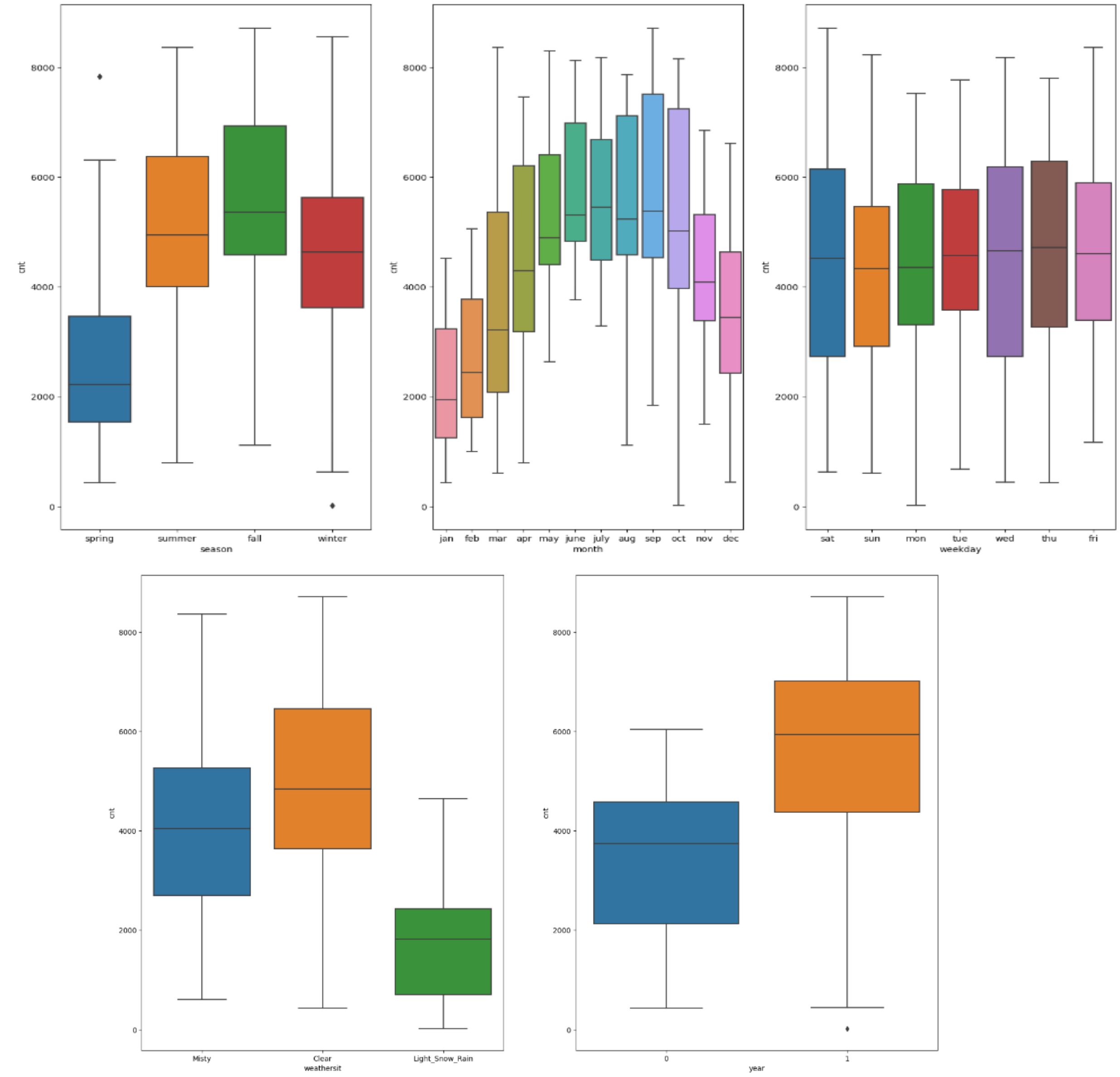
Assignment-based Subjective Questions

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

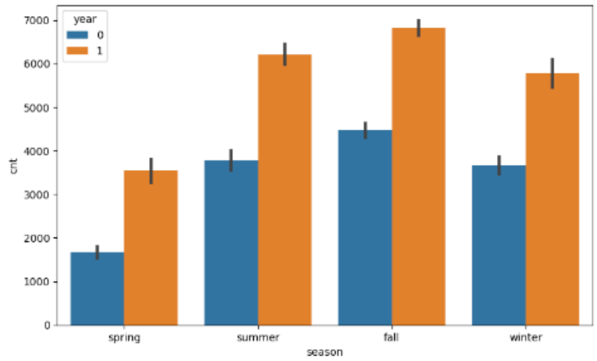
From my analysis below are the significant independent categorical variable which has an effect on the dependent target variable ‘cnt’

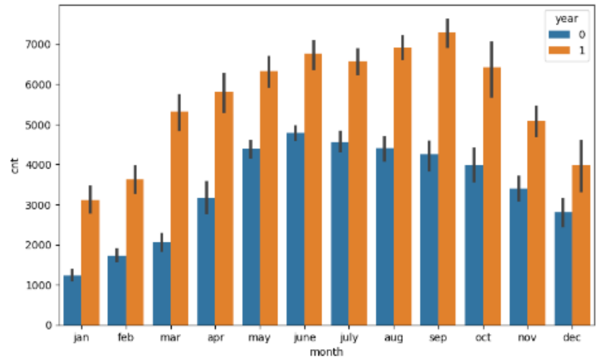
1. season
2. months
3. weekday
4. weathersit
5. year

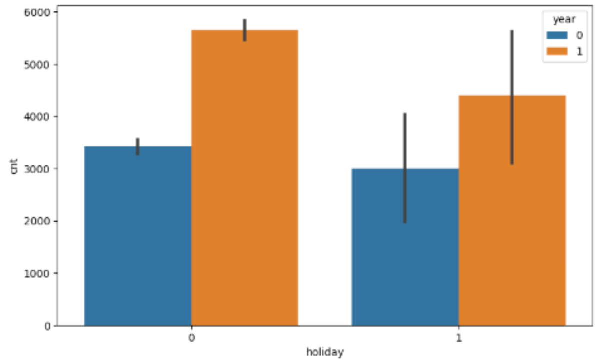
Boxplots of these categorical features with count on the y-axis

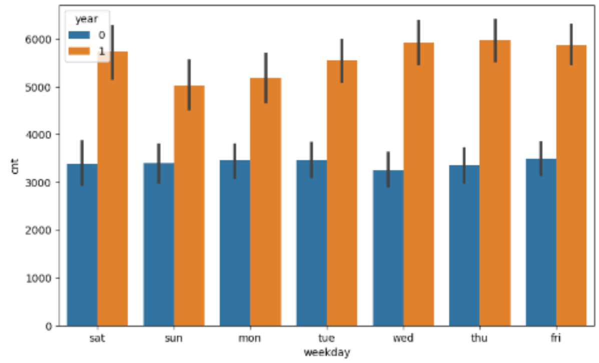


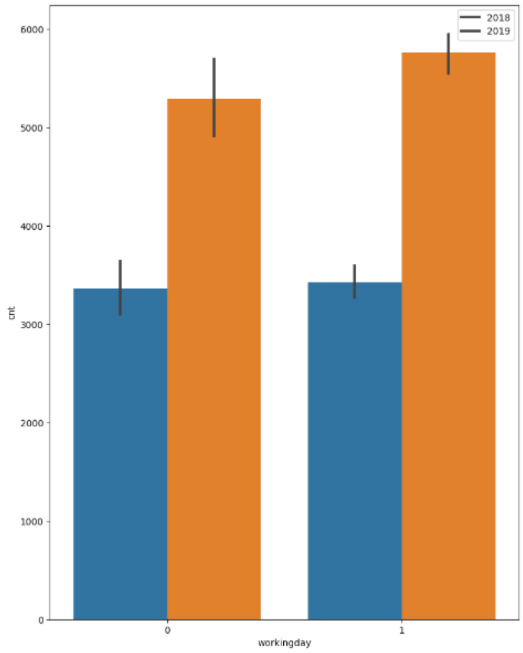
Bar plot of these variables with cnt on the y-axis

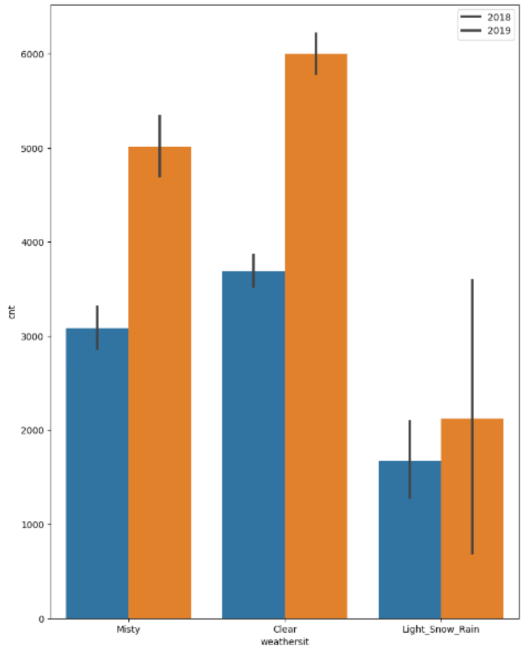












Observation:

- The count of rental bike is more during fall season of both 2018 & 2019
- The demand for rental bikes gradually increases and then decreases during 2018 and 2019
- The demand of rental bikes is more during non holiday season for both 2018 and 2019 years
- The demand for rental bikes is more during weekdays when compared to weekends
- The demand is high when the weather is Clear

After fitting a regression line, the coefficients of these categorical variables are found to be:

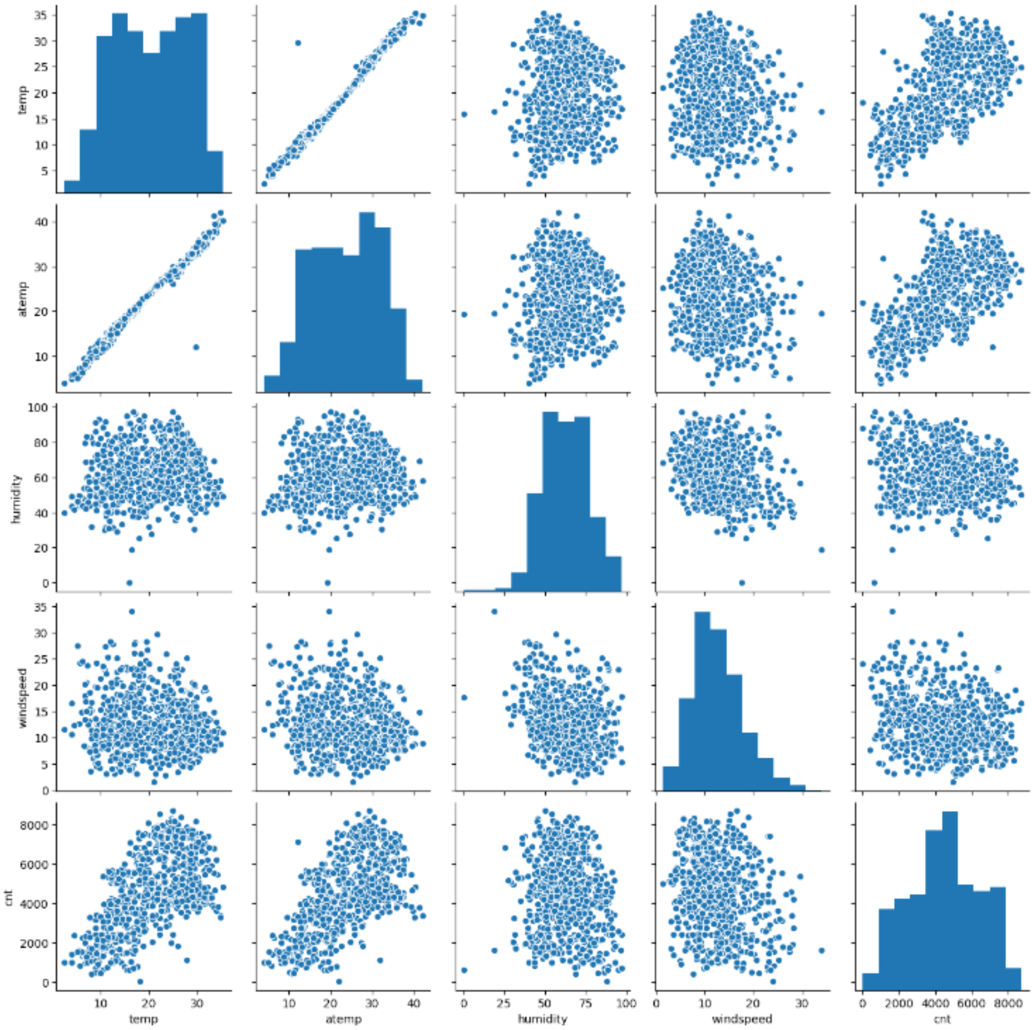
Sl.No	Feature	Coefficient	Effect
1	spring	-0.068197	Negative
2	summer	0.047885	Positive
3	winter	0.081830	Positive
4	july	-0.048253	Negative
5	sep	0.072321	Positive
6	sun	-0.044959	Negative
7	Light_Snow_Rain	-0.284654	Negative
8	misty	-0.080237	Negative
9	year	0.233876	Positive

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

In general, if there are p levels for a categorical variable then that can be converted to p-1 indicator variables. Here, p dummy levels are not required since the p-1 variables will indicate all the values of the dummy variable with one level as the base state. Having p dummy variables instead of p-1 would only add complexity but no significant indication. Hence the first level is dropped by using the argument drop_first=True will creating the dummy variables.

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

Numerical variables ‘temp’ and ‘atemp’ has highest correlation with the target variable



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

Assumptions of linear regression

Sl.NO	Assumption	Validation
1	Linearity	By visualizations
2	Distribution of residuals	Residual Analysis - By making a plot of error terms and checking If they are normally distributed
3	Multicollinearity	Variance inflation Factor – By computing the VIF values and removing the features with very high VIF value
4	Homoscedasticity	Scatter plot – By checking if the error is constant along the values of the dependent variables

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

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
A Linear regression algorithm tries to build a linear relationship between an input/independent variable and a output/dependent variable by fitting a line. Here the relationship between these variables can be a positive or negative relationship.

The equation of this line is given by

$$Y = Beta0 \ + \ Beta1 * \mathrm{X}$$

Where:
Beta0 = Intercept
Beta1 = slope/coefficient
X is the independent variable

The main assumptions of the linear regression algorithm are:

1. Linearity – There is a linear relationship between x and y
2. Distribution of residuals - The errors terms are normally distributed
3. Multicollinearity – The input variables must be independent. There should not be correlations among the independent variables.
4. Homoscedasticity – The error terms should have constant variance

In case, of simple linear regression there will be only 1 independent variable. Similarly, in case of multiple linear regression there will be multiple independent variables. Where the Beta1, Beta2, and so on will be the coefficient of the multiple linear regression model. The goal of the algorithm is to identify the best fit line while minimizing the loss function/cost function. Gradient descent is one of the popular optimization algorithm which can be used to compute the coefficients while minimizing the loss function.

----

2. Explain the Anscombe’s quartet in detail. (3 marks)
The idea behind the Anscombe's quartet is that there can be some datasets which have nearly identical descriptive statistics. But when these datasets are plotted, they can have different distributions. According to Wikipedia these datasets were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of plotting the data when analysing it, and the effect of outliers and other influential observations on statistical properties.

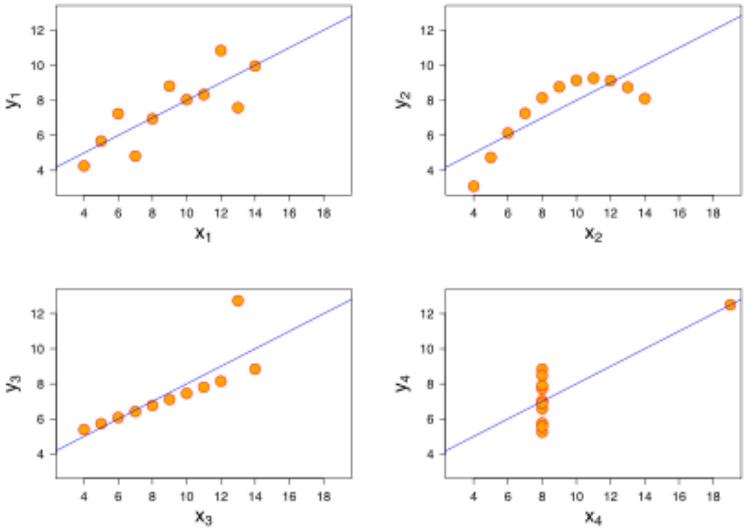


Image source: Wikipedia

- The 1st plot fits a linear regression line as its seems to have a linear relationship between x and y
- The 2nd plot doesn't show any linear relationship between x and y. This could mean they don't comply with basic linear aggression assumptions
- The 3rd plot shows some linear relationship but contains outliers which severely affects the best fit line
- The 4th plot indicates a very high correlation coefficient, which doesn't result in a significant model.

Hence, it is very important to visualize the data to understand its distribution while analysing the data.

----

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

The Pearson correlation coefficient (r) measures the strength and direction of the relationship between two variables. Its values range between –1 and 1.

Given by the formula

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

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

Image source: Wikipedia

Pearson’s R values	**Correlation**	**Example**
Between 0 and 1	Positive correlation	Human height and weight
0	No correlation	Rainfall in China and population in India
Between -1 and 0	Negative correlation	Price and Demand

----

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

Scaling: Scaling is a technique to standardize the independent features present in the data in a fixed range.

Why: When the numerical features of the dataset have very different ranges of values, it is important to bring them to a common scale for building a good model.

Benefits of scaling:

1. Ease of interpretation
2. Faster convergence of the optimization algorithm like Gradient Descent
3. Scaling just affects the coefficients and none of the parameters like t-statistic, f-statistic, p-value, R-squared etc.,

Normalization and Standardization:

Normalization also know min-max scaling is technique in which the numerical feature values are modified by using the formula:

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Image source: kdnuggest

Standardization also transforms the numerical feature values by using the formula

$$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation }(x)}$$

Image source: kdnuggest

Sl.NO	Normalization	Standardization
1	The range of values is between 0 and 1 or -1 and 1	The mean is centred at 0, but the range is not bound to a range
2	Normalization can be sensitive to outliers, since we are using min and max values for the transformation	Standardization is not affected by outliers, since we are using the min and max values in the transformation
3	Generally normalization is used when the distribution of the data in unknown	Standardization is generally used when the data is distributed normally

----

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

If VIF is infinite, then that indicates a perfect correlation between the two independent variables. In this scenario the R squared becomes 1 which then leads the value of VIF to infinity. To deal with this situation, one of the variables must be dropped, since it is causing a perfect multicollinearity. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

----

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

Q-Q Plots stands for (Quantile-Quantile plots). It is a graphical method for comparing two probability distributions by plotting their quantiles against each other. A quantile is a fraction where certain values fall below that quantile. A Q-Q plot can help us to identify if the 2 portions of the dataset have the same distribution or not .In general, a 45 degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, then the QQ plot can be more linear.

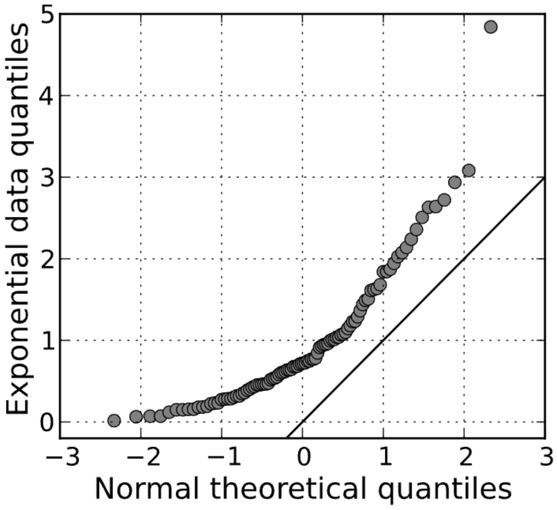


Image source: Wikipedia

Importance of QQ Plot in Linear Regression:

1. In Linear Regression the distribution of both the train set and the test set can be checked using a Q-Q plot. Ideally, both the sets should have a common distribution.
2. A Q–Q plot is used to compare properties such as location, scale, and skewness are similar or different in the two distributions.
