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)

Categorical variables found out –

- spring
- Jan
- summer
- Nov
- Sep
- windspeed
- Dec
- light precipitation
- Holiday - people will still prefer to rent a bike during holiday but that will not be major factor behind renting.
- cloudy
- 2019
-

Weather conditions are more likely to determine the rental .. spring season would be best combination for rental of bikes

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

We need to minimise the number of variables to use for prediction. Frop_first = True gives us an option to drop one column variable for sure. This is because one column value can be correctly determined by other columns value combinations.

Example - a variable season (summer, winter, spring, autumn)

summer	winter	spring	autumn
0	0	0	1
0	0	1	0
0	1	0	0
1	0	0	0

We can safely remove autumn column as summer, winter, spring any column having value 1 will mean autumn 0 and all values 0 will mean autumn value 1…

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

Temperature (atemp) has the highest correlation with cnt

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

a. The result y_test and y_pred scatter plot was almost with slope 1, with intercept around 0
b. The error distribution was close to a normal distribution with mean centered around 0

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

a. Spring season
b. January month
c. Summer season
d. November month

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
    a. It is a mechanism showing relationship between a dependent variable and one or more independent variables
    b. Example - rental sale of bikes dependency on external factors like weather, holiday, season etc
    c. Goal is to find a best fitting line that minimizes the difference between predicted and actual dependent variable value, if we formulate the dependent variable to be m times x where x is independent variable and m is the linear coefficient
    d. Steps involved in linear regression are -
        i. Data preprocessing (EDA, data cleaning etc done here)
        ii. Splitting the data to test and train dataset
        iii. Model representation
        iv. Model training
        v. Model evaluation - iterate to find the best fit model using manual or automated strategies
        vi. Model optimisation
        vii. Prediction
    e. Finally the validation is done by checking the test data values and predicted values

2. Explain the Anscombe's quartet in detail.(3 marks)
    a. It is a collection of four datasets that indicate that even if summary statistics are the same for those 4 datasets, the patterns and relationships among the variables in the datasets are very different when plotted. The quartet emphasizes on the importance of visualizing data and not solely relying on summary statistics.
    b. Consider following example (not created personally)
    Dataset I:
    x | y = 10.0 | 8.04 , 8.0 | 6.95 , 13.0 | 7.58 , 9.0 | 8.81 , 11.0 | 8.33 , 14.0 | 9.96 , 6.0 | 7.24 , 4.0 | 4.26 , 12.0 | 10.84 , 7.0 | 4.82 , 5.0 | 5.68
    Dataset II:
    x | y = 10.0 | 9.14 , 8.0 | 8.14 , 13.0 | 8.74 , 9.0 | 8.77 , 11.0 | 9.26 , 14.0 | 8.10 , 6.0 | 6.13 , 4.0 | 3.10 , 12.0 | 9.13 , 7.0 | 7.26 , 5.0 | 4.74
    Dataset III:
    x | y = 10.0 | 7.46 , 8.0 | 6.77 , 13.0 | 12.74 , 9.0 | 7.11 , 11.0 | 7.81 , 14.0 | 8.84 , 6.0 | 6.08 , 4.0 | 5.39 , 12.0 | 8.15 , 7.0 | 6.42 , 5.0 | 5.73
    Dataset IV:
    x | y = 8.0 | 6.58 , 8.0 | 5.76 , 8.0 | 7.71 , 8.0 | 8.84 , 8.0 | 8.47 , 8.0 | 7.04 , 8.0 | 5.25 , 19.0 | 12.50 , 8.0 | 5.56 , 8.0 | 7.91 , 8.0 | 6.89
    c. Although each dataset has the same means, variances, correlations, and regression lines, when plotted, they reveal very different relationships. This illustrates the limitations of relying solely on summary statistics and emphasizes the importance of visualizing the data to understand its patterns and relationships.
    d. Here, dataset I shows a roughly linear relationship, Dataset II exhibits a curvilinear relationship, Dataset III has an outlier that influences the linear regression line, and Dataset IV has a strong relationship except for an outlier that significantly impacts the line. These differences highlight the need to explore and interpret data graphically in addition to relying on numerical summaries.

3. What is Pearson's R? (3 marks)
    a. It is plainly called as correlation coefficient
    b. It indicates how one dependent variable value varies with variation in value of independent variables
    c. If the value of correlation coefficient increases with increase in value of dependent variable, coefficient is +ve
    d. If the value of correlation coefficient decreases with increase in value of dependent variable, coefficient is -ve
    e. If there is no observed change, coefficient will be close to zero
    f. the magnitude of the correlation coefficient indicates the strength of the relationship:
        i. Values close to 1 or -1 indicate a strong linear relationship.
        ii. Values close to 0 indicate a weak or no linear relationship.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
    a. Scaling refers to the transformation of variables to a specific range or distribution. It involves adjusting the values of variables to ensure that they are comparable and do not bias the analysis or the performance of certain algorithms. Scaling is performed to standardize the variables and bring them to a similar scale, making them easier to interpret and compare. Without scaling, the coefficient of higher value parameters tend to dominate the model while this actually may not be the case
    b. Normalized scaling transforms the values of variables to a specific range, typically between 0 and 1. It is done by subtracting the minimum value of the variable and then dividing by the range (maximum value minus minimum value). The formula for normalized scaling is: x_normalized = (x - x_min) / (x_max - x_min)
    c. Standardized scaling transforms the values of variables to have a mean of 0 and a standard deviation of 1. It is done by subtracting the mean of the variable and then dividing by the standard deviation. The formula for standardized scaling is: x_standardized = (x - μ) / σ
    d. Normalised scaling gives a better understanding in our current course session… If the absolute values and proportions are important, normalized scaling is often used. On the other hand, if the distribution and relative positions of the variables are of interest, standardized scaling is preferred.
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
    a. VIF will be infinite when one or more of the independent variables can be expressed as a perfect linear combination of the other independent variables in the model. This situation is known as perfect multicollinearity. In perfect multicollinearity, one or more variables can be predicted exactly using a combination of the other variables, resulting in an infinite VIF.
    b. For example, lets consider two variables X1 and X2, and the following relationship: X2 = 3*X1. In this scenario, X2 can be expressed as a perfect linear combination of X1, resulting in perfect multicollinearity. The VIF for X2 would be infinite in this case. It is important to detect and address multicollinearity issues in regression analysis to ensure reliable and meaningful results.
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
    a. The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.
    b. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. Quantile means the fraction (or percent) of points below the given value. Example, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.
    c. We can draw a 45-degree reference line. If the two sets come from a population with the same distribution, the points should fall approximately along this

reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

d. The advantages of the q-q plot are:
    i. The sample sizes do not need to be equal.
    ii. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.