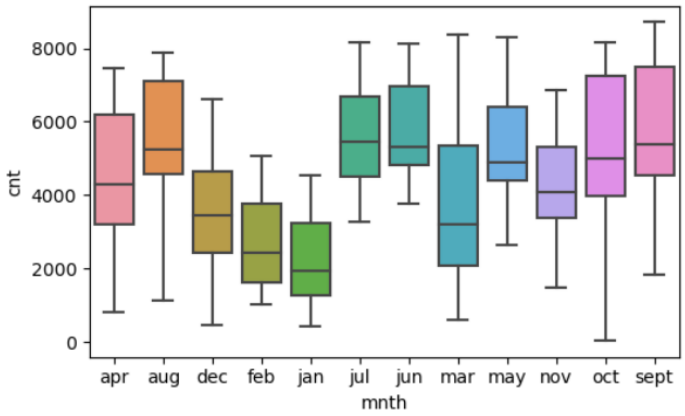
Assignment-based Subjective Questions

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

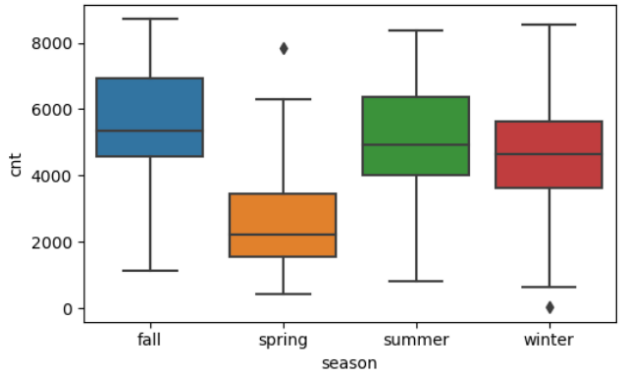
After performing statistical analysis and observing the charts, correlation values etc, following can be stated about the effect of categorical variables : mnth, season, yr , weekday, holiday, workingday, weathersit

mnth:- The following box plots show that bike rental was higher during the months of jun, jul, aug, sep. The correlation values were also positive and indicated the same effect. A few months showed negative correlation values and that is also visible in the below box plot. Bike rental service was running really low in the months of dec, jan & feb.

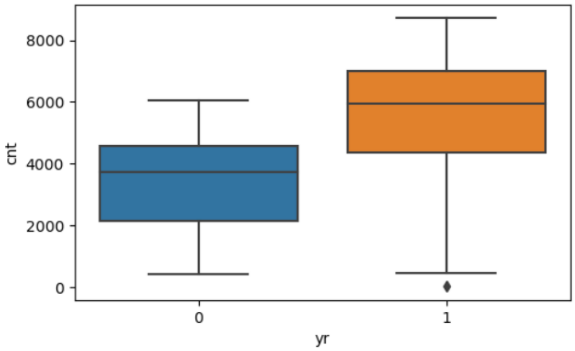


season:- Spring season is not favourable for bike rental service. It also shows up in the below graph. The correlation also has come out to be negative (-0.56).

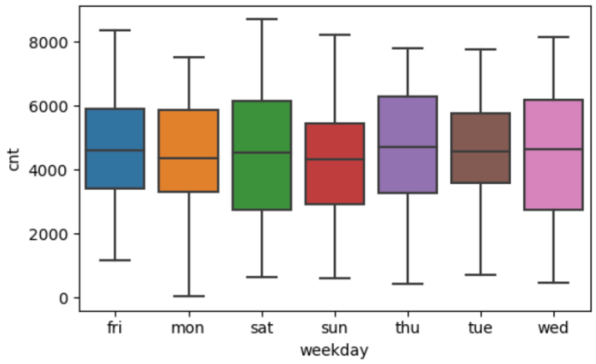
Season summer, and season fall shows positive correlation with the target variable.



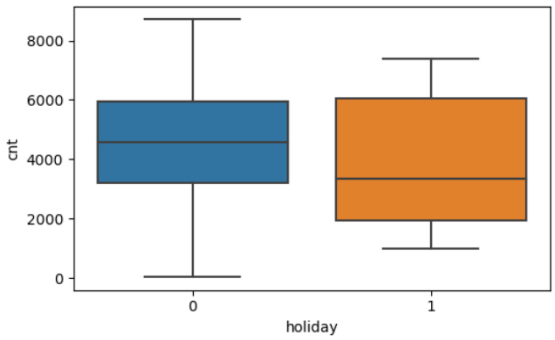
yr:- There was an increase in bike rental service in the subsequent year as compared to year 2018.



weekday:- Sunday has a small negative correlation of -0.059. Weekdays such as Thursday, Wednesday appear to be better for business. It seems there is more demand for bikes during the weekdays as compared to weekend.

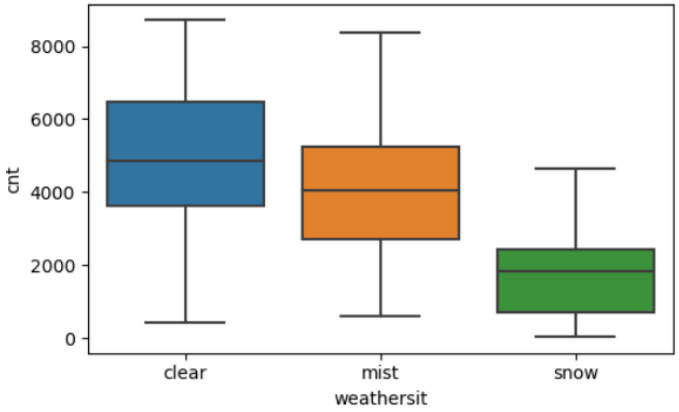


holiday:- There is a negative correlation value for holiday. It is -0.069. Thus holiday has negative impact on bike rental service.



workingday:- Workingday has a positive effect on bike rental.

weathersit:- From the box plot below, it is clear that weather situation has impact on the target variable. Good weather situation promotes bike rental. As the weather situation deteriorates, the bike rental numbers also slip. Even the correlation values for mist and rain weather situation are -0.17 & -0.24 respectively, which confirm the pattern.



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

In order to represent x different values, we can use x-1 variables. The xth variable can be represented implicitly by setting flags for all other x-1 variables to false.

When we create dummy variables, one dummy variable is created for every possible value. Thus if there are x values, x dummy variables will be created. This leads to the problem of multicollinearity as these x variables will be correlated with each other.

Hence, drop_first =True is used to drop the extra variable, and that aides in the mitigation of the problem of multicollinearity.

If we take example of the current project, we have 7 possible values for weekdays, i.e., Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday. We can easily represent these 7 values using 6 columns. If we want to represent a particular day, say Monday, we can set the value for that column to 1 and 0 for the rest of the columns. Thus we can easily represent 6 weekdays. The 7^{th} day, that is, Sunday can be represented by setting 0 in all the 6 columns.

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

temp & atemp have the highest positive correlation with the target variable.

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

We can validate the assumptions of linear regression by inspecting the outputs produced by it. Following key outputs and their numeric values were examined:

1. Residual analysis was done and the distribution was computed. The distplot was plotted and it was observed that the distribution was a normal distribution, with its mean around 0.
2. VIF values of all the predictor variables were < 5, which proved that there was no multicollinearity.
3. p-value was nearly 0 or less than 0.05 for all the predictors.
4. R-squared and Adjusted R-squared scores were found to be high nearly 80%.
5. The model was used to make predictions and the scatter plot showed that the values were within a tight range with very few outliers.
6. R-squared & Adjusted R-squared scores were also computed for the test data set and the values were compared to that obtained on training data. These values were close to each other.

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

Following features can be focused:

1. Temperature (temp):- temp had high positive coefficient.
2. Weather:- If weather is good, bike rental demand is high. If weather is bad (snow, rain, severe), the bike rental demand is poor.
3. Windspeed (windspeed):- windspeed had negative impact on bike's demand.

Seasons with good weather conditions showed high demand for bikes, e.g. Fall, Summer.

Besides, bike hiring was more in 2019 as compared to 2018. It can be assumed that bike rental would see an increase in upcoming years.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Regression is the most commonly used predictive analysis model. Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric values. This is based on the popular equation "y = mx

+ c". There is a linear relationship between the dependent variable(y) and the predictor(s)/independent variable(x).
Linear Regression is broadly divided into simple linear regression and multiple linear regression.

SLR [Simple Linear Regression] :

This is also called as Univariate Linear regression**.** This is used when the dependent variable is predicted using only one independent variable.

When we try to find out a relationship between a dependent variable (Y) and one independent (X) then it is known as Simple Linear Regression/ Univariate Linear regression.
The mathematical equation can be given as:

$$Y = \beta_0 + \beta_1 * x$$

Where

- Y is the response or the target variable
- x is the independent feature
- β_1 is the coefficient of x
- β_1 is the intercept

MLR [Multivariate Linear Regression] :

This is used when the dependent variable is predicted using multiple independent variables. The equation for multiple linear regression is similar to the equation for a simple linear equation. The formula for multiple linear regression would look like,

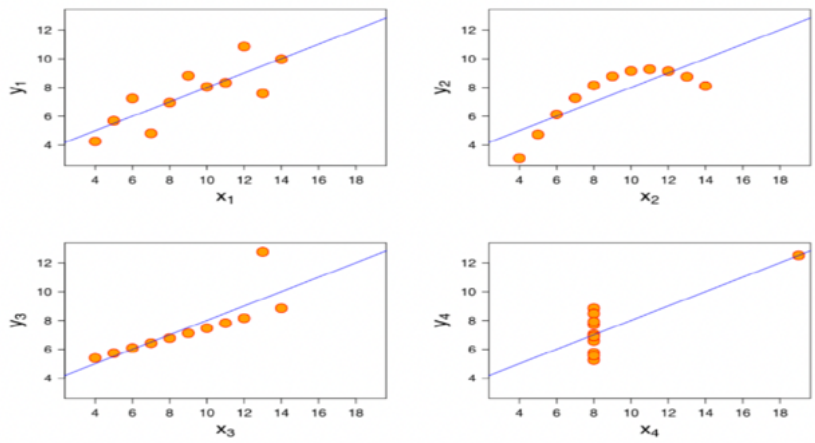
$$Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + ... + \beta n * Xn + \epsilon$$

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet was devised to illustrate how important it was to not just rely on statistical measures when analysing data. To do this he created 4 data sets which would produce nearly identical statistical measures.

There are multiple datasets which are completely different but after training, the regression model looks the same. A regression model is not always necessarily an exact one, it can also be fooled by some (smart) data! In certain cases. A group of four such datasets having identical descriptive statistics but with some peculiarities, is the Anscombe's quartet.

The 4 data sets created by Anscombe would produce nearly identical mean, standard deviation, and regression line. But there graph look entirely different as shown below:-



Inference

As seen in the above diagram, first graph shows that there is a linear regression between y & x. The second graph shows that there exists no linear relationship between the predictor and the target variable. In the third graph, there exist outliers which could not be explained by the linear model. The fourt graph on the other hand shows presence of strong correlation.

3. What is Pearson's R?

Pearson's R is used to measure linear correlation between two variables. It's value lies between -1 and 1, where the negative/positive sign shows the direction of the relation and the numeric value shows its strength.

Pearson correlation coefficient (*r*)	**Correlation type**	**Interpretation**
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the **same direction**.
0	No correlation	There is **no relationship** between the variables.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the **opposite direction**.

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

Feature Scaling is the process of **standardizing the independent features present in the data in a fixed range.** It is also known as data normalization and is considered an important data pre-processing step. Feature scaling becomes necessary when dealing with datasets containing features that have different ranges, units of measurement, or orders of magnitude. In such cases, the variation in feature values can lead to biased model performance or difficulties during the learning process. For example, if we have multiple independent variables like fuel_price, & car_mileage with their range as (90-100 Rs), (8-12 kms), respectively, feature scaling would help them all to be in the same range

Using feature scaling, the dataset's features can be transformed to a more consistent scale, making it easier to build accurate and effective machine learning models.

There are two types of scaling:

1. Normalized scaling:

Normalization is good to use when the distribution of data does not follow a Gaussian distribution. Normalization or Min-Max Scaling is used to transform features to be on a similar scale. The new point is calculated as:

$$X\_scaled = (X - X\_min)/(X\_max - X\_min)$$

Normalization is affected by the presence of outliers, so we either remove them first or we don't perform normalized scaling on them. For example, income of a few people may be too high from the normal population, but the age of people will be within a well-defined range, so we won't scale the age variable, but we can normalize income.

2. Standardized scaling:

Standardization is usually used where data follows Gaussian distribution. The new points are calculated as:

$$X\_scaled = (X - mean)/Std$$

Standardization does not produce data in a bounding range, so they can handle outlier data easily.

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

The greater the VIF, the higher the degree of multicollinearity. In the limit, when multicollinearity is perfect (i.e., the regressor is equal to a linear combination of other regressors), the VIF tends to infinity. If there is a perfect correlation between the dependent variable and independent variable(s), the R-squared value comes out to be 1. Hence VIF, which is (1/(1-R^2)) turns out to approach infinity.

In other words, can say If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity.

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

The quantile-quantile (q-q) plot is a probability plot. In this technique we generate graphs for determining if two data sets come from populations with a common distribution. Quantile-Quantile plot or Q-Q plot is created by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against. It is a graphical tool to assess if sets of data come from the same statistical distribution.

Q-Q plot is used for the following purpose:

- Determine whether two samples are from the same population.
- location, scale, and skewness are similar or different in the two distribution

Importance and benefits of Q-Q plot:

- Since Q-Q plot is like probability plot. So, while comparing two datasets the sample size need not to be equal.
- Since we need to normalize the dataset, so we don’t need to care about the dimensions of values.