Assignment-based 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)**

A: In the dataset, 'count' is the target or dependent variable which tells us the number of bike rentals at a particular time.

We see some surprising insights in the plain sight itself when we get plots.

- Bike rentals see more growth in the Fall months. And then followed by the Summer months. However, the warm Spring months throw a surprise. Because these months are a little bit warm. But why no demand? The spring months are romantic to go for cycling, aren't they? Let's see. (**count** vs **season**)
- There's a steep increase in demand for the bikes in 2019. (**count** vs **year**)
- The demand for bikes is more on the weekend (Saturday) and mid-weekdays (Wednesday and Thursday)
- Moreover, **clear weather situation** makes things perfect for biking!

Overall, the categorical variables have their strong influence on the dependent variable.

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

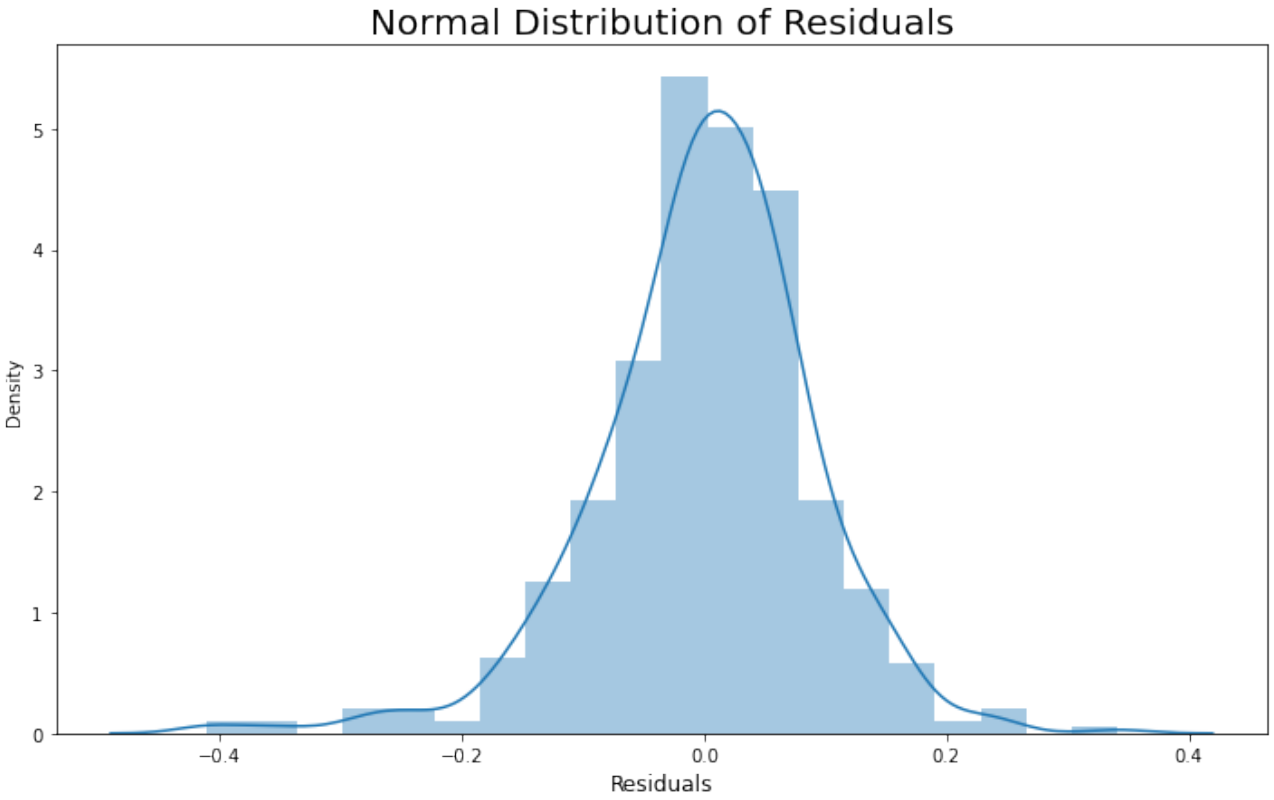
A: To put it in a straight and simple way, we drop it to avoid the multicollinearity among the dummy variables. And if there is a multicollinearity among the dummy variables, it leads to large standard errors. Hence, the inferences are less significant statistically. For that reason, it's important to use 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)**

A: Just by looking at the pair-plots, we can say that the temperature (**temp and atemp**) have the highest correlation with the target variable, **count**.

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

A: To validate the assumptions of Linear Regression after building the model, I plotted the histogram to see whether the residuals or errors follow the Normal Distribution, with mean as 'Zero.'



Secondly, I checked the values of the VIF. All variables have the VIF below 10, which is accepted value of the VIF. Strictly-speaking, except temperature, the variables have the VIF below 5 in the model we built. That means, the error terms are independent of each other.

Thirdly, there is homoscedasticity or constant variance between the Independent variables and the target variable. Since the R-squared and Adj. R-squared values are about 83 percent, the model isn't memorising the data values.

Thus, I validated the assumptions of multiple linear regression in this model.

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: Here are the top three features contributing to the demand for the shared bikes:

1. Temperature (temp)
2. Year
3. Winter

General Subjective Questions

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

A: Linear Regression algorithm is a statistical model to analyse the relationship between dependent and one or more independent variables. The linearity between the dependent and independent variables is given by the mathematical equation,

$$y = c + mX$$

Here,

- y is the dependent variable
- c is the constant or intercept
- m is the slop or the effect on the independent variable
- X is the dependent variable

Depending on the slope or the coefficient of the independent variable, the relationship between the variable will be either positive or negative.

- Positive slope: dependent variable increases as the independent variable increases.

- Negative slope: dependent variable increases as the independent variable decreases or vice versa.

Depending on the number of variables, we will classify linear regression as

1. Simple linear regression, with one independent variable
2. Multiple linear regression, with two or more independent variables

A linear regression algorithm, simple or multiple, follows this process, given as steps:

1. Understanding & Inspecting Data

2. Standardise Data by removing null, missing values and other needful things

3. Data Visualisation to graphically see the relation among the variables

4. Data Preparation to create dummy variables for categorical variables

5. Split Data into Train and Test. Usually, train size is 70 percent.

6. Building the LR Models. We drop the highly correlated values by checking the VIF values.

7. Residual Analysis. The algorithm plots the error terms whether they follow Normal Distribution. And whether they have constant variance (or, homoscedasticity).

8. Predictions on Test Data Using the Best Model. Based on the best fit model, the algorithm predicts values.

9. Model Evaluation by plotting a scatter plot between actual and predicted values.

Thus the algorithm helps to build models and at the same time test them against assumptions of linear regression.

Finally, these are the parameters that help us decide the best fit model built using linear regression algorithm.

- P-values. They should below 0.05
- VIF values. They should be below 10.
- R-squared and Adj. R-squared values

- F-statistic. It explains the best fit model and it should greater than 1.

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

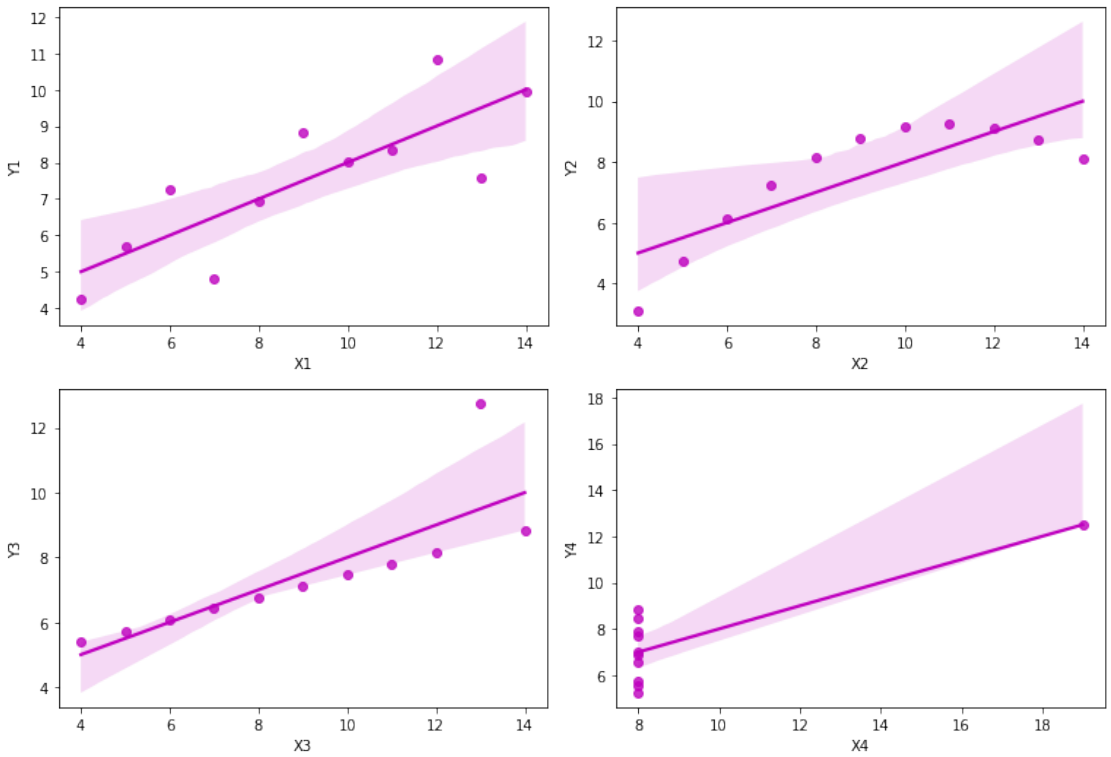
A: Anscombe's quartet consists of four datasets, with nearly identical values. Each dataset consists of eleven data points (x, y). Though these data values look identical, they are different when they are plotted on a graph.

The aim of the Anscombe's quarter is to show the importance of visualisation of data and the effect of the outliers on statistical properties. These quartets were constructed in 1973 by the famous statistician Francis Anscombe.

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

Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817



In the plain sight, the means and variances of X and of Y for all the sets are the same respectively. But we get a different view of these quartets, when we plot them on the graph.

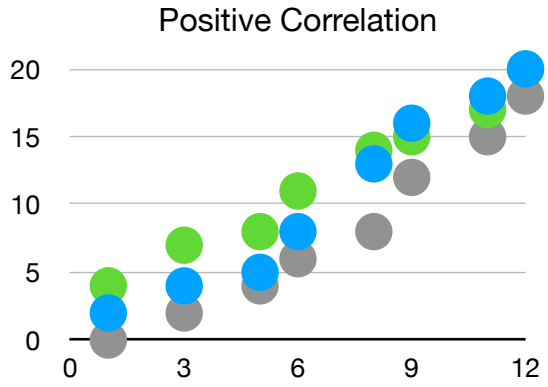
Plotting those datasets, we get these graphs. Now we can understand the importance of the data visualisation.

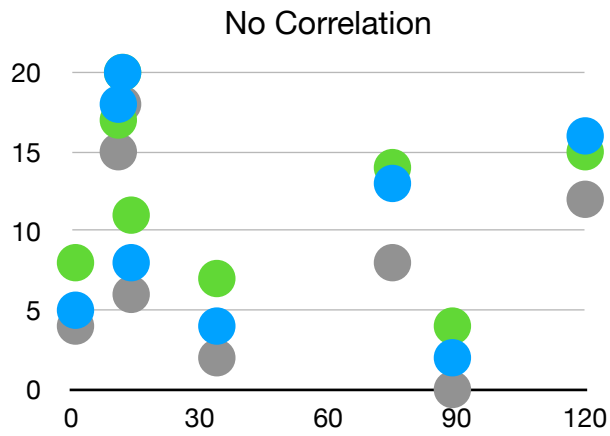
3. What is Pearson's R? (3 marks)

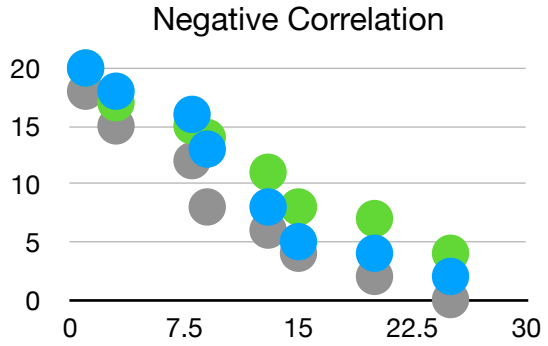
A: Person's R is a statistical coefficient that tells us how the variables are related linearly. That is if variables go up all together or do down all together. That shows positive relationship. If one variable goes up and the other goes down and the vice versa, the correlation is negative. If one variable doesn't affect the other variable at all, then there is zero correlation.

The R is between -1 and + 1.

- If positive correlation, then the R coefficient value is between 0 and +1
- If negative correlation, then the R coefficient value is between - 1 and 0
- If no correlation, then the R coefficient value is 0







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

A: Scaling is a technique to standardise the data values of the independent features. It is usually done as pre-processing thing before we start building the machine learning models.

Scaling is important because when we standardise or generalise the data points so that it will be easier for the machine learning models for faster decision-making. If the data is not scaled, then there will be higher differences between the values, which only leads to uncertainty and difficult for to train the models.

S.No.	Standardised Scaling	Normalized Scaling
1	Mean and Standard Deviation are the parameters used for scaling	Minimum and Maximum values are used here to scale features
2	Mean is 0 and Standard Deviation is 1	With mean centred at 0, it follows Normal Distribution
3	It doesn't influence the outliers	It influences outliers so that they are generalised to a range
4	There's no particular range for this scaling	The range here is between 0 and 1. Or between -1 and 1.
5	Use StandardScaler from Scikit Learn	Use MinMaxScaler from Scikit Learn

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

A: Mathematically, the Variance Inflation Factor (VIF) is given as,

$$VIF = 1/1 - R^2$$

When R^2 value is 1, the VIF equals to infinity. The R-squared value becomes one when there is a perfect linear correlation between two independent variables. That is, there is perfect multi-collinearity between them. So the solution is to drop one of the variables. The higher the VIF value of a variable, the higher the collinearity of it with other variables.

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, or Q-Q plot, is a graphical technique to see whether two datasets from populations come from the same distribution.

The main use of the Q-Q plot is to determine whether two datasets have come from populations with the same distribution. In the plot, there will be a 45-degree line. If the point from the quantiles fall close to that line, then the

datasets have come from the populations with the same distribution. The further away the points fall away from the reference line, the more clear it gets that the datasets have come from the populations that have different distributions.

The importance of Q-Q plots are to find out the estimates of the common location and scale if they have come from the population which has the same distribution. Otherwise, it will offer more insights on the differences than other test like Chi-square.