LINEAR REGRESSION – SUBJECTIVE QUESTIONS

ASSIGNMENT BASED QUESTIONS

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

   Answer: From the analysis, I was able to assess that:-

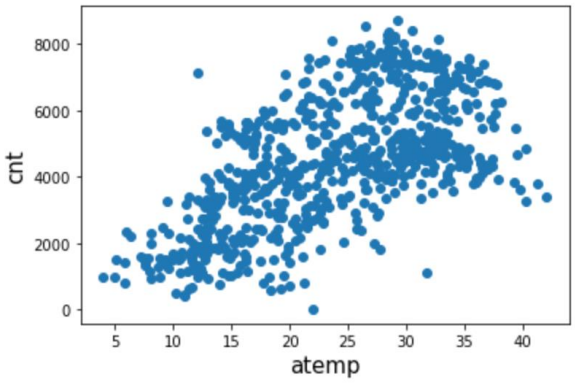
   a. **Year** 2019 performed far better than 2018 in terms of bike rentals (cnt)
   b. Analyses on the **Months** & **Seasons** variables indicated that as the climate starts getting colder, the demand starts falling before starting to revive again in Summer.
   c. Demand is noticeably lesser on **holidays.**
   d. It was observed that the day of week being **working day** or not made little difference to demand.
   e. There was no significant difference observed in demand during different **days of week.**
   f. **Weather** had clear effect on the dependent variable. As the weather got worse, so did the rental demand.

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

   Answer: The idea of dummy variable creation is to create *n-1* variables for a categorical variable with n levels. We require only *n-1* variables as the they are able to explain the last remaining variable which becomes redundant. Therefore, we use drop_first = True to remove the first of the n dummy variables that are initially created so that we are left with *n-1* variables.

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

   Answer: By looking at the pairplots, it was visually obvious that the numeric variable with the highest correlation with the dependent variable was **atemp** (Reference Image Below)

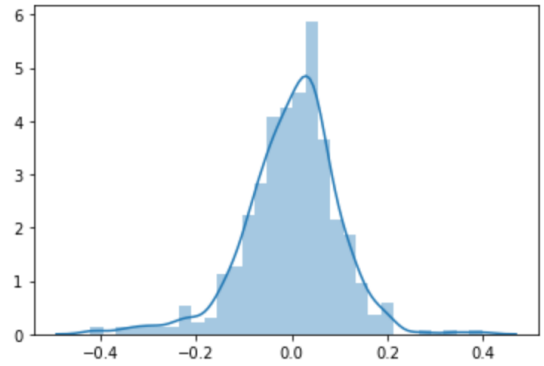


4. How did you validate the assumptions of Linear Regression after building the model on the training set?

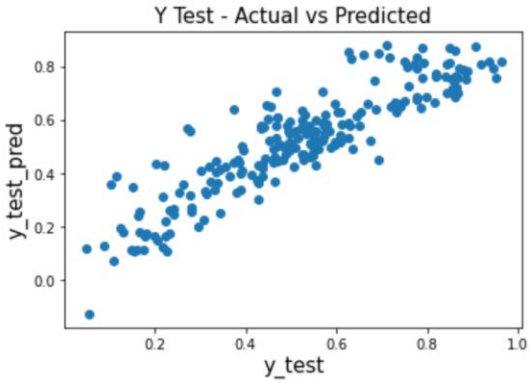
   Answer: The following steps were taken to validate the model:-

   a. Residual Analysis: Checked if the error terms were normally distributed.

```
y_train_cnt = lr_model.predict(X_train_rfe)
y_residual = y_train - y_train_cnt
sns.distplot(y_residual)
plt.show()
```



   b. Y_test vs y_test_pred: Checked if predicted vs actual y test followed a linear relation.



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

   Answer: The following 3 features were found to contribute the most to bike demand:-

   a. **atemp** is a highly significant feature in predicting demand for rentals with a positive coefficient of **0.6076**.
   b. **Weather3** is the next most significant feature with a negative coefficient of **0.2855**. Here, Weather3 refers to category = 3 defined as: *Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds*
   c. **Yr** (or year) is the 3rd most significant feature with a positive coefficient of **0.2331**.

GENERAL SUBJECTIVE QUESTIONS

1. Explain the linear regression algorithm in detail.

   Answer: Linear Regression (LR) is a machine learning algorithm that is used to predict a target value based on one or more independent variables. The algorithm is suited for predictions where the variables involved have a relationship that tends to be linear. The algorithm also makes use of Supervised Learning to learn patterns prior to making predictions. There are 2 manners in which the Linear Regression algorithm is used, namely:-

   a. Simple Linear Regression (SLR): A single independent variable is used to predict the dependent variable.
   b. Multiple Linear Regression (MLR): Multiple variables are used to predict the dependent variable.

   The resultant LR model attempts to draw a regression line or a best fit line that is able to predict the dependent variable based on inputs from independent variables. The function for this line is defined as:-

   a. For SLR: $Y= B0 + B1*X$
   b. For MLR: $Y = B0 + B1*X1 + B2*X2 + B3*X3 +...BN*XN$

   Where B1,B2...BN explain the influence of their respective X variables in predicting the dependent variable, Y. B0 is a constant that may or may not have real world interpretation.

2. Explain the Anscombe's quartet in detail.

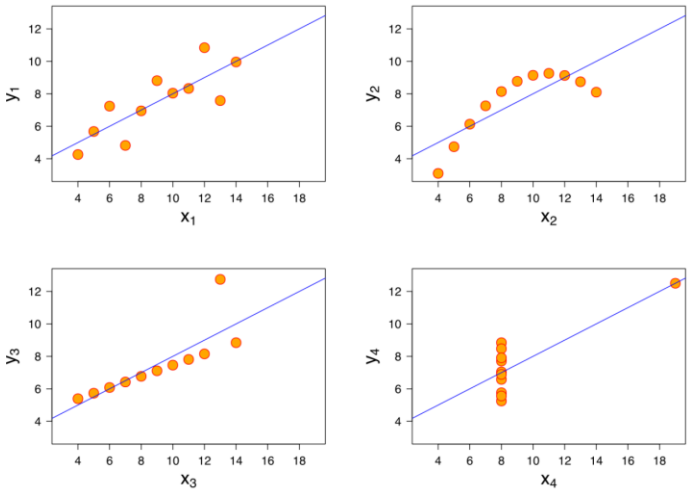
Answer: Anscombe's quartet is a set of four datasets that have nearly identical statistical properties but appear quite different when the datapoints are plotted on graphs. Each graph tells a different story even though they have similar summary statisitcs. Each dataset consists of 11 (x,y) datapoints.

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

Quartet's Summary Stats

Taking an example of 4 such datasets as described in the above image, it can be seen that all of them have the same summary stats, specifically – Sum, Average, Standard Deviation/Variance. Further analysis would also indicate that the correlation between X & Y is also quite similar.

When plotting the same data, we see that the regression lines are the same but the data in each graph tells a different story.



The takeaway from the quartet is the importance of visualizing data and to demonstrate that statistical properties only go so far in helping us understanding the nature of the data.

3. What is Pearson's R?

   Answer: Pearson's R, also known as Pearson's correlation coefficient, is a statistic that is used to measure correlation between 2 different variables. Values range from –1 to +1. Values in the negative indicate a negative correlation and positive values show a positive correlation. 0 would indicate that there is no correlation between the 2 variables.

   In practice, a best fit line is attempted to be drawn through the data of the 2 variables and Pearson's r or the correlation coefficient measures how far away the data points are from these lines.

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

   Answer: Scaling is the process of bringing multiple variables with varying units of measurement to a single comparable scale of measurement.

   Scaling is required particularly when the variables are used in building a regression model and the influence of predictor variables needs to be interpreted. The unit of measurement will then need to be same across in order to understand the impact of a variable as well as to compare impact between different variables.

   Normalized Scaling: The data is compressed between a range of 0 and 1. Any outliers will also fall within this range as the data is bounded. It is preferred for variables that do not follow a normal distribution.

   Standardized scaling: A scaling technique that centers the values of the variable around its mean with a standard deviation = 1. The centered mean will be equal to 0 and the values are measured by SDs away from mean = 0. The difference is that the variable's values are not bound within a certain range despite scaling. It is the preferred technique for variables with values that are normally distributed.

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

   Answer: An infinite VIF for a particular variable indicates that there is a perfect correlation with all other independent variables in consideration. This implies that it is a severe case of multicollinearity that requires corrective action.

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

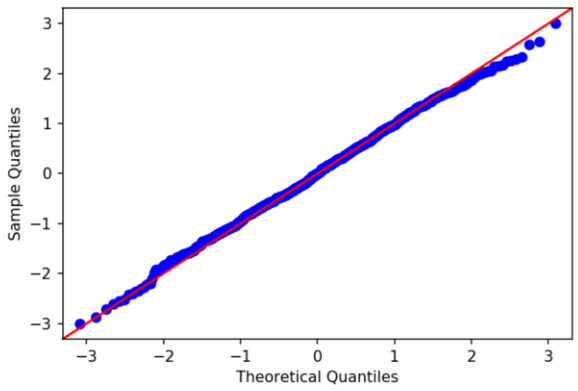
Answer: The q-q plot or the quantile-quantile plot is a graphical technique used to determine if:-

a. A dataset matches a certain type of distribution that we are interested in – normal, uniform, exponential etc.
b. two data sets come from populations with a common distribution.

A 45 degree line is plotted as a reference line. If data comes from the distribution of our interest (normal,uniform, exponential etc.) the data points should approximately fall on this reference line.

In the case of two data sets, if they come from populations with a common distribution, their respective data points will fall on the reference line.

<u>Example of Q-Q Plot where the distribution matches:-</u>



QQ plot of a normally distributed random variable

<u>Example of Q-Q Plot where distribution does not match:-</u>