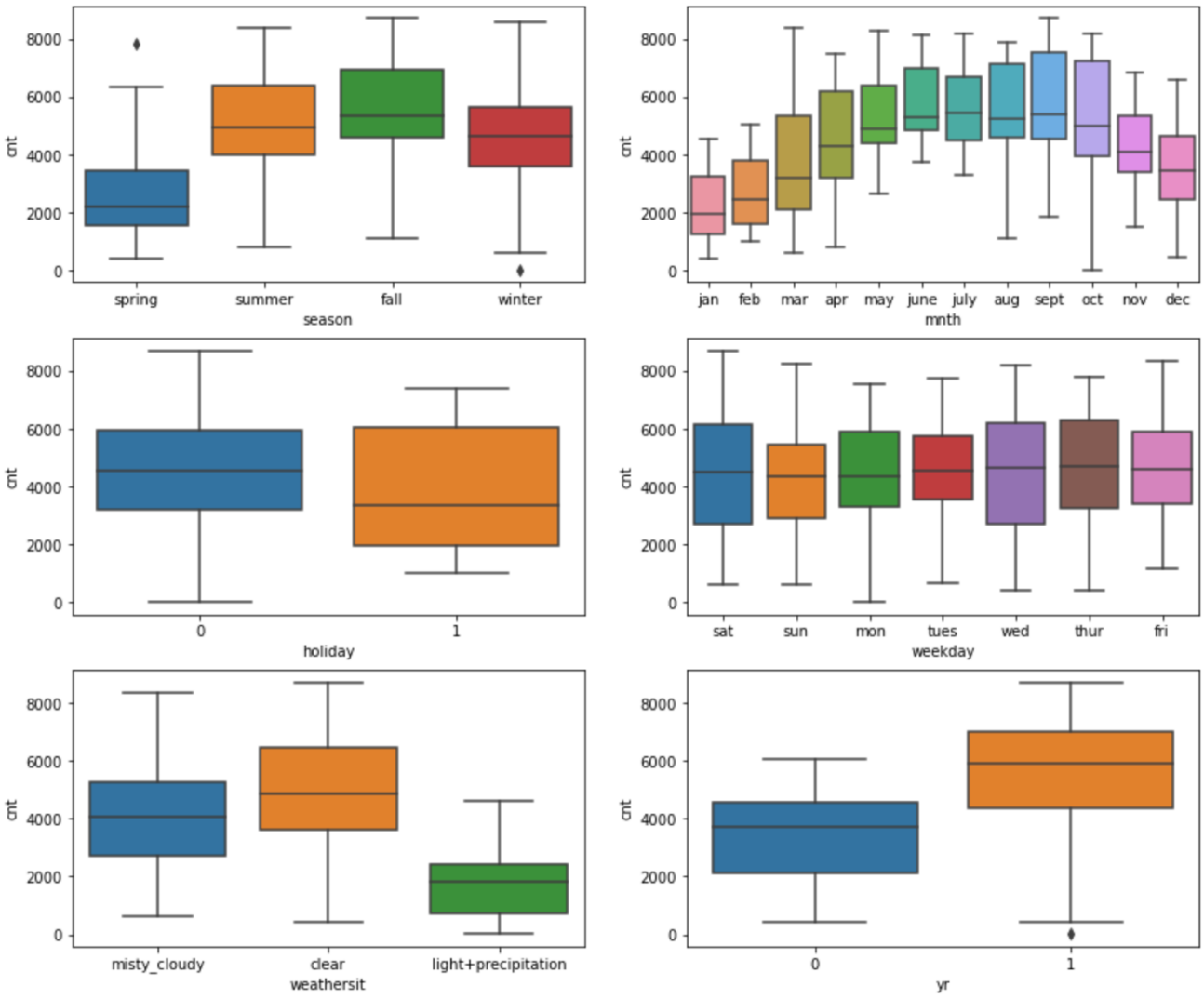
Assignment based

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

Answer:



1. Season: We can see that the bike sharing is more in the month of summer and fall whereas has dropped considerably in the month of spring
2. Month: The month of June and July has the highest number of people sharing a bike, whereas for the month of Jan it is the least
3. Holiday: On holidays the share count is less probably people are sharing bike for offices or school
4. Weekday: Day of the week does not have any impact on the count
5. Weather: If it's raining, people are not preferring bike because of the fear of getting drenched in the rain
6. Year: 2019 has considerably more number of share indicating an increase in trend

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

Answer:

When we create a dummy variable without passing drop_first=True, then it'll create the same number of dummy variables as the levels of that feature. Suppose a feature has n levels then it'll create n number of variables. But for a feature of n level the information can be conveyed with n – 1 variable. Hence by passing drop_first=True it creates only n – 1 variables

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

Answer:

Temperature and feeling temperature are the two that has the highest correlation with the target variable. Temperature and feeling temperature has the highest correlation among themselves so we can remove either of these

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

Answer:

1. Linear relationship between X and y. By plotting a pairplot we can see that the relationship between independent and dependent variables are linear. For eg relationship between temp and cnt
2. Normal distribution of errors. We plotted a histogram of residuals and found out that the residuals are normally distributed
3. Mean of residuals is zero. On calculating the mean of residuals it came out to be zero. Also it is evident from the histogram
4. Residuals are homoscedastic. We plotted a residual plot and found out that there are no visible pattern being seen. Hence we can conclude that the residuals are homoscedastic
5. No multicollinearity between variables. We calculated the VIFs and it was well below 3 for all the variables

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

Answer:

The final regression model looks like

F(cnt) = 0.2353 * yr - 0.0969 * holiday + 0.6612 * temp + 0.0224 * spring + 0.0893 * summer + 0.1564 * winter - 0.2901 * light+precipitation - 0.0710 * misty_cloudy - 0.0415 * july + 0.0999 * sept

Here we can see that the variable `temp`, `yr` and `winter` season are contributing the most in increasing order positively

General:

1. Explain the linear regression algorithm in detail.

Answer:

Linear regression is commonly used for predictive analysis. It is a supervised learning method. In this regression model the target variable value is predicted based on independent variables. The target variable is expected to be a continuous one. The regression examines mainly two things 1. Does a set of predictor variable does a good job in predicting the outcome variable 2. Which variable are significant in predicting the outcome variable and in what way they influence magnitude and sign. The simplest form of linear regression with one independent and one dependent variable is represented by

$$y = mx + c$$

where m: slope, c: intercept, x: independent variable, y: dependent variable

When there is only one independent variable it is called Simple Linear Regression, whereas when there is more than one independent variables it is called Multiple Linear Regression.

There are few assumptions for Linear Regression

a. Absence of multicollinearity
b. Homoscedasticity of residual
c. Normal distribution of residuals
d. Linear relationship between dependent and independent variable
e. Mean of residual should be 0

Example of Linear regression:

a. Relationship between advertisement and sales
b. Revenue forecast based on past performance

2. Explain the Anscombe's quartet in detail.

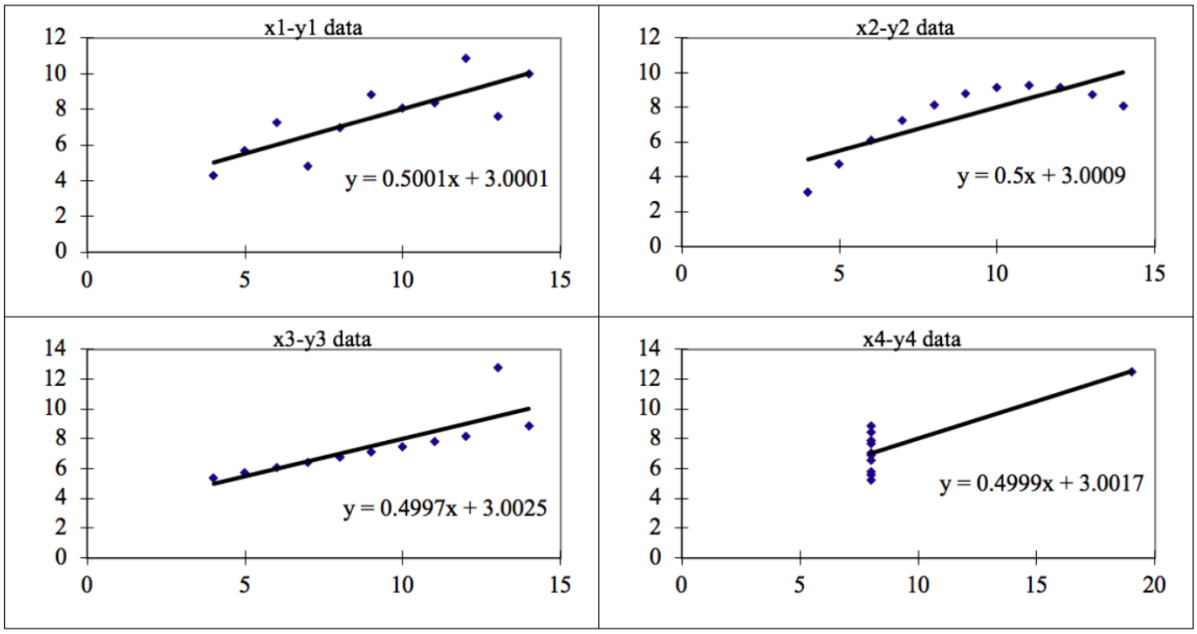
Answer:

Anscombe's quartet consist of four dataset that have almost identical descriptive statistical analysis example, sample size, mean, standard deviation, etc, yet have very different distribution and appearance when plotted on a graph. This shows a significance of plotting a graph before doing analysis and modelling. Plotting also helps in discovering anomalies in data such as outliers, etc which is not given by statistic

For eg: a dataset with its stats information looks like this

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	

When this is plotted on graph it looks like this



Reference: https://towardsdatascience.com/importance-of-data-visualization-anscombes-quartet-way-a325148b9fd2

This shows how the distribution of data varies a lot even when stats are same. This gives us the significance of plotting a distribution of data

3. What is Pearson's R?

Answer:

Correlation is a statistic measure that signifies the relationship between two variables. The value ranges between -1 to +1. A value +1 shows a perfect positive correlation between two variable, meaning if one variable is increased by positive 1 unit then the correlated variable will also increase by positive one unit and vice versa. The value -1 shows a perfect negative correlation, meaning if one variable is increase by positive 1 unit then the correlated variable will decrease by negative one unit and vice versa. The correlation with 0 signifies that there is no relationship between two variable, meaning increase or decrease in one variable will not have any effect on the other variable

Pearson's coefficient correlation is also called as Pearson's R, it also measure the relationship between two variable and its value ranges from -1 to +1. It cannot capture the nonlinear relationship between two variable

It is given as the covariance of two variable divide by the product of the standard deviation

Formula:

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

Where:

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

Σx2 = the sum of squared x scores

Σy2 = the sum of squared y scores

Reference: https://www.questionpro.com/blog/pearson-correlation-coefficient/

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

Answer:

Scaling is a data pre-processing technique where the continuous variables are normalize between certain range. This processing of scaling helps in accelerating the algorithm's performance. Scaling of variables help the optimization algorithm reaches the global minima at a faster rate. Scaling also helps the quantity to be unit agnostic.

In normalization scaling is done to bring the data between the range of 0 and 1. MinMaxScalar as a part of sklearn library can be used to implement normalization

$$\boldsymbol{MinMaxScaling}(\boldsymbol{x}) = \frac{\boldsymbol{x} - \mathbf{min}(\boldsymbol{x})}{\mathbf{max}(\boldsymbol{x}) - \mathbf{min}\ (\boldsymbol{x})}$$

In case of Standardization the data is converted to standard normal distribution having mean as 0 and standard deviation of 1. Standardization is useful when the data has outliers

$$\boldsymbol{Standardization}(\boldsymbol{x}) = \frac{\boldsymbol{x} - \boldsymbol{mean}(\boldsymbol{x})}{\boldsymbol{sd}(\boldsymbol{x})}$$

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

Answer:

A VIF of an independent variable is given as the relationship between this independent variable and other independents variables.

$$\boldsymbol{VIF} = \frac{\mathbf{1}}{\mathbf{1} - \boldsymbol{R}^{\mathbf{2}}}$$

Now when a given independent variable can be completely described by other independent variables, then in that case R2 of that variable will be 1, so the VIF would become

$$\boldsymbol{VIF} = \frac{\mathbf{1}}{\mathbf{1} - \mathbf{1}^{\mathbf{2}}} = \frac{\mathbf{1}}{\mathbf{0}} = \boldsymbol{Infinite}$$

Hence the VIF would become Infinite when a particular variable can be described completely by other independent variables

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

Answer:

A Quantile-Quantile plot is a probability plot, it is a graphical method for plotting two probability distributions by plotting quantiles against each other. It is used to determine if two data set comes from population with a common distribution or not. In this Q-Q plot a 45 degree reference line is plotted and if the two dataset comes from population with same distribution then the points should fall approximately on the line.

In linear regression we can use Q-Q plot to determine that the train and test data that we received separately are from the population with same distribution or not