Assignment-based Subjective Questions

Question-1

From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

1) **holiday**

- The demand for bikes was higher when there were no holidays, as we saw in the univariate analysis section. As a result, the boxplot displays that the lower quartile value of the number of bike rentals during the non-holiday period is roughly equivalent to the median of the bike rentals during the holiday period. Hence, in the event of holidays, there is a much lower likelihood of bike demand. It shows that this variable may influence the demand for rented bikes.

- But there is no significant difference in the higher quartile values. This shows that the overall number of bike rental requests vary much widely in just 21 holidays which were recorded over a period of two long years at different times/seasons, and hence the difference in mean and median could be due to randomness. We may need to perform hypothesis testing to ascertain that this mean difference is not due to randomness and is statiscally significant.

2) **workingday**

- The lower quartile (q1) of the total number of bike rentals on workdays in the boxplot is larger than that on the non-working day, confirming the univariate analysis's finding that demand is higher on workdays.

- But their higher quartile and median values do not differ significantly. It indicates that the amount of variation of the bike rental counts vary widely between working and non working day, resulting in small difference in mean, could be due to randomness. We may need to perform hypothesis testing to ascertain if the small difference in mean is statistcally significant.

3) **yr**

- The number of bike rentals increases very significantly from 2018 to 2019. Quartiles q1, q2, q3 - all differ very significantly over the years. This could be a strong influencer of bike demand. Since these bike-sharing systems are slowly gaining popularity, the demand for these bikes is increasing every year

4) **season**

- In the spring season, the number of rentals is the lowest. The 75th percentile value of the number of bike rents in spring season is lower than even the 25th percentile rents on remaining seasonsThis indicates potential impact of season on bike rental demand, specifically in the summer, winter and fall.

- There is a gradual change in the 25th percentiles representing number of rentals from winter, to summer, & then fall. But when we compare the medians or 75th percentile values around these seasons, they do not seem to rise steadily. Also, the distributions between them vary widely, indicating the difference between their means could be due to randomness. Maybe we need to conduct hypothesis testing to ascertain if the difference in means is statiscally significant and is not due to randomness.

5) **weathersit**

- When the weather is pleasant, the demand of bike rentals is also appreciated. But, as the climatic condition worsen, the demand descreases. We have noticed it also during univariate analysis earlier.

- Even the 25th percentile values of the total number of bike rentals when it is foggy or misty are lower than the higher quartile (q3) of the total number of bike rentals during unfavorable weather (precipitation or thunderstorm). It shows that weathersit may influence the demand for rented bikes. There was hardly any occurrences of the thunderstorm like situtations which could have caused such lower bike demands at the time.

- In foggy or misty weather, the 75th percentile count of all bike rentals appears to be substantially lower than in pleasant weather, which is also true across their lower quartiles. The differences between the bike demands' medians in favorable and foggy conditions are not as great as those between the upper and lower quartiles, representing the bike demand variances are wide enough causing the small difference in the averages due to randomness, hinting that this variable might not influence the bike rental demand. We can conduct hypothesis testing to ascertain whether the mean difference is statistically significant and not due to randomness.

Question-2

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

Dummies created from the same categorical variable are highly correlated and mutually exclusive. And therefore we drop one level to get k-1 dummies out of k levels of the categorical variable.

For example, if we have 3 categorical levels, we can drop one of them to get 2 dummies out of 3 levels. Since they are mutually exclusive, therefore if both the dummies are zeros, then it indicates the one which we drop.

Question-3:

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

Among the numerical variables, **temp** and **atemp** have the highest correlation with the target variable **cnt**

Question-4

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

Using distplot to verify if the ERROR terms are normally distributed. RESIDUAL ANALYSIS

Question-5

Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the

shared bikes?

yr, **season_spring** and **weathersit_precipitation** are the top three influencers that can potentially explain the demand of shared bikes.

General Subjective Questions

Question-1

Explain the linear regression algorithm in detail.

1). A linear regression model regresses on the variables to predict a target which is numerical in nature.

2). A simple linear regression line can be represented in point-slope form **y=mx+b** or slope intercept form **m = (y-y1)/(x-x1)**.

3). **'b'** is the intercept where the line intersects the y-axis and **'m'** is the slope the line

4). This slope could be negative, positive, zero or undefined. If its positive, then it indicates a positive correlation between variables. Similarly if its negative, indicates negative correlation between variables.

5) If the slope **m** is ZERO, then the estimated line **do not regress** the predictors and hence it fails to explain the variance of the target variable, and the expected value by the estimated line is basically the same for each value of the predictor or datapoint, which is equal to the y-intercept where the line intersects the y-

axis.

6). If the line runs parallel to the y-axis, then its **slope is undefined (tan 90 degree)**

7). We need to learn/compute the weights or coefficients optimally so we can derive a line that can fit the datapoints or explain the relationship between predictor and target variable.

8). In MLR (multiple linear regression), the line is represented as

y = b0+b1*x1+b2*x2+...

9). We determine the residual as difference between the actual and predicted value and our objective is to use the ORDINARY LEAST SQUARE method to learn the weight/coefficients such that the difference between the actual and predicted value is minimal.

10). Oridinary Least Square method offers RESIDUAL SUM OF SQUARES (RSS) as the cost function and it helps in optimization/minimization of this objective function, using gradient descent approach. This is UNCONSTRAINED minimization.

11). Assumptions of Linear Regression includes, assumption of **Exogenity**, assumptions of **NORMALITY** of Error Terms, assumptions of **INDEPENDENCE** of Error Terms, and assumptions of **HOMOSCEDASTICITY** of error terms.

Question-3

What is Person's R?

It refers to the correlation coefficient that is calculated for numerical variables. It is formulated as below:

Cor(x, y) = Covariance(x, y)/ sqrt(variance(x) * variance(y))

Question-4

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

Scaling is the process of shifting the scales or transforming the values from one scale to another. There are 2 basic advantages:

1). It helps in ease of interpretation of coefficients. If the numerical variables are in the same scale, its easier to comare their coefficients when we need to interpret them per business requirement.

2). Faster convergence of gradient descent algorithm

NORMALIZED SCALING: It compresses the values to fall in the range of 0 and 1. If the variable that is scaled has outliers, then they can be found near to 0 or 1

STANDARDIZATION: It does not compress the values to fall in any specific range. It transforms the scale of the values such that their mean is ZERO and standard deviation is 1.

Question-5

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

Yes, the phenomenon of having infinite values for the Variance Inflation Factor (VIF) in regression analysis typically occurs due to perfect multicollinearity among predictor variables. Perfect multicollinearity means that one predictor variable can be exactly predicted by a linear combination of other predictor variables.

Question-6

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

to assess whether a given dataset follows a particular probability distribution, typically the normal distribution. You can use it to check if the error terms are normally distributed at the time of RESIDUAL analysis.