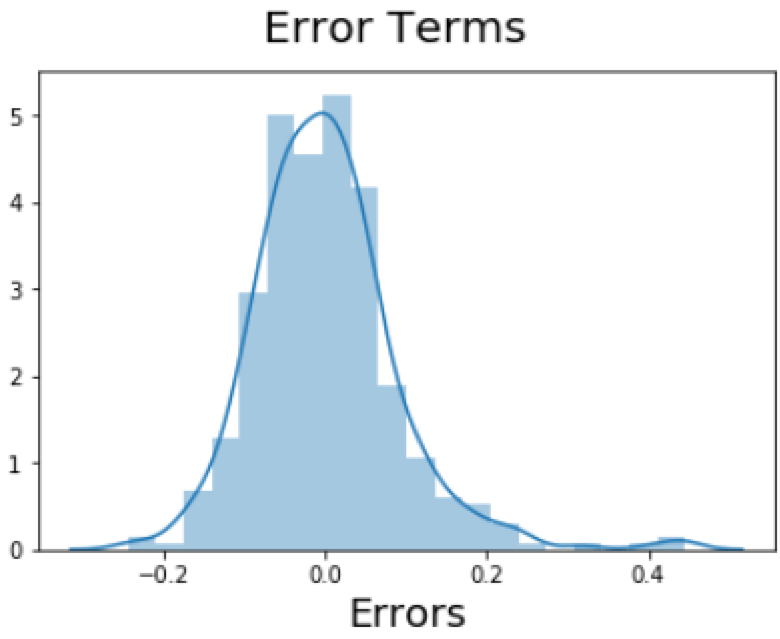
Assignment-based

1. Subjective Questions 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
    - The categorical variables I have considered are :
        - season: The median of the count is highest in fall and lowest during the springs.
        - yr: The median count in 2019 is higher than the count in 2018.
        - mnth: Although, I have dropped this column cause it represents something which is already clear from the season column. The count in months of the fal(Aug, Sep, oct) is higher.
        - holidays: count mean is less during holidays.
        - Weekdays: Very less impact on the count cause the means are close although distributions are different.
        - workingday: Again not much of impact but the count mean is higher for the workingdays than non-working.
        - weathersit: It's significantly less for the Harsh weather and highest for Good weather.
2. Why is it important to use drop_first=True during dummy variable creation?
    - Well the basic need of this command is to drop an extra column while making dummy columns. For example: if we are making the dummy columns for the 4 different seasons the number of columns/variables actually required is n-1 i.e 3 The three columns/variables are sufficient to represent the 4 values: Spring-(0,0,1),Summer-(0,1,0),fall-(1,0,0) and winter(0,0,0)

3. 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
    - temp and atemp seems to have almost equal correlation with the count.
4. How did you validate the assumptions of Linear Regression after building the model on the training set?
    - To validate the assumptions of the Linear Regression we check the distribution of the Error terms, this is done only for the training data set.
    - The Error is defined as E=f(y-train,y-train-pred), that is the actual value versus the value predicted by the module and the distribution of all the erros for different values of independent variable and he target variables should be a normal curve centred around zero.



- As we can see the mean is at 0 and the distribution is normal.

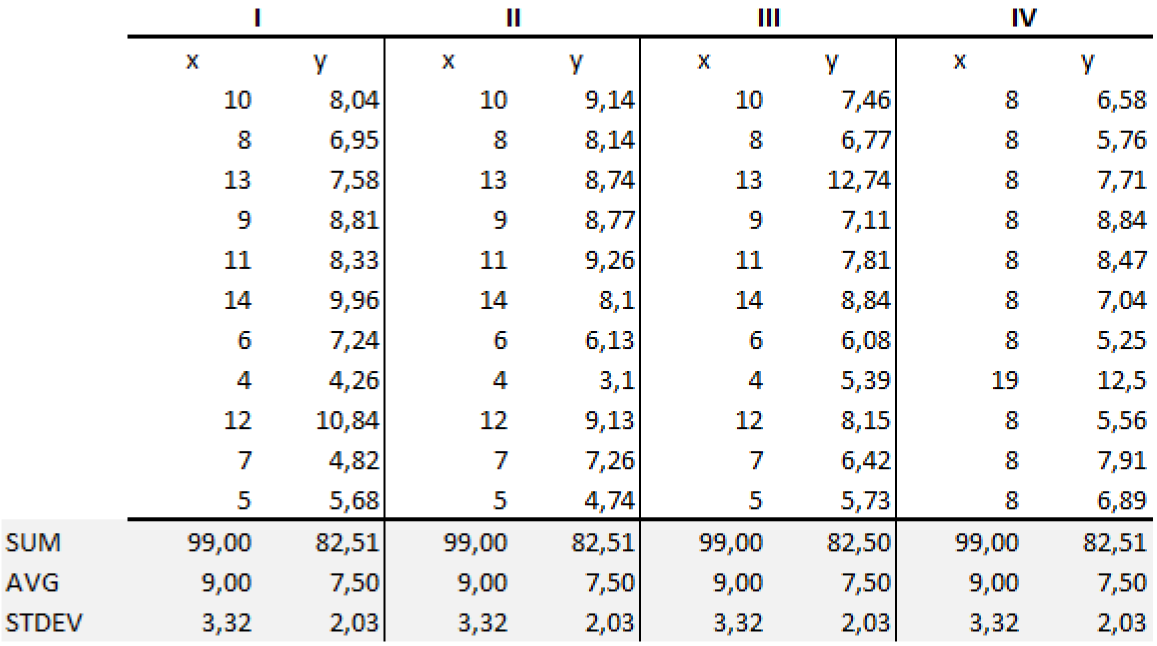
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
    - The top three features are:
        - i. temp
        - ii. harsh weather: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
        - iii. windspeed

General Subjective Questions

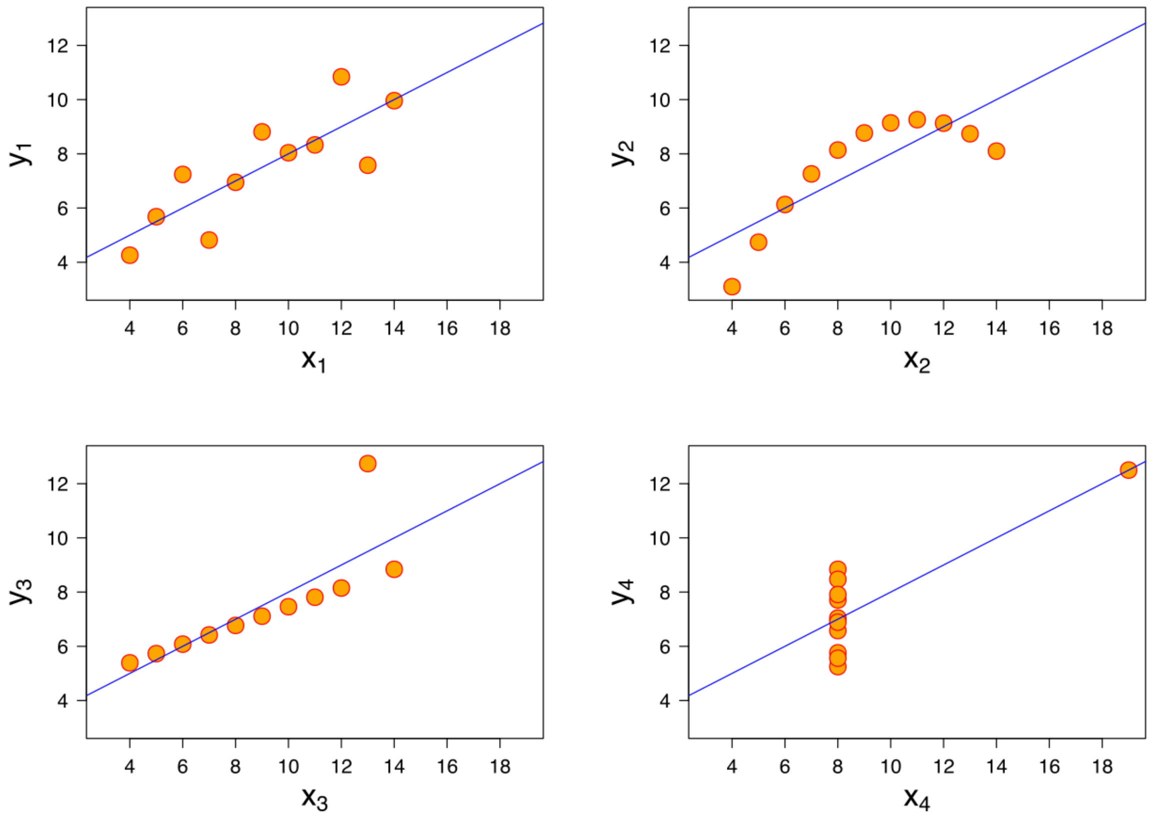
1. Explain the linear regression algorithm in detail.
    - It is an algorithm based on the supervised learning technique.
    - It comprises of a target variable also called the independent valriable and one or more independent variable.
    - Linear regression predicts dependent variable value (y) based on a given independent variable (x).by finding a linear relationship between x (input) and y(output)
    - After EDA we do the dummy conversions and split the dataset into training and test.
    - We use the training data to train a model i.e to learn the values of the coefficients and the constant (β0).
        - This includes training the modle checking the r2 value and the VIF.
        - We can either choose the top to down approach or the bottom-up approach.
    - Once the values are learnt we have the Yi and the Yi(pred) and go for te residual analysis and use the model on the test dataset to make predictions.

6. Explain the Anscombe's quartet in detail.

- It is basically four sets of data or (x,y) pairs which have exactly the same summary statistics(SUM,AVG and STDEV) but have very different plots.



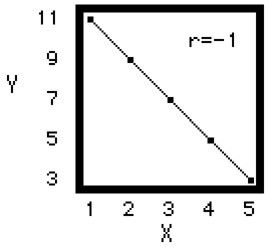
	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	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



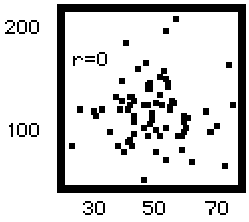
- Dataset I has a clean and well-fitting linear models.
- Dataset II does not have a normal distribution.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

2. What is Pearson’s R?
    - Pearson’s correlation is basically the measure of collinearity between variables and is measured in a limit of -1 to 1.
    - When computed in a sample it is designated as ‘r’ and is called ‘pearson’s r.
    - A correlation of +1 means that there is a perfect positive linear relationship between variables. The scatterplot shown on this page depicts such a relationship. It is a positive relationship because high scores on the X-axis are associated with high scores on the Y-axis.
    - A correlation of -1 means that there is a perfect negative linear relationship between variables. The scatterplot shown below depicts a negative relationship. It is a negative relationship because high scores on the X-axis are associated with low scores on the Y-axis.



A correlation of 0 means there is no linear relationship between the two variables. The second graph shows a Pearson correlation of 0.



Correlations are rarely if ever 0, 1, or -1. Some real data showing a moderately high correlation are shown on the next page.

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

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Example: If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

Techniques to perform Feature Scaling

Consider the two most important ones:

- Min-Max Normalization: This technique re-scales a feature or observation value with distribution value between 0 and 1.

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

- Standardization: It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

About Normalized Data

The word "normalization" is used informally in statistics, and so the term normalized data can have multiple meanings. In most cases, when you normalize data you eliminate the units of measurement for data, enabling you to more easily compare data from different places. Some of the more common ways to normalize data include:

- Transforming data using a z-score or t-score. This is usually called standardization. In the vast majority of cases, if a statistics textbook is talking about normalizing data, then this is the definition of "normalization" they are probably using.
- Rescaling data to have values between 0 and 1. This is usually called feature scaling. One possible formula to achieve this is:

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

- Standardizing residuals: Ratios used in regression analysis can force residuals into the shape of a normal distribution.
- Normalizing Moments using the formula μ/σ.
- Normalizing vectors (in linear algebra) to a norm of one. Normalization in this sense means to transform a vector so that it has a length of one.

This list is by not means all-inclusive. I've included the most common ones, but be aware there are many, many other meanings for the word normalization.

Normalization vs. Standardization

The terms normalization and standardization are sometimes used interchangeably, but they usually refer to different things. Normalization usually means to scale a variable to have a values between 0 and 1, while standardization transforms data to have a mean of zero and a standard deviation of 1.

This standardization is called a z-score, and data points can be standardized with the following formula:

$$z_i = \frac{x_i - \bar{x}}{s}$$

A z-score standardizes variables.

Where:

- xi is a data point (x1, x2…xn).
- x̄ is the sample mean.
- s is the sample standard deviation.

Z-scores are very common in statistics. They allow you to compare different sets of data and to find probabilities for sets of data using standardized tables (called z-tables). For more about z-scores, see: Z-score: Definition, Formula, and Calculation.

----

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

The user has to select the variables to be included by ticking off the corresponding check boxes. ... 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).

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. For example, we would fit the following models to estimate the coefficient of determination R1 and use this value to estimate the VIF:

X_1=C+ α_2 X_2+α_3 X_3+⋯

〖VIF〗_1=1/(1-R_1^2)

Next, we fit the model between X2 and the other independent variables to estimate the coefficient of determination R2:

X_2=C+ α_1 X_1+α_3 X_3+⋯

〖VIF〗_2=1/(1-R_2^2)

If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the 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. This would mean that that standard error of this coefficient is inflated by a factor of 2 (square root of variance is the standard deviation). The standard error of the coefficient determines the confidence interval of the model coefficients. 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. A general rule of thumb is that if VIF > 10 then there is multicollinearity. Note that this is a rough rule of thumb, in some cases we might choose to live with high VIF values if it does not affect our model results such as when we are fitting a quadratic or cubic model or depending on the sample size a large value of VIF may not necessarily indicate a poor model.

VIF	Conclusion
1	No multicollinearity
4 - 5	Moderate
10 or greater	Severe

What to do if VIF is large?
If VIF is large and multicollinearity affects your analysis results, then you need to take some corrective actions before you can use multiple regression. Here are the various options:

- One approach is to review your independent variables and eliminate terms that are duplicates or not adding value to explain the variation in the model. For example, if your inputs are measuring the weight in kgs and lbs then just keep one of these variables in the model and drop the other one. Dropping the term with a large value of VIF will hopefully, fix the VIF for the remaining terms and now all the VIF factors are within the threshold limits. If dropping one term is not enough, then you may need to drop more terms as required.
- A second approach is to use principal component analysis and determine the optimal set of principal components that best describe your independent variables. Using this approach will get rid of your multicollinearity problem but it may be hard for you to interpret the meaning of these “new” independent variables.
- The third approach is to increase the sample size. By adding more data points to our model, hopefully, the confidence intervals for the model coefficients are narrower to overcome the problems associated with multicollinearity.
- The fourth approach is to transform the data to a different space like using a log transformation so that the independent variables are no longer correlated as strongly with each other.
- Finally, you can use a different type of model call ridge regression that better handles multicollinearity.

In conclusion, when you are building a multiple regression model, always check your VIF values for your independent variables and determine if you need to take any corrective action before building the model.

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

In statistics, a Q–Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.[1] First, the set of intervals for the quantiles is chosen. A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate). Thus the line is a parametric curve with the parameter which is the number of the interval for the quantile.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the 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 $y = x$. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions. Q–Q plots can be used to compare collections of data, or theoretical distributions. The use of Q–Q plots to compare two samples of data can be viewed as a non-parametric approach to comparing their underlying distributions. A Q–Q plot is generally a more powerful approach to do this than the common technique of comparing histograms of the two samples, but requires more skill to interpret. Q–Q plots are commonly used to compare a data set to a theoretical model.This can provide an assessment of "goodness of fit" that is graphical, rather than reducing to a numerical summary. Q–Q plots are also used to compare two theoretical distributions to each other.Since Q–Q plots compare distributions, there is no need for the values to be observed as pairs, as in a scatter plot, or even for the numbers of values in the two groups being compared to be equal.

The term "probability plot" sometimes refers specifically to a Q–Q plot, sometimes to a more general class of plots, and sometimes to the less commonly used P–P plot. The probability plot correlation coefficient plot (PPCC plot) is a quantity derived from the idea of Q–Q plots, which measures the agreement of a fitted distribution with observed data and which is sometimes used as a means of fitting a distribution to data.

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

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. 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.

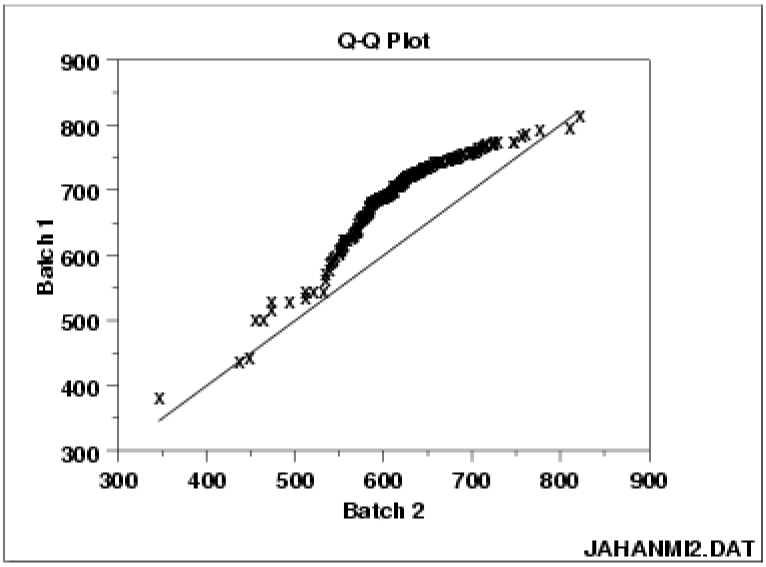
A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

The advantages of the q-q plot are:

1. The sample sizes do not need to be equal.
2. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.

The q-q plot is similar to a probability plot. For a probability plot, the quantiles for one of the data samples are replaced with the quantiles of a theoretical distribution.

Sample Plot



This q-q plot of the JAHANMI2.DAT data set shows that

1. These 2 batches do not appear to have come from populations with a common distribution.
2. The batch 1 values are significantly higher than the corresponding batch 2 values.
3. The differences are increasing from values 525 to 625. Then the values for the 2 batches get closer again.

Definition: Quantiles for Data Set 1 Versus Quantiles of Data Set 2

The q-q plot is formed by:

- Vertical axis: Estimated quantiles from data set 1
- Horizontal axis: Estimated quantiles from data set 2

Both axes are in units of their respective data sets. That is, the actual quantile level is not plotted. For a given point on the q-q plot, we know that the quantile level is the same for both points, but not what that quantile level actually is.

If the data sets have the same size, the q-q plot is essentially a plot of sorted data set 1 against sorted data set 2. If the data sets are not of equal size, the quantiles are usually picked to correspond to the sorted values from the smaller data set and then the quantiles for the larger data set are interpolated.

Questions

The q-q plot is used to answer the following questions:

- 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 behavior?

Importance: Check for Common Distribution

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.