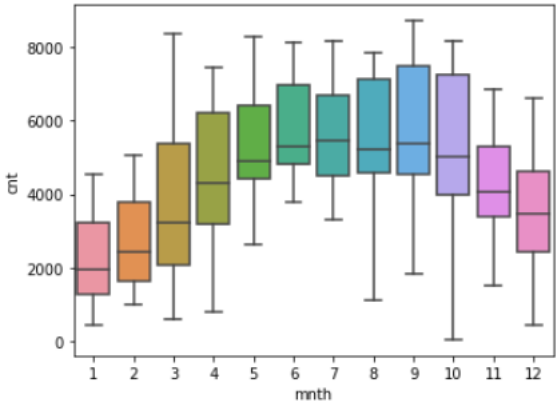
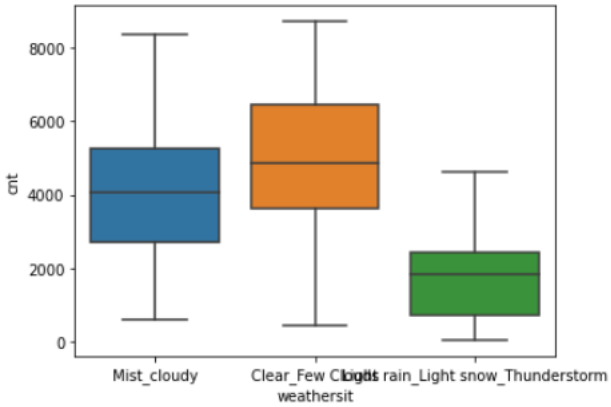
Assignment-based Subjective Questions.

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

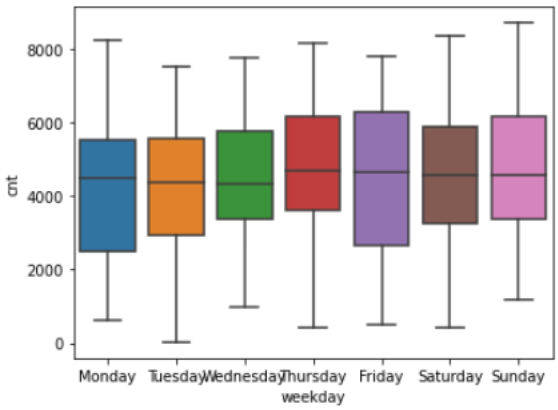
Solution:



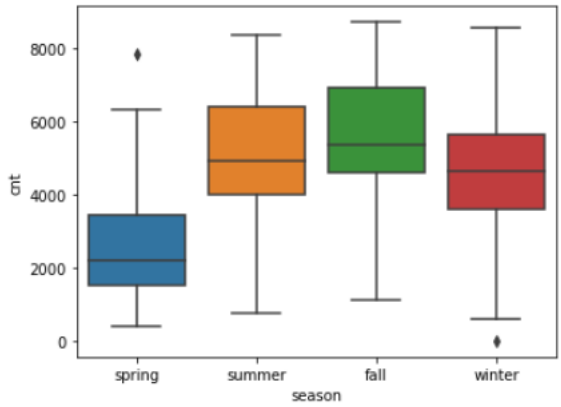
Mnth – Number of bikes rented goes high in mid-year.



Weather – Number of bikes rented are high when the sky is clear, few clouds.



Weekday – Number of bikes goes high in the mid of the week.



Season – Fall is the top season in which the number of bikes is rented.

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

Solution:

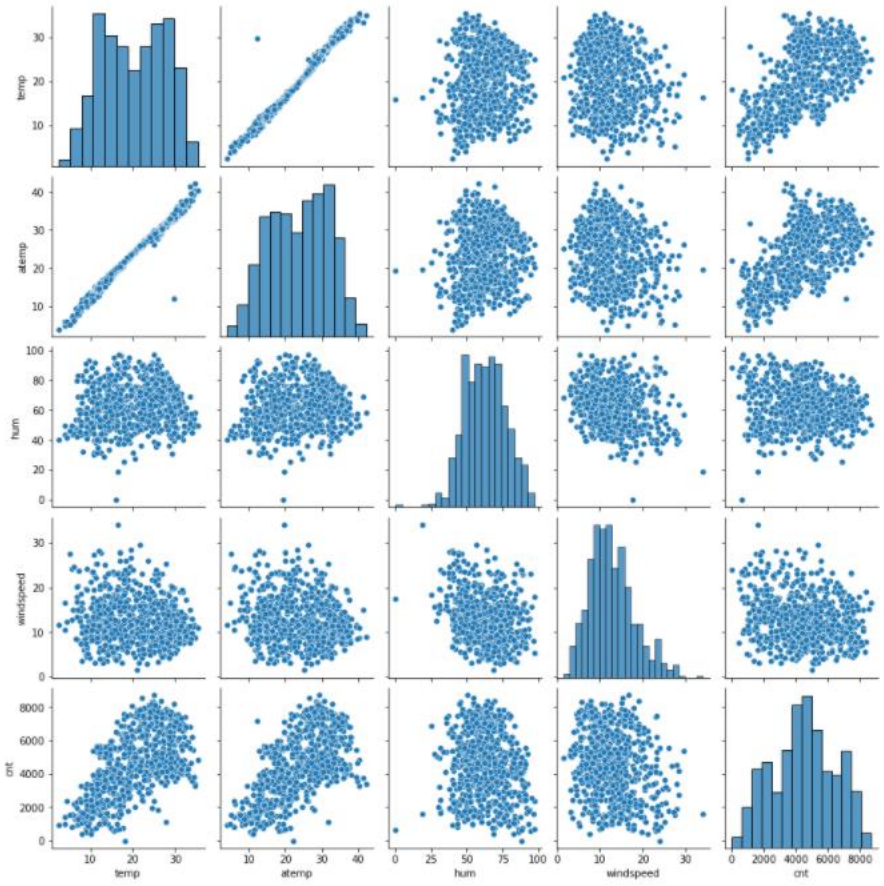
In the dummy variables, if you don't drop the first column then it can affect the model adversely. The effect is stronger when the cardinality is small.

Hence, Drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.
For example, if we have 3 variables then after creating dummies, we only require 2 of them to give the values of the third one. If one variable is not furnished and semi furnished, then it is obviously unfurnished.

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

Solution:

From the graph below, we can easily identify that there is a very high correlation between temp and a temp.



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

Solution:

While performing EDA, I visualized the relationship between the categorical variables and the target variable. Let's consider the effect of the categorical variable 'weathersit' on the target variable 'cnt'.

It was seen that during the weather 'Clear, few clouds, partly cloudy', a high number of bike rentals were made, with the median being 50,000 approximately.

I saw a significant growth in the value of R-squared and adjusted R-squared during model building on inclusion of categorical features such as yr, season etc.

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

Solution:

After building the final linear regression model, it is clear that weathersit, mnth and weekday columns explain the contribution towards demand of shared bikes.

Bad weather conditions really affect the demand just like in the real life.

Mnth in which the the temperature is not very high or low.

Demand increases in the mid-week.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Solution:

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

Example:

Let's say you are running a sales promotion and expecting a certain number of count of customers to be increased now what you can do is you can look the previous promotions and plot if over on the chart when you run it and then try to see whether there is an increment into the number of customers whenever you rate the promotions and with the help of the previous historical data you try to figure it out or you try to estimate what will be the count or what will be the estimated count for my current promotion this will give you an idea to do the planning in a much better way about how many numbers of stalls maybe you need or how many increase number of employees you need to serve the customer. Here the idea is to estimate the future value based on the historical data by learning the behaviour or patterns from the historical data.

Linear regression is used to predict a quantitative response Y from the predictor variable X.

Mathematically, we can write a linear regression equation as:

$$y = a + bx$$

Where a and b given by the formulas:

$$b\,(slope) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a\,(intercept) = \frac{n\sum y - b(\sum x)}{n}$$

Here, x and y are two variables on the regression line.
b = Slope of the line.
a = y-intercept of the line.
x = Independent variable from dataset
y = Dependent variable from dataset

Use Cases of Linear Regression:

Prediction of trends and Sales targets – To predict how industry is performing or how many sales targets industry may achieve in the future.
Price Prediction – Using regression to predict the change in price of stock or product.
Risk Management- Using regression to the analysis of Risk Management in the financial and insurance sector.

2. Explain the Anscombe’s quartet in detail.

Solution:

Anscombe’s quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties.

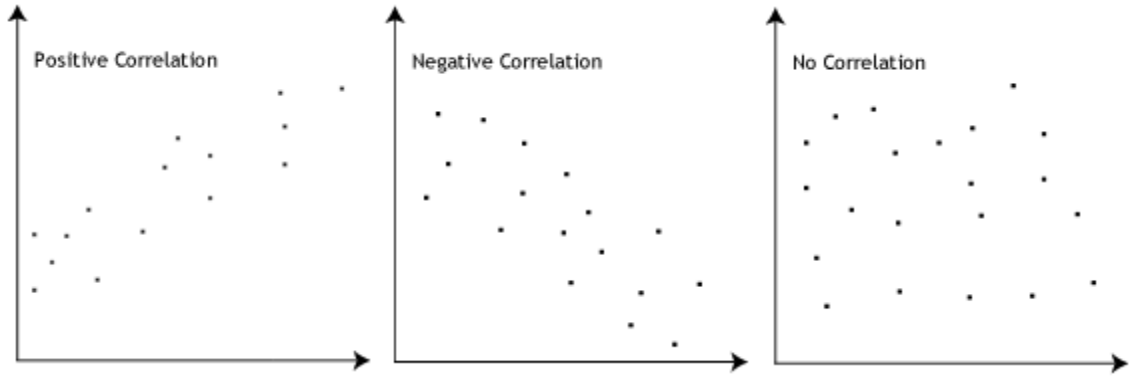
Once Francis John “Frank” Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points.

3. What is Pearson’s R?

Solution:

In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

The Pearson's correlation coefficient varies between -1 and +1 where:
r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
r = 0 means there is no linear association
r > 0 < 5 means there is a weak association
r > 5 < 8 means there is a moderate association
r > 8 means there is a strong association



Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

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

Solution:

Scaling data is the process of increasing or decreasing the magnitude according to a fixed ratio, in simpler words you change the size but not the shape of the data.

It is not mandatory to use feature scaling but it definitely is a good practice. It helps handling disparities in units. During long processes it definitely helps reduce computational expenses. In the machine learning eco space, it helps improve the performance of the model and reducing the values/models from varying widely.

Normalization usually means to scale a variable to have a values between 0 and 1, while standardization transforms data to have a mean of zero and a standard deviation of 1.

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

Solution:

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).

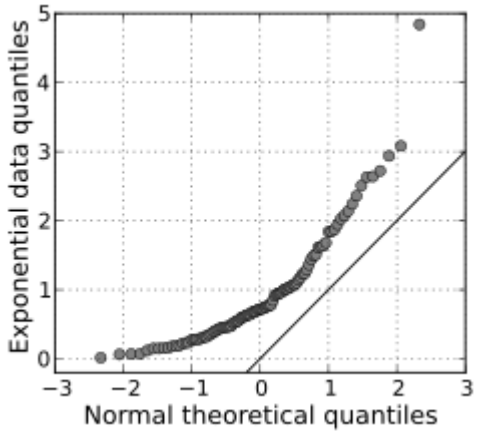
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Solution:

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:



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.