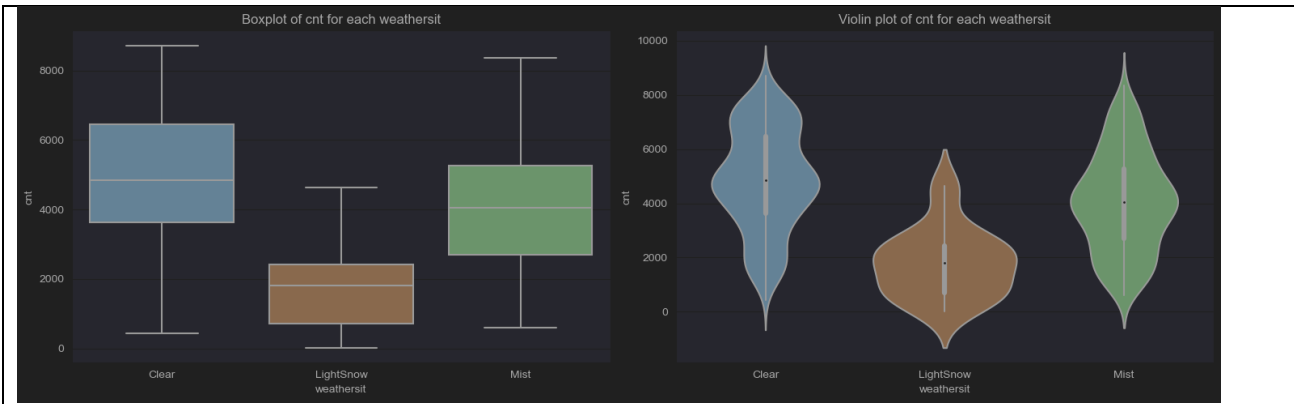
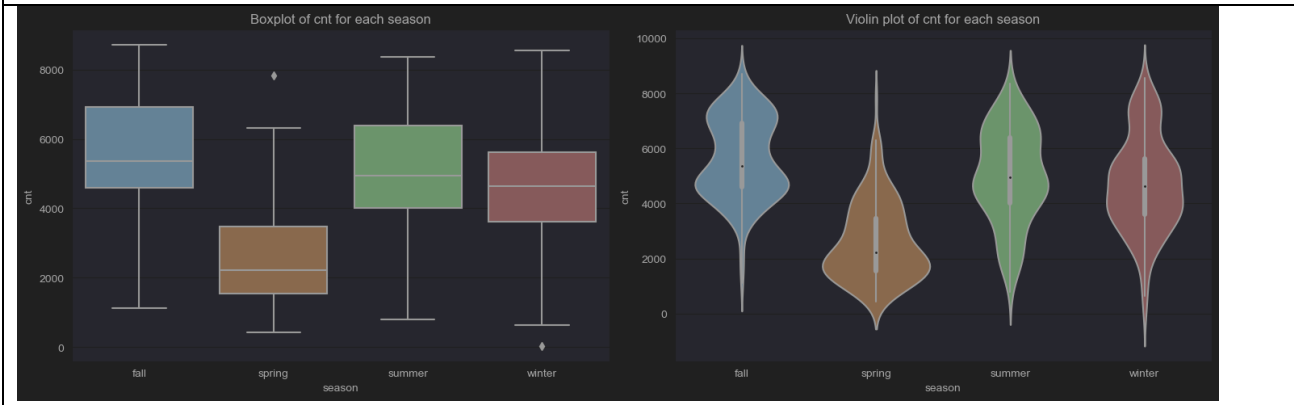
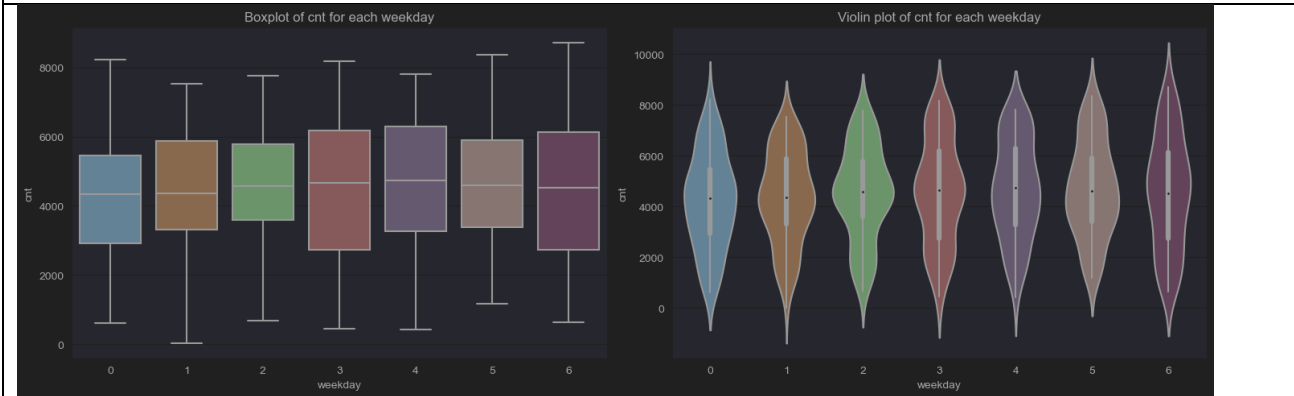
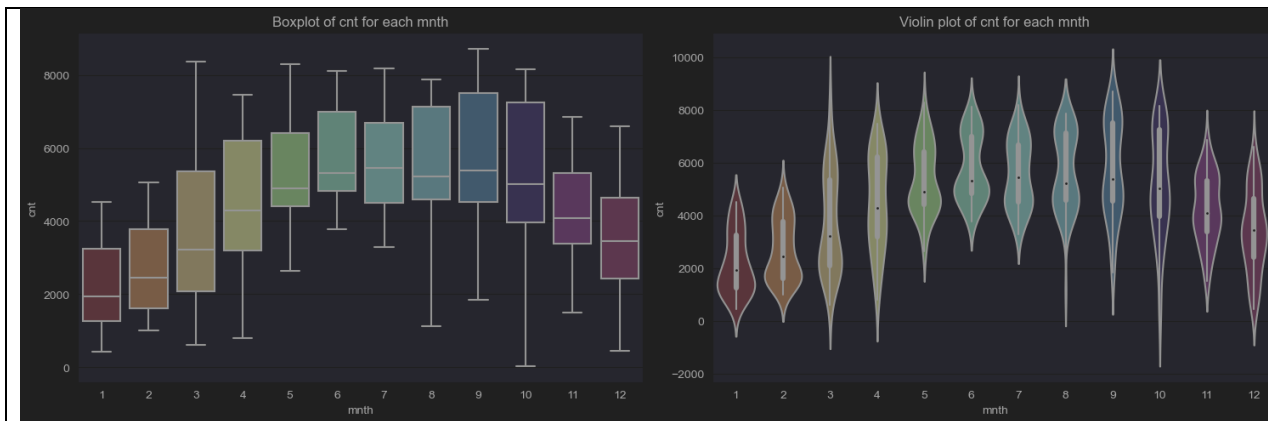


Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

<div><div>Boxplot of cnt for each yr</div><p>A boxplot comparing the count of rides (cnt) for the years 2018 and 2019. The y-axis is labeled 'cnt' and ranges from 0 to 8000. For 2018, the median is approximately 3800, the first quartile is around 2200, and the third quartile is around 4600. For 2019, the median is approximately 6000, the first quartile is around 4400, and the third quartile is around 7100. The whiskers extend from approximately 500 to 6000 for 2018 and 500 to 9000 for 2019.</p></div> <div><div>Violin plot of cnt for each yr</div><p>A violin plot showing the distribution of the count of rides (cnt) for the years 2018 and 2019. The y-axis is labeled 'cnt' and ranges from 0 to 10000. The 2018 violin is wider at the lower end (around 2000-4000), while the 2019 violin is wider at the higher end (around 6000-8000), indicating a shift towards higher counts in 2019.</p></div>	<p>1. Count of rides increases with each year</p>
<div><div>Boxplot of cnt for each holiday</div><p>A boxplot comparing the count of rides (cnt) for non-holiday (0) and holiday (1) periods. The y-axis is labeled 'cnt' and ranges from 0 to 8000. For non-holiday, the median is approximately 4500, the first quartile is around 3200, and the third quartile is around 5900. For holiday, the median is approximately 3300, the first quartile is around 2000, and the third quartile is around 6000. The whiskers extend from 0 to 9000 for non-holiday and 1000 to 7500 for holiday.</p></div> <div><div>Violin plot of cnt for each holiday</div><p>A violin plot showing the distribution of the count of rides (cnt) for non-holiday (0) and holiday (1) periods. The y-axis is labeled 'cnt' and ranges from 0 to 10000. The non-holiday violin is wider at the lower end (around 2000-4000), while the holiday violin is wider at the higher end (around 4000-6000), indicating a shift towards higher counts during holidays.</p></div>	<p>1. Median and first quartile of count of rides is lower during holidays even though the third quartile is same. 2. Hence the spread of the rides at the lower range is more during holidays</p>
<div><div>Boxplot of cnt for each workingday</div><p>A boxplot comparing the count of rides (cnt) for non-workingday (0) and workingday (1) periods. The y-axis is labeled 'cnt' and ranges from 0 to 8000. For non-workingday, the median is approximately 4500, the first quartile is around 2800, and the third quartile is around 5900. For workingday, the median is approximately 4600, the first quartile is around 3300, and the third quartile is around 6000. The whiskers extend from 500 to 9000 for non-workingday and 0 to 8500 for workingday.</p></div> <div><div>Violin plot of cnt for each workingday</div><p>A violin plot showing the distribution of the count of rides (cnt) for non-workingday (0) and workingday (1) periods. The y-axis is labeled 'cnt' and ranges from 0 to 10000. The non-workingday violin is wider at the lower end (around 2000-4000), while the workingday violin is wider at the higher end (around 4000-6000), indicating a shift towards higher counts on workingdays.</p></div>	<p>1. During a workingday the spread at the lower quartile is less and more count are towards the higher range.</p>

 <p>Boxplot of cnt for each weathersit</p> <p>Violin plot of cnt for each weathersit</p>	<ol style="list-style-type: none"> Highest count is during when the weather is Clear followed by Mist and LightSnow
 <p>Boxplot of cnt for each season</p> <p>Violin plot of cnt for each season</p>	<ol style="list-style-type: none"> Highest count is during the fall followed by summer, winter and spring
 <p>Boxplot of cnt for each weekday</p> <p>Violin plot of cnt for each weekday</p>	<ol style="list-style-type: none"> Median for weekday 2,3,4,5 are almost the top 4 ones. Spread of count for weekday 2 is the lowest. Count is at the same range on this day and so usage of rides is highly probable for this day. Spread is the highest on weekday 6 and so the variations are depended on other factors a lot for this weekday.

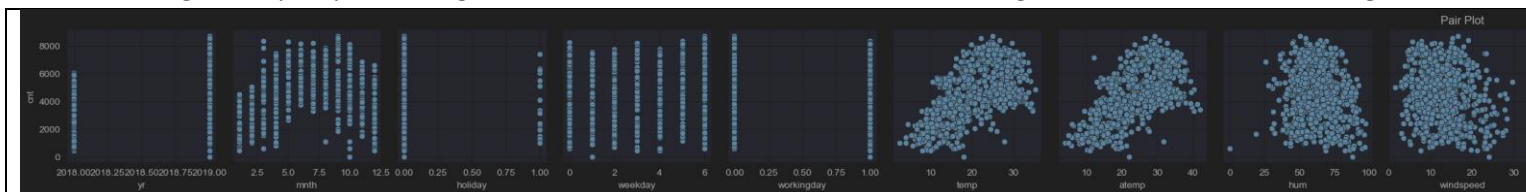


1. Highest usage is for months no 6,7,8,9,10 for top 5
2. Lowest is for the month 12 followed by month 1.
3. Highest spread is for the month 10 followed by month 3 and hence predictability based on month is less for these nos.

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

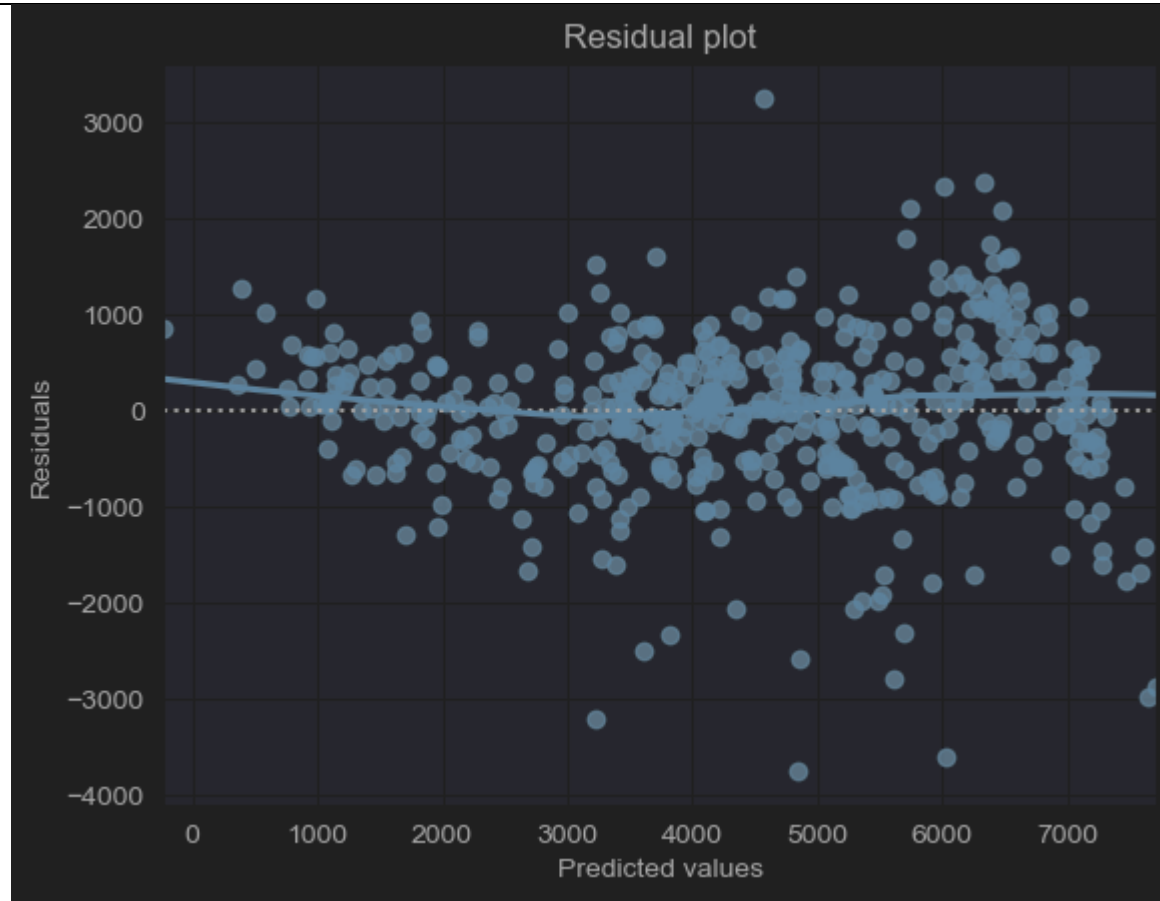
1. The parameter `drop_first=True` in the `get_dummies` function in pandas is used to avoid a situation known as multicollinearity, which can occur when one variable can be linearly predicted from others.
2. When we use `get_dummies` to turn categorical variables into a series of 0s and 1s, using `drop_first=True` effectively removes one column of information to ensure that there are not interdependencies between the columns.
3. In some types of models, particularly linear regression and logistic regression, multicollinearity can lead to coefficients that are difficult to interpret and a model that overfits the data, as the model can attribute changes to the target variable that result from one feature to another correlated feature.

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

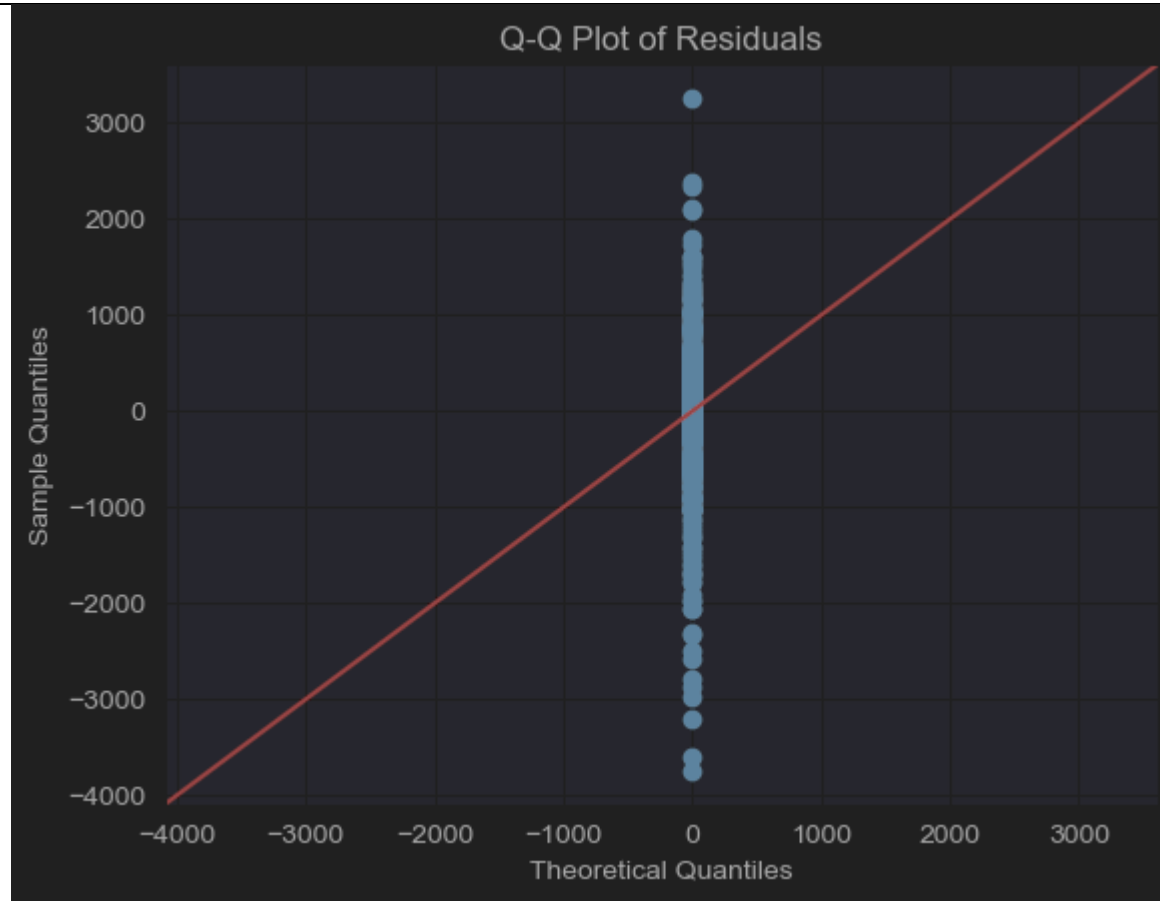


1. As temp reaches 30 the rides are higher and then it dips down after this
2. As humidity is more than 40 rides increase and then decrease after reaching 75
3. As windspeed increases from 15 the rides count start to get decreasing

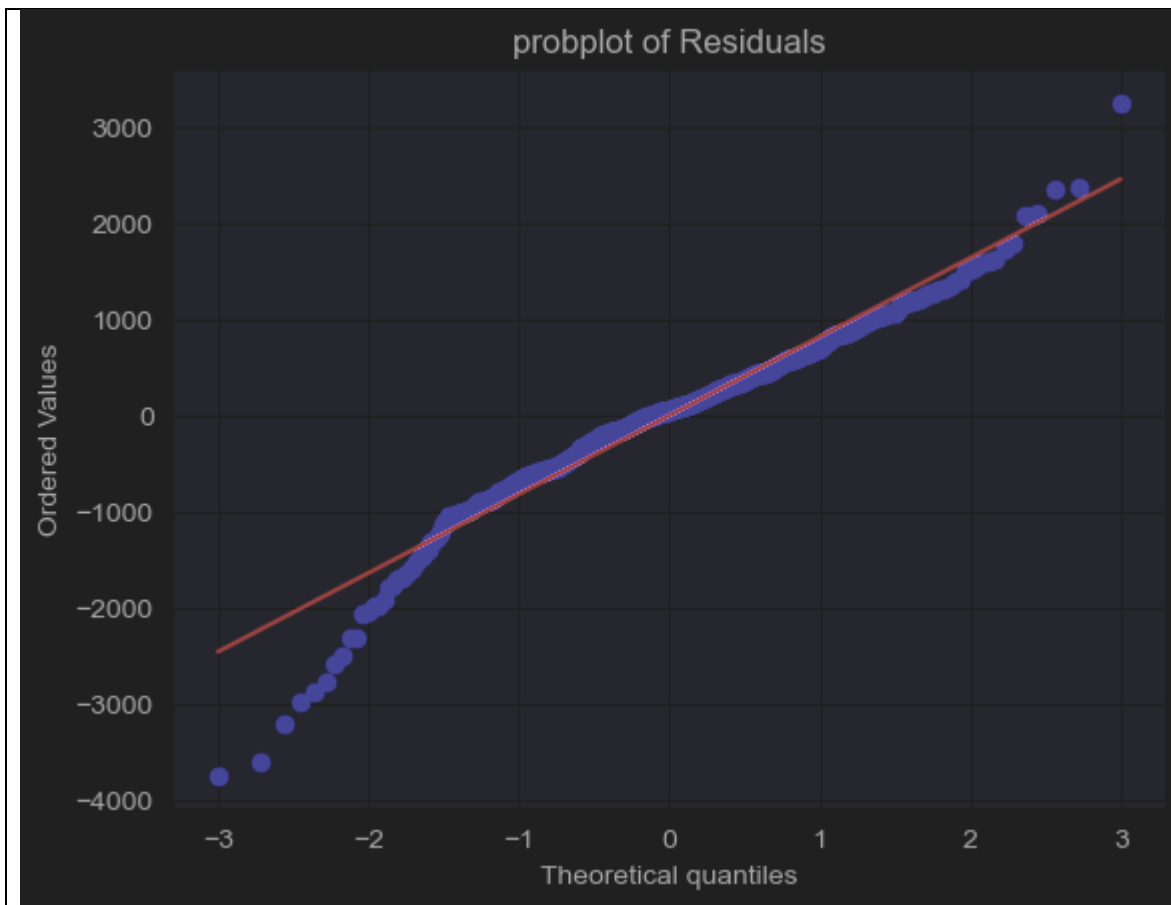
5. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

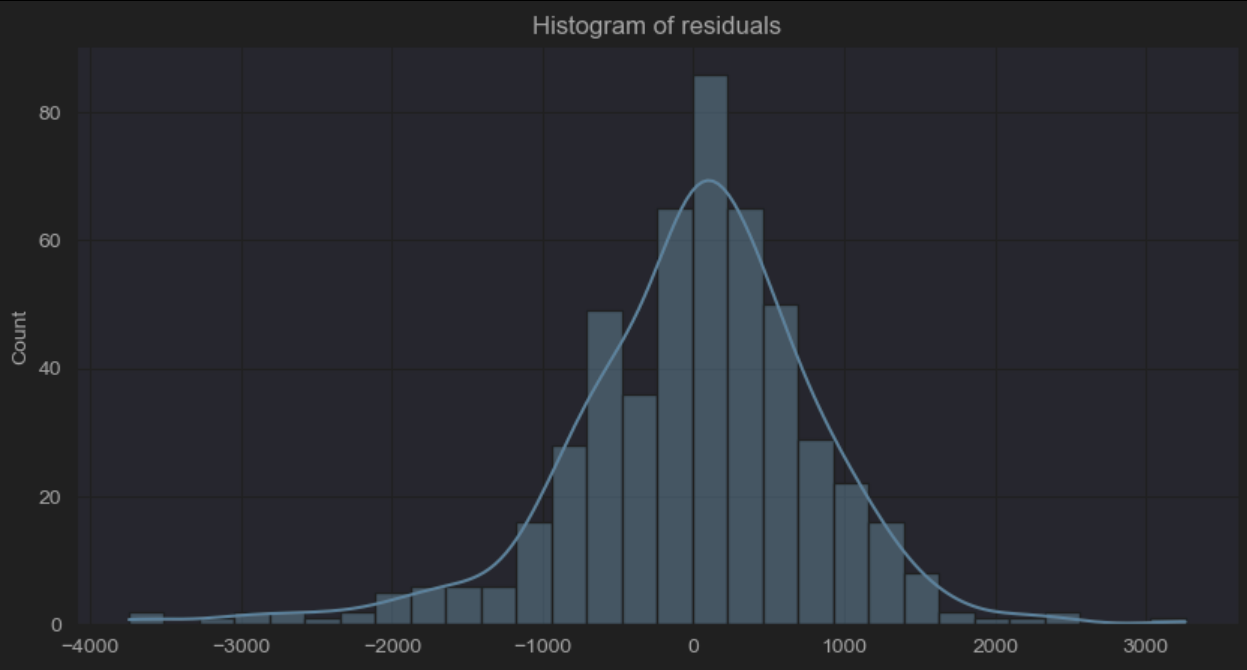


1. The scatter plot of residuals does not display any discernible pattern or correlation along the x-axis, suggesting that the observations are randomly distributed. Therefore, the assumption of linearity needed for linear regression appears to hold true.
2. The distribution of residuals across the predictor values is even, forming a rectangular shape rather than a funnel. This indicates that the residuals have a constant variance, fulfilling the condition of homoscedasticity instead of Heteroscedasticity. Thus, the assumption of homoscedasticity, crucial for linear regression, is also satisfied.



1. QQ Plot is not as expected as the dotted line is almost deviating from the line plotted 45 degrees. so the data is either heavily skewed or have outliers or non-linear transformation to be considered.
2. Probplot is as expected with most of the dotted line between head and tail are on the line and follows the assumption of linear regression with residuals being normally distributed



	<ol style="list-style-type: none"> 1. Residuals for histplot mentions that residuals follows normally distributed in a bell curve
<p>➔ Shapiro-Wilk test</p> <p>Statistics=0.958, p=0.000</p> <p>Sample does not look Gaussian (reject H0)</p>	<ol style="list-style-type: none"> 1. p value is not greater than 0.05 and so it indicates sample data is normally distributed and so consideration of non-linear transformation or further outlier correction to be done 2. However the histogram and qqplot indicates its normal & there are no outliers that we can see in initial violin plots for whole dataset too. So we will ignore this statistical warning for now
<p>➔ Durbin_Watson Test value: 1.9728283777141598</p>	<ol style="list-style-type: none"> 1. Durbin_Watson value close to 2 suggests that there is no autocorrelation

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

<p>Top features are as below:-</p> <ol style="list-style-type: none"> 1. Temp 2. weathersit_LightSnow -> Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds 3. yr
--

General Subjective Questions

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

Linear Regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method used to fit a linear model to the observed data. The goal of Linear Regression is to find the best fitting straight line through the data points.

The best fitting line is called the regression line.

The generic equation of a straight line is:

$$y = mX + c$$

Where,

y is the dependent variable (output/target),

X is the independent variable (input/feature),

m is the slope of the line and

c is the y-intercept.

M and C are two constants representing the slope and intercept of the line respectively. Linear regression finds the optimal values for m (also called weights) and c (also called bias).

Linear Regression Process:

Model representation: The model is represented with the linear equation ($y = mX + c$), where 'm' and 'c' are the parameters of the model.

Goodness of fit (Loss function): The goodness of fit of the model is computed using a loss function which measures the discrepancy between the predicted and actual outcomes. For linear regression, Mean Squared Error (MSE) is commonly used as the loss function.

Best fit line (Learning Algorithm): The learning algorithm tries to find the best values for model parameters ('m' and 'c') that minimize the loss function. This is typically achieved through a process called Gradient Descent.

Model Evaluation: Once the model is trained, its performance is evaluated on unseen data using various metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R-Squared, Adjusted R-Squared etc.

Assumptions of Linear Regression:

Linear Relationship: Linear regression assumes that the relationship between the input and output is linear.

Independence: Observations are assumed to be independent of each other.

Homoscedasticity: The error term (residuals/error) is the same across all levels of the independent variables.

Normality: For any fixed value of X, Y is normally distributed.

No Multicollinearity: The independent variables should not be too highly correlated with each other.

Linear regression is a good choice for simple and quick modelling tasks where interpretability is important. It's used for both simple and multiple regression and forms the basis for many forms of machine learning, including logistic regression, poisson regression and so on. However, it may not be a good choice for datasets with non-linear relationships or if there's no linear correlation between the independent and dependent variables.

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

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. The quartet was created by the statistician Francis Anscombe to demonstrate two points:

1. Graphs are essential in statistical analysis, not just because they conveniently summarize the data but they can also help us spot unusual observations, patterns, or deviances that may not be apparent from just looking at the statistics.
2. The importance of checking the underlying assumptions (like linearity, homoscedasticity, etc.) of the statistical methods, especially when working with small datasets. In the case of Anscombe's quartet, all four datasets appear to be similar when using basic summary statistics, but vary considerably when graphed.

All four sets are identical when examined using simple summary statistics (mean, variance, correlation, and linear regression line), but when graphed, it becomes clear that the first set is a simple linear relationship, the second set is not linear, the third set is linear but with an outlier skewing the regression line, and the fourth set has x constant for all but one point (which is an outlier).

Let's look at the common statistics:

- Mean of x in each case is 9
- Variance of x in each case is 11
- Mean of y in each case is 7.50 to 2 decimal places
- Variance of y in each case is 4.12 to 2 decimal places
- Correlation between x and y in each case is 0.816
- Linear regression (least squares) line in each case is $y = 3.00 + 0.500x$
- Coefficient of determination of the linear regression $R^2 = 0.67$

Anscombe's quartet is a powerful demonstration of why visualization is an important complement to summary statistics, as it can reveal aspects of the data that may be hidden by the statistics.

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

Pearson's R, also known as Pearson's correlation coefficient, is a statistic that measures linear correlation between two variables X and Y. It ranges from -1 to 1, inclusive, and reflects the degree of linear dependence between the variables.

The formula for Pearson's R is:

$$r = \frac{\sum[(x_i - \text{mean}(x))(y_i - \text{mean}(y))]}{(n-1) * \text{std_dev}(x) * \text{std_dev}(y)}$$

Where:

- x_i and y_i are the individual sample points indexed with i
- $\text{mean}(x)$ and $\text{mean}(y)$ are the mean values of X and Y
- $\text{std_dev}(x)$ and $\text{std_dev}(y)$ are the standard deviation of X and Y
- n is the total number of data points
- Σ is the sum of the products from $i=1$ to n

The key characteristics of Pearson's R are:

1. It is symmetric, meaning that the correlation from X to Y is equal to the correlation from Y to X.
2. It is scale invariant, meaning it does not change if the units of measurement of the quantities or their order are changed.
3. A positive correlation coefficient means that as the value of one variable increases, the value of the other variable also increases; they move in tandem.
4. A negative correlation coefficient indicates that as one variable increases, the other decreases.
5. A correlation of zero means that no relationship exists between the variables.
6. If the correlation coefficient is 1, the variables are perfectly positively correlated; if it's -1, they are perfectly negatively correlated.

Correlation does not imply causation. A correlation between two variables does not necessarily mean that one causes the other.

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

Scaling in the context of data preprocessing is a step in which numerical features in the dataset are transformed to have certain helpful properties, often to have values in a defined range. The necessity of this process can be due to the algorithms used in machine learning, which will often not perform optimally when given raw, unprocessed data.

Why is Scaling Performed?

Scaling is performed to standardize the range of features of input data set. Some machine learning algorithms can perform better when numerical input variables are scaled to a standard range. This includes algorithms that use a weighted sum of inputs like linear regression, and algorithms that use distance measures like k-nearest neighbours.

What is Normalized Scaling and Standardized Scaling?

Normalized Scaling (Min-Max Scaling): This method rescales the features to a fixed range between 0 to 1. It's useful when the algorithms expect the input features in the same positive scale. However, it doesn't handle outliers well. The formula used for normalization is:

$$X_{\text{norm}} = (X - X_{\text{min}}) / (X_{\text{max}} - X_{\text{min}})$$

- Where X_{max} and X_{min} are the maximum and minimum values of feature X respectively.

Standardized Scaling (Z-score normalization): In this approach, the features are rescaled so that they have the properties of standard normal distribution with mean=0 and standard deviation (std)=1. It is useful in cases where the results required should be not impacted by the magnitude of the features. The formula used for standardization is:

$$Z = (X - \mu) / \sigma$$

- Where μ is the mean and σ is the standard deviation of feature X.

By choosing normalization or standardization, we're basically transforming the original feature so that it fits within a certain scale, like 0–100 or 0–1, but they do it in slightly different ways.

Your choice of normalization or standardization will depend on the specific requirements of Machine Learning algorithm and the nature of data. For example, if we are dealing with a dataset with significant outliers, we might want to consider standardization.

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

The Variance Inflation Factor (VIF) is a measure of multicollinearity in a multiple regression. It quantifies how much the variance is increased because of multicollinearity, i.e., correlation among predictors (independent variables).

VIF is defined as the ratio of the variance of the estimated regression coefficient of a particular predictor to the variance if the predictor was not correlated with any other predictors. Mathematically, it is calculated as:

$$\text{VIF} = 1 / (1 - R^2)$$

- Where R^2 is the coefficient of determination from the regression of one predictor on the rest of the predictors.

Sometimes, you may observe an infinite value of VIF. This can happen in a couple of scenarios:

1. Perfect Multicollinearity: This occurs when one predictor can be perfectly predicted by a linear equation of other predictors. In such a situation, the denominator ($1 - R^2$) becomes 0 making the VIF infinity. In plain terms, one of your variables is a perfect mix of other variables, so its unique variance is 0 which causes an infinite VIF.
2. Near-Perfect Multicollinearity: Similar to perfect multicollinearity but in this case, a predictor can be approximately (but not perfectly) represented by a linear equation of other predictors. Even in this situation, R^2 can be so close to 1 that it causes $(1 - R^2)$ to round to 0 due to limited floating point precision, making the VIF extremely large or infinite.

It's important to note that perfect or near-perfect multicollinearity is a sign of serious issues with data. If the VIF is infinite, we should revise our model, perhaps dropping one of the redundant variables or more, or combining them into one. This can help alleviate issues of multicollinearity and improve the stability and interpretability of model.

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

A Q-Q (Quantile-Quantile) plot is a type of graphical tool that helps us assess if a dataset follows a theoretical distribution. It does this by plotting the quantiles of the observed data against the quantiles of the chosen theoretical distribution.

If the dataset follows the chosen distribution, the points in the Q-Q plot will approximately lie on the line $y = x$ (the 45-degree line). If the dataset does not follow the distribution, the points will deviate from this line.

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

In the context of linear regression, Q-Q plots are typically used to check the assumption of normality of the errors (residuals). Here's why this is important:

1. Checking the Normality Assumption

When using a Q-Q plot to check the assumption of normality for the residuals from a linear regression, the following steps are typically taken:

Fit your linear regression model and calculate the residuals (the differences between the observed and predicted values).

Generate a Q-Q plot using these residuals. The x-axis should represent the quantiles from a standard normal distribution, and the y-axis should represent the quantiles of your residuals.

If the residuals are normally distributed, the points in the Q-Q plot should approximately form a straight line along the line $y = x$. They don't have to fall exactly on the line, as there might be some deviation due to randomness. A curve away from the line at the ends often indicates skewness in the data.

If the points largely deviate from this line (especially in a systematic way), it suggests that the residuals may not be normally distributed. This could thus violate the normality assumption of linear regression.

2. Identifying Outliers

Outliers in the data may cause significant deviations from the $y = x$ line in a Q-Q plot:

If your data follows the chosen distribution except for a few points, those few points will appear far away from the $y = x$ line in the Q-Q plot. These points are probably outliers.

The outliers might not follow the normal distribution even if the majority of your data does. By looking for points that deviate significantly from the $y = x$ line, you can identify potential outliers in your data.

Once you identify potential outliers, it's important to further investigate these data points to determine if they are "real" (e.g. a valid extreme value), or result from errors or anomalies in the data collection process. Depending how these outliers are treated could have implications for your final regression model.

In both cases, the Q-Q plot provides a valuable diagnostic tool, but conclusions often rely on the analyst's subjective judgment. Quantitative tests for normality and outlier detection can complement the visual inspection of a Q-Q plot.

3. Homoscedasticity Check:

Q-Q plot can get an insight about homoscedasticity, though it's not the primary tool for this purpose.

Homoscedasticity is one of the core assumptions of linear regression models, which means that the variance of the errors is constant across all levels of the independent variables.

Checking Homoscedasticity using Q-Q Plot

In the Q-Q plot, if the residuals follow a pattern where they remain close to the line at the mean of the distribution, but they spread out towards the extremes (both ends of the distribution), this could suggest a violation of the homoscedasticity assumption. In other words, a "funnel" shape or a "bow-tie" shape on a Q-Q plot.

The points on a Q-Q plot should be evenly distributed around the line without any obvious patterns. If residuals are homoscedastic, they should randomly and evenly depart from the line. If residuals are heteroscedastic, they will tend to depart from the line more in one part of the plot than the other.

Note: A Q-Q plot would never be used as a conclusive test for homoscedasticity. It can only provide minor evidence and is not considered a robust way to detect heteroscedasticity. There are other more appropriate techniques to formally test the assumption of homoscedasticity, such as plots of residuals vs predicted values, Breusch-Pagan test or Cook-Weisberg test, among others.

The Q-Q plot is primarily a tool for checking the normality of residuals, with assessment of homoscedasticity as a secondary purpose and more indicative in nature, rather than definitive.