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?**

   Visualized continuous variables with pairplot and categorical variables in boxplot, Following are the results.

   - Summer and fall, has the highest count
   - June, July, August, September, October has the highest count
   - If there is no holiday count is more.
   - Year 2019 had more count.
   - Temp and a temp may have a perfect collinearity.

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

   In the column **season** there will be four categories namely(clear, misty, light, heavy). So if we encode these categories in binary, it will result in creation of 4 new columns.

   Ex. 0 1 0 0 is for clear, 0 0 1 0 is for misty and so on. So this identification can also be done if we remove the first variable. Therefore clear will be denoted by 1 0 0 and so on.

   This will reduce any possible chance of correlation because of creating dummies.

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

   Temp variable has the highest correlation with the target variable cnt.

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

   Error terms are following normal distribution, No multicollinearity among the selected features, R-squared and adjusted R-squared also indicates that the model is significant, Homoscedasticity: there is no such visual variation in the plot.

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

   Windspeed, workingday, winter are the top 3 variables which are contributing significantly in demand of shared bikes.

General Subjective Questions

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

   Linear Regression is an algorithm which analyses the relation between the dependent variable and one or more independent variables.

   Linear regression is like drawing a best fitting from a set of points on the graph. Suppose we have some data points that represent two variables: one we want to predict (Y) and another variable that you think affects it (X). For example, you might want to predict someone's weight (Y) based on their height (X).

   Linear regression helps in finding the best-fitting line that represents the relationship between these two variables. The line has an equation: Y = b0 + b1X, where b0 is the starting point on the Y-axis (where the line crosses it) and b_1 is the slope (how steep the line is).

   There are two types of Linear regression models.

   Simple linear regression: It involves a single independent variable (X) and a dependent variable (Y). It models the linear relationship between X and Y using a straight line. The equation for simple linear regression is Y = b0 + b1X, where b0 is the intercept and b1 is the slope of the line.

   Multiple Linear Regression: Multiple linear regression is an enhancement to simple linear regression. It includes multiple independent variables (X1, X2, ..., Xn) that may affect the

dependent variable (Y). The equation for multiple linear regression is Y = b0 + b1X1 + b2X2 + ... + bnXn. It allows for the analysis of the combined effects of multiple variables on the dependent variable.

There are several assumptions to be made.

Linearity: There should be a linear relationship between the variables. If the relationship is nonlinear, linear regression may not be appropriate, and other regression models may be more suitable.

Homoscedasticity: The residual values should not make any patterns in the model. If there is a visual pattern, then the model may not be the accurate one.

Normality: The residual values should follow a normal distribution. Any deviation in normality can impact the accuracy of hypothesis tests and the reliability of statistical inferences.

No multicollinearity: The independent variables should not be highly correlated with each other. High multicollinearity can make it difficult to determine the individual effects of the independent variables on the dependent variable. VIF will help to determine the multicollinearity.

2. Explain the Anscombe's quartet in detail.

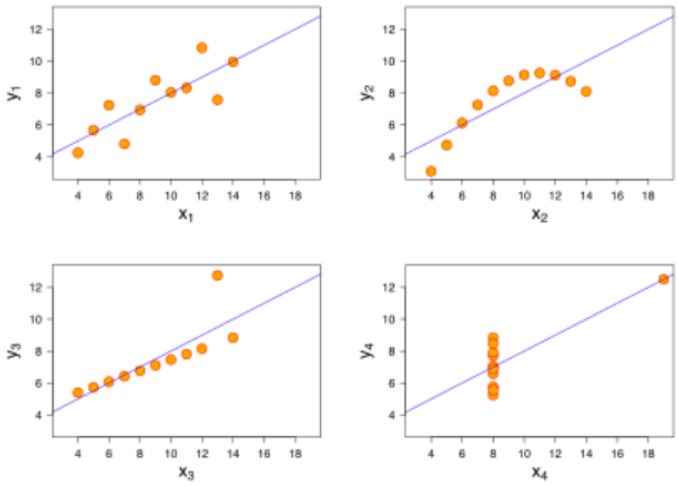
Anscombe's quartet is a set of four datasets that were created by the statistician Francis Anscombe.

Each dataset consists of 11 points and has two variables: x and y.

The quartet has different values, but the statistical properties of these datasets are similar.

Anscombe's quartet

I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



Despite having different patterns when plotted, all four datasets have nearly identical values for statistical measures such as mean, variance, correlation, and the coefficient of determination. This highlights the importance of visualizing data and understanding the underlying patterns before drawing conclusions based solely on summary statistics.

Anscombe's quartet indicates that statistical analysis should be complemented with exploratory data visualization to have a better understanding of the data and to avoid making errors in conclusions based on summary statistics.

3. What is Pearson’s R?

Pearson's R, is also called Pearson's correlation coefficient.

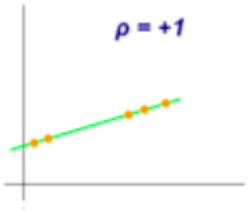
It is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables.

It has values in the range of -1 and +1.

The value of Pearson's correlation coefficient (r) can go to -1 when there is a perfect negative linear relationship between the two variables.

all the data points fall perfectly along a straight line with a negative slope.

The value of Pearson's correlation coefficient (r) can go to +1 when there is a perfect negative linear relationship between the two variables.



all the data points fall perfectly along a straight line with a positive slope.

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

Scaling refers to the process of transforming the values of variables to a specific range or distribution. It is performed to ensure that variables are on a comparable scale, which can be beneficial for various data analysis techniques and models.

If there are variables with values ranging from 0 to 5 and others ranging from 10,000 to 20,000 in a dataset, directly modeling on this data without scaling could lead to issues. The model may consider the high-value variables as more prominent and give them more weight or importance compared to the low-value variables. This can result in an imbalance and potentially biased predictions.

Normalized Scaling	**Standardized Scaling**
Also known as MinMax scaling	Also known as ZScore scaling
Rescales the values to a specified	Scales the values to have a mean

range, typically between 0 and 1	of 0 and a standard deviation of 1.
Formula = (x - min(x)) / (max(x) - min(x))	Formula = (x - mean(x)) / std(x)
Used when original values need to be preserved.	Used when comparing variables with different means and standard deviations

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

VIF is associated with multicollinearity. Sometimes there is a perfect multicollinearity between variables. I.e some variables will have exact linear relationships.

When there is a perfect multicollinearity the VIF will be infinite.

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

A QQ plot is a graph that helps us see if our data follows a particular pattern or shape, like a smooth curve. It compares our data to what we would expect if it followed a certain pattern, like a straight line.

The QQ plot is important because it helps us understand if our data follows a certain pattern or shape. This is useful because many statistical methods and models rely on certain assumptions about how our data is distributed.