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?

- Booking is drastically increased in 2019
- Weekends seems to have a greater number of bookings
- Based on the barplots large number of bookings done in clear weather
- Based on the barplots Thursday, Friday, Saturday has more number of bookings
- Based on the barplots large number of bookings is done at Clear weather situations
-

2. Why is it important to use drop_first=True during dummy variable creation?
Ans: When we create dummy variables, creates additional columns, hence to delete such additional columns drop_first is used
Ex:

```
df_bike =
pd.get_dummies(data=df_bike,columns=["season","mnth","weekday","weathersit"],drop_first=
True)
```

3.Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Ans: Temp variable, same has been proved in the note book

4. How did you validate the assumptions of Linear Regression after building the model on the training set
Ans:
- Linear relationship validation
- Multicollinearity Validation

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

Ans:
- Temperature
- Season
- Year

General Subjective Questions

1) Linear Regression

Is a statistical regression method used for predictive analysis which shows relationship between independent variable(X) and the dependent variable (Y).

Simple linear regression:
If there is a single input variable (x), such linear regression is called simple linear regression.
multiple linear regression
if there is more than one input variable, such linear regression is called multiple linear regression.
Example: bike sale prediction based on holiday, season temperature etc

Linear regression can be visualised using the sloped line

Mathematically these sloped lines follow the following equation,

Y = m*X + b

Where X = dependent variable (target)

Y = independent variable

m = slope of the line

2) Explain the Anscombe's quartet in detail. (3 marks)

It consists of 4 datasets, each contains 11, x and y pairs.

This has been identified by Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers.

3) What is Pearson's R?

This explains linear association between variables.
Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, where:

Advantages:

1. It helps in knowing how strong the relationship between the two variables is.
2. We can identify correlation between 2 variable. Negative or positive

Disadvantages:

1. This method takes much time to arrive at the results.
2. Slope information about the line won't be identifies with this approach.

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

Refers to putting the features values to same range.
In few cases we might wanted to give the same weightage for both the features in such case scaling is applied.

In normalization, we map the minimum feature value to 0 and the maximum to 1. Hence, the feature values are mapped into the [0, 1] range:
In standardization, we don't enforce the data into a definite range. Instead, we transform to have a mean of 0 and a standard deviation of 1
In general, standardization is preferred option as compared to normalization.