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: Here are some of the inferences I made from my analysis of categorical variables from the dataset on the dependent variable (Count)

1. Fall has the highest median, which is expected as weather conditions are most optimal to ride bike followed by summer.

2. Median bike rents are increasing year on as year 2019 has a higher median then 2018, it might be due the fact that bike rentals are getting popular and people are becoming more aware about environment.

3. Overall spread in the month plot is reflection of season plot as fall months have higher median.

4. People rent more on non-holidays compared to holidays, so reason might be they prefer to spend time with family and use personal vehicle instead of bike rentals.

5. Overall median across all days is same but spread for Saturday and Wednesday is bigger may be evident that those who have plans for Saturday might not rent bikes as it a non-working day.

6. Working and non-working days have almost the same median although spread is bigger for non-working days as people might have plans and do not want to rent bikes because of that

7. Clear weather is most optimal for bike renting, as temperate is optimal, humidity is less, and temperature is less.

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

Ans: Dummy variables will be correlated if you don't remove the first column (redundant). This may have a negative impact on some models, and the effect is amplified when the cardinality is low. Iterative models, for example, may have difficulty convergent, and lists of variable importance may be distorted. Another argument is that having all dummy variables results in multi collinearity between them. We lose one column to keep everything under control.

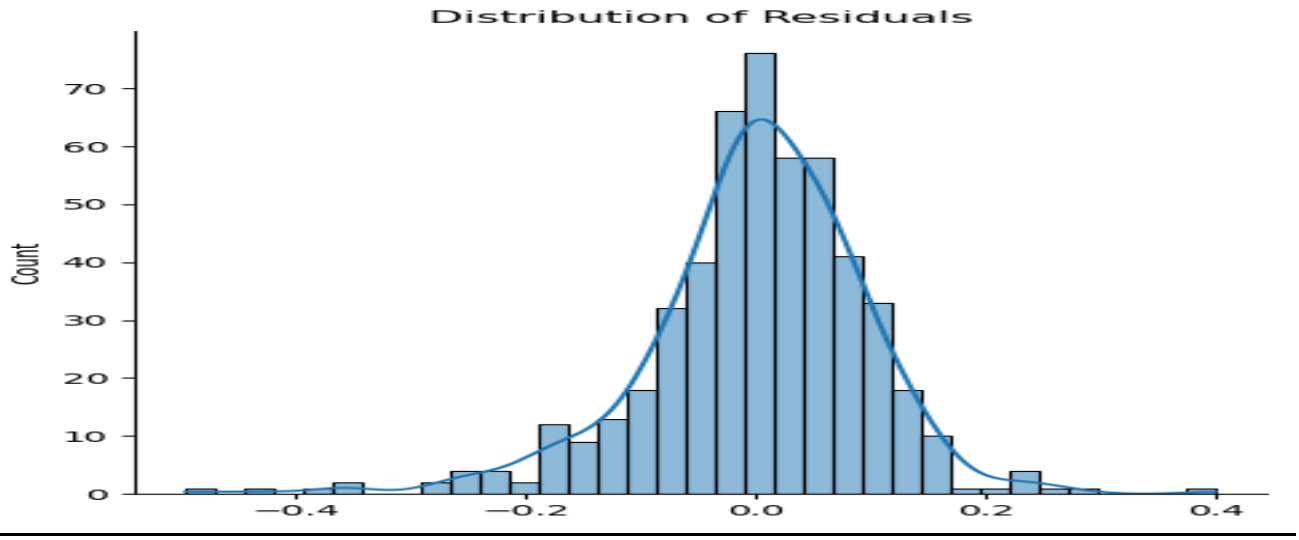
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans: “temp” and “atemp” has the highest correlation (0.63) with the target variable

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

Ans: The distribution of residuals should be normal and centred around 0. (The mean is 0). We test this residuals assumption by producing a distplot of residuals to see if they follow a normal distribution or not.

The residuals are scattered around mean = 0 as seen in the diagram below.



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

Ans: The Following are the top 3 features contributing significantly towards explaining the demands of the shared bikes:

1. temp (0.3821)
2. weathersit_drizzle (-0.3200)
3. year (0.2407)

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

   Ans:

   Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric values. Linear Regression is the most basic form of regression analysis. Regression is the most commonly used predictive analysis model. Linear regression is based on the popular equation

“y=mx+c".

It assumes that there is a linear relationship between the dependent variable(y) and the predictor(s)/independent variable(x). In regression, we calculate the best fit line which describes the relationship between the independent and dependent variable. Regression is performed when the dependent variable is of continuous data type and Predictors or independent variables could be of any data type like continuous, nominal/categorical etc. Regression method tries to find the best fit line which shows the relationship between the dependent variable and predictors with least error. In regression, the output/dependent variable is the function of an independent variable and the coefficient and the error term.

Regression is broadly divided into simple linear regression and multiple linear regression.

1. **Simple Linear Regression:** SLR is used when the dependent variable is predicted using only one independent variable.
2. **Multiple Linear Regression:**MLR is used when the dependent variable is predicted using multiple independent variables.

The equation for MLR will be:

$$y_i=\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_p x_{ip}+\epsilon$$

where, for i=n observations:

yi=dependent variable

xi=explanatory variables

*β*0=y-intercept (constant term)

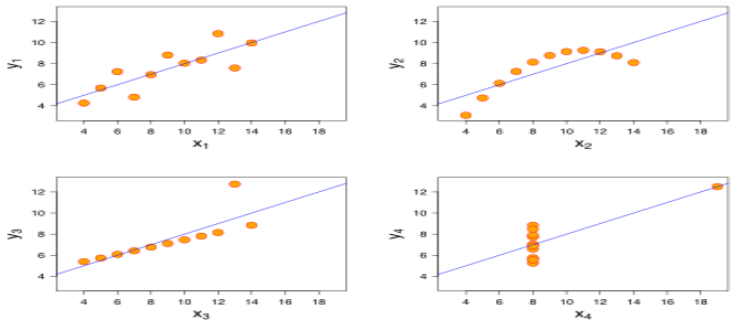
βp=slope coefficients for each explanatory variable

ϵ=the model’s error term (also known as the residuals)

2. Explain the Anscombe’s quartet in detail.

Ans:

Anscombe's quartet comprises of four data sets that have nearly identical simple descriptive statistics but have quite different distribution when visualized graphically. The simple statistics consist of mean, sample variance of x and y, correlation coefficient, linear regression line and R-Square value. Anscombe's Quartet shows that multiple data sets with many similar statistical properties can still be vastly different from one another when graphed. The graphs are shown below:



Statistical Properties:

- The first scatter plot (top left) appears to be a simple linear relationship.
- The second graph (top right) is not distributed normally; while there is a relation between them it's not linear.
- In the third graph (bottom left), the distribution is lincar, but should have a different regression line The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

3. What is Pearson's R?

Ans: Pearson's R measures the strength of association of two variables. It is the covariance of two variables divided by the product of their standard deviation. It has a value from +1 to -1.

• A value of 1 means a total positive linear correlation. It means that if one variable increase then other will also increase

• A value of 0 means no correlation

• A value of -1 means a total negative correlation. It means that if one variable increase then other will decrease

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

Ans:

Scaling of a variable is performed to keep a variable in certain range. Scaling is a pre-processing step in linear regression analysis. The reason we scale a variable is to make the computation of gradient descent faster. The step size of gradient descent are generally low for accuracy, if the data has some small variables (values in the range of 0-1) and some big variables (values in the range of 0-1000) than the time taken by the gradient descent algorithm will be huge.

Normalised Scaling	Standardized scaling
Called min max scaling, scales the variable such that the range is 0-1	Values are centred around mean with a unit standard deviation
Good for non- gaussian distribution	Good for gaussian distribution
Value id bounded between 0 and 1	Value is not bounded
Outliers are also scaled	Does not affect outliers

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

Ans: VIF - the variance inflation factor: - The VIF indicates how much collinearity has increased the variance of the coefficient estimate. (VIE) is equal to 1/(1-Ri^2). VIF-infinity if there is perfect correlation. Where R-1 denotes the R-square value of the independent variable for which we want to see how well it is explained by other independent variables.

If an independent variable can be completely described by other independent variables, it has perfect correlation and has an R-squared value of 1. As a result, VIF-1/(1-1) provides VIF-1/0, which is "infinity."
Basically, if R square is 1 than VIF becomes infinite. It means that there is perfect correlation between the features.

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

Ans: A Q-Q plot is a scatter plot of two sets of quantiles against each other. Its purpose is to check if the two sets of data came from the same distribution. It is a visual check of data. If the data is from same source than the plot will appear as a line.
The quantiles of the first data set are plotted against the quantiles of the second data set in a q-q graphic. It's a tool for comparing the shapes of different distributions. A scatterplot generated by plotting two sets of quantiles against each other is known as a Q-Q plot. Because both sets of quantiles came from the same distribution, the points should form a line.
That's a fairly straight line.
The 4-4 plot is used to answer the following questions:

- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behaviour?