Exam 1

Cody Frisby

10/2/2017

1

What is the total number of citations for all criminals on the criminal activity sheet?

```
## [1] 5796
```

Using the 2x standard deviation method, how many criminals, if any, have total outstanding fines that would be considered serious outliers?

```
## [1] 11
```

List all three measures of central tendency and one measure of dispersion for the age at first felony for the 30 to 34-year-old age demographic.

The mean is 21.013245.
The median is 22.
The mode is 23. This is the most frequent value for our dataset.
The standard deviation is 2.4521739.

2)

There appears to be significant correlation between a criminal's voilence index and other criminality measures. The correlations are displayed below.

Num_of_Citations	-0.8033640
Num_of_Arrests	0.9507214
Violence_Index	1.0000000
Outstanding_Fines	0.9328241
Age_First_Felony	-0.1075831
Age.in.Years	0.3701642

There is strong positive correlation with **Number of Arrests** as well as with **Outstanding Fines**. There is strong negative correlation with **Number of Citations**.

3)

The value for the largest correlation is 0.8005077. The two variables are **Outstanding Fines** and **Violence Index**. Kendall would argue that his method is superior becuase it is a rank-based calculation and doesn't rely too much on any underlying assumptions of linearity between the two variables.

4)

Using kmeans, we can create 3 clusters from the data. The three cluster centroids are displayed below.

Triage.Score	Length.of.Stay..Days.	Age.in.Years	Total.Bill
2.114286	41.34286	44.31429	10037.143
1.906250	32.65625	45.53125	5046.875
1.757576	41.45455	47.78788	14239.394

For an individual in cluster 1, generally they would have the shortest length of stay and the lowest overall bill. These people would also be the youngest, on average. For one individual in this group, their triage score was 1, they were 20 years old, their total bill was 4600, and they stayed for 36 days.

For one in cluster 2, a triage score of 3, age was 23, total bill was 10200, and they stayed for 52 days.

And finally for an individual in cluster 3. They had a triage score of 3, they were 38 years old, their bill was 14000, and their length of stay was 45 days.

5) Assuming a 75% confidence limit, are there any indicators of fraud risk that are so frequently related to one another that you would consider their relationship to be a rule? If so, identify them (reduce your Min. Criterion Value all the way) and list both their support and confidence percentages. (Hint: Import your data set with binominal data types). Write the the Support and Confidence Percentages between unverified_phone and unverified_address.

I used R for this problem (all of the actually). The top rules based on confidence (set at 0.75 with supp = 0.05) did not contain any rules that had one of **unverified_phone** on either LH or RH and the other (**unverified_address**) on the other side. But the rule {unverified_addres, unverified_phone, surveillance_data} → {international_passport} has a support of 0.054 and a confidence of 0.8709677.

6)

```
##
##   3  4  6  8
##   2 72 39 69
```

The most common number of arrests is 4 for ages greater than 35. The output above is from R where we count how many times each value occurs. As can bee seen, 4 has the largest value.

7)

The mean of all the values is 5.4666667.

We then split the data here where all values above and below are assigned group membership. We then calculate the mean for each group. This is **centroid**. We then count how many elements are in each group. This is **size**. Centorid at 7.57 has size 14. Centroid at 3.625 has size 16.

8)

Business/Org Understanding → Data Understanding → Data Preperation → Modeling → Evaluation → Deployment. And then repeat.