General Subjective Questions: -

1. Explain Linear Regression Algorithm in detail

Ans – Linear Regression is a supervised fundamental algorithm that enables to establish a linear relationship between one output variable (may be called as dependant variable) and single/ multiple input variables (may be called as Independent variables)

1. Concept: It fits a line (or hyperplane) to data to predict the dependent variable based on independent variables.

2. Mathematics:

- Simple Linear Regression: $Y = \beta_0 + \beta_1 X + \epsilon$
- Multiple Linear Regression: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$

3. Objective: Find coefficients that minimize the difference between predicted and actual values using the Mean Squared Error (MSE).

4. Optimization: Coefficients are estimated using Ordinary Least Squares (OLS) or Gradient Descent.

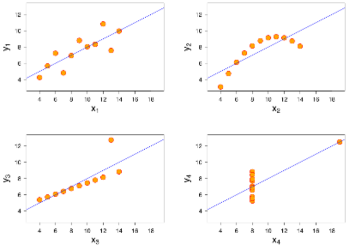
5. Assumptions: Linearity, independence, homoscedasticity, and normality of errors.

6. Evaluation: Metrics include R-squared, Adjusted R-squared, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE).

7. Applications: Used for predicting outcomes, understanding relationships, and feature selection.

2. Explain the Anscombe's quartet in detail.

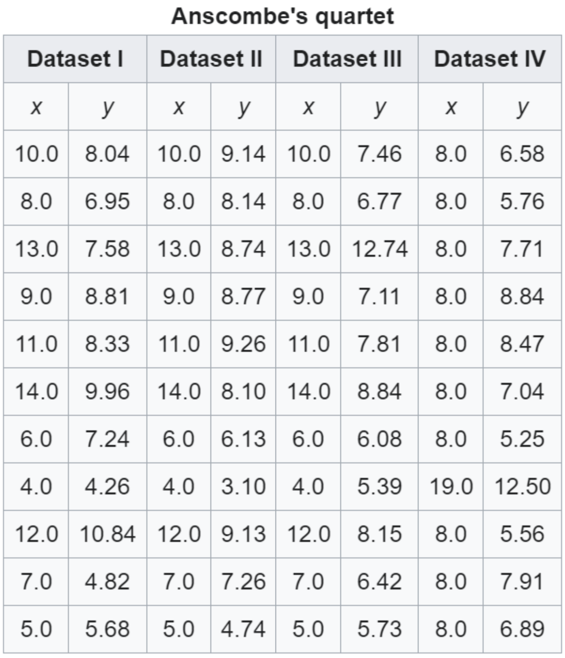
Ans- Anscombe's quartet is a set of four datasets that was created by the statistician Francis Anscombe in 1973 to illustrate the importance of graphical analysis in statistics. Despite having nearly identical statistical properties (mentioned below), each dataset reveals different underlying patterns when plotted. Here is a detailed breakdown:



- Anscombe's quartet consists of four datasets, each with 11 pairs of (x,y)(x, y)(x,y) values. The datasets are designed to demonstrate that summary statistics like mean, variance, and correlation can be misleading if we don't consider the data's graphical representation as well.
- All four datasets have the following common statistical properties:
  - Mean of xxx: 9
  - Mean of yyy: 7.5
  - Variance of xxx: 11
  - Variance of yyy: 4.12
  - Correlation between xxx and yyy: 0.816

- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two correlated variables, where *y* could be modelled as gaussian with mean linearly dependent on *x*.
- For the second graph (top right), while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). 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.

The dataset is as follows :-



Anscombe's quartet

Dataset I		Dataset II		Dataset III		Dataset 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

3. **What is Pearson's R?**

Ans- The Pearson correlation coefficient, also known as Pearson's r, is a widely used statistic that quantifies the strength and direction of the linear relationship between two quantitative variables. Here are the key points:

- Range: It takes values between -1 and 1.
- Interpretation:
  - Positive correlation: When one variable changes, the other variable changes in the same direction (e.g., baby length and weight).
  - No correlation: When there is no relationship between the variables (e.g., car price and width of windshield wipers).
  - Negative correlation: When one variable changes, the other variable changes in the opposite direction (e.g., elevation and air pressure).
- Visualizing: Think of it as how close the observations are to a line of best fit; it also indicates whether the slope of the line is positive or negative.
- Inferential: It can be used to test whether there is a significant relationship between two variables

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

Ans-

- It is a step of data pre - processing technique which is applied on independent variables & used to normalize the data into some specific ranges . It also helps speeding up the calculations of an algorithm .

- Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we must do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

- Normalized Scaling is also known as Min-Max scaling. It always ranges between 0 & 1 . So, it can tackle outliers as well in the dataset. But The standardization scaling doesn't compress the data within any particular range like Normalization. So, Standardization scaling will affect the value of dummy variables in such a way that the mean of the dataset becomes 0 & SD becomes 1. So, it will clearly distort the dummy variables as some of them can become negative.

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

Ans- The infinite value of the Variance Inflation Factor (VIF) occurs due to perfect multicollinearity among independent variables in a regression model. When two or more variables are linearly dependent, one can be entirely predicted by the another. This leads to an infinite VIF value. In other words, when the R-squared approaches 1, it symbolizes that a set of independent variables explains nearly all the variability in another variable, the VIF becomes infinite. VIFs help detect multicollinearity, which can cause problems for regression models. Keep in mind that a VIF of 1 indicates no correlation between predictor variables in the model.

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

Ans -

- Q-Q plot is Quantile-Quantile plot. This is graphical tool used to assess if a dataset follows a specific theoretical distribution, commonly the normal distribution. In linear regression, it helps check whether the residuals (differences between observed and predicted values) are normally distributed, which is an important assumption for valid statistical inference.
  - If the points in the Q-Q plot fall roughly along a straight line, it suggests that the residuals are normally distributed.
  - Deviations from a straight line indicate that the residuals may not be normally distributed. For instance, a curve might suggest skewness or heavy tails.

- Linear regression assumes that the residuals are normally distributed, especially for the purposes of hypothesis testing and confidence intervals. A Q-Q plot helps in visually checking this assumption.
  - Normality of residuals ensures that the model is appropriately specified and that the inferences made from the model (such as confidence intervals and significance tests) are valid. If the normality assumption is violated, it might indicate issues with the model fit or the need for a different modelling approach.
  - If the Q-Q plot reveals deviations from normality, you might need to consider transformations of the dependent variable or explore other types of models that better capture the distribution of the residuals.

So, in a nutshell , we can say that the Q-Q plot is essential for validating the assumptions of linear regression and ensuring the reliability of the model's results.

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: -

We can come to these conclusions:-

- Demand is continuously growing each month till June ; It is a bit down in July ; Again highest in September & then it is decreasing again .
- When weather is good demand is highest ; As Weather is becoming bad , Demand is decreasing .
- Fall season has highest demand for rental bikes; followed by summer , winter & spring .
- Demand is quite higher in 2019 than 2018 .
- Demand is higher if its not a holiday.
- From working day plots, nothing is conclusive.

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

- To achieve n-1 dummy variables for n no of distinct values of a particular column (More than 2 distinct values preferably) as we can delete extra columns . N-1 dummy variables are enough to indicate the value of the other column . Also, after creating dummy variables we can delete the actual column of the data frame as well as the n-1 dummy variables are enough to represent the actual column as well.
    - For Ex, If Weather column, if Good weather is represented by 1-0 , Moderate is represented by 0-1 & then automatically 0-0 indicates Bad weather .
- It is also used to reduce collinearity among the dummy variables.

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

Ans :- From the pair plots , We can see that atemp & temp columns are having highest correlation with the target variable (count) & corr value is 0.63 for both the columns.

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

- The main validation is Normal distribution of error terms as I have plotted the same towards the end of our analysis in .ipybnb file .
- We can find well valued VIFs of all the variables (All are less than 2 whilst the max permissible value is 5) . This indicates less multicollinearity .
- Constant variance of errors that we can observe in the built regression model .
- R-squared value of both Train dataset & test data set are quite close & value is also quite higher (Both are around 0.8 apprx)

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

The top 3 features are :-

- Temperature
- Light rain-snow
- September Month