General Subjective Questions:

1. Explain the linear regression algorithm in detail?

Ans:

In simple terms, Linear Regression is a method of finding best straight-line fitting to the given data i.e., finding best linear relationship between dependent and independent variables. In technical terms, Linear Regression is a machine learning algorithm that finds the best linear-fit relationship on any given data, between independent and dependent variables. It is mostly done by the Sum of Squared Residuals Method.

Assumptions in Linear Regression:

- Linearity Assumption: it is assumed that there is a linear relationship between dependent and independent variables.
- Assumptions about residuals:
    1. Normality assumption: It is assumed that the error terms are normally distributed.
    2. Zero mean assumption: It is assumed that the residuals have a mean value of zero, i.e., the error terms are normally distributed around zero.
    3. Constant variance assumption: It is assumed that the residual terms have the same (but unknown) variance, σ2. This assumption is also known as the assumption of homogeneity or homoscedasticity.
    4. Independent error assumption: It is assumed that the residual terms are independent of each other, i.e., their pair-wise covariance is zero.

There are two types of Linear Regressions

1. Single Linear Regression: when there is one dependent and one independent variable.
2. Multiple Linear Regression: when there is one dependent and more than one independent variables.

Linear Regression algorithm follows below steps:

1. Cleaning and making data ready for Linear Regression
2. Splitting the data into train and test sets
3. Building a model by using train set (using statsmodels or sklearn modules)
4. Check whether the model is fit or not by seeing p-values, F-static and VIF values.
5. Checking whether our assumptions hold true for the model.
6. Using the model for predicting values for test set and checking whether the r2 values for train and test set matches or not, if those values matches then we have build a model which predicts same for test and train set.

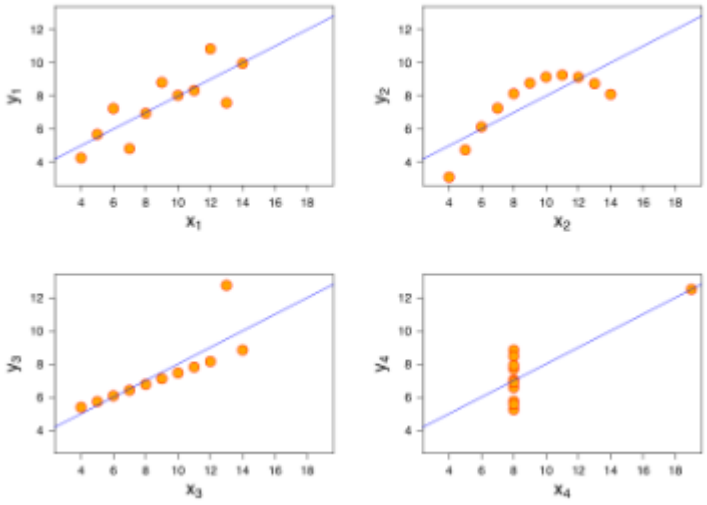
2.Explain the Anscombe's quartet in detail?

Ans:

When we develop a linear regression model, we should not simply relay on the statistics rather we need to visualize the data to get the relationship between them because there are some short comings in the linear regression like

1. It is sensitive to outliers.
2. It models the linear relationships only.
3. A few assumptions are required to make the inference.

These phenomena can be best explained by Anscombe's quartet.



As we can see, all the four linear regressions are exactly the same. But there are some peculiarities in the datasets which have fooled the regression line. While the first one seems to be doing a decent job, the second one clearly shows that linear regression can only model linear relationships and is incapable of handling any other kind of data. The third and fourth images showcase the linear regression model's sensitivity to outliers. Had the outlier not been present, we could have gotten a great line fitted through the data points. So, we should never ever run a regression without having a good look at our data.

3. What is Pearson's R?

Ans: Pearson's r, commonly referred to as Pearson's correlation coefficient, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It is a standardized measure, meaning it doesn't have units, and it ranges from -1 to 1.

- r=1: A perfect positive linear correlation. This means that as one variable increases, the other variable increases proportionally, following a perfect straight-line relationship.
- r= -1: A perfect negative linear correlation. This indicates that as one variable increases, the other variable decreases proportionally, following a perfect inverse straight-line relationship.
- -1<r<0: A negative linear correlation. As r approaches -1, the negative linear relationship becomes stronger.
- r=0: No linear correlation. This means that there is no linear relationship between the variables. However, it's important to note that there could still be other types of relationships present.
- 0<r<1: A positive linear correlation. As r approaches 1, the positive linear relationship becomes stronger.

Pearson's r coefficient is a widely used tool in statistics, data analysis, and scientific research to quantify and understand the degree of association between two variables. It's important to remember that Pearson's r specifically measures linear relationships and may not capture non-linear patterns or other complexities in the data. Additionally, correlation does not imply causation, so a high correlation between two variables does not necessarily mean that changes in one cause changes in the other.

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

Ans:

Scaling in the context of data preprocessing refers to the process of transforming your data so that it fits within a specific range. It involves adjusting the values of your variables to ensure they are on a comparable scale. Scaling is important for various machine learning algorithms that are sensitive to the magnitude of input features.

There are two common types of scaling: normalized scaling and standardized scaling.

Normalized Scaling:

Normalized scaling, also known as min-max scaling, In normalized the data is transformed to fit within a specific range, usually between 0 and 1. The formula for normalized scaling is as follows for each data point:

$$x_{\text{normalized}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

Where:

- x is the original value of the data point.
- X min is the minimum value in the dataset.
- X max is the maximum value in the dataset.

Normalized scaling is useful when you want to ensure that all your features have the same scale and are within a specific range. It is sensitive to outliers since extreme values can affect the scaling range.

Standardized Scaling:

In standardized scaling, also known as z-score normalization, the data is transformed such that it has a mean of 0 and a standard deviation of 1. The formula for standardized scaling is as follows for each data point:

$$x_{\text{standardized}} = \frac{x-\mu}{\sigma}$$

Where:

- x is the original value of the data point.
- μ is the mean of the dataset.
- σ is the standard deviation of the dataset.

Standardized scaling is particularly useful when you want to compare variables that have different units or when you're applying algorithms that assume normally distributed data. It centers the data around 0 and gives it a consistent spread. Unlike normalized scaling, standardized scaling is less affected by outliers since it is based on the mean and standard deviation.

Difference between Normalized Scaling and Standardized scaling:

1. **Range:** Normalized scaling transforms data to a specific range (usually 0 to 1), while standardized scaling centers data around 0 with a standard deviation of 1.
2. **Sensitivity to Outliers:** Normalized scaling can be sensitive to outliers as they can affect the scaling range. Standardized scaling is less sensitive to outliers since it uses the mean and standard deviation.
3. **Interpretation:** Normalized scaling preserves the original distribution's shape, while standardized scaling transforms the data distribution to have a mean of 0 and a standard deviation of 1.
4. **Use Cases:** Normalized scaling is commonly used when features should be in a specific range. Standardized scaling is preferred when comparing variables with different units or when applying algorithms that assume normally distributed data.

Choosing between normalized and standardized scaling depends on the characteristics of your data and the requirements of your analysis or machine learning algorithm.

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

Ans:

VIF, or Variance Inflation Factor, is a measure used in regression analysis to assess multicollinearity among predictor variables. Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated, which can lead to issues in interpreting the coefficients and their significance. VIF helps identify the extent to which multicollinearity exists.

A common formula to calculate the VIF for a predictor variable Xi is:

$$VIF(X_i) = \frac{1}{1-R^2_{X_i}}$$

Where RXi2 is the coefficient of determination (R-squared) obtained from regressing Xi on all other predictor variables in the model.

VIF values should always be greater than or equal to 1. A VIF of 1 implies that there is no multicollinearity, as the variance of the estimated coefficient for Xi is not inflated due to correlations with other predictors. VIF values greater than 1 indicate some level of multicollinearity, with larger values indicating stronger correlations.

However, it's possible to encounter situations where the VIF is extremely high or even infinite. This happens when one or more predictor variables are perfectly collinear with each other. Perfect collinearity means that one predictor variable is a linear combination of one or more other predictor variables. In this case, the correlation between the variables is so strong that it results in perfect multicollinearity.

When one variable is a perfect linear combination of others, it means that its VIF would theoretically be infinite according to the formula. This is because the denominator in the formula becomes zero, leading to a division by zero and an infinite VIF value.

Perfect multicollinearity is a problem in regression analysis because it prevents the regression algorithm from finding a unique solution for the coefficients. In other words, the model becomes numerically unstable and can't provide reliable coefficient estimates. In practice, this situation is detected by software that performs regression analysis, and infinite VIF values may be treated as indicators of perfect multicollinearity.

To address the issue of multicollinearity, it's important to identify and handle highly correlated predictor variables before constructing a regression model. This can involve removing one of the correlated variables, transforming variables, or using dimensionality reduction techniques.

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

Ans:

A Quantile-Quantile (Q-Q) plot is a graphical tool used in statistics to assess whether a dataset follows a particular theoretical distribution, such as the normal distribution. It compares the quantiles of the observed data against the quantiles of the expected theoretical distribution. This comparison helps to visually determine if the data and the theoretical distribution have a similar pattern of spread and central tendency.

In the context of linear regression, a Q-Q plot is often used to check the assumption of normality of residuals. Residuals are the differences between the observed values and the predicted values of the dependent variable in a linear regression model. The normality assumption of residuals is important because many statistical tests and methods, including linear regression, assume that the residuals are normally distributed.

Here's how the Q-Q plot is used and its importance in linear regression:

1. **Assumption Checking**: Linear regression assumes that the residuals are normally distributed with a mean of zero. This assumption is crucial for the validity of statistical inference and hypothesis testing associated with the regression model. If the assumption is violated, it can lead to incorrect parameter estimates, confidence intervals, and hypothesis tests.
2. **Creating a Q-Q Plot**: To create a Q-Q plot for linear regression, you plot the quantiles of the observed residuals against the quantiles that would be expected from a normal distribution. If the residuals follow a normal distribution, the points in the Q-Q plot should lie approximately along a straight line.
3. **Interpreting the Q-Q Plot**: If the points in the Q-Q plot deviate significantly from a straight line, it suggests that the residuals do not follow a normal distribution. Deviations could include heavy tails, skewness, or other departures from normality.
4. **Impact on Inference**: If the Q-Q plot indicates a departure from normality, it could affect the reliability of statistical inference based on the regression model. In such cases, it might be necessary to consider alternative regression methods or transformations of the variables to satisfy the normality assumption.
5. **Model Improvement:** Identifying and addressing deviations from normality is important for building a more accurate and robust linear regression model. By identifying the specific issues causing the departure from normality, you can make necessary adjustments to improve the model's performance.

In summary, a Q-Q plot is a valuable diagnostic tool in linear regression that helps assess the assumption of normality for residuals. It provides a visual assessment of how well the residuals fit a normal distribution, which in turn affects the validity of statistical inference and the reliability of the regression model's results. If deviations from normality are detected, it prompts further investigation and potentially adjustments to the model or the data.

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 the data dictionary we can see that there are six categorical variables in the dataset they are

Season, year, month, weekday, working day, holiday. By plotting boxplot, we can infer their effect on dependent variable

- Season: we can infer that fall season has more bike bookings, followed by summer, winter and very less bookings in winter
- Year: we can infer that there are two years 2018 and 2019 bike bookings are increasing in 2019 as compared to 2018
- Month: we can infer that there are more number of bookings in August, September and October months
- Weekday: weekday doesn't effect any bookings as there are same on the every day of the week
- Working day: we can infer that there are more number of booking on holiday than on working day
- Holiday: we can infer that there are more bookings in holiday

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

Ans:

When creating dummy variables for categorical data in a machine learning or statistical analysis context, setting drop_first=True is a common practice that has several important reasons and benefits:

- **Avoiding Multicollinearity**: Multicollinearity occurs when two or more independent variables in a regression model are highly correlated with each other. In the context of dummy variables, if you include all levels of a categorical variable (i.e., create dummies for all categories), one category's information can be perfectly predicted from the others. This leads to multicollinearity, which can cause problems in some statistical models, such as linear regression, because it becomes difficult to separate the effects of individual dummy variables from the overall constant (intercept) term. By dropping one level, you avoid this issue.
- **Simplifying Interpretability:** When you set drop_first=True, the dropped category becomes the reference category (baseline), and the coefficients for the remaining dummy variables represent the change from the baseline. This simplifies the interpretation of the model coefficients. Without dropping one category, the coefficients for all levels would represent the change from the mean of all levels combined, making it harder to interpret the effects of individual categories.
- **Reducing Dimensionality**: Including dummy variables for all levels of a categorical variable can increase the dimensionality of your dataset, especially when you have categorical variables with many levels. High-dimensional data can lead to increased computational complexity, slower

training times, and the curse of dimensionality, which can negatively affect model performance. Dropping one category helps reduce dimensionality.

- **Enhancing Model Generalization**: In some cases, including dummy variables for all categories might lead to overfitting. Overfit models perform well on training data but poorly on unseen data. By reducing the number of dummy variables through drop_first=True, you can help your model generalize better to new data.
- **Conforming to Model Assumptions**: Some statistical models, like linear regression, assume that the residuals (the differences between predicted and actual values) are normally distributed with constant variance (homoscedasticity). Including all dummy variables can violate these assumptions, as multicollinearity can lead to unstable coefficient estimates and non-constant variance. Dropping one category can help conform to these assumptions.

In summary, setting drop_first=True during dummy variable creation is a practical approach to handle categorical data efficiently, reduce multicollinearity, simplify model interpretation, and improve the overall performance and stability of statistical and machine learning models.

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

Ans:

By seeing at pair-plot we can infer that temp and atemp variables have highest correlation with the 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 is a crucial step to ensure that your model's predictions are reliable and that the underlying assumptions of linear regression are met. Here are some common methods and techniques for validating these assumptions:

- **Residual Analysis:**
    1. **Residual Plot**: Plot the residuals (the differences between the observed and predicted values) against the predicted values. The plot should show a random scatter of points with no clear patterns. If you see a pattern, such as a funnel shape or a curve, it may indicate heteroscedasticity (non-constant variance) or non-linearity in the data.
    2. **Normality of Residuals**: Create a histogram or Q-Q plot of the residuals to assess whether they follow a roughly normal distribution. You can also use statistical tests like the Shapiro-Wilk test or Anderson-Darling test to check for normality. Deviations from normality may suggest a need for transformation or a different modeling approach.
- **Homoscedasticity:**

1. **Residuals vs. Fitted Values Plot**: Check the residuals vs. fitted values plot for evidence of heteroscedasticity. Ideally, the spread of residuals should be relatively consistent across the range of predicted values. Heteroscedasticity may require model transformation or the use of robust regression techniques.

- **Linearity**:
    1. **Partial Residual Plots**: Create partial residual plots for each predictor variable to check for linearity. These plots show the relationship between a single predictor and the residuals while holding other variables constant. If the relationship is not approximately linear, you may need to consider transformations or polynomial terms.
- **Outlier Detection**:

    Identify and investigate potential outliers in the residuals. Outliers can strongly influence the regression model and may indicate issues with the data or the model. Tools like Cook's distance or studentized residuals can help in outlier detection.
- **Multicollinearity**:

    Check for multicollinearity among predictor variables using correlation matrices or variance inflation factor (VIF) values. High multicollinearity can make it challenging to interpret individual coefficients and may require variable selection or transformation.

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

Ans:

Based on the final model we can say top 3 features contributing significantly towards explaining the demand of the shared bikes year, temp and windspeed

- As year increases the shared bike demand increases at the rate of 0.235 i.e., for a unit increase in the year the bike demand increases by 0.235.
- As temp increases the shared bike demand increase at the rate of 0.449 i.e., for a unit increase in the temp the bike demand increase by 0.449.
- As light snow in weather increases the shared bike demand decreases at the rate of 0.286 i.e., for a unit increase in windspeed the bike demand decreases by 0.286.