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)

In the dataset, we examined a total of 7 categorical variables. Some variables had binary values (0,1), while others had categories represented by numbers. We encoded these variables with categorical labels and created dummy variables for further analysis. Here are the key findings from the univariate and bivariate analysis:

Season: Among the four seasons, fall initially appeared to have higher bike bookings. However, after model building and analysis, it was concluded that summer and winter seasons have higher bike demands compared to spring and fall.
Yr: The variable "Yr" indicates a consistent increase in bike rental demands over the years, suggesting a growing trend in the popularity of bike rentals.
Mnth: The specific months of the year did not show a significant impact on bike rentals.
Temp: Temperature has a positive effect on bike rental demand, meaning that as the temperature increases, the demand for bike rentals tends to rise.
Holiday: Bike rental demand decreases during holidays, indicating that people are less likely to rent bikes on these occasions.
Weekday: Bike rentals exhibit higher demand on weekdays compared to weekends. Interestingly, Wednesday stands out as the weekday with the highest demand for bike rentals.
Workingday: Working days have higher demand for bike rentals compared to non-working days, implying that commuters or individuals with work-related purposes are the primary customers for bike rentals.
Weathersit: Clear weather conditions have a positive impact on bike rental demand, leading to more bookings. Conversely, light rain, cloudy, or misty weather conditions are associated with decreased demand for bike rentals.

Overall, based on the analysis, it is recommended to focus on increasing bike availability and promotional efforts during the summer and winter seasons, as they show higher bike rental demands. Additionally, taking into account temperature variations and planning marketing strategies accordingly can help maximize bike rental bookings. On holidays and weekends, special incentives or marketing campaigns could be designed to attract more customers. Finally, monitoring weather conditions and adjusting rental operations accordingly can contribute to better customer satisfaction and increased rental bookings.

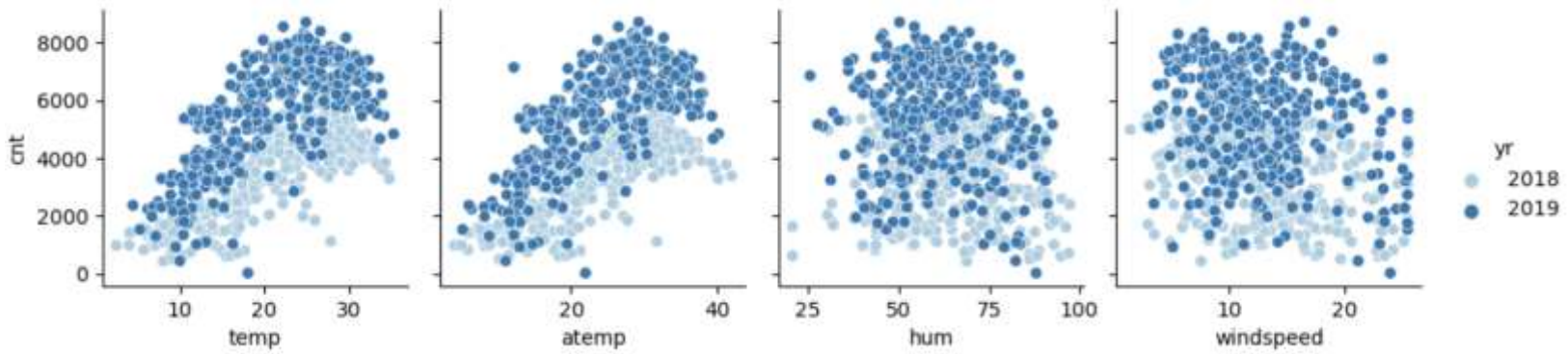
2. Why is it important to use drop_first=True during dummy variable creation?

It is important to use drop_first=True during dummy variable creation to avoid multicollinearity issues in regression models. When creating dummy variables, dropping the first category helps to remove perfect collinearity between the variables.

For instance, consider the following example: When encoding gender (with male represented as 1 and female as 0, or vice versa), if we don't set drop_first = True, a set of n dummy variables will be generated. These predictor variables (n dummy variables) will be correlated with each other, leading to a phenomenon called multicollinearity or the Dummy Variable Trap.

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

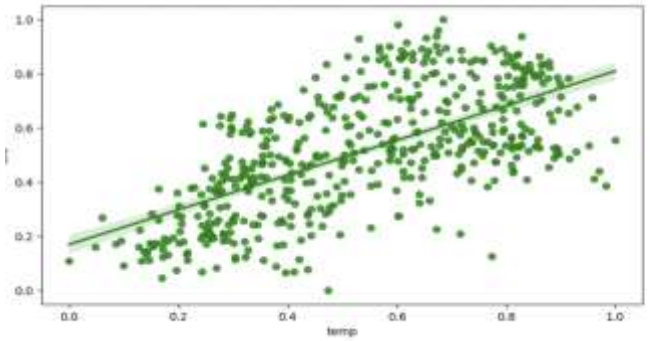
Looking at the pair plot from the analysis we can say that temp/atemp has highest positive correlation with the target variable which is **0.63%.**

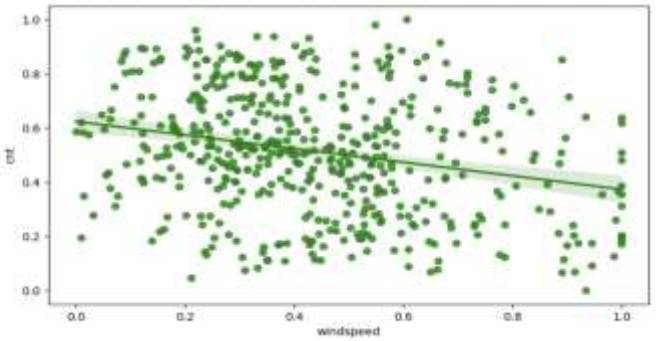


4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Linearity: Created scatter plots of the dependent variable against each independent variable to visually inspect if there is a linear relationship.

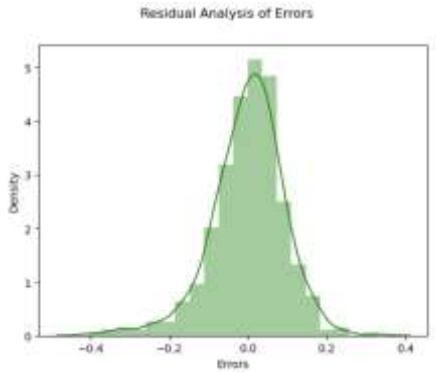
Observed a Linear relationship





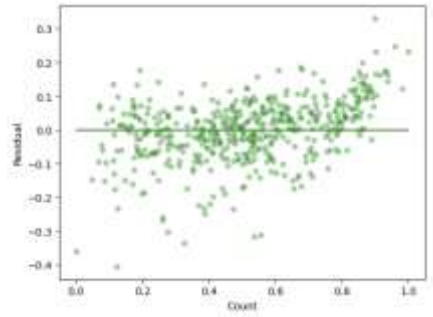
Normality of residuals: Plotted a histogram or a Q-Q plot of the residuals to assess if they follow a normal distribution.

The error terms were normally distributed in the model



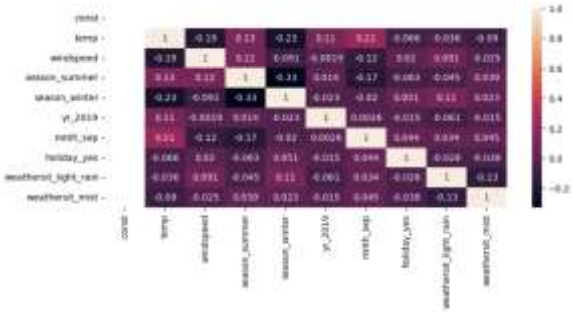
Homoscedasticity: Plotted the residuals against the predicted values to check if the spread of the residuals is consistent across all levels of the predicted values.

The spread of residuals is consistent across all predicted values



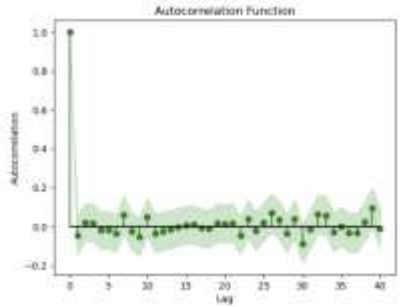
Multicollinearity: Used heatmap to identify the correlation between variables.

As per the heatmap the final model posses a weak or no multi collinearity



Residual autocorrelation: Plotted the residuals (to find correlation or pattern.

There is no autocorrelation identified as the plotted residual does not show any pattern or correlation.



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

Based on the Final Model the

Temp: temperature with a high coefficient of 0.548443 is the top variable that has the 1st place in explaining the demand of shared bikes.
Yr_2019 : Year 2019 with a high coefficient of 0.232747 is the 2nd top variable in explaining the demand of shared bikes.
weathersit_3(or light rain) : Weather situation 3 or light rain with a negative coefficient of - 0.283237 is the 3rd top variable in explaining the demand of shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a supervised learning algorithm used for predicting a continuous numerical output based on one or more input features. It assumes a linear relationship between the independent variables (features) and the dependent variable (target).

1.Data Preparation & EDA: Gather a dataset consisting of input features (X) and corresponding target values (y). Perform Data cleaning pre-processing and EDA on the Dataset, Split the dataset into a training set and a test set for model evaluation.
2.Model Representation: Linear regression represents the relationship between the features (X) and the target variable (y) using a linear equation of the form: **y = b0 + b1*x1 + b2*x2 + ... + bn*xn** where y is the predicted output, b0 is the y-intercept (bias), b1, b2, ..., bn are the coefficients (weights) associated with the input features x1, x2, ..., xn.
3.Training the Model: The goal is to find the optimal values for the coefficients (weights) that minimize the difference between the predicted values and the actual values. This is done by minimizing a cost function, such as the Mean Squared Error (MSE) or the Ordinary Least Squares (OLS) method. The model adjusts the weights iteratively using optimization algorithms like Gradient Descent or Normal Equation.
4.Model Evaluation: Once the model is trained, evaluate its performance using the test set. Common evaluation metrics for linear regression include the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared value (coefficient of determination).
5.Prediction: Finally, the trained linear regression model can be used to make predictions on new, unseen data by applying the learned coefficients to the input features.

Key Assumptions of Linear Regression:

- **Linearity**: Assumes a linear relationship between the independent variables and the target variable.
- **Independence**: Assumes that the observations are independent of each other.
- **Homoscedasticity**: Assumes that the variance of the errors (residuals) is constant across all levels of the independent variables.
- **Normality**: Assumes that the errors follow a normal distribution.

Linear regression is widely used for tasks such as sales forecasting, price prediction, trend analysis, and more, where there is a need to understand and model the relationship between variables in a linear fashion.

2. Explain the Anscombe’s quartet in detail.

Anscombe’s quartet comprises four datasets that have nearly identical simple descriptive statistics , yet have very different distributions and appear very different when graphed.

Anscombe's quartet refers to a set of four datasets that have nearly identical statistical properties but exhibit vastly different patterns when visualized.

The purpose of Anscombe's quartet is to demonstrate the importance of data visualization in understanding the underlying patterns and relationships within a dataset. It highlights the limitations of relying solely on summary statistics, such as mean, variance, and correlation, as they can be misleading without visual inspection. Anscombe's quartet emphasizes the need for exploratory data analysis and the critical role of data visualization in gaining insights from data.

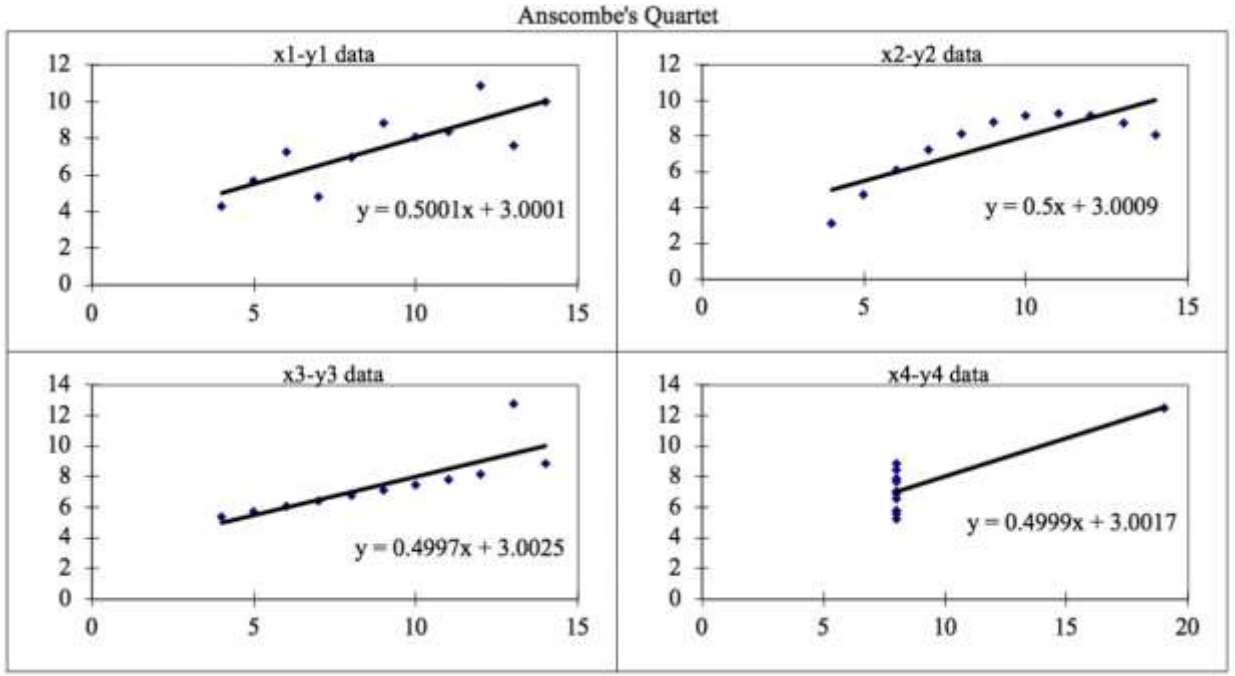
The four datasets can be described as:

1.Dataset 1: this **fits** the linear regression model pretty well.

2.Dataset 2: this **could not fit** linear regression model on the data quite well as the data is non-linear.

3.Dataset 3: shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

4.Dataset 4: shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model



3. What is Pearson’s R?

Pearson's R, also known as Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to 1, where -1 indicates a perfect negative linear relationship, 0 indicates no linear relationship, and 1 indicates a perfect positive linear relationship. It is widely used to assess the degree of association between variables in many fields of research.

Assumptions of Pearson’s R:

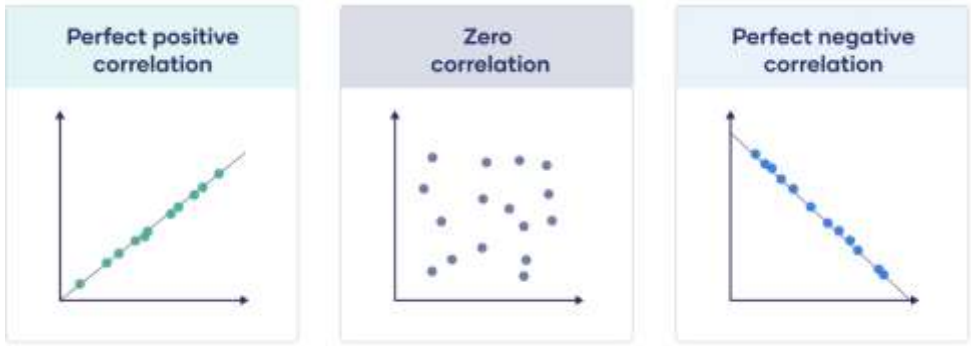
Linearity: The relationship between the two variables should be linear. Pearson's R is specifically designed to measure the strength of linear relationships.

Normality: The variables should follow a normal distribution. Deviations from normality can affect the accuracy of the correlation coefficient.

Homoscedasticity: The variability of the data points should be similar across all levels of the variables. This means that the spread of the data points should be constant.

Independence: The observations should be independent of each other. Autocorrelation or dependence between observations can affect the reliability of Pearson's R.

Correlation coefficient value	Correlation type	Meaning
1	Perfect positive correlation	When one variable changes, the other variables change in the same direction.
0	Zero correlation	There is no relationship between the variables.
-1	Perfect negative correlation	When one variable changes, the other variables change in the opposite direction.



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

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in 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 have to 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

Normalization/Min-Max Scaling:
It brings all of the data in the range of 0 and 1. ***sklearn.preprocessing.MinMaxScaler*** *helps to implement normalization in python.*

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:
*Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ**) zero and standard deviation one (**σ**)*

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

sklearn.preprocessing.scale *helps to implement standardization in python.*

One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers

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

In certain cases, the value of VIF can become infinite. This occurs when there is perfect multicollinearity among the independent variables in the regression model.

Perfect multicollinearity refers to a situation where one or more independent variables in a regression model can be exactly predicted from a linear combination of the other independent variables. In other words, there is a perfect linear relationship among the independent variables.

When perfect multicollinearity exists, it means that the coefficient of determination (R^2) for the affected variables is equal to 1. The VIF is calculated using the formula $VIF = 1 / (1 - R^2)$. When R^2 is equal to 1, the denominator in the formula becomes zero, resulting in an undefined division. Mathematically, dividing by zero is undefined, so the VIF value is said to be infinite.

In practical terms, this infinite VIF value indicates that the variable experiencing perfect multicollinearity can be fully explained by a linear combination of the other independent variables in the model. It signifies that the variable is redundant and does not contribute unique information to the regression analysis.

To resolve this issue, it is necessary to remove one of the variables involved in the perfect multicollinearity. By doing so, the VIF value for the remaining variables will no longer be infinite, and the model can be estimated correctly.

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

A Q-Q plot, or quantile-quantile plot, is a graphical tool used to assess the similarity between a dataset's distribution and a theoretical distribution. In linear regression, a Q-Q plot is crucial for evaluating the assumption of normality in residuals. By plotting observed quantiles against expected quantiles from a normal distribution, it helps visually identify departures from normality. A well-behaved Q-Q plot indicates that residuals follow a normal distribution, validating a key assumption in linear regression. Detecting deviations prompts investigation into potential model issues, allowing researchers to refine their analysis and ensure the reliability of their regression results.

How to Draw Q-Q plot

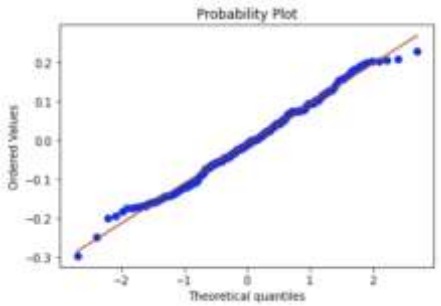
- Collect the data for plotting the quantile-quantile plot.
- Sort the data in ascending or descending order.
- Draw a normal distribution curve.
- Find the z-value (cut-off point) for each segment.
- Plot the dataset values against the normalizing cut-off points.

Advantages of Q-Q plot

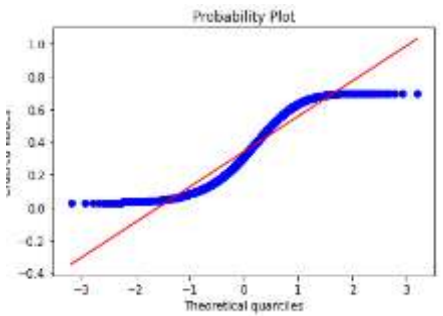
- Since Q-Q plot is like probability plot. So, while comparing two datasets the sample size need not to be equal.
- Since we need to normalize the dataset, so we don’t need to care about the dimensions of values.

Types of Q-Q plots

Shows normally distributed plot



For Left-tailed distribution



For the uniform distribution: