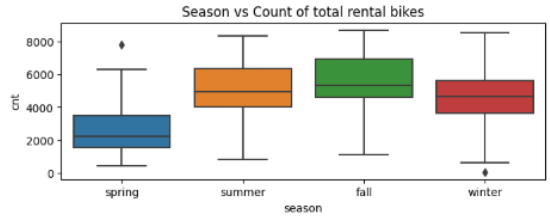
Joshy P J
ML C48 Batch
Upgrad – III-T Bangalore.

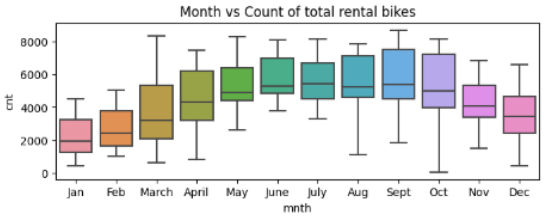
Assignment-based Subjective Questions

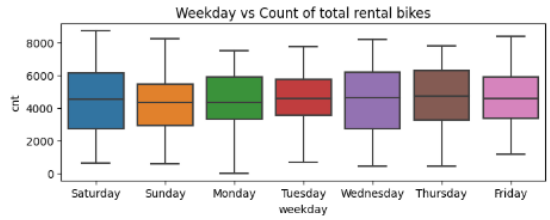
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

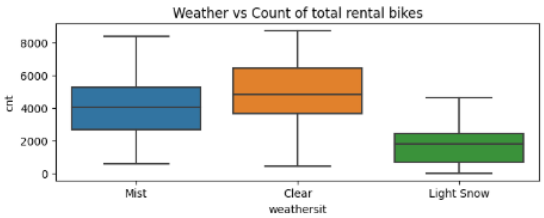
Ans:

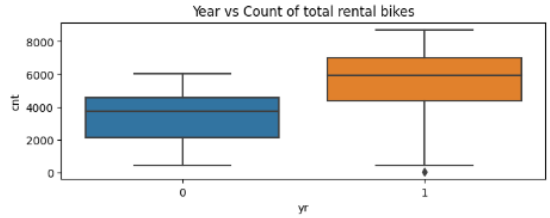
As part of the categorical analysis, independent categorical variables and plotted against dependent count variable to find the insight from these. Below screenshot shows the pictorial view of the same.

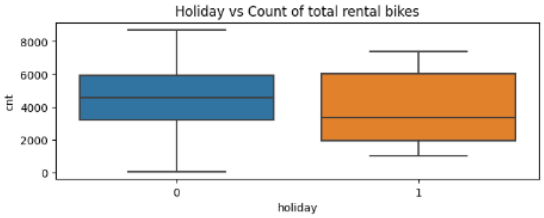


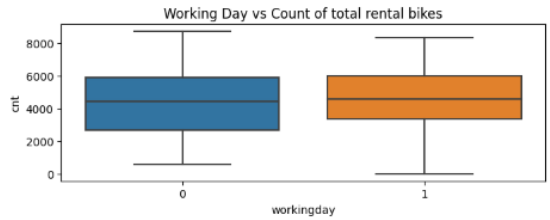












Observations:

- More bookings are done during the Fall season then summer, bookings are dropping from winter to spring. From this observation its very clear that opting bike is very much dependent up on the season, internally the climate temperature.
- Bike bookings slowly increasing from March, peaking at June to September and from Nov onwards its going down. Less number of bookings are in the month of Jan. So this is directly connected to the seasons variations. December/January are the winter season plus holiday , so very less bookings.
- More bookings are happening in the order of Saturday, Thursday and Wednesday, and slight drop-in other days in the week.
- When the weather is clear, more booking is reported. People prefer two-wheeler ride when the weather is clear.

- Year on Year bike bookings are increasing, means, slowly people are preferring to change.
- More bookings are reported on non- holiday and working day. So on holiday people prefer travel with family or be at home. May be during this time bachelors prefer two wheeler.

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

Ans:

Machine learning algorithms expect all the values be in numerical, so as part of the data preparation, categorical values have to be converted to numerical. Pandas provides methods to convert those to numerical, which is called dummy variable creation. This converts values to 0,1 combinations.
By default, this method creates 'n' columns to articulate 'n' values in the column. But if we use 'drop_first = True' is helps in reducing the extra column created while dummy variable creation. When everything is marked as '0', that represents last column. Hence it reduces the number of columns, increases the efficiency of the model, reduces the complexity, very easy to interpret , improves the model training time.

Syntax -
drop_first: bool, default False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level.

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

Ans:

By looking at the pair plot, temp showing slightly higher correlation with the target variable 'cnt'
Temp is the temperature in celsius and atemp is the feeling temperature.
Both atemp and temp showing almost same correlation with cnt and atemp and temp and highly correlated too.

```
print(bike_data.atemp.corr(bike_data.cnt))
print(bike_data.temp.corr(bike_data.cnt))
```

```
0.6306853489531039
0.6270440344135154
```

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

Ans:

Here are the assumptions of Linear regression

- Linear relationship between independent and target variable
    - Using the pair plot and correlation metrics, validated the linear relation ship with target and some of the independent variables
- Error Term should be normally distributed
    - Based on the resdisual (y_pred – y_test) plot , error terms are normally distributed with mean 0
- Error term should be independent of each other
    - No visible relation across the error terms
- Error variance should be constant or shouldn't follow any pattern
    - No visible pattern in residual patterns

- Along with above assumptions, also address below points
    - Multicollinearity
    - R Square
    - Adjusted R Square
    - Prob(F-Statistics)
    - P -Value

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

Ans :

Based on the coefficients of the features, below are the top features which positively or negatively affecting the bike booking

- temp with a coefficient of 0.547
    - Based on the feeling temperature, people will decide to opt for bike riding
- weathersit_Light Snow with a coefficient of - 0.2883
    - Based on the climate situation, people opt bike riding, if it's a snow fall or light snaw, people may not opt for bike riding
- yr with a coefficient of 0.2328
    - Year on year, the demand for bike increases

The best fitted line equation is

*Count of Bike Booking (cnt) = 0.1344 + 0.2328 * yr - 0.1067* holiday + 0.5471* temp - 0.1531* windspeed + 0.0878* season_summer + 0.1311seaon_winter + 0.0994 * mnth_Sept - 0.0498 weekday_Sunday - 0.2883 * weathersit_Light Snow - 0.0806 * weathersit_Mist*

General Subjective Questions

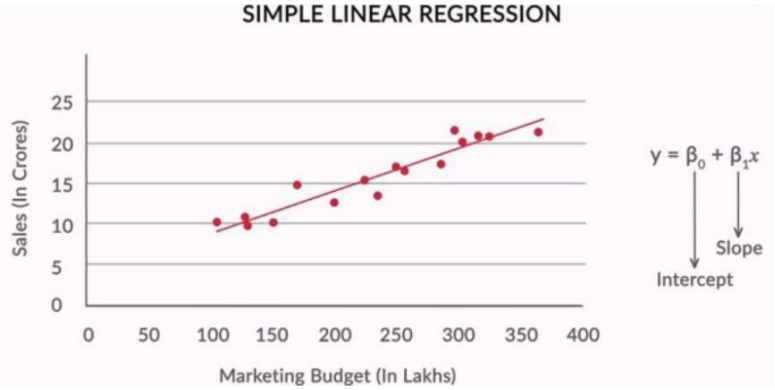
1. Explain the linear regression algorithm in detail. (4 marks)

Ans:

A linear regression model describes the relationship between a dependent variable, y, and one or more independent variables, X. The dependent variable is also called the response variable. Independent variables are also called explanatory or predictor variables. Continuous predictor variables are also called covariates, and categorical predictor variables are also called factors. The matrix X of observations on predictor variables is usually called the design matrix.

Linear relation can be positive or negative, in case of positive , if one value increases , other value also increases , in the case of negative , it behaves in the opposite way.

The standard equation of the regression line is given by the following expression: $Y = \beta_o + \beta_1 X$



Based on the number of predictor variables, Linear Regression can be of two types.

1. Simple Linear Regression
   simple linear regression which explains the relationship between a dependent variable and one independent variable using a straight line.
   The equation for simple linear regression model is $Y = \beta_o + \beta_1 X$

2. Multiple Linear Regression
   Multiple linear regression is a statistical technique to understand the relationship between one dependent variable and several independent variables (explanatory variables).
   The equation for simple linear regression model is $Y = \beta_o + \beta_1 X_1 + \beta 2 X_2 + \ldots \beta_n X_n$

Assumptions of Linear Regression

1. Linear relationship between predictor variable (x) and target variable (Y)
2. Error terms are normally distributed (not X, Y)
3. Error terms are independent of each other
4. Error terms have constant variance (homoscedasticity)

The linear regression model can be measure in terms of the R^2 value. R-Squared (R^2 or the coefficient of determination) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable.

$R^2 = 1 – RSS/TSS$

Where TSS = Sum of errors of the data from the mean , RSS - Residual sum of squares.

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

Ans :

Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

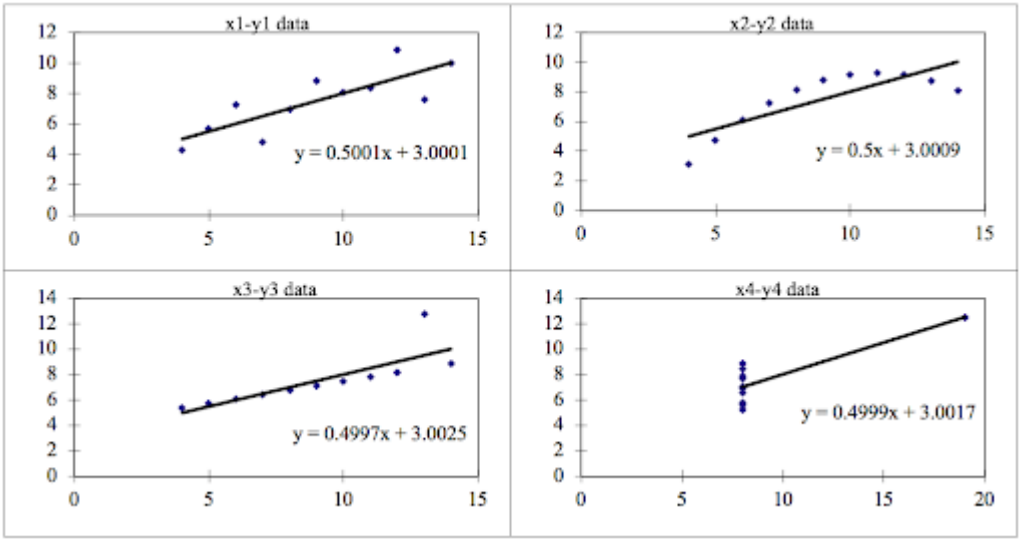
We can define these four plots as follows:

Anscombe's Data									
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89

The statistical information for these four data sets are approximately similar. We can compute them as follows

Anscombe's Data									
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

However, when these models are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm, as you can see below:



ANSCOMBE'S QUARTET FOUR DATASETS

Data Set 1: fits the linear regression model pretty well.

Data Set 2: cannot fit the linear regression model because the data is non-linear.

Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model.

Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

As you can see, Anscombe's quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

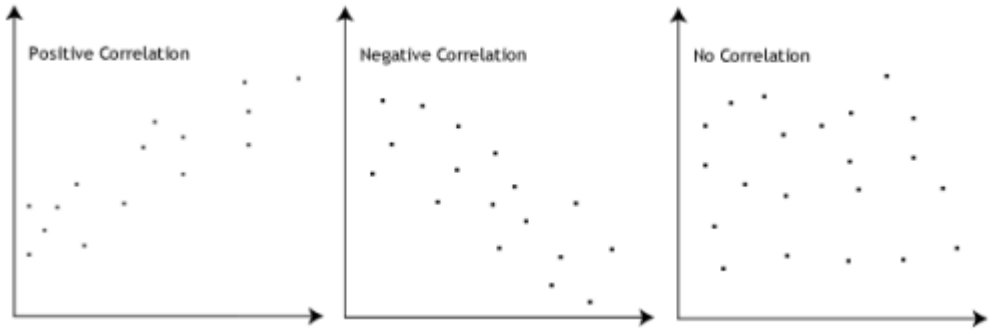
3. What is Pearson's R? (3 marks)

Ans :

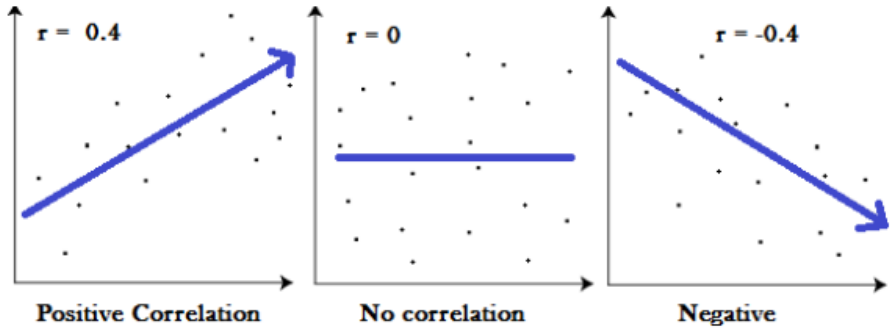
A correlation or simple linear regression analysis can determine if two numeric variables are significantly linearly related. A correlation analysis provides information on the strength and direction of the linear relationship between two variables, while a simple linear regression analysis

estimates parameters in a linear equation that can be used to predict values of one variable based on the other.
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.
Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, r, indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit).

The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.



Pearson correlation coefficient formula.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\,n\Sigma x^2 - (\Sigma x)^2\,]\,[\,n\Sigma y^2 - (\Sigma y)^2\,]}}$$

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

ans :
Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data pre-processing step.

Methods for Scaling
Now, since you have an idea of what is feature scaling. Let us explore what methods are available for doing feature scaling. Of all the methods available, the most common ones are:

Normalization
Also known as min-max scaling or min-max normalization, it is the simplest method and consists of rescaling the range of features to scale the range in [0, 1]. The general formula for normalization is given as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

max(x) and min(x) are the maximum and the minimum values of the feature respectively.
Normalization rescales the values into a range of [0,1]. also called min-max scaled.

Standardization
Feature standardization makes the values of each feature in the data have zero mean and unit variance. The general method of calculation is to determine the distribution mean and standard deviation for each feature and calculate the new data point by the following formula:

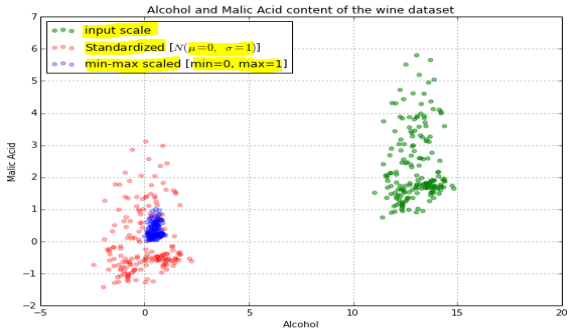
$$x' = \frac{x - \bar{x}}{\sigma}$$

Here, σ is the standard deviation of the feature vector, and $\bar{x}$ is the average of the feature vector.

Standardization rescales data to have a mean (μ) of 0 and standard deviation (σ) of 1.So it gives a normal graph.

Normalization vs Standardization
standardization may be especially crucial to compare similarities between features based on certain distance measures. Another prominent example is the Principal Component Analysis, where we usually prefer standardization over Min-Max scaling, since we are interested in the components that maximize the variance (depending on the question and if the PCA computes the components via the correlation matrix instead of the covariance matrix.

One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.
Normalization is good to use when the distribution of data does not follow a Gaussian distribution. It can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbours.
Standardization can be helpful in cases where the data follows a Gaussian distribution. Though this does not have to be necessarily true. Since standardization does not have a bounding range, so, even if there are outliers in the data, they will not be affected by standardization



S.NO.	Normalized scaling	Standardized scaling
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

Scaling of same data using Normalization vs Standardization

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

Ans :
A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model. This can adversely affect the regression results. Thus, the variance inflation factor can estimate how much the variance of a regression coefficient is inflated due to multicollinearity
The formula for VIF is
$VIF_i = 1/1-R_i^2$ R^2 is coefficient of the determination.

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 R^2 =1, which lead to $1/(1-R^2)$ 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).

Its advisable to remove any predictor variable whose VIF is greater than 5 to address the multicollinearity.

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

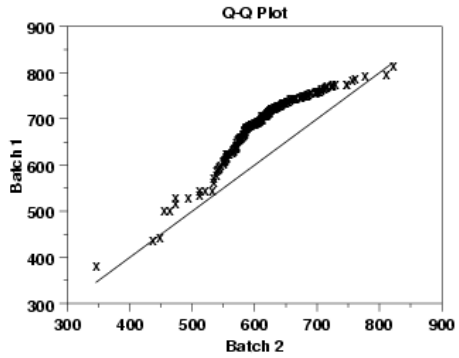
Ans :

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

Use of Q-Q plot:
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence

for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q plot:
When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.



Observation -
These 2 batches do not appear to have come from populations with a common distribution.
The batch 1 values are significantly higher than the corresponding batch 2 values.
The differences are increasing from values 525 to 625. Then the values for the 2 batches get closer again.
The q-q plot is formed by:
Vertical axis: Estimated quantiles from data set 1
Horizontal axis: Estimated quantiles from data set 2