# Assignment questions

-Kuldeep Parmar

# Assignment-based Subjective Questions

**Que1:** From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

After analyzing the categorical columns using box plots and bar plots. Based on the visualization, the following observations can be made:

- Most of the bookings have been done during the month of May, June, July, August, September, and October. The trend shows an increase from the beginning of the year until the middle of the year, followed by a decrease towards the end of the year.
- Thu, Fri, Sat, and Sun have a higher number of bookings compared to the start of the week.
- The fall season seems to have attracted more bookings. Additionally, the booking count has increased significantly from 2018 to 2019 in each season.
- When it's not a holiday, the number of bookings seems to be lower, which is reasonable as people may prefer to spend time at home and enjoy with their families during holidays.
- Clear weather conditions seem to have attracted more bookings, which is expected.
- The number of bookings appears to be roughly the same on both working and non-working days.
- In 2019, there was a higher number of bookings compared to the previous year, indicating positive progress in terms of business.

# Que 2: Why is it important to use drop_first=True during dummy variable creation?

Using **drop_first=True** during dummy variable creation helps to address multicollinearity, improve interpretability, and enhance the efficiency of the model.

However, it's important to carefully consider the specific context and requirements of the analysis before deciding whether to drop the first category or not.

**Let's consider a simple example** where we have a categorical variable "Color" with three categories: "Red", "Blue", and "Green". We want to create dummy variables for these categories.

**Without drop_first=True:**

This introduces perfect multicollinearity.

**If we create dummy variables with drop_first=True**

we avoid perfect multicollinearity and make the interpretation of the coefficients more straightforward.

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

Looking at pair-plot among numerical columns, ‘temp’ variable has the highest correlation with the target variable.

‘Temp’ variable shows correlation of 0.63 with respect to counts variable.

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

During the validation of the Linear Regression Model, the following assumptions have been assessed:

**1. Normality of error terms:** The error terms should follow a normal distribution.

**2. Multicollinearity check:** It is important to verify that there is no significant multicollinearity among the variables.

**3. Validation of linear relationship:**The presence of a linear relationship should be observed among the variables.

**4. Homoscedasticity:**There should be no discernible pattern in the residual values, indicating homoscedasticity.

**5. Independence of residuals:** Auto-correlation should be absent in the residuals.

By evaluating these assumptions, the validity of the Linear Regression Model can be established.

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

Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes :–

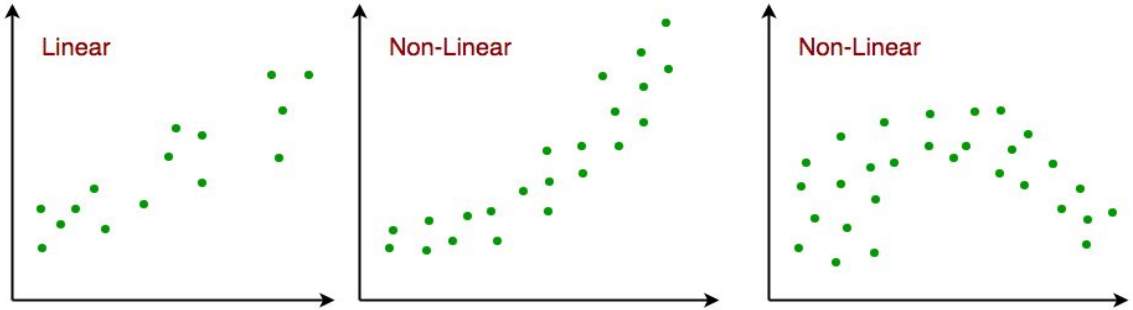
- temp
- winter
- sep

# General Subjective Questions

# **QUE 1:** Explain the linear regression algorithm in detail.

Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features. When the number of the independent feature, is 1 then it is known as Univariate Linear regression, and in the case of more than one feature, it is known as multivariate linear regression.

The goal of the algorithm is to find the best linear equation that can predict the value of the dependent variable based on the independent variables. The equation provides a straight line that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable(s).



# Explanation of the linear regression algorithm:

**Data Preparation:**

- Gather the dataset consisting of the independent variables (also called features or predictors) and the dependent variable (also known as the target variable or response variable).
- Perform any necessary data preprocessing steps, such as handling missing values, outlier detection, and feature scaling.

**Model Representation:**

- Linear regression represents the relationship between the dependent variable (Y) and the independent variables (X) using a linear equation of the form: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \varepsilon$ Here, Y is the predicted value, $\beta_0$ is the intercept term, $\beta_1$, $\beta_2$, ..., $\beta_n$ are the coefficients (also known as weights or parameters) corresponding to the independent variables $X_1$, $X_2$, ..., $X_n$, and $\varepsilon$ is the error term (residual).

**Cost Function:**

- The goal of linear regression is to find the optimal values for the coefficients that minimize the difference between the predicted values and the actual values of the dependent variable.
- This is achieved by defining a cost function, often the mean squared error (MSE), which quantifies the average squared difference between the predicted values and the actual values.
- The task is then to minimize this cost function.

**Parameter Estimation:**

- The coefficients $\beta_0$, $\beta_1$, $\beta_2$, ..., $\beta$□ are estimated using various methods. The most common method is Ordinary Least Squares (OLS), which calculates the coefficients that minimize the sum of squared residuals.
- OLS involves solving a system of equations to find the optimal values for the coefficients.

**Model Evaluation:**

- Once the coefficients are estimated, the model's performance needs to be evaluated.
- Common evaluation metrics for linear regression include the R-squared value, which indicates the proportion of the variance in the dependent variable explained by the model, and the Root Mean Squared Error (RMSE), which measures the average prediction error.
- Additionally, residual analysis and diagnostic plots are used to assess the assumptions of linear regression, such as the normality of residuals and homoscedasticity.

**Prediction and Inference:**

- The trained linear regression model can be used for making predictions on new, unseen data.
- The coefficients obtained from the model can also provide insights into the strength and direction of the relationship between the independent variables and the dependent variable.

# Assumption for Linear Regression Model

1. **Linearity**: The independent and dependent variables have a linear relationship with one another. This implies that changes in the dependent variable follow those in the independent variable(s) in a linear fashion.
2. **Independence**: The observations in the dataset are independent of each other. This means that the value of the dependent variable for one observation does not depend on the value of the dependent variable for another observation.
3. **Homoscedasticity**: Across all levels of the independent variable(s), the variance of the errors is constant. This indicates that the amount of the independent variable(s) has no impact on the variance of the errors.
4. **Normality**: The errors in the model are normally distributed.
5. **No multicollinearity**: There is no high correlation between the independent variables. This indicates that there is little or no correlation between the independent variables.

# **QUE 2:** Explain the Anscombe’s quartet in detail.

Anscombe's quartet is a famous statistical demonstration that highlights the importance of data visualization and the limitations of relying solely on summary statistics. It consists of four datasets that have nearly identical descriptive statistics but exhibit vastly different patterns when visualized.

The quartet was created by the statistician Francis Anscombe in 1973 to emphasize the significance of exploring data visually. By examining the quartet, we can better understand the importance of graphical analysis in understanding the underlying structure and relationships in datasets.

This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

# **QUE 3:** What is Pearson’s R?

Pearson's R, also known as the Pearson correlation coefficient or Pearson's correlation, is a measure of the linear relationship between two continuous variables. It quantifies the strength and direction of the linear association between the variables.The Pearson correlation coefficient is denoted by the symbol "r" and takes values between -1 and +1. Here's an explanation of the interpretation of different values of r:

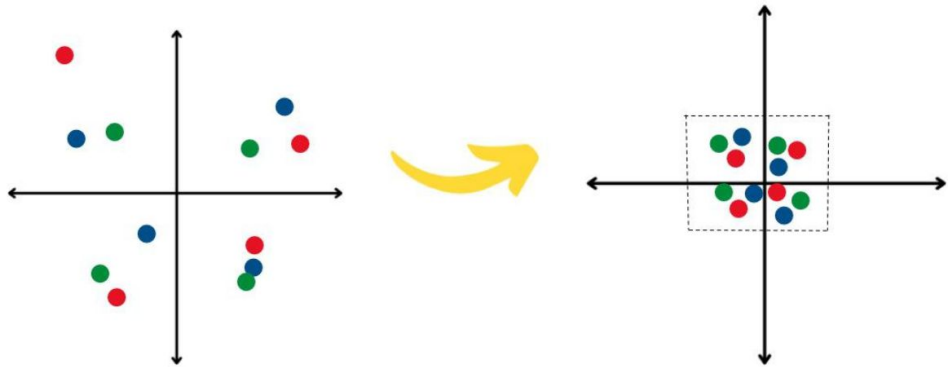
- **r = +1:** A perfect positive linear relationship. It indicates that as one variable increases, the other variable increases proportionally.
- **r > 0:** A positive linear relationship. It suggests that as one variable increases, the other variable tends to increase, but not necessarily at a constant rate.
- **r = 0:** No linear relationship. There is no systematic linear association between the variables.
- **r < 0:** A negative linear relationship. It indicates that as one variable increases, the other variable tends to decrease, but not necessarily at a constant rate.
- **r = -1:** A perfect negative linear relationship. It suggests that as one variable increases, the other variable decreases proportionally.

Pearson's correlation coefficient is calculated by dividing the covariance of the two variables by the product of their standard deviations. It is widely used inEDA,assessing the strength of relationships, feature selection, and assessing the performance of predictive models.

# **QUE 4:** What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

**Scaling,** refers to the process of transforming the values of variables to a specific range or distribution. It is performed to ensure that the variables are on a comparable scale, which can be beneficial for various reasons in data analysis and modeling.

It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.



# The difference between normalized scaling and standardized scaling:

**Normalized Scaling:**

- Normalization scales the values of a variable to a specific range, typically between 0 and 1.
- The formula for normalization is: X_normalized = (X - X_min) / (X_max - X_min)
- In this scaling method, the minimum value of the variable is subtracted from each value, and the result is divided by the range (maximum value minus minimum value).

**Standardized Scaling:**

- Standardization transforms the values of a variable to have zero mean and unit variance.
- The formula for standardization is: X_standardized = (X - X_mean) / X_std
- In this scaling method, the mean of the variable is subtracted from each value, and the result is divided by the standard deviation.

The main difference between normalized scaling and standardized scaling is in the range and distribution of the transformed values. Normalization scales the values to a specific range (e.g., 0 to 1), while standardization centers the values around zero with unit variance.

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

The occurrence of **infinite** values for the Variance Inflation Factor **(VIF) typically arises due to perfect multicollinearity in the dataset.** Perfect multicollinearity refers to a situation where one or more predictor variables can be perfectly predicted using a linear combination of other predictor variables.

The VIF measures the extent to which multicollinearity exists between predictor variables in a regression model. Specifically, it quantifies the inflation of the variance of the estimated regression coefficients due to multicollinearity.

The formula to calculate the VIF for a particular predictor variable is: $VIF = 1 / (1 - R^2)$

If perfect multicollinearity exists, it means that one or more of the independent variables can be expressed exactly as a linear combination of the other independent variables. In this case, the coefficient of determination ($R^2$) for the regression model is 1, which leads to a division by zero in the VIF formula: $VIF = 1 / (1 - 1) = 1 / 0$

This division by zero results in an infinite VIF value.

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

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess the distributional similarity between a sample of data and a theoretical distribution. It compares the quantiles of the sample data against the quantiles of the chosen theoretical distribution. The Q-Q plot is commonly used in statistics and data analysis to assess whether a dataset follows a particular distribution.

## **Uses of a Q-Q plot in linear regression:**

- **Assessing normality assumption:** The Q-Q plot helps determine whether the residuals in a linear regression model follow a normal distribution.
- **Detecting departures from normality:** Departures from normality, such as skewness or heavy tails, can be identified by examining deviations from the expected straight line pattern in the Q-Q plot.
- **Outlier detection:** Q-Q plots can be useful in identifying outliers by highlighting data points that significantly deviate from the expected straight line.

## Importance of a Q-Q plot in linear regression:

- **Model validity and reliability:** The Q-Q plot is a crucial tool for assessing the adequacy of the linear regression model. It helps ensure that the model's assumptions, such as normality of residuals, are met.
- **Diagnostic tool:** Q-Q plots serve as diagnostic tools for linear regression models. By visually examining the Q-Q plot, analysts can identify potential issues, such as non-normality, skewness, or outliers, that may affect the interpretation and inference from the model.
- **Decision-making:** Q-Q plots provide valuable information for decision-making in linear regression analysis.
- **Assumption checking:** Linear regression relies on various assumptions, including normality of residuals.

In summary, the Q-Q plot is a useful tool in linear regression analysis. It aids in assessing the normality assumption, detecting departures from normality, identifying outliers, and diagnosing the adequacy of the model. By examining the patterns and deviations in the Q-Q plot, analysts can make informed decisions about model refinement, assess model validity, and improve the reliability of regression results.

# Thank you