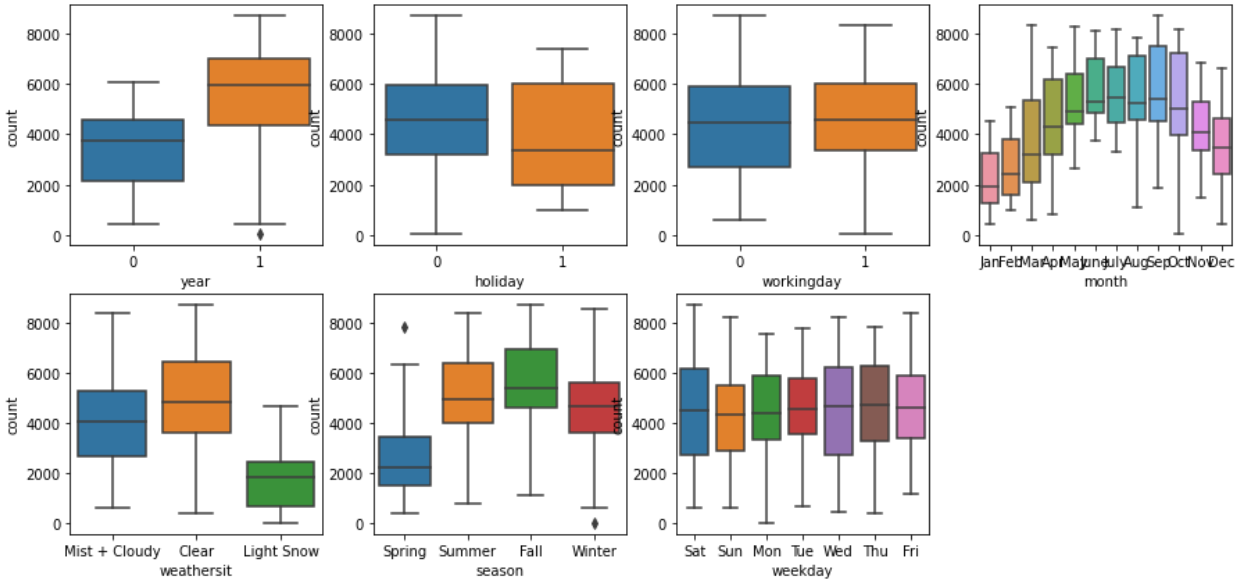
Assignment-based Subjective Questions

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



As per the above plots, the figures of Bike Rentals are more :

- During the Fall season and then in summer
- In the year 2019 compared to 2018
- In partly cloudy weather
- On Saturday, Wednesday and Thursday

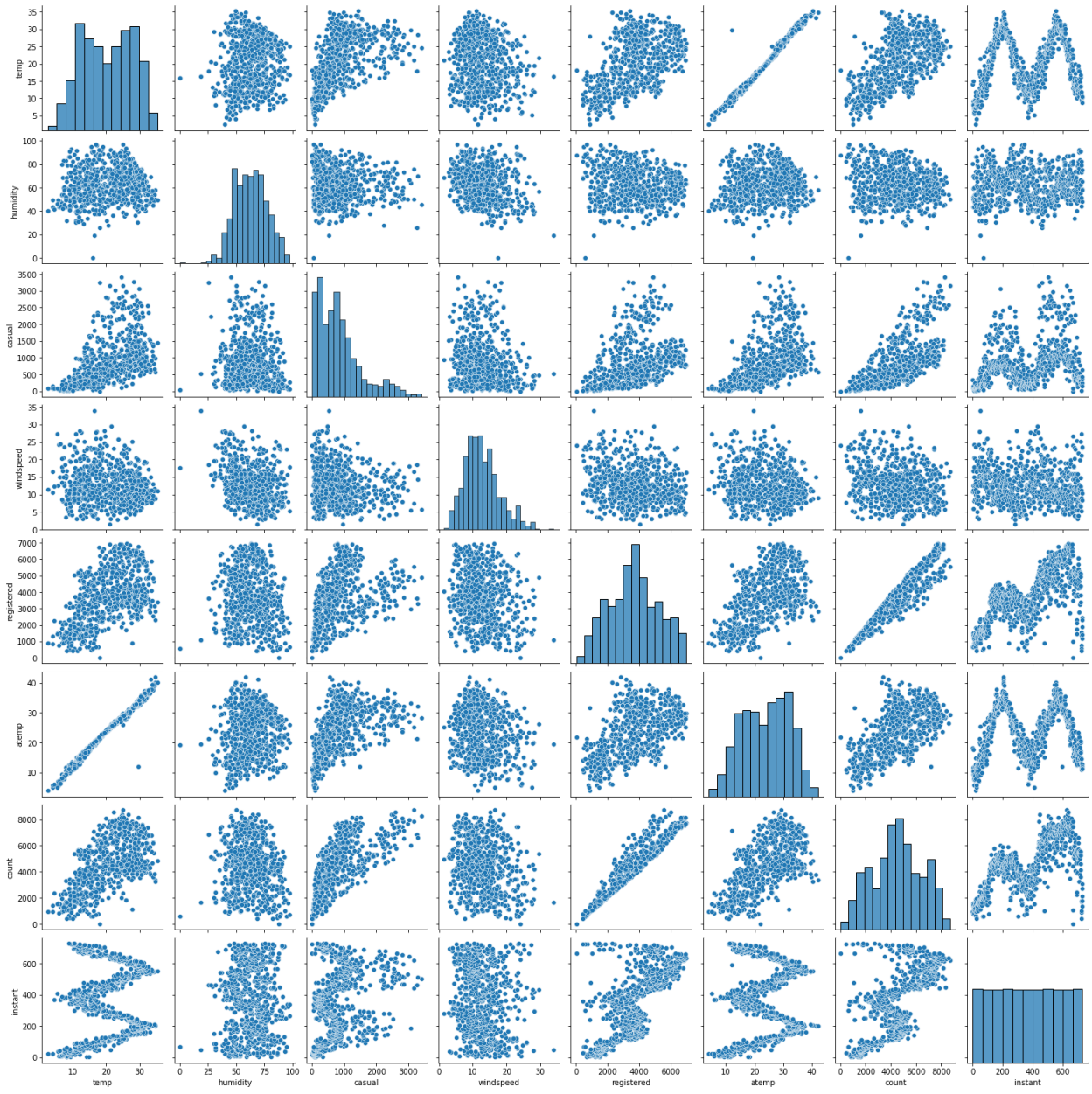
Also, after performing univariate analysis on the variables, we were able to conclude that:

- Bike are rented more during the Fall season and then in summer
- Bike are rented more in partly cloudy weather
- Bike are rented more in the year 2019 compared to 2018
- Bike Rentals are observed at higher temperatures
- Bike rentals more at high humidity

2. Why is it important to use *drop_first=True* during dummy variable creation? (2 mark)
Answer: Dummy variable is used while performing one-hot encoding on the dataset. Thus we use the get_dummies function to convert categorical variables into dummy or indicator variables.
While performing this operation use ***drop_first=True***, as this value assists in reducing the extra column created during dummy variable creation. Which in turn, reduces the correlations created among dummy variables.

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

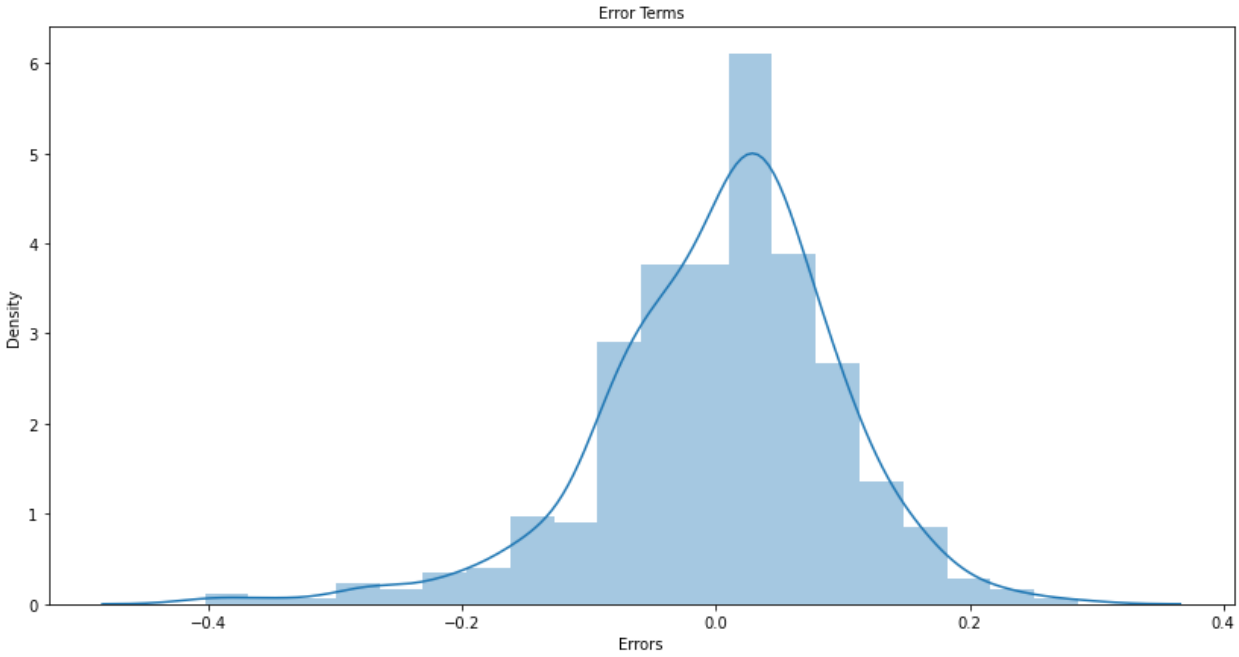
Answer:



As per the pairplot, *temperature* is the most correlated variable to the bike rentals.

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

Answer: The assumptions of the linear regression model after training it on the provided data is done by performing Residual Analysis of the train data. Please find the histogram plot of the error terms. We can analyze that the Error terms are normally distributed.



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

Answer: As per our final model with the VIFs and p-values both are within an acceptable range and with RFE support columns, the top three features would be:

- The *temperature* variable is with the highest coefficient 0.6918, which means if the temperature increases by one unit the number of bike rentals increases by 0.6918 units.
- Second would be *Light Snow* with coefficient -0.2829. A negative coefficient suggests that as the independent variable increases, the dependent variable tends to decrease.
- Third would be *year* with coefficient 0.2369

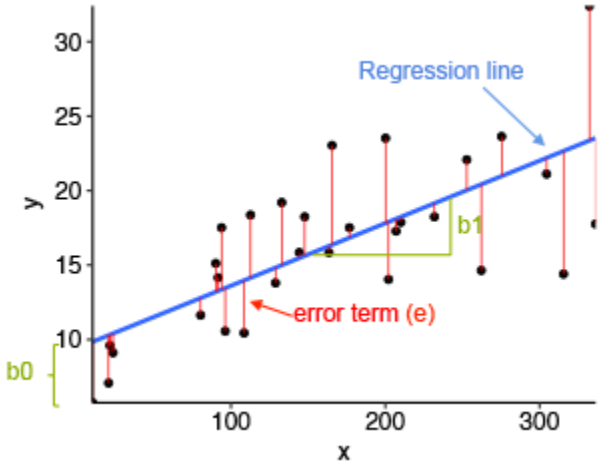
General Subjective Questions

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

Answer: Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

Before attempting to fit a linear model to observed data, a modeler should first determine whether or not there is a relationship between the variables of interest. This does not necessarily imply that one variable causes the other (for example, higher SAT scores do not cause higher college grades), but that there is some significant association between the two variables. A scatterplot can be a helpful tool in determining the strength of the relationship between two variables. If there appears to be no association between the

proposed explanatory and dependent variables (i.e., the scatterplot does not indicate any increasing or decreasing trends), then fitting a linear regression model to the data probably will not provide a useful model. A valuable numerical measure of association between two variables is the correlation coefficient, which is a value between -1 and 1 indicating the strength of the association of the observed data for the two variables.

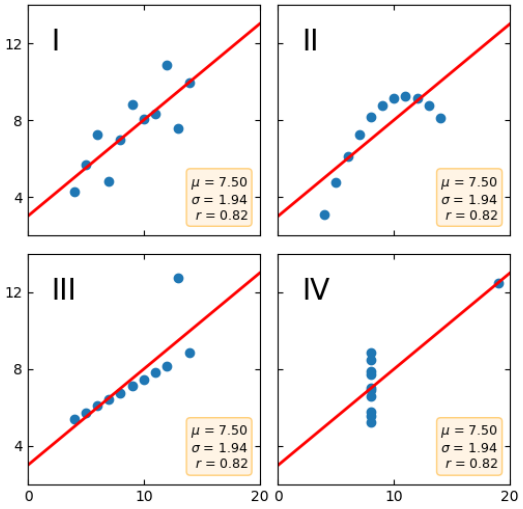


A linear regression line has an equation of the form Y = a + bX, where X is the explanatory variable and Y is the dependent variable. The slope of the line is b, and a is the intercept (the value of y when x = 0).

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

Answer: Anscombe's quartet is a group of datasets (x, y) that have the same mean, standard deviation, and regression line, but which are qualitatively different.

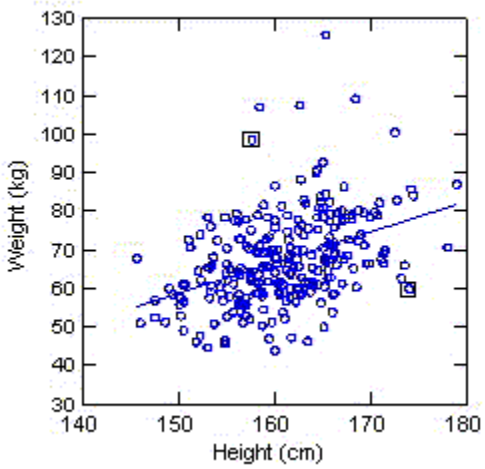
It is often used to illustrate the importance of looking at a set of data graphically and not only relying on basic statistic properties.



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

Answer: Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

"Tends to" means the association holds "on average", not for any arbitrary pair of observations, as the following scatterplot of weight against height for a sample of older women shows. The correlation coefficient is positive and height and weight tend to go up and down together. Yet, it is easy to find pairs of people where the taller individual weighs less, as the points in the two boxes illustrate.



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

Answer:

- **Scaling:** It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
- **Importance :** It is important to have all the variables on the same scale for the model to be easily interpretable. We can use standardization or normalization so that the units of the coefficients obtained are all on the same scale.
- **Difference between normalized scaling and standardized scaling:** The major difference is between the range the data is scaled, such as:
    - *Normalisation* (Min-Max scaling): Between 0 and 1
    - *Standardisation* : Between mean-0, sigma-1

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

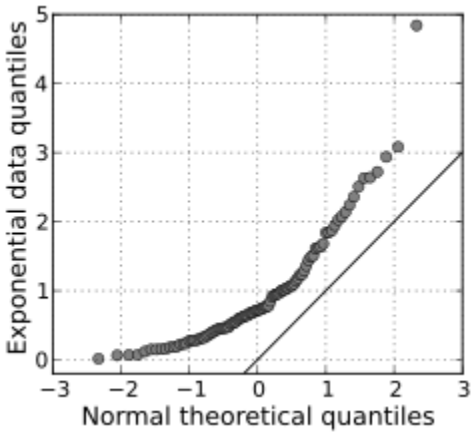
Answer: If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To

solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
Answer: Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q-Q plot showing the 45 degree reference line:



Importance of a Q-Q plot in linear regression

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.