.) **weathersit** : Almost 67% of the bike booking were happening during ‘weathersit1 with a median of close to 5000 booking (for the period of 2 years). This was followed by weathersit2 with 30% of total booking. This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable. Holiday: Almost 97.6% of the bike booking were happening when it is not a holiday which means this data is clearly biased. This indicates, holiday CANNOT be a good predictor for the dependent variable.

4.) **weekday**: weekday variable shows very close trend (between 13.5%-14.8% of total booking on all days of the week) having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor. I will let the model decide if this needs to be added or not.

5.) **workingday** : Almost 69% of the bike booking were happening in ‘workingday’ with a median of close to 5000 booking (for the period of 2 years). This indicates, workingday can be a good predictor for the dependent variable.

2.)

- **drop_first=True** is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

- 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 furnished and semi_furnished, then It is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished.

3.)

The numerical variable ‘registered’ has the highest correlation with the target variable ‘cnt’, if we consider all the features. But after data preparation, when we drop registered due to multicollinearity the numerical variable ‘atemp’ has the highest correlation with the target variable ‘cnt.

4.)

Linear regression captures only linear relationship. This was validated by plotting a pair plot between the features and the target.

The Pair-Plot tells us that there is a LINEAR RELATION between 'temp','atemp' and 'cnt'

Multicollinearity is a state of very high inter-correlations or inter-associations among the independent variables. It is therefore a type of disturbance in the data if present weakens the statistical power of the regression model. Pair plots and heat maps (correlation matrix) can be used for identifying highly correlated features.

The heatmap clearly shows which all variable are multicollinear in nature, and which variable have high collinearity with the target variable.

We refered this heatmap back-and-forth throughout the project while building the linear model so as to validate different correlated values along with VIF & p-value, for identifying the correct variable to select/eliminate from the model.

5.)

As per our final Model, the top 3 predictor variables that influences the bike booking are:

- **Temperature (temp)** - A coefficient value of ‘0.5636’ indicated that a unit increase in temp variable increases the bike hire numbers by 0.5636 units.
- **Weather Situation 3 (weathersit_3)** - A coefficient value of ‘-0.3070’ indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.3070 units.
- **Year (yr)** - A coefficient value of ‘0.2308’ indicated that a unit increase in yr variable increases the bike hire numbers by 0.2308 units.

So, it's suggested to consider these variables utmost importance while planning, to achive maximum Booking

The next best features that can also be considered are

season_4: - A coefficient value of ‘0.128744’ indicated that w.r.t season_1, a unit increase in season_4 variable increases the bike hire numbers by 0.128744 units.

windspeed: - A coefficient value of ‘-0.155191’ indicated that, a unit increase in windspeed variable decreases the bike hire numbers by 0.155191 units.

General Subjective Questions

1.)

Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It's used to predict values within a continuous range, (e.g. sales, price) rather than trying to classify them into categories (e.g. cat, dog). There are two main types:

Simple regression

Simple linear regression uses traditional slope-intercept form, where mm and bb are the variables our algorithm will try to "learn" to produce the most accurate predictions. X represents our input data and Y represents our prediction.

y=mx+b

Multivariable regression

A more complex, multi-variable linear equation might look like this, where ww represents the coefficients, or weights, our model will try to learn.

f(x,y,z)=w1x+w2y+w3zf(x,y,z)=w1x+w2y+w3z

The variables x,y,zx,y,z represent the attributes, or distinct pieces of information, we have about each observation. For sales predictions, these attributes might include a company's advertising spend on radio, TV, and newspapers.

Sales=w1Radio+w2TV+w3News

2.)

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

This tells us about the importance of visualising 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. Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

3.)

In statistics, the Pearson correlation coefficient also known as **Pearson's r**, the Pearson product-moment correlation coefficient (PPMCC), the bivariate correlation or colloquially simply as the correlation coefficient is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation. As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1.

4.)

Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like **t-statistic, F-statistic, p-values, R-squared**, etc.

Normalization (also called, Min-Max normalization) is a scaling technique such that when it is applied the features will be rescaled so that the data will fall in the range of [0,1]

Normalized form of each feature can be calculated as follows:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

The mathematical formula for Normalization

Here ‘x’ is the original value and ‘x`’ is the normalized value.

Standardization (also called, **Z-score normalization**) is a scaling technique such that when it is applied the features will be rescaled so that they’ll have the properties of a standard normal distribution with **mean,*μ=0* and standard deviation, *σ=1***; where *μ* is the mean (average) and *σ* is the standard deviation from the mean.

Standard scores (also called ***z** scores*) of the samples are calculated as follows:

$$z = \frac{x - \mu}{\sigma}$$

The mathematical formula for Standardization.

This scales the features in a way that they range between [-1,1].

In clustering analyses, **standardization** may be especially crucial in order to compare similarities between features based on certain distance measures. Another prominent example is the Principal Component Analysis, where we usually prefer standardization over normalization since we are interested in the components that maximize the variance.

However, this doesn’t mean that **normalization** is not useful at all! A popular application is image processing, where pixel intensities have to be normalized to fit within a certain range (i.e., 0 to 255 for the RGB color range). Also, a typical neural network algorithm requires data on a 0–1 scale.

5.)

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

6.)

Quantile-Quantile (**Q-Q**) **plot**, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps 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.

Q Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. 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. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.