Mark Eagle

May 11, 2021

GEOG 5113

Dr. Greene

Computer Portion of Final

**Please show all your work (e.g., the relevant R output), and clearly state and identify all appropriate hypotheses, assumptions, and conclusions. Be also sure to state your final results in a complete non-technical sentence(s). (125 Total Points**)

1. Using the wheat\_yield\_comparison files (you will need both the ANOVA and t-test wheat yield comparison files) to answer the following:
   1. Produce and analyze the 95% confidence intervals for the 15 Feb and 31 January predictions. Perform a test to determine if the estimates are statistically different from one another. (5 points)

Null hypothesis: The means of the yields for 15 February and 31 January are not statistically different.

Alternative hypothesis: The means of the yields for 15 February and 31 January are statistically different.

Call:

aov(formula = X31.Jan ~ X15.Feb, data = wheat\_yield\_comparison\_ANOVA)

Terms:

|  |  |  |
| --- | --- | --- |
|  | 15 February | Residuals |
| Sum of Squares | 3673.909 | 5095.011 |
| Deg. of Freedom | 1 | 30 |

Residual standard error: 6.497108

Estimated effects may be unbalanced

97 observations deleted due to missingness

> summary(anova3)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
| 15 February | 1 | 3674 | 3674 | 21.63 | 6.23e-05 \*\*\* |
| Residuals | 30 | 5095 | 170 |  |  |

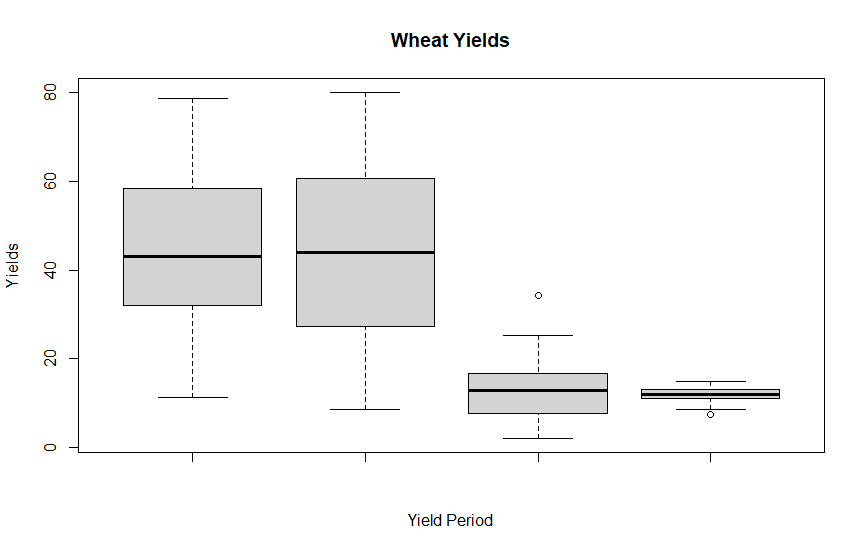
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

97 observations deleted due to missingness

The 95% CI for the 31 January was 6.0638. The 95% CI for 15 February was 6.9104. To determine whether the estimates were statistically different I ran an ANOVA to determine the variance between the two periods. The mean square for the two periods was equal to 3,674 with a residual square mean equal to 5,095. The F-value for the ANOVA was 21.63. The probability of getting the F-value is 6.23e-05, which means that there is an extremely low possibility of randomly getting the F-value. In conclusion we would reject the null hypothesis because the two periods of wheat harvest are statistically significant.

* 1. Produce box plots for 15 Feb, 31 Jan, 28 Feb, and 15 March and analyze your results. (10 points)

Analyzing the box plots I created displaying the wheat yields of the different time periods there are some statistically different variances between the four periods. The null hypothesis would be that the wheat yields over time all the means are not different. The alternate hypothesis would be that the wheat yields over time all the means are different. The first two yield periods as we learned in the question above are statistically significant which is why we rejected the null hypothesis. The third and fourth time periods have very different observations which is why just by looking at them in the boxplots we can assume that they are statistically significant from the other variables and therefore we would also reject the null hypothesis.



* 1. Perform a test to determine if the mean predictions for 15 Feb, 31 Jan, 28 Feb, and 15 March are statistically different from one another. (10 points)

Null hypothesis: The means of the yields for 15 February, 31 January, 28 February and 15 March are not statistically different.

Alternative hypothesis: The means of the yields for 15 February and 31 January, 28 February and 15 March are statistically different.

Call:

aov(formula = X31.Jan ~ Wheat.Yields, data = wheat\_yield\_comparison\_ttest)Terms:

|  |  |  |
| --- | --- | --- |
|  | Wheat.Yields | Residuals |
| Sum of Squares | 3673.909 | 5095.011 |
| Deg. of Freedom | 1 | 30 |

Residual standard error: 13.03203

Estimated effects may be unbalanced

33 observations deleted due to missingness

> summary(anova4)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
| Wheat.Yields | 1 | 3674 | 3674 | 21.63 | 6.23e-05 \*\*\* |
| Residuals | 30 | 5095 | 170 |  |  |

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

In order to perform a test on the four different yield periods for the wheat harvest I ran an ANOVA test on the yield for the 31st of January and compared it to the “wheat yields paired” variable. The wheats yield paired variable included the wheat yields of 15 February and 28 February included within it already. The mean square for the two variables was the same as answer one. The mean square for the two periods was equal to 3,674 with a residual square mean equal to 5,095. The F-value for the ANOVA was 21.63. The probability of getting the F-value is 6.23e-05, which means that there is an extremely low possibility of randomly getting the F-value. In conclusion we would reject the null hypothesis because the two variables of wheat harvest are statistically significant.

(my second answer for how to solve this problem because I was having difficulty with it. My actual submission is the first paragraph but I was wondering if either attempt was correct and to get some feedback please.)

> # Independent 2-group t-test (with defaults)

> t.test(Wheat.Yields1 ~ Time, data = wheat\_yield\_comparison\_ttest)

Welch Two Sample t-test

data: Wheat.Yields1 by Time

t = 8.3581, df = 39.117, p-value = 3.098e-10

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

22.85190 37.44185

sample estimates:

mean in group 1 mean in group 2

43.13437 12.98750

In order to perform a test on the four different yield periods for the wheat harvest I ran an independent two group T-test analyzing the Wheat paired Yields to the time periods. The mean for the paired wheat yields was equal to 43.1344. The mean for the second group, time period was 12.9875. The t-value was equal to 8.3581. The degrees of freedom were equal to 39.117. The P-value was equal to 3.098e-10. My confidence interval range was 22.85-37.44, which meant that there was an extremely low chance that the mean value of X would randomly occur and is statistically significant. Therefore we would reject the null hypothesis.

This question was really difficult for me. Initially I did the ANOVA for the first part of the question and it worked well and made sense to me. I could not figure out how to compare the multiple variables though required for this problem. My original solution was to compute the ANOVA for the first two yield periods January and February 15. Then I would compute the ANOVA variables for the third and fourth yield periods February 28 and March 15. Then I wanted to compare them to each other but was not sure how it was possible.

* 1. Perform a hypothesis test that the mean estimated value for 15 Feb is different than 48 (5 points)

I used a one sample T-test with two sides to estimate that the value for 15 February was not equal to 48. The T-value was equal to -1.436. The degrees of Freedom was equal to 31. The P-value was equal to 0.161. The mean of X is equal to 43.1344. My confidence interval range was 36.22-50.04, which meant that there was a 16.1% chance that the mean value of X would randomly occur and is not statistically significant. Therefore we would accept the null hypothesis.

1. Use the Titanic dataset to answer the following
   1. Perform tests to examine the relationship between class and survival and gender and survival (20 points).

Null hypothesis: There is no relationship between class and survival and gender and survival.

Alternative hypothesis: There is a relationship between class and survival and gender and survival.

Titanic Paid class and Survival:

|  |  |  |
| --- | --- | --- |
| Paid Class | Did Not Survive | Survived |
| 1 | 123 | 200 |
| 2 | 158 | 119 |
| 3 | 528 | 181 |

|  |  |  |
| --- | --- | --- |
| X-Squared | Degrees of Freedom | P-value |
| 127.86 | 2 | 2.2e-16 |

Titanic Gender and Survival:

|  |  |  |
| --- | --- | --- |
| Gender | Did Not Survive | Survived |
| Female | 127 | 339 |
| Male | 682 | 161 |

|  |  |  |
| --- | --- | --- |
| X-Squared | Degrees of Freedom | P-value |
| 363.62 | 1 | 2.2e-16 |

In conclusion we can see that there is a relationship between gender and paid class and those who survived the sinking of the Titanic. Therefore, we would reject the null hypothesis stating that there is no relationship between paid class and survival and between gender and survival. We can determine this due to the extremely low p-values for both tests. Both tests had p-values of 2.2e-16, which means that there is almost no possibility of these results occurring randomly. This leads us to conclude that there is a relationship between gender, paid class, and survival. This does not mean that there is correlation between the observations, but that there is simply a relation and should be considered more.

1. Using the Fitbit dataset answer the following
   1. Find the correlation between steps and calories burned. Does this suggest that a linear model may be appropriate? Explain. (10 points)

Pearson's product-moment correlation

data: Fitbit\_Data$CaloriesBurned and Fitbit\_Data$Steps

t = 22.102, df = 176, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.8128278 0.8919993

sample estimates:

cor

0.8574058

Looking at the correlation between steps and calories burned we can see that there is a strong correlation between the two. The correlation summary states that there is an 85.74% correlation between steps and calories burned. With a 95% confidence interval most values will range between 0.8128 to 0.8920. The p-value states that the probability of an observation randomly occurring within the confidence interval range is extremely low 2.2e-16 which means that we would reject the null hypothesis because there is a significant relation between the two.

* 1. Create a scatterplot and describe the association (5 points)

The scatterplot I created depicting the trend of calories burned by the number of steps taken clearly shows a relation between increasing steps and increasing the number of calories burned. This is a positively skewed linear model with a few outliers within the data.



* 1. Create a linear model. (5 points)

Call:

lm(formula = CaloriesBurned ~ Steps, data = Fitbit\_Data)

Residuals:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Min | 1Q | Median | 3Q | Max |
| -456.12 | -102.76 | -15.98 | 88.17 | 794.52 |

Coefficients:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| Intercept | 1.778e+03 | 3.143e+01 | 56.59 | <2e-16 \*\*\* |
| Steps | 7.204e-02 | 3.260e-03 | 22.10 | <2e-16 \*\*\* |

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 163.3 on 176 degrees of freedom

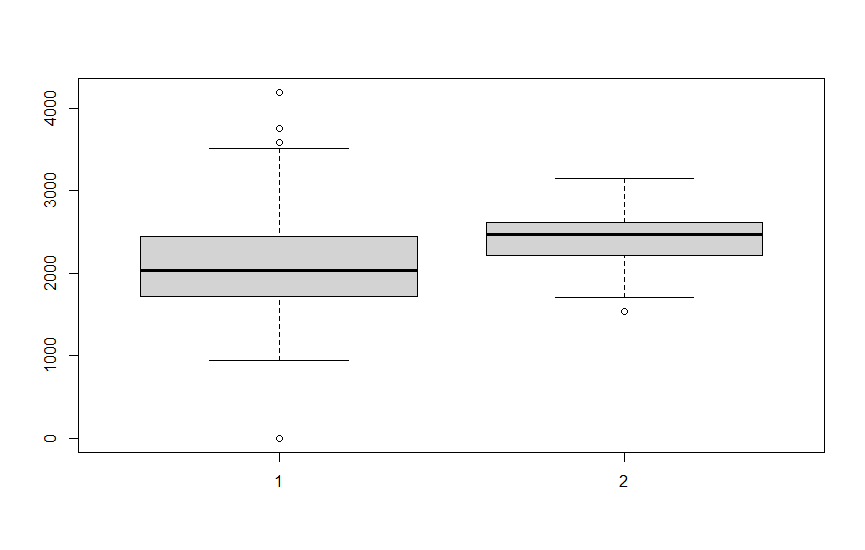
Multiple R-squared: 0.7351, Adjusted R-squared: 0.7336

F-statistic: 488.5 on 1 and 176 DF, p-value: < 2.2e-16 

After creating a linear model of steps and calories burned the intercept was identified as 1.778e+03 and the intercept for steps was 7.204e-02. This means that as calories burned increases by one, steps increase by 7.204e-02. This means that there is a positive correlation in the data. The multiple R-squared is the explained variability which means that 73.51% of the observations follow the best fit line. The adjusted R-squared is 73.36%.

* 1. Explain why a linear model may not be appropriate. Show statistical evidence to support your statement (5 points)

A linear model might not be appropriate in comparing steps to calories burned because of their distribution. If we looked at a box plot of the two variables, we would be able to see the normal distribution of the data and how most observations fall within the first and third quartiles. Looking at the five-number summary we can clearly see where the values fall and how they fit into the confidence intervals. A linear model helps us to compare the correlation between the steps and calories burned but is not the only statistical method for us to use.



1: Calories In

2: Calories Burned

* 1. Perform a hypothesis test to determine if there is a difference between the calories in and calories burned. (10 points)

Null hypothesis: The means of calories in and calories burned are not statistically different.

Alternative hypothesis: The means of calories in and calories burned are statistically different.

Call:

aov(formula = CaloriesIn ~ CaloriesBurned, data = Fitbit\_Data)

Terms:

|  |  |  |
| --- | --- | --- |
|  | Calories Burned | Residuals |
| Sum of Squares | 2017938 | 57610063 |
| Deg. of Freedom | 1 | 176 |

Residual standard error: 572.1275

Estimated effects may be unbalanced

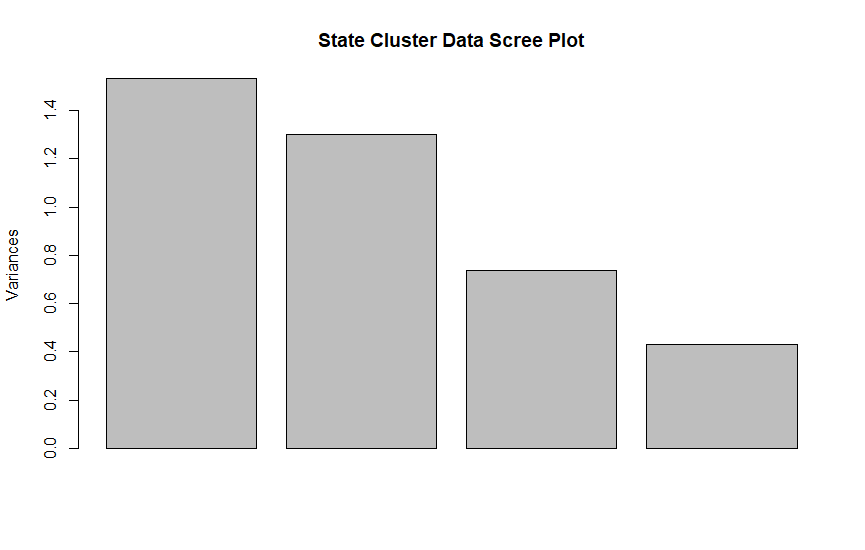
> summary(anovaFB)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
| CaloriesBurned | 1 | 2017938 | 2017938 | 6.165 | 0.014 \* |
| Residuals | 176 | 57610063 | 327330 |  |  |

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

To determine whether there is a difference between calories in and calories burned, I ran an ANOVA test between the two variables. An ANOVA allows me to calculate if there is a variance between the two variables and what that variance might be. The mean square value between calories in and calories burned is 2,017,938. The mean square value of the error is 327,330 which is the residuals within calories in and calories burned. The F-value is 6.165 and the p-value is 0.014. The F-value means that the pattern between calories in and calories burned is not random. The p-value means that there is 1.4% chance of the values to randomly and that there is a significant relation between the two variables. In conclusion we would reject the null hypothesis that the means of calories in and calories burned are not statistically different.

1. Using the world\_data file or state\_cluster\_data, perform a principal component analysis of what you think are relevant variables.
   1. Identify the appropriate number of retained components (10 points)



> summary(spc)

Importance of components:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | PC1 | PC2 | PC3 | PC4 |
| Standard Deviation | 1.2376 | 1.1401 | 0.8593 | 0.6558 |
| Proportion of Variance | 0.3829 | 0.3250 | 0.1846 | 0.1075 |
| Cumulative Proportion | 0.3829 | 0.7079 | 0.8925 | 1.0000 |

> spc

Standard deviations (1, .., p=4):

[1] 1.2376211 1.1400818 0.8592884 0.6558437

Rotation (n x k) = (4 x 4):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | PC1 | PC2 | PC3 | PC4 |
| Modern Dance | 0.1164427 | 0.67050402 | -0.72931727 | -0.07043992 |
| Xbox | -0.5388063 | -0.43297161 | -0.53138357 | 0.48974977 |
| Ice Fishing | 0.4285726 | -0.59748921 | -0.43030900 | -0.52360892 |
| Sickness | -0.7158592 | 0.07724372 | 0.02370644 | -0.69355394 |

I analyzed the state cluster data and focused on the types of hobbies people report within the data. I decided to focus on four ways people spend their time in each state, modern dance, Xbox, ice fishing and sickness. After running a principal component analysis (pca) on the four selected variables, it became apparent that only three of the four were relevant for my analysis. While selecting which variables might be relevant, I was hesitant in selecting whether to include sickness in my initial selection or not. I was not sure if being sick was a valid way of spending free time, and it did not really fit in with my other selection of possible hobbies. However, I decided to include it, in hopes of identifying any relativity between location, hobbies and illness in certain states. I thought, that might help to identify any interesting correlations to state and illness. Ice fishing was another variable I thought might struggle with relativity within the data, due to the lack of accessibility to ice fish in certain states. If not, every state can ice fish equally then it will not be as statistically significant compared to modern dance and Xbox, which anyone can do anywhere.

* 1. Describe the significant variables in each component via an analysis of the component loadings (10 points)

I choose four variables from the state cluster dataset that I believed could be statistically relevant to each other in determining how people in various states spend their free time. The four variables I chose were modern dance, Xbox, ice fishing and sickness. Sickness was the variable that was proven to not be significant in my analysis with an eigenvalue of 0.6558. The most significant variable was modern dance with an eigenvalue of 1.2376. The second most significant was Xbox with an eigenvalue of 1.1401. Ice fishing was a bit difficult to determine whether it was relative or not because of its eigenvalue of 0.8593. An eigenvalue greater than a 1.0 is the normal cut-off for determining relevancy in principal component analysis. I decided to include it in my list of final variables for relevancy due to its relative proximity to a value of 1.0 and that there were not many other relevant eigenvalues.

1. Use the output from your PCA above to perform a cluster analysis
   1. Identify the appropriate number of clusters and explain why you picked that value (10 points)

For my cluster analysis of the state cluster dataset, I chose to create three clusters of types of ways people spend time within the United States, modern dance, Xbox, and ice fishing. I chose three clusters based off my principal component analysis I performed in the previous step. Only three of the variables had eigenvalues which were statistically relevant to each other.

* 1. Describe the general distinguishing characteristics of each cluster (10 points)

> sc

Call:

hclust(d = sd)

Cluster method : complete

Distance : euclidean

Number of objects: 51

> skm

K-means clustering with 4 clusters of sizes 9, 17, 1, 24

Cluster means:

|  |  |  |  |
| --- | --- | --- | --- |
|  | PC1 | PC2 | PC3 |
| 1 | 1.2237220 | -1.5427823 | -0.5091994 |
| 2 | 0.7856798 | 0.8489182 | 0.2582500 |
| 3 | -0.1020590 | 2.9707365 | -4.1027903 |
| 4 | -1.0111665 | -0.1465544 | 0.1789723 |

Clustering vector:

[1] 4 1 4 4 2 2 2 4 2 2 4 4 4 4 4 1 4 4 4 1 2 2 4 1 4 4 1 4 4 2 2 2 2 2 1 4 4 2 2 2 4 1 4 4 3 2 2 4 4 1 1

Within cluster sum of squares by cluster:

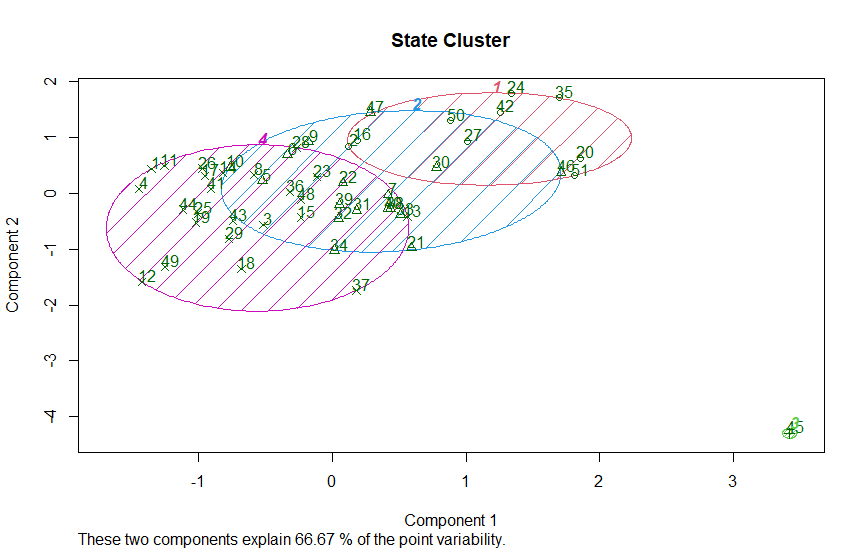
[1] 12.11300, 21.50791, 0.00000, 32.26914

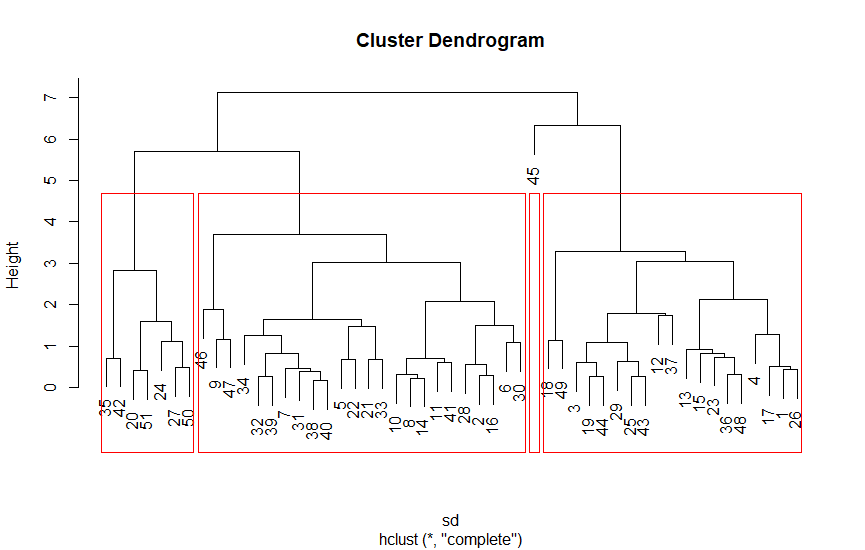
(between\_SS / total\_SS = 63.1 %)

Available components:

[1] "cluster" "centers" "totss" "withinss" "tot.withinss" "betweenss" "size"

[8] "iter" "ifault"





In my cluster analysis of principal component analysis based on the state cluster dataset I computed the cluster means and created two different cluster charts as visual aids. The “within cluster sums of squares by cluster” range was 63.1%. The three groups broken down were 9, 17, 1 and 24. This means that 63.1% of the possible values fell within the range of the first three groups of state cluster data. The k-means graph I created demonstrates how 66.67% of the point variability fall within the range. I did see that there was one outlier located within my data and that was observation number 45 which is the state of Texas. The state of Texas had an eigenvalue of -4.1028 for ice fishing and an eigenvalue of 2.9707 for Xbox. This leads me to believe that there is something interesting occurring in Texas that might be worth digging deeper into. In summary looking at the cluster graphs I did of the Fitbit data there are strong similarities within the data. The analysis proves that 66.67% of the values fall within the first four groups of my pca of state cluster data focusing on how people within the United States spend their free time.