Assignment-based Subjective Questions and Answers

- Chaithra R

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:

- There is an increase in the bike rental count in spring and summer seasons , and then a decrease in the bike rental count in fall and winter season.
- The demand for rental bikes increased in the year 2019 when compared with the year 2018.
- Month Jun to Sep is the period when bike demand is high. The Month Jan is the lowest demand month.
- Bike demand is less in holidays in comparison to when not being holidays.
- The demand for rental bikes is almost similar throughout the weekdays.
- There is no significant change in bike demand with working days and non working days.
- During clear, partly cloudy weather, the bike rental count is the highest, second-highest during misty cloudy weather, and followed by 3rd highest, during light snow and light rain weather.

---

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

Ans:

drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

---

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans:

- The temp and atemp variables are highly positively correlated to each other, it means that both are carrying the same information.
- The total_count,casual and registered are highly positively correlated to each other.

---

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans: We used R-squared , or Coefficient of Determination , which is 0.81 on average in our case . It means that the predictor is only able to predict 81% of the variance in the target variable which is contributed by independent variables.

---

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans:

- Temperature
- Month_8(august)
- Year

General Subjective Questions and Answers

1. Explain the linear regression algorithm in detail. (4 marks)

Ans: Linear regression is one of the very basic forms of machine learning where we train a model to predict the behavior of your data based on some variables. In the case of linear regression , as you can see the name suggests linear, that means the two variables which are on the x-axis and y-axis should be linearly correlated.
Mathematically, we can write a linear regression equation as:

$$y = a + bx$$

Where a and b given by the formulas:

$$b(slobe) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a(intercept) = \frac{n\sum y - b(\sum x)}{n}$$

Here, x and y are two variables on the regression line.
b = Slope of the line
a = y-intercept of the line
x = Independent variable from dataset
y = Dependent variable from dataset

---

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

Ans: Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fool the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly the same statistical observations, which

provide the same statistical information that involves variance, and mean of all x,y points in all four datasets.

---

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

Ans: In statistics, the Pearson correlation coefficient ¯ also known as Pearson's R, the Pearson product-moment correlation coefficient, the bivariate correlation, or colloquially simply as the correlation coefficient ¯ is a measure of linear correlation between two sets of data.

Formula:

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where:

- N = number of pairs of scores
- Σxy = sum of the products of paired scores
- Σx = sum of x scores
- Σy = sum of y scores
- Σx^2 = sum of squared x scores
- Σy^2 = sum of squared y scores

---

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans:

- Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm**.**
- Most of the times, the collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then the algorithm only takes magnitude in account and not units hence incorrect modeling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. Scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

- Normalized / Min-Max scaling vs Standardized scaling:
  Normalized: It brings all of the data in the range of 0 and 1.
  sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

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**). **sklearn.preprocessing.scale** helps to implement standardization in Python.

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

---

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans: If there is perfect correlation, then VIF (Variance Inflation Factor) = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which leads 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. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Ans:

- 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 fall on that reference line.
