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? (3 marks)

Response-1:

I inferred the following from my analysis of the categorical variables,

- Demand of bikes is as follows,
    - Highest in the **Fall** season and takes a dip in the **Spring**
    - There is a year wise increase in demand from 2018 to **2019**, indicating demand is growing year on year
    - Demand peaks in **Jun, Jul, Aug and Sep** and then start to declines
    - Demand is highest during **Clear weather** and it drops as weather situation deteriorates
- There is not much difference of demand during any day of a week or if it is a working day or not

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

Response-2: There are two major reason it is important to use drop_first=True during dummy variable creation

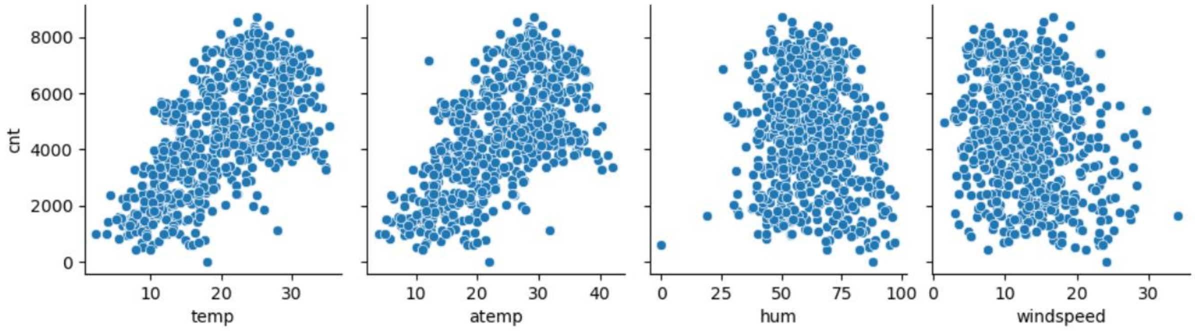
i. **To Prevent Multicollinearity**: When two or more independent variables in a regression model are highly correlated then it can lead to unstable estimates and make it difficult to interpret the model's results. For example, for a categorical variable with n categories, n-1 dummy variables are created. This is because the nth category can be inferred from the absence of the other n-1 categories

ii. **To Avoid Redundancy**: The information contained in the nth category (in point-1) can be inferred from the other n-1 categories and hence one column can be reduced

Kindly note that in a couple of cases, drop_first=True is not applicable as all values of the column are equally important, so dropping any of the columns would not be right. Hence, drop_first=False is the default parameter in Python.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

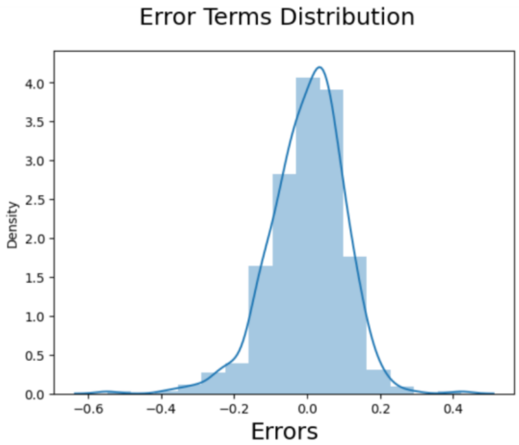
Response-3: With 0.646475, atemp has the highest correlation with the target variable, closely followed by temp at 0.643517. Pair plot is shown below for quick reference



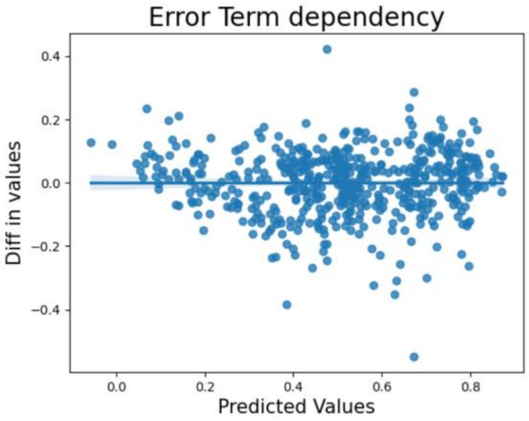
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Response-4: I validated all three assumptions of the error terms after building the model on the training set as below,

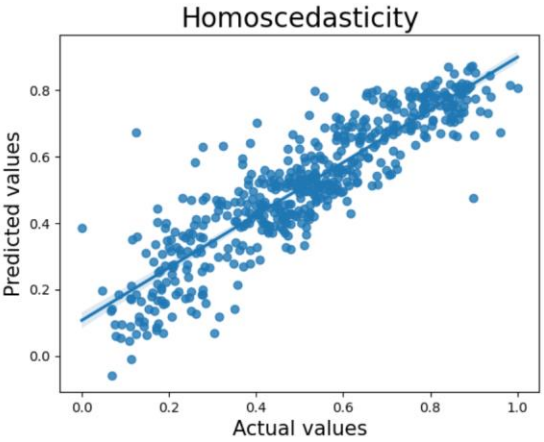
i. **Assumption-1**: Error terms are also normally distributed – I plotted the histogram and found the values are normally distributed around the mean zero, inline with the expectations



ii. **Assumption-2**: Error terms are independent of each other - I plotted the regplot and found no relationship in the Predicted values and residuals, inline with the expectations



iii. **Assumption-3**: Error terms have constant variance (Homoscedasticity) - I plotted the regplot and found that variance of the values are almost same across all the datapoints, inline with the expectations



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Response-5: The top 3 features contributing significantly towards explaining the demand of the shared bikes based on the coefficient of the final model are,

i. LightSnow_LightRain
ii. yr
iii. Spring

General Subjective Questions

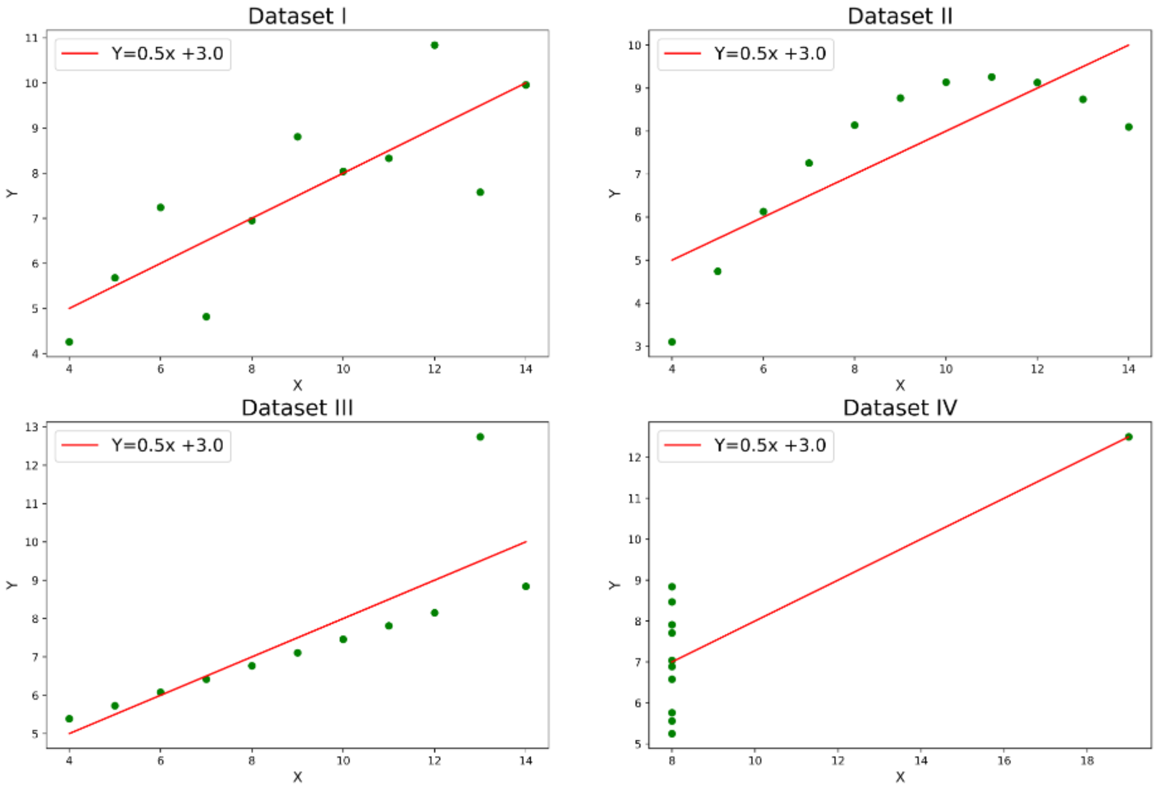
1. Explain the linear regression algorithm in detail. (4 marks)

Response-1: Linear regression algorithm is as follows,

- It is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It is a type of supervised Machine learning
- It assumes a linear relationship between the variables and aims to find the best-fitting line that minimizes the distance between the predicted values and the actual values
- Equation of linear regression is $y = \beta_0 + \beta_1 x1 + \ldots . + \beta nxn + \varepsilon$ where,
    - y is the dependent variable
    - β_0 is the intercept
    - β_1 is the coefficient for the first feature
    - β_1n is the coefficient for the nth feature
    - ε is the error term
- The aim of linear regression is to find the values of β_0, β_1 and so on that minimize the sum of squared errors (SSE)
- It involves multiple steps like
    - Step 1: Reading and Understanding the Data
    - Step 2: Visualising the Data
    - Step-3: Creating Dummy Variables
    - Step 4: Splitting the Data into Training and Testing Sets
    - Step 5: Preparing the model
    - Step 6: Building model using statsmodel
    - Step 7: Residual Analysis of the train data
    - Step 8: Making Predictions on the Test dataset
    - Step 9: Model Evaluation using r2_score etc
    - It also involves cross verifying assumptions of the error terms as below
        - Assumption-1: Error terms are also normally distributed
        - Assumption-2: Error terms are independent of each other
        - Assumption-3: Error terms have constant variance (Homoscedasticity)

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

Response-2: Anscombe’s quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph. The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.



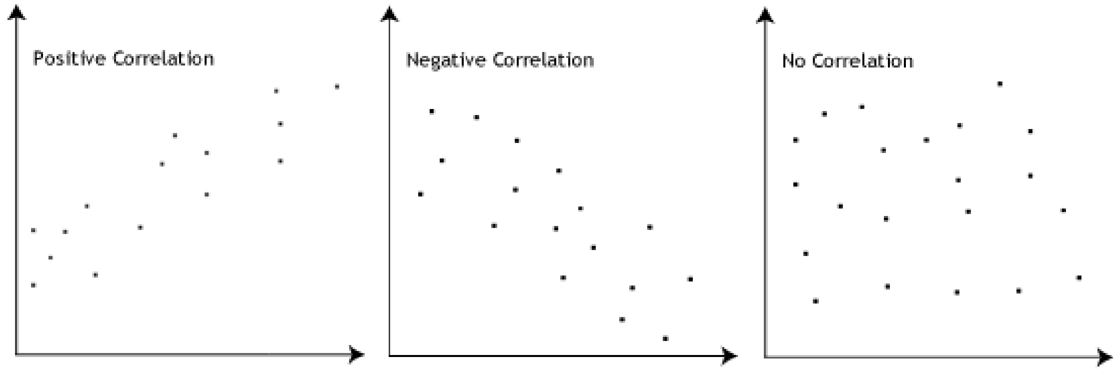
Source: https://www.geeksforgeeks.org/anscombes-quartet/

Conclusion: While the descriptive statistics of Anscombe's Quartet may appear uniform, the accompanying visualizations reveal distinct patterns, showcasing the necessity of combining statistical analysis with graphical exploration for robust data interpretation

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

Response-3: The Pearson product-moment correlation coefficient (or Pearson correlation coefficient, for short) is a measure of the strength of a linear association between two variables and is denoted by r. Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, r, indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit).

The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below



Source: ***https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php***

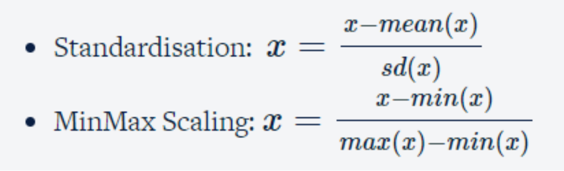
Below table can be used to interpret the r value

r value	Interpretation
r=1	Perfect positive linear correlation
1>r≥0.81	Strong positive linear correlation
0.8>r≥0.40	Moderate positive linear correlation
0.4>r>0	Weak positive linear correlation
r=0	No correlation
0>r≥−0.40	Weak negative linear correlation
−0.4>r≥−0.8	Moderate negative linear correlation
−0.8>r>−1	Strong negative linear correlation
r=−1	Perfect negative linear correlation

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

Response-4: Scaling is a process of transforming the values of variables to a specific range. This is done to ensure that all variables have a comparable impact on the regression model. Scaling can help prevent certain variables from dominating the model due to their larger magnitude. For example, in a model where there are two independent variables as “AGE” and “Income in INR”. Here income will always have large values while age will be in single or double digit

Two major methods are employed to scale the variables: standardisation and MinMax scaling. Standardisation brings all the data into a standard normal distribution with mean 0 and standard deviation 1. MinMax scaling, on the other hand, brings all the data in the range of 0-1. The formulae used for each of these methods are as given below;



- Standardisation: $x = \frac{x - mean(x)}{sd(x)}$
- MinMax Scaling: $x = \frac{x - min(x)}{max(x) - min(x)}$

Scaling just affects the coefficients and none of the other parameters, such as t-statistic, F-statistic, p-values and R-squared. Also, the interpretation of the coefficients in linear regression is not affected by scaling. If we scale a variable, its coefficient will change accordingly to reflect the scaled units, but the interpretation remains the same

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

Response-5: Infinite VIF is a Sign of Perfect Multicollinearity. Infinite VIF means that the variable is a perfect linear combination of the other independent variables in the model.

Causes of Infinite VIF are,

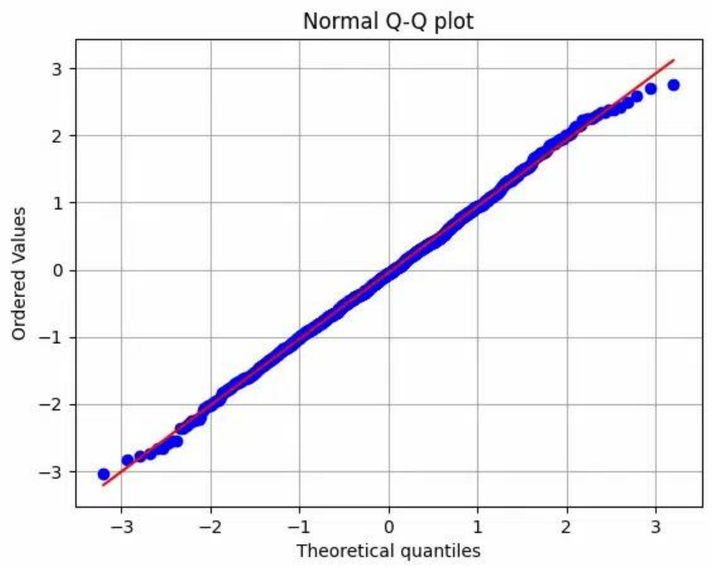
- **Redundant Variables**: Including two or more variables that are perfectly correlated or linearly dependent. For example, including both "distance in meters" and " distance in kilometers"
- **Dummy Variable Trap**: Including all possible dummy variables for a categorical variable without dropping one reference level

Infinite VIF can be addressed by identifying and removing the redundant variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Response-6: The QQ plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a normal or exponential.

Use: A QQ plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.



Source: https://www.geeksforgeeks.org/quantile-quantile-plots/

Importance:

i. **Flexible Comparison**: Q-Q plots can compare datasets of different sizes without requiring equal sample sizes

ii. **Dimensionless Analysis**: They are dimensionless, making them suitable for comparing datasets with different units or scales

iii. **Visual Interpretation**: Provides a clear visual representation of data distribution compared to a theoretical distribution

iv. **Sensitive to Deviations**: Easily detects departures from assumed distributions, aiding in identifying data discrepancies

v. **Diagnostic Tool**: Helps in assessing distributional assumptions, identifying outliers, and understanding data patterns