CERTIFICATE COURSE IN MACHINE LEARNING AND NLP

ASSIGNMENT: BIKE RENTAL CASE STUDY

ANSWERS TO SUBJECTIVE QUESTIONS BY SIDDHARTHA GHOSH

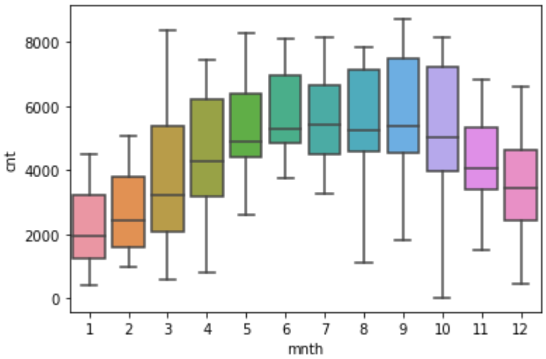
DECEMBER 20, 2022

ASSIGNMENT-BASED QUESTIONS

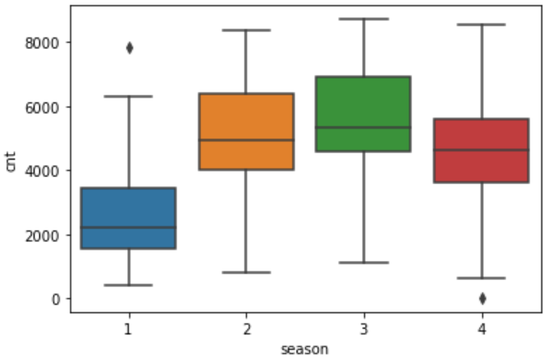
From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Here are my inferences based on the Box Plots:

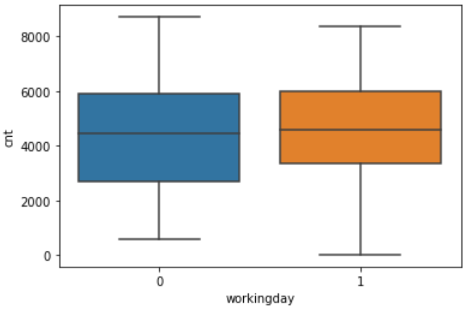
1. Demand seems cyclic. It goes up during the months of better weather (March to September) and comes down thereafter, coinciding with the months of Rain, Snow / Sleet.



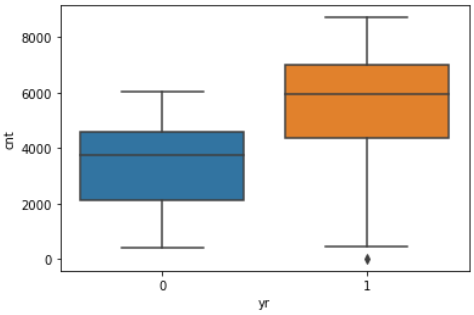
2. Season does seem to play a role, and obviously month (notice the similarity of the trajectories of the Box Plots)



3. Working and Non-Working Days seem to have very similar means in the Total Count of Bookings, and both types of days seem to be important for business. There seems to be greater variation in rental numbers on Non-Working Days. But Working and Non-Working Days seem to have very similar means.



4. Business has been going up, year-on-year (2018-19).



Why is it important to use drop_first=True during dummy variable creation?
drop_first=True helps in reducing the extra column created during dummy variable creation, and in the process, reduces the correlations created among dummy variables. When we have a categorical variable with K mutually exclusive categories, we actually only need K – 1 new dummy variables to encode the same information. This is because if all of the existing dummy variables equal 0, then we know that the value should be 1 for the remaining dummy variable. For example, if region_North == 0, and region_South == 0, and region_West == 0, then region_East must equal 1. This is implied by the existing 3 dummy variables, so we don't need the 4th. The extra dummy variable literally contains redundant information. Hence, it is a common convention to drop, for the sake of efficiency, the dummy variable for the first level of the categorical variable that we are encoding.

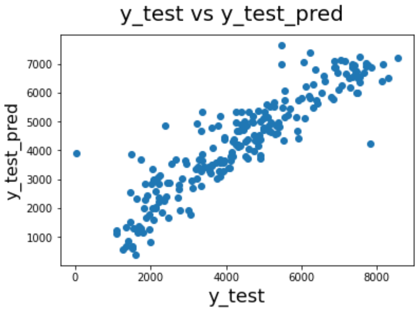
Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable
Temperature / Average Temperature with a Correlation Coefficient of 0.63 with the Target Variable, 'cnt'.

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

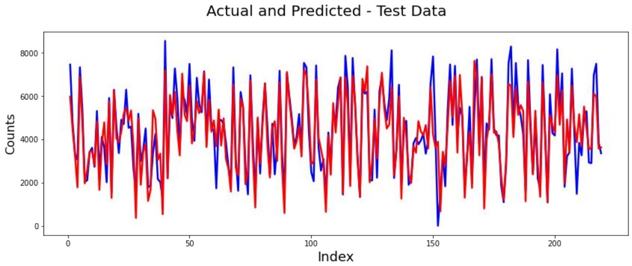
The assumption of linearity has been tested out using:

- The Adjusted R^2 value of 0.815 (which meant that the variables chosen could explain 81.5% of the result.

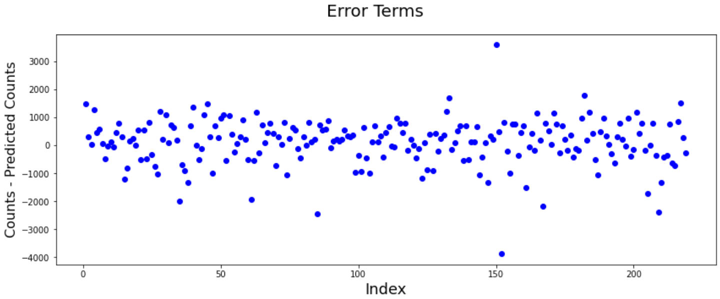
- Checking for (the absence of) Heteroskedasticity using a Scatter Plot.



- Visually checking for the proximity of Actual and Test Data



- Visually checking the Error terms of the Test Data



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

1. Temperature (an obvious positive influence), as mentioned earlier on. After "constant", has the highest coefficient value. And a t-score of 13.761.
2. The Year. At least for this data set. It is more of an indicator of the good service during the inaugural year (2018) that has led to a higher rental count during the second year (2019). The high t-score of 27.002 bears this out.
3. Weathersit (The Weather Situation). The worse it is, the less are the rentals (notice the difference in negative coefficients between Moderate and Bad Weather). And the negative t-score of -10.360.

GENERAL QUESTIONS

Explain the linear regression algorithm in detail

In its most generic form, a linear regression algorithm is represented in the following way:

Y(cap) = a + b_1X_1 + b_2X_2+ ……. b_nX_n + c where
Y(cap) = estimated value corresponding to the dependent variables
a= Y- intercept
X_1, X_2,… X_n = values of the independent variables
b_1, b_2… b_n,, = slopes associated with X_1, X_2,… X_n respectively
c = constant

In Simple regression, the estimating Equation Y(cap) = a + b_1X_1 + b_2X_2 + c describes the relationship between 2 independent variables, X_1 and X_2, with the dependent variable, Y.

For example, what would be the expected production cost of a table (dependent variable), depending on the unit costs (X_1 and X_2) and quantities (b_1, b_2) needed of the two main inputs, wood and labour.

It basically measures, how much would Y be expected to change for every unit of X_1, X_2, b_1, b_2
If $b_1X_1 + b_2X_2 = 0$, then the intercept a represents the expected value of Y without the intervention of any of the independent variables X_1 and X_2.

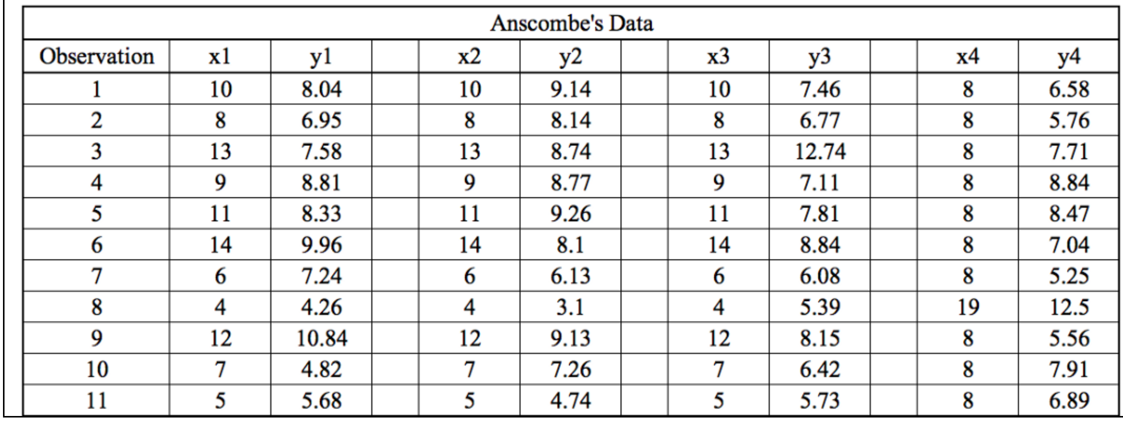
A Simple Regression Equation is represented by a line on a graph.

In case of Multiple Regression, the number of independent variables exceeds 2 (hence the equation extends to Xn). In our example, we could add the unit cost and amount of varnish, nails and screws. The equation of a Multiple Regression is represented as a plane.

Explain the Anscombe's quartet in detail

Anscombe's Quartet is a group of four data sets which are nearly identical in simple descriptive statistics (that involves variance, and mean of all x,y points in all four datasets) but have peculiarities in the dataset. They have different distributions and appear differently when plotted on scatter plots.

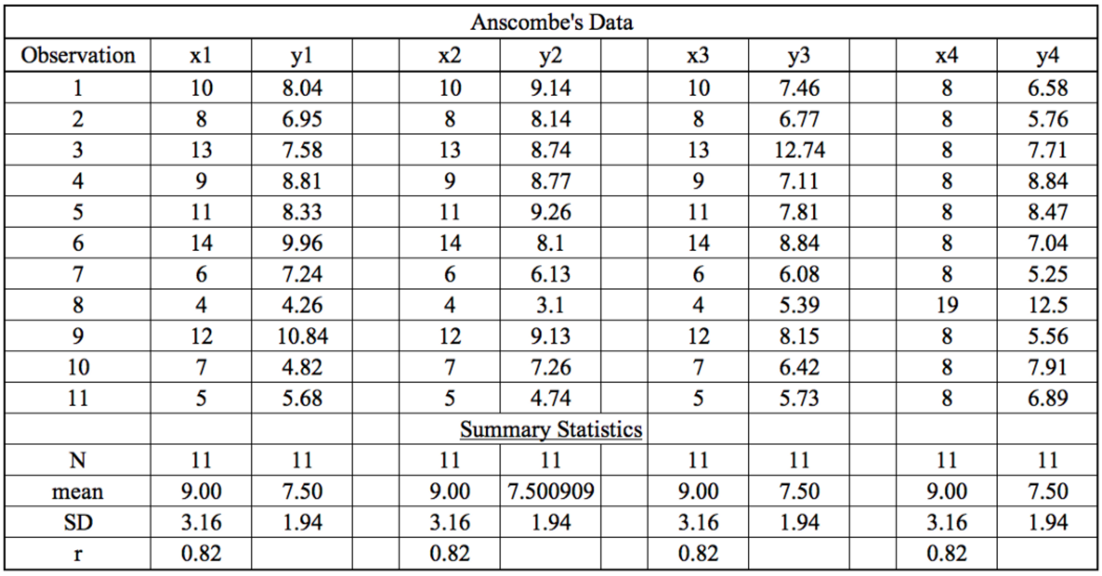
For example, the following 4 data sets:



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

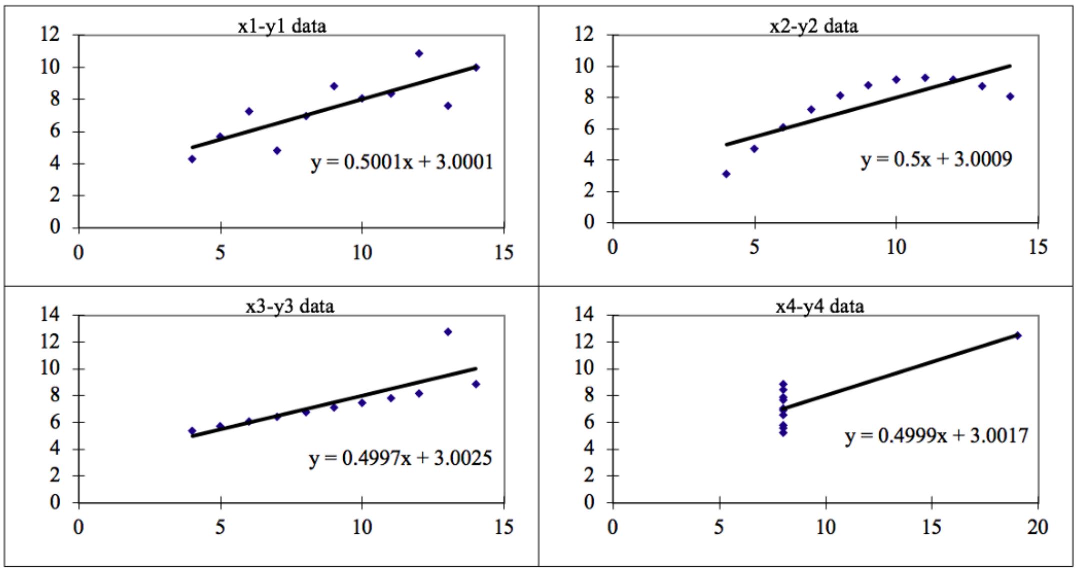
They have very similar statistical information (Mean, St Dev and r):



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	

And yet throw up very different Scatter Diagrams that may not be correctly interpreted by the resulting regression line (please especially refer to the line representing the x4, y4 data):



They were constructed in 1973 statistician Francoise Anscombe, to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough." (Wikipedia)

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

What is Pearson's r?

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation between two numeric variables, say x and y. Correlation refers to the degree of association between variables.

For example, let (x_1,y_1), (x_2,y_2), (x_3,y_3).... (x_n,y_n) be n pairs of observations on two variables, x and y. The Correlation Coefficient between x and y, denoted by the symbol r is:

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

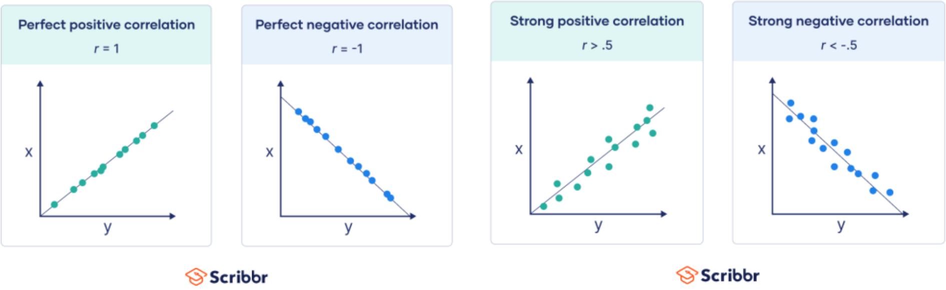
(Image sourced from the Internet)

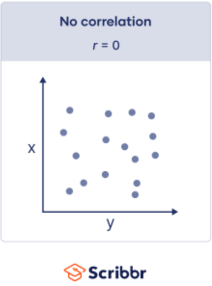
The Correlation Coefficient lies between –1 and 1 that measures the strength and direction of the relationship between the two variables.

In laymen's terms, a positive correlation means that the value of y would increase with the value of x, while a negative correlation means that the value of y would decrease with the value of x

Usually, a correlation coefficient of greater than +0.5 is taken as an indicator of a reasonably strong positive correlation between two variables. And a correlation coefficient of less than -0.5 is taken as an indicator of a positive correlation between two variables. Two values may also be zero-correlated.

Positive, Negative and Zero correlations are represented via the following scatter diagrams (sourced from the Internet):





The Pearson correlation coefficient works best when all of the following conditions are true:

- Both variables are quantitative.
- The variables are normally distributed

- The data have no outliers
- The relationship is linear, ie, the relationship between the two variables can be described reasonably well by a straight line.

However, one must also be careful about the following and exercise judgment:

- A high value of r does not automatically imply a causal relationship between variables. For example, high ice cream sales and high incidence of drowning in swimming pools. The high correlation could be coincidental with the summer season.
- And as we had seen while discussing Anscombe's Quartet, similar r values could have different distributions and appear differently when plotted on scatter plots. Therefore it is important to check out scatter plots before arriving at conclusions !

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

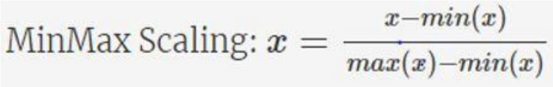
Scaling is a process applied while pre-processing independent variables to normalize the data within a particular range.

Often, collected data sets contain features that vary highly in magnitudes, units and range. If scaling is not performed, the resulting algorithm only takes magnitude in account and not units of measurement, leading to incorrect modelling. Scaling addresses this issue, bringing all the variables to the same level of magnitude.

Scaling is required because some models are sensitive to the order of magnitude of the features. If a feature has an order of magnitude equal to 1000, for example, and another feature has an order of magnitude equal to 10, some models may "think" that the first feature is more important than the second one. It is obviously a bias, because the order of magnitude does not provide any information about the predictive power. Hence, the need to remove this bias by transforming the variables to give them the same order of magnitude.

Of the two types of Scaling, Normalised and Standardised:

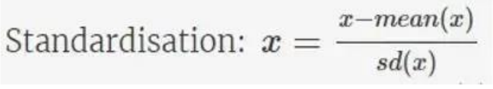
Normalised Scaling brings all of the data in the range of 0 and 1. It is calculated in the following manner:



$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Lasso, Ridge and Elastic Net regressions are powerful models that require normalization. Similarly, neural networks are very sensitive to the order of magnitude of the features. The activation functions always require normalized data

On the other hand, **Standardization** replaces the values by their Z scores. It brings all the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ). It not only helps with scaling but also centralizes the data.



$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

In general, standardization is considered to be more suitable than normalization in most cases. One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables.

If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. Thus, the standard error of this coefficient is inflated by a factor of 2. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity.

If there is perfect correlation, then VIF takes on the value of infinity. In any case, in case of large VIF scores and hence multi-collinearity, corrective action would be necessary.

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

A Q-Q Plot (Quantile-Quantile Plot) is a probability plot, for comparing two probability distributions by plotting their quantiles against each other.
It is a graphical technique for determining if two data sets come from populations with a common distribution.

In simple terms, a QQ Plot tries to answer:

- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behaviour?

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the identity 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.

Why is the QQ Plot important?

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.