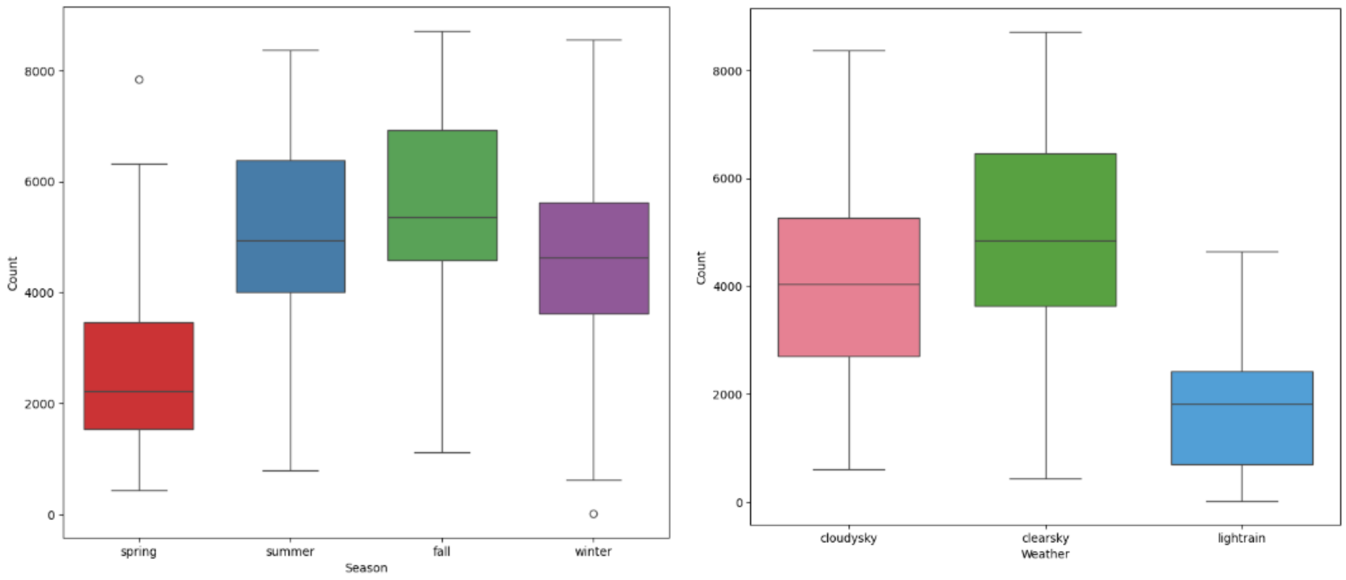
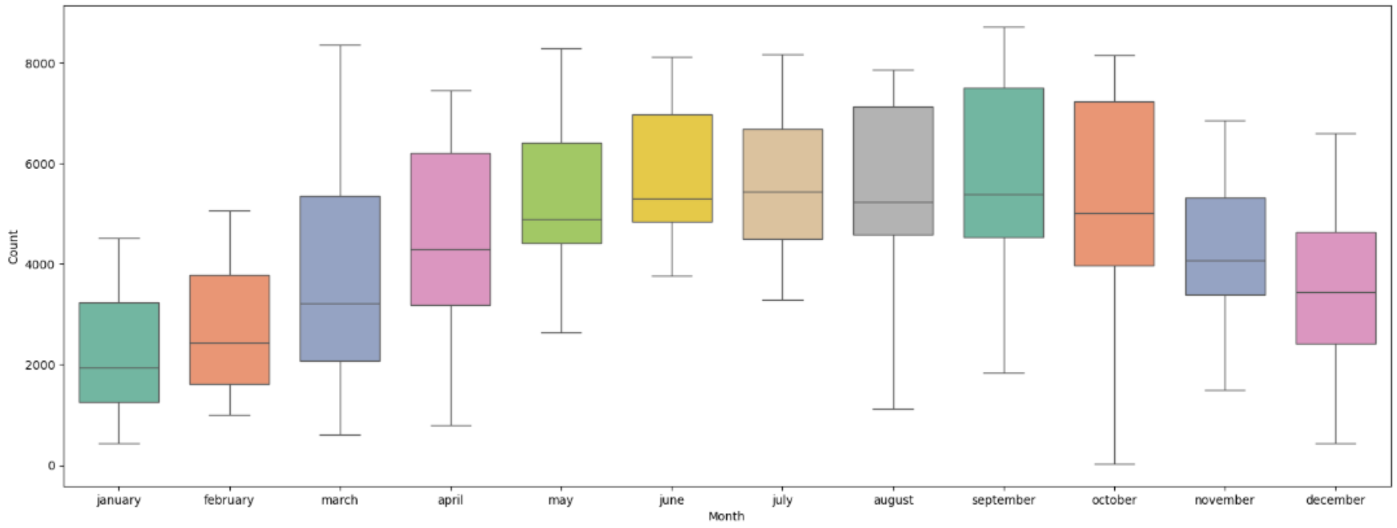
Assignment-Based Subjective Questions

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

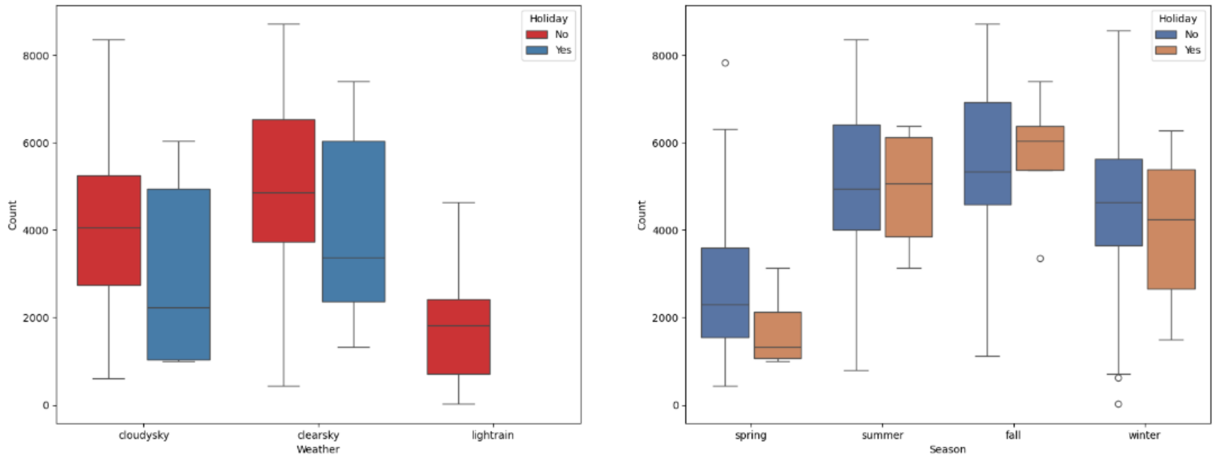
Count vs Season/Weather plot

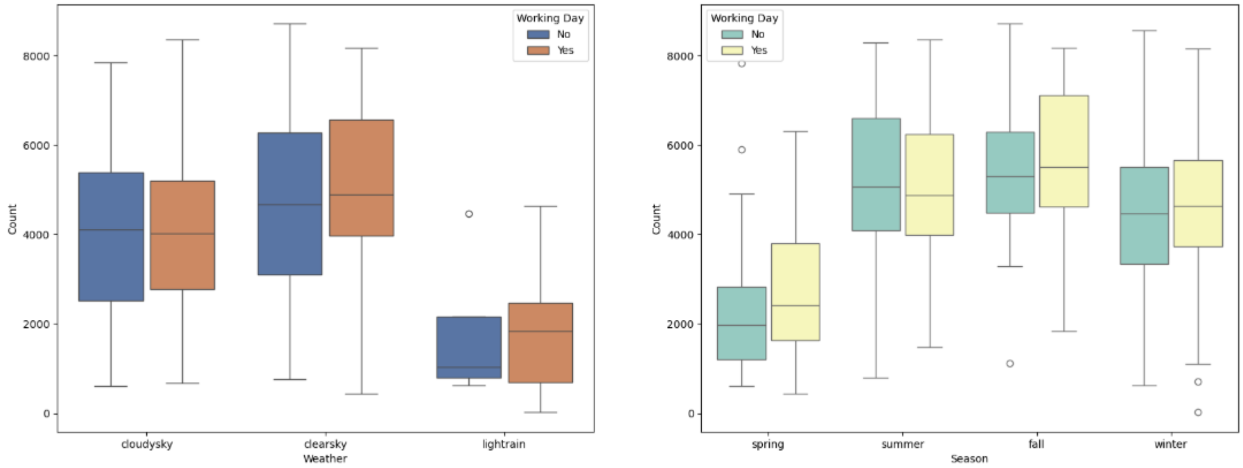


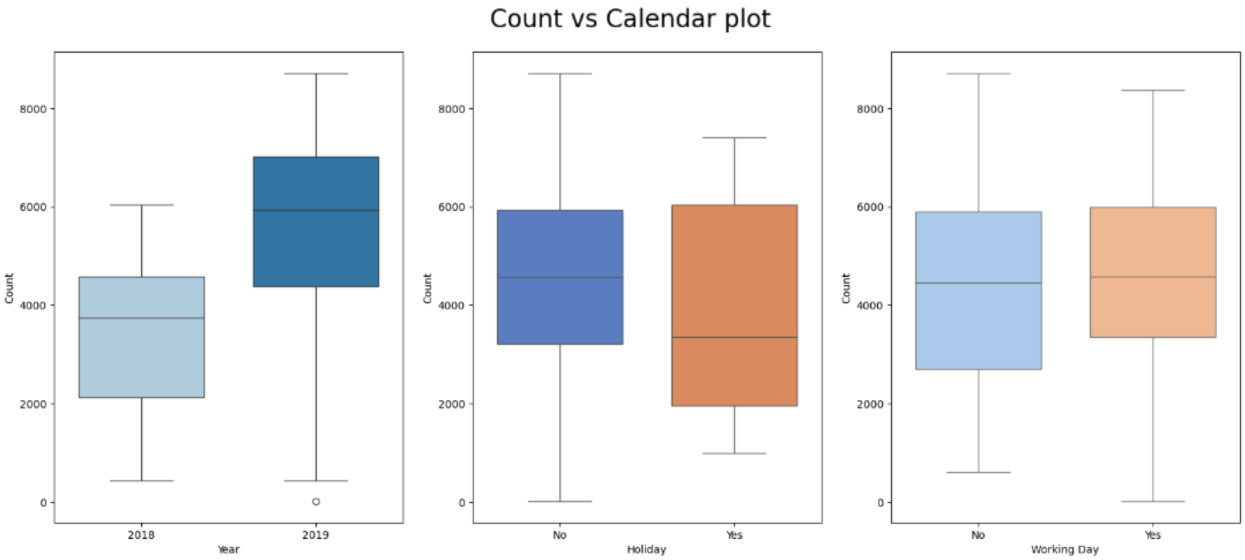
Count vs Month/Weekday plot



Count vs Season/Weather plot against Holiday/Workingday







Inferences:

- Median count decreases considerably in spring and on days with light rain
- Median count increases as the year progresses and starts decreasing again towardsthe end of the year
- Median count is considerably higher on non-holidays irrespective of weather but it varies based on the season.
- Median count is higher on working days when weather is having clear sky.
- Median count has increased significantly from 2018 to 2019
- Median count decreases significantly over holidays but 75thpercentile remains almost the same
- Median count remains almost the same across all weekdays

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

1. Without dropping the first level, there's a risk of perfect multicollinearity. This occurs when one variable can be perfectly predicted from others, leading to unstable and unreliable model estimates.
2. Dropping the first level simplifies model interpretation. The remaining dummy variables represent the difference between each category and the reference category (the one that was dropped)
3. Dropping the first level can reduce the computational burden, especially in models with many categorical variables

<u>Example:</u>

Consider a categorical variable "color" with three categories: "red", "green", and "blue". Without drop_first=True, there would be three dummy variables:

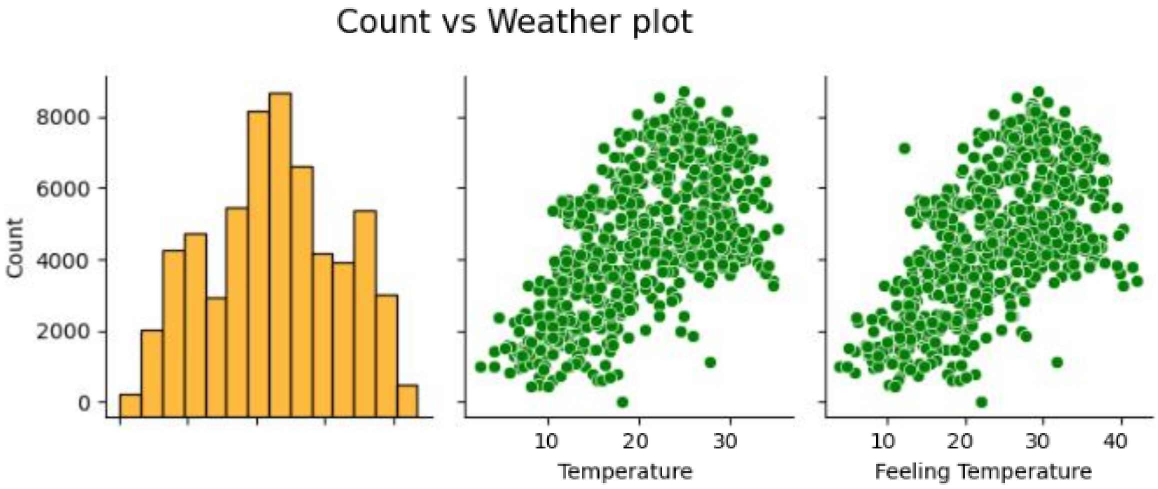
- color_red
- color_green
- color_blue

However, these three variables are linearly dependent. If color_red and color_green are both 0, then color_blue must be 1. By setting drop_first=True, there would be only two dummy variables:

- color_green
- color_blue

color_red is implicitly represented as the reference category (when both color_green and color_blue are 0)

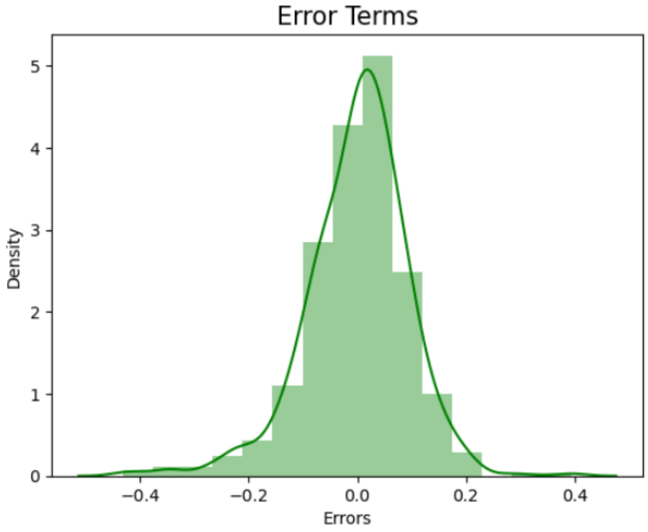
3. Looking at the pair-plot among the numerical variables, which one has the highest correlationwith the target variable?



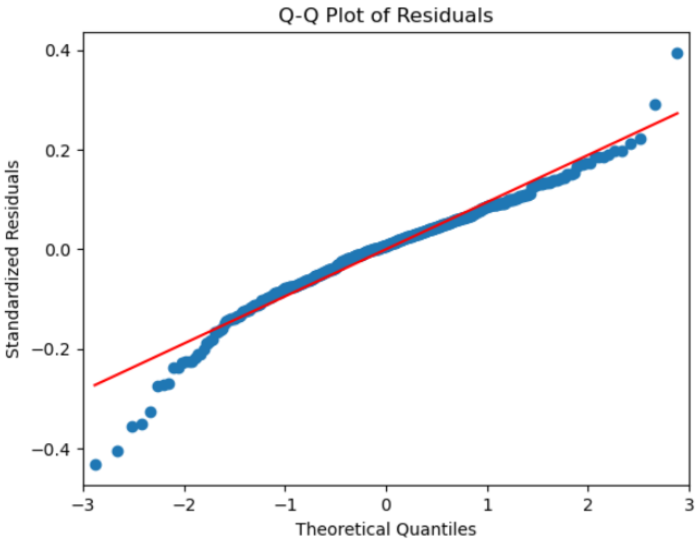
Looking at the pair-plot among Numerical Variables, Temperature (temp) and Feeling Temperature (atemp) has highest correlation with Target Variable (cnt) .

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

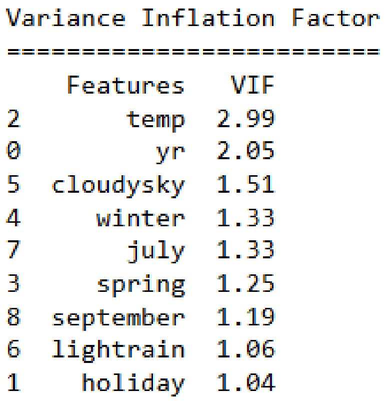
1) Created a scatter plot of residuals against predicted values and looked for patterns, such as heteroscedasticity (non-constant variance) or non-linearity



2) Created a Q-Q plot to check for normality of residuals. A straight line indicates normality

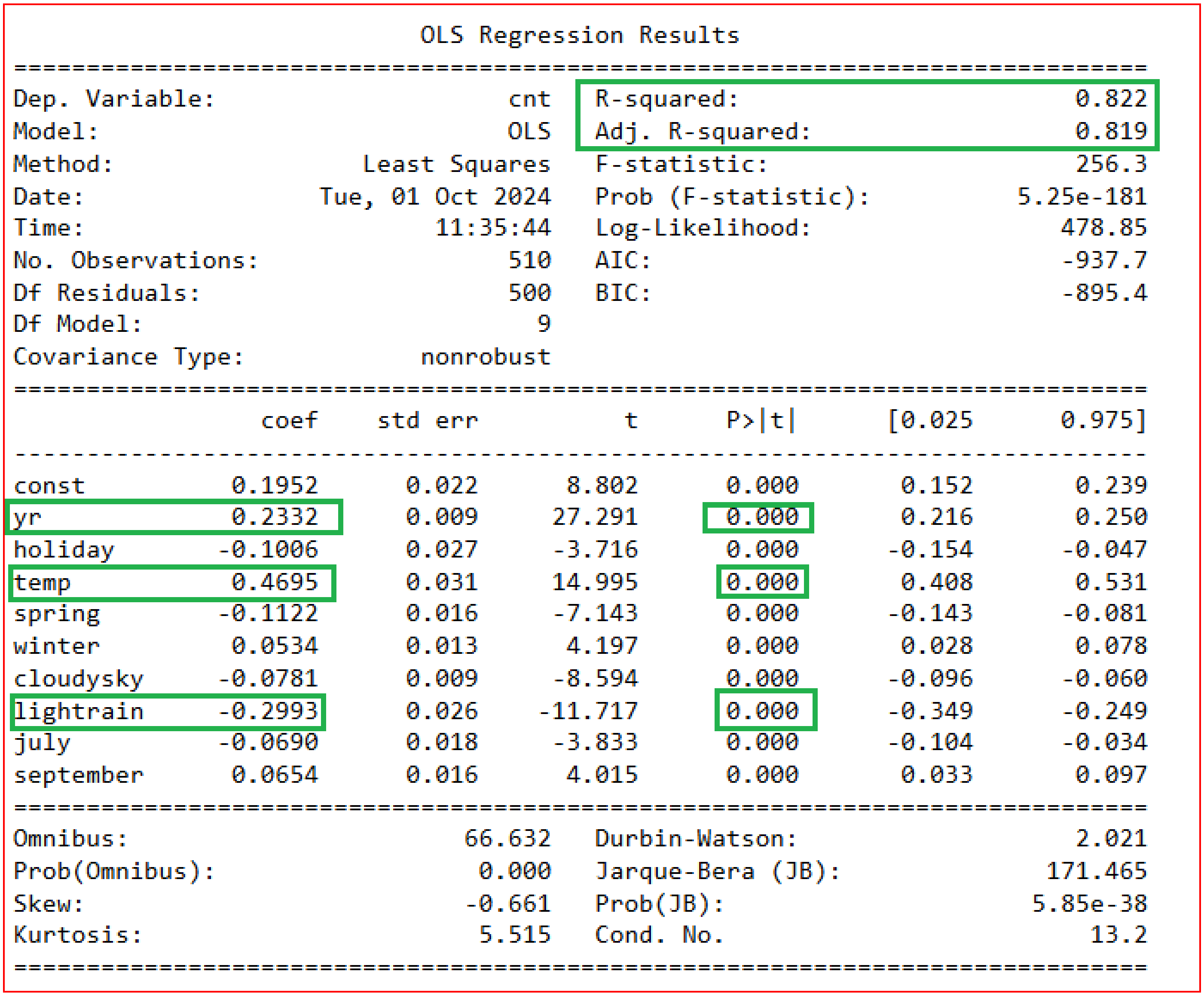


3) A measure of how much the variance of a regression coefficient is inflated due to multicollinearity. A VIF greater than 5 is often considered indicative of a problem.



```
Variance Inflation Factor
=========================
    Features   VIF
2       temp  2.99
0         yr  2.05
5  cloudysky  1.51
4     winter  1.33
7       july  1.33
3     spring  1.25
8  september  1.19
6  lightrain  1.06
1    holiday  1.04
```

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



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.822
Model:                            OLS   Adj. R-squared:                  0.819
Method:                 Least Squares   F-statistic:                     256.3
Date:                Tue, 01 Oct 2024   Prob (F-statistic):          5.25e-181
Time:                        11:35:44   Log-Likelihood:                 478.85
No. Observations:                 510   AIC:                            -937.7
Df Residuals:                     500   BIC:                            -895.4
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.1952      0.022      8.802      0.000       0.152       0.239
yr             0.2332      0.009     27.291      0.000       0.216       0.250
holiday       -0.1006      0.027     -3.716      0.000      -0.154      -0.047
temp           0.4695      0.031     14.995      0.000       0.408       0.531
spring        -0.1122      0.016     -7.143      0.000      -0.143      -0.081
winter         0.0534      0.013      4.197      0.000       0.028       0.078
cloudysky     -0.0781      0.009     -8.594      0.000      -0.096      -0.060
lightrain     -0.2993      0.026    -11.717      0.000      -0.349      -0.249
july          -0.0690      0.018     -3.833      0.000      -0.104      -0.034
september      0.0654      0.016      4.015      0.000       0.033       0.097
==============================================================================
Omnibus:                       66.632   Durbin-Watson:                   2.021
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              171.465
Skew:                          -0.661   Prob(JB):                     5.85e-38
Kurtosis:                       5.515   Cond. No.                         13.2
==============================================================================
```

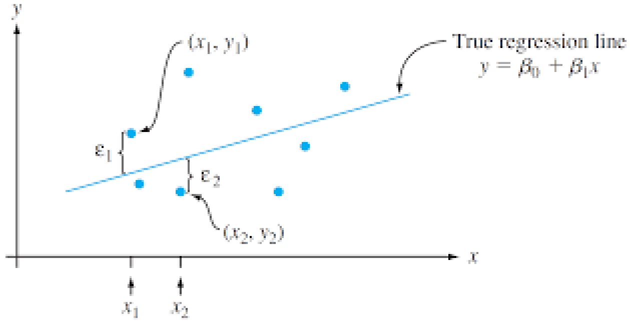
From the above Regression results and VIF, following are the features contributing significantly towards demand of shared bikes:

1. Year (yr) has the 2nd highest positive coefficient and as per business too, there is a year-on-year increase in the bike rental service
2. Temperature (temp) is a clear factor driving bike rentals evident from highest coefficient in the formula
3. Light Rain or Light Thunderstorm has the highest negative coefficient decreasing the demand for shared bikes and discouraging people from opting for rental

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The goal is to find a linear equation with the best straight line fitting to the given data, allowing us to predict the dependent variable based on the values of the independent variables.



It is of two types:

- Simple Linear Regression
  When there is only one independent variable. Equation takes the form:
  $y = mx + b$
- Multiple Linear Regression
  When there are multiple independent variables. Equation takes the form:$y = b_0 + b_1x_1 + b_2x_2 + \ldots\ldots + b_nx_n$

The most common method to find the best-fitting line is the least squares method. It minimizes the sum of the squared differences between the predicted values (from the line) and the actual values.

Linear regression makes several assumptions:

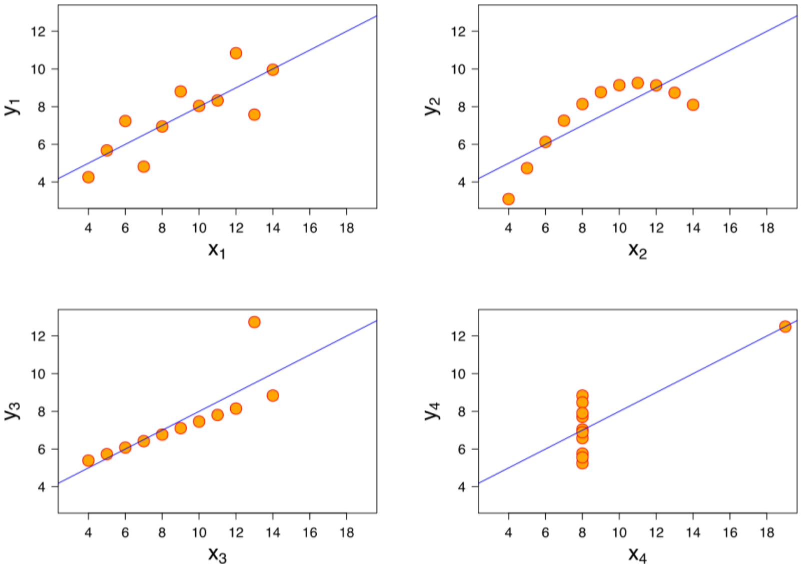
1. The relationship between the dependent and independent variables is linear
2. The observations are independent of each other
3. The variance of the residuals (errors) is constant across all levels of the independent variable(s), also known as Homoscedasticity
4. The residuals are normally distributed

Linear regression is widely used in various fields, including:

1. Predicting economic indicators like GDP, inflation, and interest rates.
2. Forecasting stock prices, bond yields, and risk.
3. Analyzing the relationship between marketing expenditures and sales.
4. Modeling physical phenomena like temperature, pressure, and stress.
5. Studying relationships between social variables like education, income, and crime rates.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a set of four datasets that, despite having nearly identical statistical properties (mean, median, variance, correlation), visually represent four very different relationships between two variables.



For all the four datasets represented above:

- Mean of x = 9
- Sample variance of x = 11
- Mean of y = 7.5
- Sample variance of y = 4.125
- Correlation between x and y = 0.816

What are the four datasets representing:

1. Plot 1 - Appears to be a simple linear relationship
2. Plot 2 - Relationship between the two variables exists, but it's not linear
3. Plot 3 - Relationship is linear, but is being offset by one outlier
4. Plot 4 - Relationship is linear, but is being offset by one outlier

3. What is Pearson's R?

Pearson's correlation coefficient (r) is a statistical measure that quantifies the linear relationship between two variables. It might not be a good measure if the relationship between variables is non-linear. It ranges from -1 to 1, where

- r = 1 means perfect positive correlation, the variables increase or decrease together perfectly
- r = -1 means perfect negative correlation, one variable increases as the other decreases perfectly
- r = 0 means no correlation between the variables

Pearson's R, $\boxed{r_{XY} = \frac{Covariance(\boldsymbol{X},\boldsymbol{Y})}{s_X s_Y}}$ where S is standard deviation

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

Scaling is used in data preprocessing to transform numerical data to a common range or scale. Scaling only affects the coefficients of variables and none of the other parameters like t-statistic, F-statistic, p-values etc.

Scaling is performed to:

- Bring features to a comparable scale, preventing features with larger magnitudes from dominating the learning process
- Scaled features can make the model's coefficients more interpretable.
- Improves the convergence rate of optimization algorithms
- Some algorithms, like K-Nearest Neighbors or Support Vector Machines, require features to be on a similar scale

There are basically two types of scaling methods:

1. Normalization

   Also known as Min-Max scaling, brings all the data in the range of 0 to 1. It’s mostly used when we want to preserve the relative differences between values and have a clear understanding ofthe minimum and maximum possible values. Formula to scale:

   $x_{iscaled} = (x_i - min(x)) / (max(x) - min(x))$

2. Standardization

   It brings all of the data into a standard normal distribution with mean zero and standard deviation as one. It’s mostly useful for algorithms that assume a normal distribution. Formula to scale:

   $x_{iscaled} = (x_i - mean(x)) / sd(x)$

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

Infinite VIF (Variance Inflation Factor) indicates a perfect multicollinearity between the independent variables in a regression model. This means that one or more independent variables can be perfectly predicted from the others, making it impossible to estimate their unique effects on the dependent variable.

Formula for VIF is, VIF = 1 / (1-R^2)

- If R^2 = 0, VIF = 1
- If R^2 = 1, VIF = Infinity

Following can be some of the causes for the same:

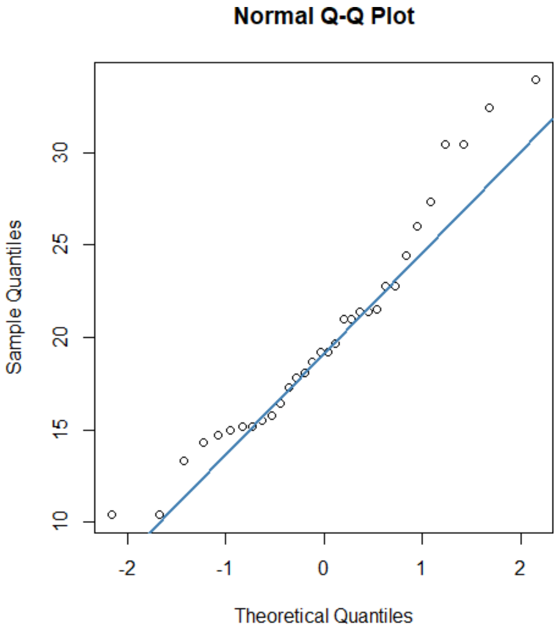
- Redundant Variables: Including the same variable twice or in different forms (e.g., cost and cost_in_millions) can lead to perfect multicollinearity
- Linear Combinations: If one variable is a linear combination of others (e.g., "total_income" as the sum of "salary" and "bonus")

Some of the ways of addressing infinite VIF are as follows:

- Dropping Variables: If a variable has an infinite VIF, it should be removed from the model. If there are multiple such variables pick the business interpretable variable
- Create new variables using the interactions of the older variables, and drop the original variables

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

A quantile-quantile (Q-Q) plot is a graphical method used to compare two probability distributions or assess how well a set of data fits a theoretical distribution. It's a diagnostic tool in various statistical analyses, including linear regression.



To draw a Q-Q plot, follow the outlined steps:

1. Collect the Data: Gather the dataset and ensure that the data are numerical and represent a random sample from the population
2. Sort the Data: Arrange the data in either ascending or descending order
3. Choose a Theoretical Distribution
4. Calculate Theoretical Quantiles: Compute the quantiles for the chosen theoretical distribution
5. Plotting:
    a. Plot the sorted dataset values on the x-axis
    b. Plot the corresponding theoretical quantiles on the y-axis
    c. Each data point (x, y) represents a pair of observed and expected values
    d. Connect the data points to visually inspect the relationship between the dataset and the theoretical distribution

A Q-Q plot can be interpreted as follows:

- If the points on the plot fall approximately along a straight line, it suggests that the dataset follows the assumed distribution
- Deviations from the straight line indicate departures from the assumed distribution, requiring further investigation