Assignment Based Subjective Questions

1. The inference that We could derive were:

a) 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, seaso n can be a good predictor for the dependent variable.

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

c) weathersit: Almost 67% of the bike booking were happening during 'weathersit1 with a median of close to 5000 b ooking (for the period of 2 years). This was followed by weathersit2 with 30% of total booking. This indicates, weat hersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.
d) holiday: Almost 97.6% of the bike booking were happening when it is not a holiday which means this data is clea rly biased. This indicates, holiday CANNOT be a good predictor for the dependent variable.
e) weekday: weekday variable shows very close trend (between 13.5%-14.8% of total booking on all days of the wee k) having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence to wards the predictor. I will let the model decide if this needs to be added or not.
f) workingday: Almost 69% of the bike booking were happening in 'workingday' with a median of close to 5000 bo oking (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 creatio n. Hence it reduces the correlations created among dummy variables.

3.Temp has the highest correlation with the target variable that is cnt .

4. After building the model of the training set I came to a assumption of good linear regression model by considering the values of R square , VIF and P-values of the model .
The P values for all the variables in the model is very less and R square of the model is high . VIF values of the varia bles are low as per the standards . This gave a conclusion that the model is good .

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

a) Temperature (temp) - A coefficient value of '0.5644' indicated that a unit increase in temp variable increases the b ike hire numbers by 0.5644 units.

b) Weather Situation 3 (weathersit_3) - A coefficient value of '-0.3071' indicated that, w.r.t Weathersit1, a unit incre ase in Weathersit3 variable decreases the bike hire numbers by 0.3071 units.

c)Year (yr) - A coefficient value of '0.2303' indicated that a unit increase in yr variable increases the bike hire numb ers by 0.2303 units. So, it's suggested to consider these variables utmost importance while planning, to achive maxi mum Booking The next best features that can also be considered are

d)season_4: - A coefficient value of '0.1292' indicated that w.r.t season_1, a unit increase in season_4 variable incre ases the bike hire numbers by 0.1292 units. windspeed: - A coefficient value of '-0.1542' indicated that, a unit increa se in windspeed variable decreases the bike hire numbers by 0.1542 units.

General Subjective Questions

1. Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Re

gression models a target prediction value based on independent variables.
It is mostly used for finding out the relationship between variables and forecasting. Different regression models diffe r based on – the kind of relationship between dependent and independent variables, they are considering and the num ber of independent variables being used.
Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output).
Hence, the name is Linear Regression. y = B1 + B2.x
B1 - Coefficient or point where the straight line passes the Y axis
B2 - Slope of the sraight line

While training the model we are given :
x: input training data (univariate – one input variable(parameter))
y: labels to data (supervised learning)

When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best B1 and B2 values.

Once we find the best B1 and B2 values, we get the best fit line. So when we are finally using our model for predicti on, it will predict the value of y for the input value of x.

2. Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive st atistics, 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.

3. Pearsons R is a statistical measure of the linear correlation between two variables .Its value ranges from -1.0 to 1.0 .

4. Scaling is used to bring the variables in the dataset to a similar scale and reduce the difference in the values betwe en the Numerical variables .

5. If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent varia bles. 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.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 f all on that reference line.