Subjective Questions

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?

Explanations:

a. Season: Fall has highest count followed by summer and winter.
b. Weather sit: Count is max in Clear weather
c. Working day: Count avg is same on working day
d. Holiday: Count is less on holidays and vice-versa
e. Month: May, June, July, August, September months have highest no of bookings.
f. Year: Significant Increase in 2019

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

Explanations:

drop_first=True should be used during dummy variable creation as it reduces extra columns created during dummy variable creation. Hence reducing the correlation between the dummy variables.

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

Explanations:

The **temp** and **atemp** variables have highest correlation with the target variable which comes up to be **0.63.**

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

Explanations:

After building the model on the training dataset the model, assumptions were compared with the test data set. They were found to be close. This is how we validate the assumptions.

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

Explanations:

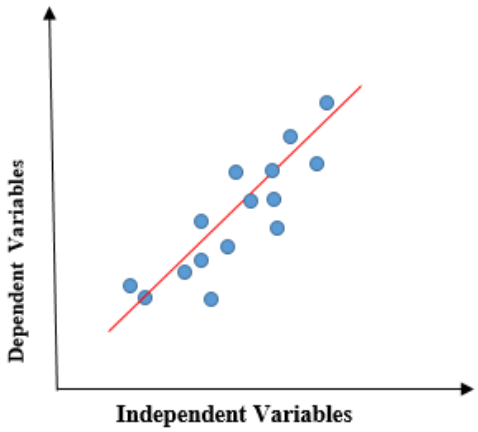
The top 3 features contributing significantly towards explaining the demand of the shared bikes are:
a. Temperature
b. Weather
c. Months (mid-year may to sept)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Explanations:

a. Linear Regression is a Machine Learning Algorithm which predicts a dependent variable value.
b. It is based on Supervised Learning.
c. It predicts a dependent variable value based on the independent variable value.



d. Here dependent var VS independent vat, the red line is the line of regression.
e. Basically it finds out the linear relationship between the two variables

2. Explain the Anscombe's quartet in detail.

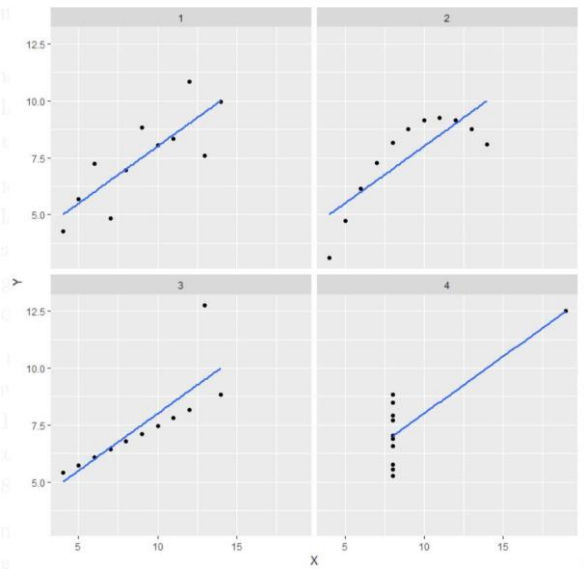
Explanations:

a. Anscombe's quartet comprises of four dataset that have nearly same dataset but appear different when ploted.
b. For example, data is

Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817

c. Graph can be,



d. In some cases, there appears a linear relationship between the variables but in some cases the there is no relation
e. It is used to see the data with different perspectives.

3. What is Pearson's R?

Explanations:

a. Pearson Correlation Coefficient, the Pearson product-moment correlation coefficient (PPMCC).
b. It measures the linear correlation between two variables.
c. It is used to identify patterns and the strength of the modal.
d. The correlation coefficient will be positive, If the variables go up and down together.
e. If one of the variable go too down or too up then, the correlation coefficient will be negative.

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

Explanations:

a. Scaling is a step applied to independent data to normalise the data within a range. It also helps to speedup calculations.
b. Most data is having varying data values. If modelling is done on the data set the modelling will be done wrong and we will have a bad model. To solve this issue we do the scaling.
c. It does not accept any other coefficients.
d. In Normalization scaling the data is rescaled in the range of 0 and 1
e. In Standardized scaling the data is rescaled has a mean of 0 and SD of 1.

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

Explanations:

a. If there is a Perfect correlation. VIF is infinity.
b. Mathematically, Perfect correlation means R=1 and VIF = 1/(1-R) which brings the infinity value.
c. If we need to solve this issue we can drop the column with is creating perfect correlation.

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

Explanations:

a. Q-Q plot is plotting two quantiles against each other.
b. It is used to compare the shapes of distribution, provide graphical view, scale, skewness are similar or different in 2 distribution.
c. It can be used to sample size.
d. It is used to check; if they have similar behaviour, similar distribution shapes, come from population with a common distribution.