**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)

As part of my analysis, from the dataset provided, following variables are taken as Categorical - Season, Holiday and Working day.

1. Season – This categorical variable, especially “FALL” has a positive impact on the dependent variable “cnt” which is helping/driving more no. of Bike Sharing possibilities.
2. Holiday – This categorical variable, has a positive impact on the dependent variable “cnt” which is helping/driving more no. of Bike Sharing possibilities. Though this variable has not been taken up as part of final model building, based on its impact
3. Workingday - This categorical variable has a positive impact on the dependent variable “cnt” which is helping/driving more no. of Bike Sharing possibilities. We see an increase in “cnt” on non-working day

2. Why is it important to use drop\_first=True during dummy variable creation? (2 mark)

Using **drop\_first=True** during dummy variable creation is important to avoid the "dummy variable trap" or "multicollinearity" in regression analysis. The dummy variable trap is a situation where the presence of highly correlated dummy variables in a regression model can lead to unstable coefficient estimates and difficulties in interpreting the model.

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

**Temperature**

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

Based on the output values of “P”, “R-Squared” and Coefficient values of independent variables we could conclude that linear regression assumptions are moderate.

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

**Temperature, Year , Month and Season (Fall)**

**General Subjective Questions**

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

Linear regression is a widely used statistical method and machine learning algorithm for modeling the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. It's particularly useful for predicting continuous numeric values. Following are the important aspects of this algorithm:

**The Linear Equation:**

At the core of linear regression is the linear equation: **y = mx + b**

In the context of linear regression:

* y is the dependent variable (the one you want to predict).
* x is the independent variable (the one used for prediction).
* m is the slope or coefficient, which represents the change in y for a unit change in x.
* b is the intercept, representing the value of y when x is zero.

In a multiple linear regression, where you have more than one independent variable, the equation extends to: **y = b0 + b1\*x1 + b2\*x2 + ... + bn\*xn**

Here, b0 is the intercept, and b1, b2, ..., bn are the coefficients for the independent variables x1, x2, ..., xn.

**Training the Model:**

The goal of linear regression is to find the best-fitting line (or hyperplane in the case of multiple linear regression) that minimizes the difference between the predicted values and the actual data points. This is achieved by finding the optimal values for the coefficients m and b (or b0, b1, b2, ..., bn) that minimize the error. Typically, the least squares method is used to calculate the sum of squared differences between the observed and predicted values, and the coefficients are adjusted iteratively to minimize this sum.

**Evaluation and Predictions:**

Once the model is trained and the coefficients are determined, you can use it to make predictions on new data. Simply plug the values of the independent variables into the equation, and the model will give you a predicted value for the dependent variable.

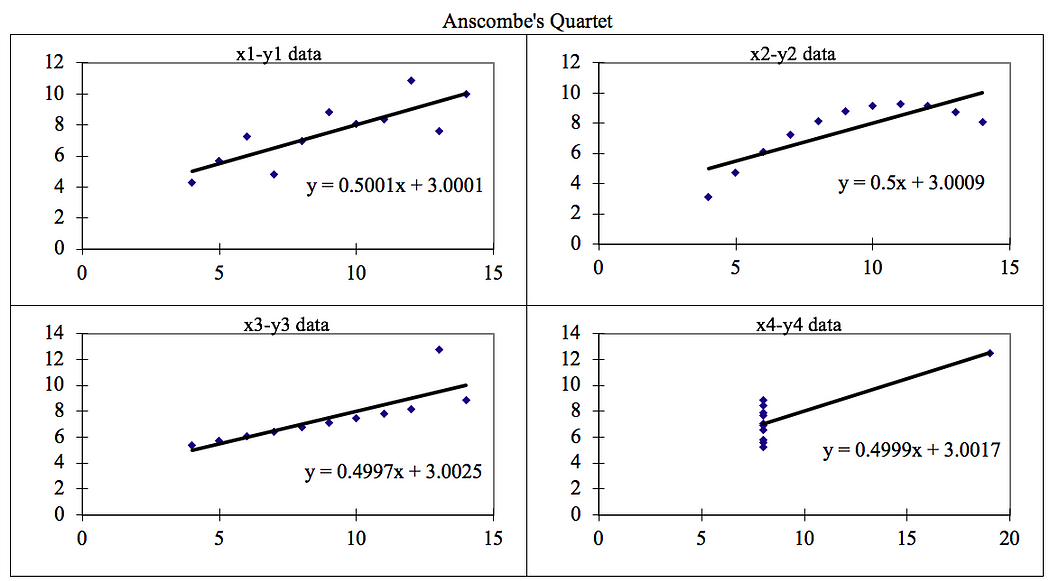
**Model Assumptions:** Linear regression makes several key assumptions:

* Linearity: It assumes that the relationship between independent and dependent variables is linear.
* Independence: It assumes that the errors or residuals (the differences between observed and predicted values) are independent of each other.
* Homoscedasticity: It assumes that the variance of the residuals is constant across all levels of the independent variable(s).
* Normality: It assumes that the residuals follow a normal distribution.

**Variants of Linear Regression:**

* Simple Linear Regression: When there is only one independent variable.
* Multiple Linear Regression: When there are multiple independent variables.
* Polynomial Regression: When the relationship between variables is better represented by a polynomial equation.
* Ridge Regression and Lasso Regression: Variations of linear regression that help address multicollinearity and overfitting issues by adding regularization terms to the equation.

Linear regression is a powerful and interpretable tool for modeling relationships between variables, but it may not perform well if the underlying assumptions are violated or if the relationship between variables is not truly linear. In such cases, other regression techniques or more complex models may be more appropriate.

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

**Anscombe’s Quartet**can be defined as a group of four data sets which are **nearly identical in simple descriptive statistics**, but there are some peculiarities in the dataset that **fools the regression model**if built. They have very different distributions and **appear differently**when plotted on scatter plots. It was constructed in 1973 by statistician **Francis Anscombe** to illustrate the **importance**of **plotting the graphs**before analyzing and model building, and the effect of other **observations on statistical properties**. There are these four data set plots which have nearly **same statistical observations**, which provides same statistical information that involves **variance**, and **mean**of all x, y points in all four datasets. This tells us about the importance of visualizing the data before applying various algorithms out there to build models. When these models are plotted on a scatter plot, all datasets generates a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows:

A collage of graphs

Description automatically generated

The four datasets can be described as: Anscombe's quartet is a famous dataset in statistics that consists of four small datasets, each with two variables. What makes Anscombe's quartet intriguing is that despite their very different data points, these four datasets have nearly identical simple descriptive statistics. This paradox serves as a compelling reminder of the importance of data visualization and the limitations of summary statistics**.**

1. **Similar Summary Statistics:** When you calculate summary statistics (mean, variance, correlation, linear regression parameters, etc.) for each of the four datasets, you'll find that they are remarkably similar. This might lead one to assume that the datasets are almost identical.

2. **Different Data Patterns:** However, if you plot these datasets, you'll discover that they have vastly different patterns. Dataset 1 appears roughly linear, Dataset 2 resembles a curved relationship, Dataset 3 has an outlier, and Dataset 4 is dominated by a single outlier.

3. **Importance of Visualization:** Anscombe's quartet emphasizes the importance of data visualization. Summary statistics alone may not reveal the true nature of the data. It illustrates how visualizing data can help uncover patterns, outliers, and relationships that may not be apparent through statistics alone.

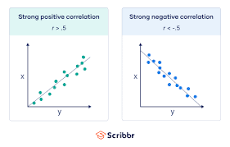
4 . Cautious Interpretation: It also underscores the importance of not relying solely on summary statistics when analyzing data. An understanding of the underlying data structure, context, and relationships between variables is crucial.

**Conclusion:**

*We have described the four datasets that were intentionally created to describe the importance of data visualisation and how any regression algorithm can be fooled by the same. Hence, all the important features in the dataset must be visualised before implementing any machine learning algorithm on them which will help to make a good fit model.*

**3. What is Pearson’s R? (3 marks)**

Pearson's correlation coefficient, often denoted as "r," is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It assesses how closely the data points in a scatterplot cluster around a straight line. The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables. Although interpretations of the relationship strength (also known as effect size) vary between disciplines, the table below gives general rules of thumb: The coefficient ranges from -1 to 1, with specific interpretations:



* Positive Correlation (r > 0): When r is positive, it indicates a positive linear relationship between the two variables. As one variable increases, the other tends to increase as well. The closer r is to 1, the stronger the positive correlation.
* No Correlation (r = 0): An r value of 0 suggests no linear relationship between the variables. Changes in one variable do not predict changes in the other.
* Negative Correlation (r < 0): When r is negative, it signifies a negative linear relationship. As one variable increases, the other tends to decrease. The closer r is to -1, the stronger the negative correlation.

## **Visualizing the Pearson correlation coefficient**

Another way to think of the Pearson correlation coefficient (r) is as a measure of how close the observations are to a [line of best fit](https://www.scribbr.com/statistics/simple-linear-regression/#how-to-perform-a-simple-linear-regression). The Pearson correlation coefficient also tells you whether the slope of the line of best fit is negative or positive. When the slope is negative, r is negative. When the slope is positive, r is positive.

* When r is 1 or –1, all the points fall exactly on the line of best fit:
* When r is greater than .5 or less than –.5, the points are close to the line of best fit:
* When r is between 0 and .3 or between 0 and –.3, the points are far from the line of best fit:
* When r is 0, a line of best fit is not helpful in describing the relationship between the variables:

**When to use the Pearson correlation coefficient**

The Pearson correlation coefficient (r) is one of several [correlation coefficients](https://www.scribbr.com/statistics/correlation-coefficient/#types-of-correlation-coefficients) that you need to choose between when you want to measure a correlation. The Pearson correlation coefficient is a good choice when **all**of the following are true:

* **Both variables are**[**quantitative**](https://www.scribbr.com/methodology/types-of-variables/#quantitative-vs-categorical)**:** You will need to use a different method if either of the variables is [qualitative](https://www.scribbr.com/methodology/qualitative-research/).
* **The variables are**[**normally distributed**](https://www.scribbr.com/statistics/normal-distribution/)**:** You can create a histogram of each variable to verify whether the distributions are approximately normal. It’s not a problem if the [variables](https://scribbr.com/methodology/types-of-variables/) are a little non-normal.
* **The data have** **no**[**outliers**](https://www.scribbr.com/statistics/outliers/)**:** Outliers are observations that don’t follow the same patterns as the rest of the data. A scatterplot is one way to check for outliers—look for points that are far away from the others.
* **The relationship is linear:** “Linear” means that the relationship between the two variables can be described reasonably well by a straight line. You can use a scatterplot to check whether the relationship between two variables is linear.

### Pearson vs. Spearman’s rank correlation coefficients

[Spearman’s rank correlation coefficient](https://www.scribbr.com/statistics/correlation-coefficient/#spearmans-rho) is another widely used correlation coefficient. It’s a better choice than the Pearson correlation coefficient when **one or more**of the following is true:

* The variables are [ordinal](https://www.scribbr.com/statistics/ordinal-data/).
* The variables aren’t [normally distributed](https://www.scribbr.com/statistics/normal-distribution/).
* The data includes outliers.
* The relationship between the variables is non-linear **and**monotonic.

## **Calculating the Pearson correlation coefficient**

Below is a formula for calculating the Pearson correlation coefficient (r):

\begin{equation*} r = \frac{ n\sum{xy}-(\sum{x})(\sum{y})}{% \sqrt{[n\sum{x^2}-(\sum{x})^2][n\sum{y^2}-(\sum{y})^2]}} \end{equation*}

The formula is easy to use when you follow the step-by-step guide below. You can also use software such as R or Excel to calculate the Pearson correlation coefficient for you.

**Significance and Uses of Pearson's Correlation Coefficient:**

* Quantifying Relationships: Pearson's r is a valuable tool for quantifying and summarizing the strength and direction of linear relationships between two variables. It provides a numerical representation of how closely the data points align with a linear trend.
* Hypothesis Testing: It is used in hypothesis testing to determine whether the observed correlation is statistically significant or if it could have occurred by random chance. The significance is typically assessed through a hypothesis test, where the null hypothesis assumes no correlation (r = 0).
* Modeling and Prediction: In predictive modeling, Pearson's r helps assess which variables might be good predictors of each other. For example, in finance, it can be used to analyze the relationship between two stock prices to inform investment decisions.
* Quality Control: It is used in quality control and process improvement to study the relationship between process parameters and product quality, identifying factors that may impact product performance.
* Assumption Checking: In linear regression analysis, Pearson's r is used to check the assumption of linearity. If the correlation between the dependent and independent variables is strong, linear regression may be an appropriate modeling technique.
* Data Exploration: In data exploration and visualization, scatterplots with correlation coefficients are commonly used to quickly assess the relationship between variables in a dataset.

It's important to note that Pearson's correlation coefficient is sensitive to outliers and assumes a linear relationship between variables. If the relationship is nonlinear or if there are outliers present, alternative correlation measures or data transformations may be more appropriate. In summary, Pearson's r is a valuable statistical tool for quantifying and assessing linear relationships between two continuous variables. Its significance lies in its ability to provide insights into the nature and strength of these relationships, aiding in hypothesis testing, modeling, and data exploration.

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

Scaling is a data preprocessing technique used in statistics and machine learning to transform and standardize the range or distribution of variables. The primary goal of scaling is to make the data suitable for analysis, modeling, and comparison by ensuring that all variables are on a similar scale. Here's why scaling is performed and the difference between normalized scaling and standardized scaling:

**Why Scaling is Performed:**

1. **Equalize Variable Scales:** Variables often have different units of measurement and magnitudes. Scaling brings them to a common scale, ensuring that no single variable dominates the analysis or model due to its larger range.
2. **Enhance Convergence:** Many machine learning algorithms, especially those based on distance or gradient descent, perform better when the variables are on a similar scale. Scaling helps these algorithms converge more quickly.
3. **Improve Model Performance:** Scaling can lead to improved model performance and stability. Models like support vector machines and k-nearest neighbors are sensitive to variable scales.
4. **Interpretability:** Scaling makes it easier to interpret the coefficients or feature importances in linear models. It ensures that the coefficients are comparable across variables.

**Normalized Scaling (Min-Max Scaling):**

Normalized scaling, also known as Min-Max scaling, transforms the data into a specific range, typically between 0 and 1. Here's how it works:

* For each variable, the minimum value (min) is subtracted from each data point.
* The result is divided by the range (max - min) of the variable.

The formula for Min-Max scaling is: ***Xs*​*caled*=*Xm*​*ax*−*Xm*​*inX*−*Xm*​*in***​  
​Normalized scaling is useful when you want to preserve the relative relationships between data points but bring them into a common range.

**Standardized Scaling (Z-Score Scaling):**

Standardized scaling, often referred to as Z-score scaling or standardization, transforms the data such that it has a mean (average) of 0 and a standard deviation of 1. Here's how it works:

* For each variable, the mean (μ) is subtracted from each data point.
* The result is divided by the standard deviation (σ) of the variable.

The formula for standardization is: ***Xs*​*tandardized*=*σX*−*μ*​**

Standardized scaling is useful when you want to center the data around a mean of 0 and have a standard deviation of 1. This makes it easier to compare variables and interpret the relative importance of their values.

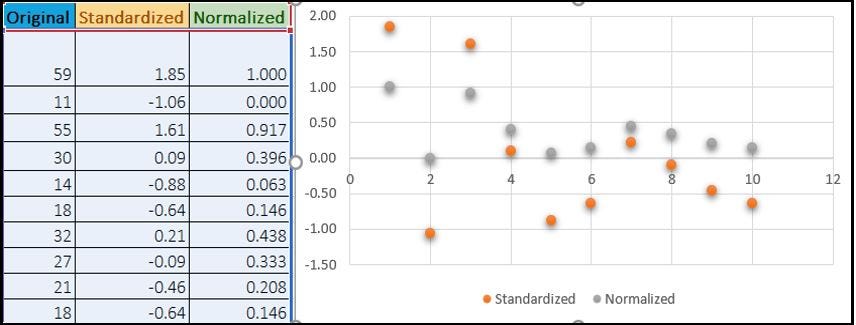
**Key Differences:**

1. **Range:** Normalized scaling (Min-Max) transforms data into a specific range (typically 0 to 1), while standardized scaling (Z-score) centers the data around 0 with a standard deviation of 1.
2. **Preservation of Relationships:** Normalized scaling preserves the relative relationships between data points, while standardized scaling does not.
3. **Outliers:** Standardized scaling is less sensitive to outliers because it uses the mean and standard deviation, which are robust statistics. Normalized scaling can be affected by outliers, as they can stretch the range.
4. **Interpretability:** Standardized scaling makes it easier to interpret coefficients or feature importances in linear models because the variables are centered around a mean of 0.

The choice between normalized and standardized scaling depends on the specific requirements of your analysis or modeling task, as well as the characteristics of your data.Top of Form

***Example:***

*Below shows example of Standardized and Normalized scaling on original values.*



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

A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model. This can adversely affect the regression results. Thus, the variance inflation factor can estimate how much the variance of a regression coefficient is inflated due to multicollinearity. A variance inflation factor is a tool to help identify the degree of multicollinearity. Multiple regression is used when a person wants to test the effect of multiple variables on a particular outcome. The dependent variable is the outcome that is being acted upon by the independent variables—the inputs into the model. Multicollinearity exists when there is a linear relationship, or correlation, between one or more of the independent variables or inputs.

The formula for VIF for an independent variable in a regression model is as follows: **VIFi​=1/1−R2​1​**

Where:

* *VIFi*​ is the Variance Inflation Factor for the ith independent variable.
* 2*Ri*2​ is the coefficient of determination (R-squared) for a regression model where the ith independent variable is regressed against all the other independent variables.

Now, the reason VIF can become infinite is related to the nature of R-squared (�2*R*2) and how it's calculated. When two or more independent variables are highly correlated, the regression model will have difficulty distinguishing the individual effects of each variable. As a result, the R-squared value for the regression of one of those variables against the others can become extremely close to 1, indicating that almost all of the variance in that variable can be explained by the other variables.

In the formula for VIF, when ��2*Ri*2​ is very close to 1 (approaching 1 from above), the denominator becomes very close to 0. Since you cannot divide by 0, the VIF for that variable becomes mathematically infinite.

In practical terms, this means that for an independent variable with an infinite VIF, there is an extremely high degree of multicollinearity between that variable and the other independent variables in the model. In such cases, it's often necessary to address the multicollinearity issue by either removing one of the correlated variables, combining them into a single variable, or using regularization techniques like ridge regression to mitigate the impact of multicollinearity on the model's stability and interpretability.

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

A Quantile-Quantile (Q-Q) plot is a graphical tool used in statistics and data analysis to assess whether a dataset follows a particular theoretical distribution, typically the normal distribution. The Q-Q plot compares the quantiles of the observed data against the quantiles of the expected distribution, such as the normal distribution. It helps in visualizing how well the data fits a theoretical distribution and identifies departures from that distribution.

Here's how a Q-Q plot works:

1. Data Sorting: The first step is to sort the data in ascending order.
2. Quantile Calculation: For each data point, the corresponding quantile value (proportion or percentile) is calculated. For instance, if you have N data points, the ith data point's quantile would be (i-0.5)/N.
3. Theoretical Quantiles: Similarly, quantiles for the chosen theoretical distribution (e.g., the normal distribution) are calculated.
4. Plotting: The quantiles of the observed data are plotted on the y-axis, while the quantiles of the theoretical distribution are plotted on the x-axis. This creates a scatterplot.

A Q-Q plot that forms a straight line with a 45-degree slope (y = x) suggests that the observed data closely follows the theoretical distribution. Departures from this straight line indicate deviations from the expected distribution.

Use and Importance of a Q-Q Plot in Linear Regression:

1. Normality Assumption: 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 is a valuable tool for assessing this assumption. If the residuals are normally distributed, the points on the Q-Q plot should closely follow the 45-degree line.
2. Detecting Departures from Normality: If the Q-Q plot deviates from a straight line, it indicates potential departures from the normality assumption. For example, if the Q-Q plot shows an upward or downward curve, it suggests non-normality in the data. This departure might prompt further investigation or the need for data transformation.
3. Identifying Outliers: Outliers in the data can disrupt the normality assumption. A Q-Q plot can help detect outliers as they tend to appear as points far away from the expected line in the plot.
4. Model Evaluation: Q-Q plots can be used to evaluate the residuals from a linear regression model. If the residuals exhibit a strong departure from normality, it may affect the model's reliability and the validity of associated statistical tests and confidence intervals.
5. Model Diagnostics: Q-Q plots are part of a set of tools used for model diagnostics in regression analysis. Alongside other diagnostic plots like scatterplots of residuals, they provide insights into the overall health of the regression model.

In summary, a Q-Q plot is a useful visualization tool for assessing the normality assumption of residuals in linear regression. It helps identify departures from normality, outliers, and potential issues with the regression model. If the Q-Q plot indicates deviations from normality, further investigation or data transformation may be necessary to meet the assumptions of linear regression.

**Top of Form**