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 –

From our analysis of categorical variables from the dataset, we would infer below point s-

1. Count is drastically more for year 2019 than for 2018
2. Fall season seems to have more renting of bike as compare to other seasons
3. Renting of bike is more on working day rather on holiday as yes
4. Renting of bike is more when weather is Good and moderate
5. Renting of bike is highest in September month although mean of all month are having only slight variation.
6. Bike rental business is very much more in 2019 rather than 2018

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

Ans –

In Python, when creating dummy variables for categorical features in linear regression modeling, setting drop_first=True is important to avoid multicollinearity issues.

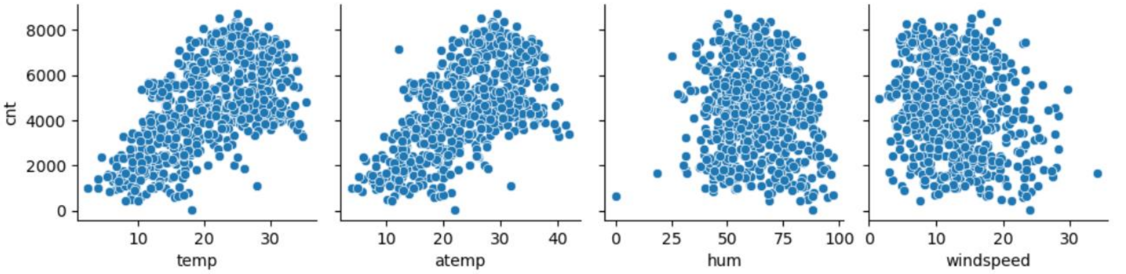
1.Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated. In the context of dummy variables, if you include all levels of a categorical variable as dummy variables, it creates perfect multicollinearity because one level can be perfectly predicted from the others. This leads to unstable estimates of the coefficients.

2.By dropping the first level of each categorical variable when creating dummy variables, you essentially remove one redundant dummy variable. This leaves n−1 dummy variables for n levels of the categorical variable, which is sufficient to represent all levels without redundancy.

So, setting drop_first=True in dummy variable creation helps to mitigate multicollinearity issues and eliminates redundancy in the regression coefficients.

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

Ans – temp and atemp has the highest correlation with target variable



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

Ans –

Validating the assumptions of linear regression after building the model on the training set typically involves several diagnostic checks.

1. Normality of Residuals: Check whether the residuals are approximately normally distributed. You can visually inspect a histogram. Departure from normality suggests that the model assumptions may not hold.

2. Independence of Residuals: Check for independence of residuals. This means that the residuals should not exhibit any pattern over time or across observations. You can plot residuals against time or against the order of observations to detect any patterns.

3. Linearity: Assess the linearity assumption by plotting the observed values against the predicted values. The relationship should be approximately linear. You can also use partial residual plots or component plus residual plots to assess linearity.

4. Collinearity: Check for multicollinearity among the independent variables. Calculate variance inflation factors (VIF) to assess the extent of multicollinearity. VIF values above a certain threshold (usually 5 or 10) indicate problematic multicollinearity.

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

Ans –

Top 3 features that significantly contributing towards explaining the demand of shared bikes are –

1. Temperature
2. Wind speed
3. Season - Spring,Summer and Winter
4. Year - 2019

General Subjective Questions

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

Ans –

Linear regression is a fundamental statistical method used for modeling the relationship between a dependent variable and one or more independent. It assumes that the relationship between the dependent variable and the independent variables is linear. The goal of linear regression is to find the best-fitting linear equation that describes the relationship between the variables.

1. **Model Representation** - Linear regression represents the relationship between the independent variables XX and the dependent variable YY using a linear equation: $Y=\beta 0+\beta 1X1+\beta 2X2+\ldots+\beta nXn+\epsilon$Y=β0+β1X1+β2X2+…+βnXn+ϵ where
    - YY is the dependent variable.
    - $X1,X2,\ldots,Xn$X1,X2,…,Xn are the independent variables.
    - $\beta 0,\beta 1,\ldots,\beta n$β0,β1,…,βn are the coefficients (parameters) representing the slope of the line for each independent variable.
    - ϵϵ is the error term, representing the difference between the observed and predicted values.
2. **Fitting the Model:**
    - The goal is to find the values of the coefficients $\beta 0,\beta 1,\ldots,\beta n$β0,β1,…,βn that minimize the difference between the observed values of YY and the values predicted by the linear equation.
    - This is typically done using a method called Ordinary Least Squares (OLS), which minimizes the sum of the squared differences between the observed and predicted values.
3. **Coefficient Estimation:**
    - The coefficients are estimated using the training data, typically through optimization techniques like gradient descent or analytical solutions.
    - For simple linear regression (with only one independent variable), the coefficients can be calculated analytically using formulas.
    - For multiple linear regression (with multiple independent variables), optimization algorithms are used to find the coefficients that minimize the error.
4. **Model Evaluation:** After fitting the model, it's important to evaluate its performance. This can involve various metrics such as:
    - R-squared: Measures the proportion of variance in the dependent variable that is explained by the independent variables.
    - Mean Squared Error (MSE): Measures the average squared difference between the observed and predicted values.
    - Residual Analysis: Examining the residuals to ensure they meet the assumptions of linear regression.

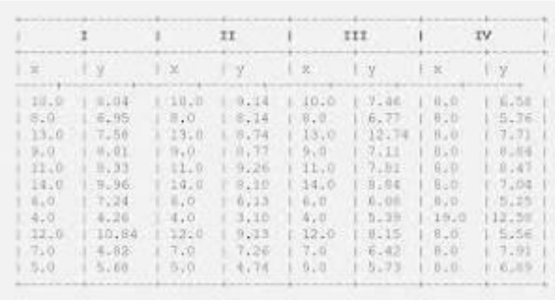
5. **Prediction**: Once the model is trained and evaluated, it can be used to make predictions on new or unseen data by substituting the values of the independent variables into the linear equation to calculate the predicted value of the dependent variable.

Overall, linear regression is a powerful technique for modeling the relationship between variables when the relationship is assumed to be linear. However, it's important to note that it has certain assumptions, such as linearity, independence of errors, constant variance of errors, and normality of errors, which should be assessed and validated during model building.

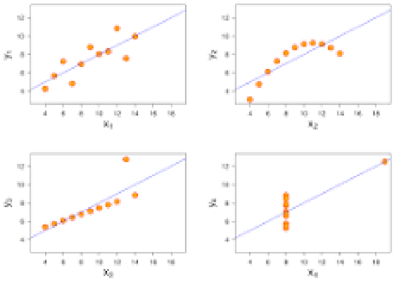
2. **Explain the Anscombe's quartet in detail.** (3 marks)

Ans –

Anscombe's quartet is a set of four datasets that have nearly identical statistical properties but differ significantly when graphed. It was created by the statistician Francis Anscombe in 1973 to illustrate the importance of visualizing data before drawing conclusions based solely on summary statistics.



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



It tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

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

Ans –

Pearson's correlation coefficient, often denoted as r, it is a measure of the linear relationship between two continuous variables. It quantifies the strength and direction of the linear association between two variables. Pearson's correlation coefficient ranges from -1 to +1, where:

r=+1: Perfect positive linear relationship

r=−1: Perfect negative linear relationship

r=0: No linear relationship

It can be calculated using formula -

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \cdot \sum (Y_i - \bar{Y})^2}}$$

- Xi and Yi are the individual data points.
- $X^{-}$and Y^{-} are the means of the respective variables.
- The denominator is the product of the standard deviations of the two variables.

Below are some key points about this –

- Strength of Association: The magnitude of r indicates the strength of the linear relationship. Larger values of $|r|$ indicate stronger linear relationships, while values closer to 0 indicate weaker relationships.

- Direction of Association: The sign of r indicates the direction of the relationship. A positive r indicates a positive linear relationship (as one variable increases, the other also tends to increase), while a negative r indicates a negative linear relationship (as one variable increases, the other tends to decrease).

- Assumes Linearity and Homoscedasticity: Pearson's correlation coefficient assumes that the relationship between variables is linear and that the variance of the variables is constant across all levels of the variables (homoscedasticity).

- Sensitive to Outliers: Pearson's r is sensitive to outliers, meaning that extreme values can disproportionately influence the correlation coefficient.

Pearson's correlation coefficient is widely used in various fields such as statistics, social sciences, finance, and many others to assess the strength and direction of relationships between variables.

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

Ans –

Scaling is a preprocessing technique used in data analysis and machine learning to transform the features of a dataset to a similar scale. It involves adjusting the range of values of the features to a common scale, which can be beneficial for various algorithms and analyses. Scaling is performed to ensure that all features contribute equally to the analysis, prevent certain features from dominating the model, and improve the performance of machine learning algorithms.

Scaling is performed because of :

- Features with larger scales or magnitudes may dominate the learning algorithm, leading to biased results. Scaling ensures that all features contribute equally to the analysis by bringing them to a similar scale.

- Many machine learning algorithms, such as gradient descent-based optimization algorithms, converge faster when the features are on a similar scale. Scaling helps in speeding up the convergence process and reaching the optimal solution more efficiently.

- Regularization techniques, such as L1 and L2 regularization, penalize large coefficients. Scaling ensures that the regularization term applies uniformly to all features, preventing certain features from being unfairly penalized due to their scale.

- Distance-based algorithms, such as K-nearest neighbors (KNN) and clustering algorithms, are sensitive to the scale of features. Scaling ensures that the distances between data points are calculated accurately and prevents features with larger scales from dominating the distance calculations.

There are two common methods for scaling data: normalized scaling and standardized scaling.

Normalized Scaling:

- Also known as Min-Max scaling.
- Involves scaling the features to a fixed range, typically between 0 and 1.
- The formula for normalized scaling is:

  $$X_{\text{normalized}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$
-
- X min and X max are the minimum and maximum values of the feature, respectively.
- Normalized scaling preserves the relative relationships between the original values and is suitable when the distribution of the data is not Gaussian.

Standardized Scaling:

- Also known as Z-score normalization or standardization.
- Involves scaling the features to have a mean of 0 and a standard deviation of 1.
- The formula for standardized scaling is:

  $$X_{\text{standardized}} = \frac{X - \mu}{\sigma}$$

- µ is the mean of the feature, and σ is the standard deviation.
- Standardized scaling transforms the features to have a Gaussian distribution with a mean of 0 and a standard deviation of 1.
- Standardized scaling is robust to outliers and is often preferred when the data distribution is Gaussian or unknown.

Scaling is performed to ensure that all features contribute equally to the analysis and to improve the performance of machine learning algorithms. Normalized scaling and standardized scaling are two common methods used for scaling data, each with its own advantages and suitable scenarios.

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

Ans –

Yes, it's possible for the Variance Inflation Factor (VIF) to become infinite in certain situations. The VIF measures the extent of multicollinearity in a regression model by quantifying how much the variance of an estimated regression coefficient is inflated due to multicollinearity with other predictors. When the VIF is infinite for a particular predictor variable, it indicates an extremely high degree of multicollinearity with one or more other predictor variables in the model.

Here are some scenarios where the VIF might become infinite:

- When one or more predictor variables in the regression model can be perfectly predicted by a linear combination of other predictor variables, perfect multicollinearity occurs. In this case, the VIF for the predictor variable(s) that can be perfectly predicted becomes infinite because the variance of the estimated coefficient cannot be calculated.

- If the predictors are linearly dependent, meaning one predictor can be expressed as a linear combination of the others, it can lead to very high VIF values, potentially approaching infinity.

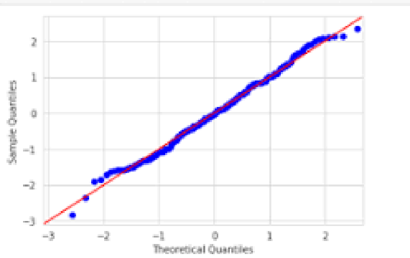
- In an under parameterized model where the number of predictors is greater than the number of observations, it's possible to encounter infinite VIF values. This occurs because the model cannot be uniquely estimated, leading to undefined or infinite VIF values.

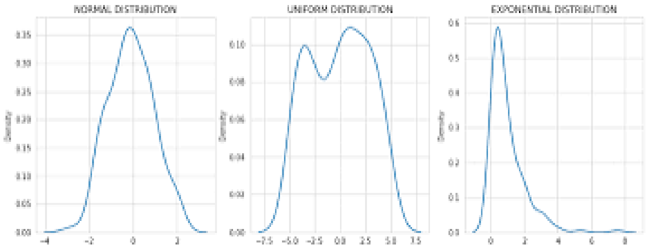
- In some cases, numerical precision issues or computational errors may lead to seemingly infinite VIF values. These issues can arise due to limitations in floating-point arithmetic or the numerical methods used to calculate VIF.

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

Ans –

A Q-Q plot, known as quantile-quantile plot, is a graphical tool used to assess whether a dataset follows a particular probability distribution or if two datasets have similar distributions. It compares the quantiles of the dataset against the quantiles of a theoretical distribution, typically the normal distribution.





The use and importance of a Q-Q plot in linear regression include:

- Normality Assumption Checking: In linear regression, one of the key assumptions is that the residuals (the differences between observed and predicted values) are normally distributed. A Q-Q plot of the residuals allows us to visually assess whether this assumption holds. If the residuals closely follow a straight line on the Q-Q plot, it indicates that they are approximately normally distributed. Deviations from a straight line suggest departures from normality.

- Identification of Distributional Differences: Q-Q plots can also be used to compare the distribution of residuals from different models or datasets. By comparing the quantiles of the residuals from different models against each other or against a theoretical distribution, we can assess whether the distributions are similar or if there are significant differences.

- Outlier Detection: Q-Q plots can help in identifying outliers or extreme observations in the dataset. Outliers may appear as points that deviate significantly from the expected straight line pattern on the plot.

- Model Assessment: Q-Q plots are useful for evaluating the adequacy and fit of a linear regression model. A well-fitting model should have residuals that closely follow a straight line on the Q-Q plot, indicating that the model captures the underlying patterns in the data adequately.

Overall, Q-Q plots are valuable diagnostic tools in linear regression analysis, providing insights into the normality of residuals and the adequacy of the regression model. They help researchers and analysts make informed decisions about the validity and reliability of their regression analyses.