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)

Answer
the categorical variables from the dataset are season, month, weekday, and weathersit

Season like spring has least count and fall has high count of users.

Count of bike rentals increased and became popular in year 2019 than 2018

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

Answer

Because there is no reason to have let's say n variables for n number of categories if the dummy variable can be interpreted with the first variable. That is only n-1 number of variables are required to interpret the dummy variable.
And this is done to avoid multicollinearity and redundant features
Example:
dummy variables
a=pd.get_dummies(day["season"], drop_first = True)
b=pd.get_dummies(day["month"], drop_first = True)
c=pd.get_dummies(day["weekday"], drop_first = True)
d=pd.get_dummies(day["weathersit"], drop_first = True)

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

Answer

the target variable) has high correlation with temperature (temp).

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

Answer

Calculate the r square for test
comparing with R_squared = 0.830

Following assumptions are validated mathematically and by plotting charts.

Linear relationship between X and y.

Normal distribution of error terms.

Independence of error terms.

Constant variance of error terms.

cnt = const×0.1259+year× 0.2329 -holiday×0.0987 + temp×0.5480 - windspeed×0.1532 + summer×0.0881 + winter×0.1293 + sep × 0.1012 -Light_snowrain ×0.2829 - Misty×0.0784

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

Answer
the target variable came out to be the function of following independent variables

cnt = const×0.1259+year× 0.2329 -holiday×0.0987 + temp×0.5480 - windspeed×0.1532 + summer×0.0881 + winter×0.1293 + sep × 0.1012 -Light_snowrain ×0.2829 - Misty×0.0784

So the top 3 features contributing significantly towards explaining the demand of the shared bikes
1.temp
2. year
3. const

General Subjective Questions

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

Linear regression models can be classified into two types depending upon the number of independent variables:

- **Simple linear regression:** This is used when the number of independent variables is 1.
- **Multiple linear regression:** This is used when the number of independent variables is more than 1
- The equation of the best fit regression line $Y = \beta_0 + \beta_1 X$ can be found by minimising the cost function (RSS in this case, using the ordinary least squares method), which is done using the following two methods:
    - **Differentiation**
    - **Gradient descent**
- The strength of a linear regression model is mainly explained by R^2, where $R^2 = 1 - (RSS/TSS)$.
    - **RSS:** Residual sum of squares
    - **TSS:** Total sum of squares

Assumptions of simple linear regression

- Linear relationship between X and y.
- Normal distribution of error terms.
- Independence of error terms.
- Constant variance of error terms

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

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 **fools the**

regression model if built. They have very different distributions and **appear differently** when plotted on scatter plots.

Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

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

Correlation measures the strength of association between two variables as well as the direction. There are mainly three types of correlation that are measured. One significant type is Pearson's correlation coefficient. This type of correlation is used to measure the relationship between two continuous variables.

In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

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

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

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

- Scaling of data is converting and fitting the data in some particular range

- without scaling there will be large as false variances in coefficient which will be a detriment for any model.

Types

1. Normalization
2. Standardizatio

1. Minimum and maximum value of features are used for scaling Mean and standard deviation is used for scaling.
2. It is used when features are of different scales. It is used when we want to ensure zero mean and unit standard deviation

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

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead 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
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.A Q Q plot showing the 45 degree reference line. - If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions. - 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