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?

Ans: In our bike dataset the categorical variables are**:** season , month , weathersit , weekday.

Inferences:

1. The renting of bike increases in the Fall season and is lowest in the sring season.
2. There is increase in bike renting in 2019 as compared to 2018.
3. We could observe bike rented are more for sep in 2019 and jun in 2018
4. Bike rentals are high on thursday,Friday. i.e they are high on working days.
5. on days with clear weather renting of bikes were more

---

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

Ans: drop_first=True is important as it reduces the extra column that is created while dummy variable creation. It reduces the corelations between the variables.

Eg. We have month jan to dec, here if month is not feb,march,apr…dec then it implies that it is jan.

Another example of furnished ,semi furnished, un furnished, if the flat is not semi furnished and un furnished it implies that it is furnished.

---

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

Ans: Looking at the pair plot, temp and atemp shows the highest positive correlation with the target variable count.

---

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

Ans: For validating the Model we check R squared value and adjs. R Squared value of train and test data set.

The value lies between 0 to 1. Higher the value greater the variance.

The difference between train and test data set’s R squared value and adjs. R Squared value should be less than 0.05 or 5%.

We also check for F – statistics: Greater the value of F -statistics and lover the value of prob(F-stats) signifies that the model is more significant

---

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

Ans: The three most significant variables affecting the demand for shared bikes are :

Temperature

Year 2019

Winter

as these features are having positive coefficients and an increase in them is going to result into an increase in the demand for shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: It s a machine learning algorithm based on supervised learning. It finds the best linear-fit relationship for give target variable and predictor variable.

Linear regression is of the 2 types:

Simple Linear Regression: It explains the relationship between a dependent variable and only one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points.

Formula for the Simple Linear Regression:

$$Y=\beta 0+\beta 1X1 +\epsilon$$

Multiple Linear Regression: It shows the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X. It fits a 'hyperplane' instead of a straight line.

Formula for the Multiple Linear Regression:

$$Y=\beta 0+\beta 1X1+\beta 2X2+...+\beta pXp+\epsilon$$

---

2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quarter comprises of four data set that have nearly identical statistical properties yet look very different when graphed. The dataset consists of 11 data points and they all share the same descriptive statistics(mean, variance, standard deviation etc) but different graphical representation.

As studied :

Data-set I — consists of a set of (x,y) points that represent a linear relationship with some variance.

Data-set II — shows a curve shape and is non-linear relationship.

Data-set III — looks like a tight linear relationship between *x* and *y*, except for one large outlier.

Data-set IV — looks like the value of *x* remains constant, except for one outlier as well.

Summary statistics can be mis leading sometimes and thus using graphical representations can help in identifying such datasets. This illustrates the use of visualization to revisit our summary statistics and re-contextualize them as needed.

---

3. What is Pearson's R?

Ans: Pearson's R or Pearson's Corelation coefficient is the measure of strength of relationship between two variables. That is, it calculate the effect of change in one variable if the other variable change.

The Pearson's r range between -1 to 1, where -1 shows negative linear corelation, 0 shows no liner corelation and 1 shows positive corelation.

Mathematically, Pearson's correlation coefficient is denoted as the covariance of the two variables divided by the product of their standard deviations.

Eg. Relation between temperature and ice cream sale.

Relation between age and height.

---

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

Ans: Scaling is applied to independent variable to normalize the data in particular range. Most of the times data collected varies in unit and range, this may result into incorrect modelling. So we use scaling to bring them all to a single range or unit.

There are two types of scaling:

1. Normalized /Min-Max scaling.
2. Standardized Scaling.

Normalized / Min -Max Scaling:

It brings all of the data in the range of 0 and 1.

Standardized Scaling:

*Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ**) zero and standard deviation one (**σ**).*

---

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

Ans: The value of VIF is calculated as:

VIF : 1/(1-R2)

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. This shows a perfect corelation between the variables.

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

Ans: Q-Q plot are known as Quantile – Quantile plots, and plots the quantile of sample distribution against the quantile of the theoretical distribution. It helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential. It determines how many values in a distribution are above or below a certain limit(quantile).

Eg. The median is a quantile where 50% of the data fall below that point and 50% lie above it.

The purpose is to find if two set of data comes from similar distribution.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.