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? (3 marks)

Ans:

The categorical variables from the dataset are

“season”, “workingday”, “weathersit”, “yr”, and “mnth”.

1. Based on the dataset, Fall and summer seasons are the favourable seasons for biking.
2. Booking is equal on working or non-working day
3. Clean weather attracts more booking
4. Based on the given 2 year data set , we could see that there is increase in the bike from 2018 -19.
5. During May, June,July,Aug and Sept , most of the bookings are done.

---

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

Ans:

This is important to use drop_first=True,as it helps in reducing the extra column creation during dummy variable creation.

---

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

Ans:

The numerical variable which has the highest correlation with the target variable is “Temp”

---

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

Ans:

1.Error Terms – Normally distributed.

2.Error Terms – Constant variance

3.Error terms - Independent.

4. Linear relationship – Dependent and Independent variables

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

Ans:

Based on the final model , the top 3 features contributing significantly towards explaining the demand of the shared bikes as follows:

- Temp
- Year (yr)
- Season

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans:

It is one of the machine learning algorithms and it is a statistical method which is used to do predictive analysis. Linear regression makes predictions for continuous / real or numeric variables.

Linear regression algorithm shows a linear relationship between a dependent and one or more independent variables, so, it is called as linear regression. Since it is a linear regression, it finds how the value of dependent variable is changing according to the value of independent variable.

There are 2 types of linear regression:

1. Simple Linear regression – Single Independent variable is used
2. Multiple Linear regression – Multiple independent variables are used

Linear regression Line : A linear line which shows the relationship between the dependent and independent variables is called as Linear regression line.This line can show 2 types of relationship.

1.Positive Linear Relationship: If dependent variable value increases , then independent variable value also increase

2.Negative Linear Relationship:if the dependent variable decreases , then independent variable

Value increases.

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

Ans:

Anscombe's quartet says about the importance of visualizing data before applying various algorithms to build models. This suggest that features must be plotted to see the distribution of samples that can help us to identify the anomalies present in the data.

By introducing a group of data sets that are identical with statistical properties to illustrate facts

---

3. What is Pearson's R? (3 marks)

Ans:

The Pearson's R measures the strength between the different variables and the relation with each other. This returns value between -1 and 1.

1. -1 coefficient – Strong inversely proportional relationship
2. 0 coefficient – No relationship
3 1 coefficient – strong directly proportional relationship

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

Ans:

1. It is a data preparation step applied to an independent variable to normalize data within the range.

2. Most of the times, collected data set contains features highly varying in magnitude, units and range. If Scaling is not done, then the algorithm only magnitude in account and not units which results in incorrect modelling. To solve this problem, scaling should be performed to bring all variables to the same level of magnitude.

SNO	Normalised Scaling	Standardized scaling
1	Scale values between [0,1] or [-1,1]	There is no certain range
2	Much Affected by outliers	Less affected by Outliers
3	Max and Min values of features are used for scaling	Mean and Std deviation are used for scaling

5. You might have observed that sometimes the value of VIF is

infinite. Why does this happen?

Ans:

VIF = 1/(1-R2)

If R2 (R- Square) is 1 , then value of VIF will be infinite. If there is a perfect correlation between two independent variables , then value of R2 to be 1.

---

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

Ans:

Q-Q Plots – Quantile -Quantile plot, is a graphical tool. It is used in helping assess a data which came from Normal, exponential or uniform distribution. It helps in determining the data sets come from populations with common distribution.

Let us say if we have received training and test data separately. We can use Q-Q plot to confirm the sets are from population with same distribution.

Advantages:

1. It is used to check if two data sets came from populations with common distribution
2. It is used to check if two data sets have common location and sale
3. It is used to check if two data sets have similar distributional shapes
4. It can be used with sample sizes as well.