CS 361 Spring 2018
Homework 1

Nathaniel Murphy (njmurph3)

1.1: Show that $\text{mean}(\{kx\}) = k\text{mean}(\{x\})$ by substituting into the definition.

Solution:

$$\text{mean}(\{x\}) = \frac{1}{N}\sum_{i=1}^{N} x_i \Rightarrow \text{mean}(\{kx\}) = \frac{1}{N}\sum_{i=1}^{N} kx_i$$

$$\frac{1}{N}\sum_{i=1}^{N} kx_i = \frac{1}{N}(kx_1 + kx_2 + \ldots + kx_N) = k\Big(\frac{1}{N}(x_1 + x_2 + \ldots + x_N)\Big) = k\Big(\frac{1}{N}\sum_{i=1}^{N} x_i\Big) = k\text{mean}(\{x\})$$

□

1.2: Show that $\text{mean}(\{x + c\}) = \text{mean}(\{x\}) + c$ by substituting into the definition.

Solution:

$$\text{mean}(\{x\}) = \frac{1}{N}\sum_{i=1}^{N} x_i \Rightarrow \text{mean}(\{x + c\}) = \frac{1}{N}\sum_{i=1}^{N}(x_i + c)$$

$$\frac{1}{N}\sum_{i=1}^{N}(x_i + c) = \frac{1}{N}\big((x_1 + c) + (x_2 + c) + \ldots + (x_N + c)\big)$$

$$= \frac{1}{N}(x_1 + x_2 + \ldots + x_N + N \cdot c)$$

$$= \frac{1}{N}(x_1 + x_2 + \ldots + x_n) + \frac{1}{N}(N \cdot c) = \left(\frac{1}{N}\sum_{i=1}^{N} x_i\right) + c = \text{mean}(\{x\}) + c$$

□

1.3: Show that $\sum_{i=1}^{N}(x_i - \text{mean}(\{x\})) = 0$ by substituting into the definition.

Solution:

$$\sum_{i=1}^{N}\Big(x_i - \text{mean}(\{x\})\Big) = \big((x_1 - \text{mean}(\{x\})) + (x_2 - \text{mean}(\{x\})) + \ldots + (x_N - \text{mean}(\{x\}))\big)$$

$$= (x_1 + x_2 + \ldots + x_N) - N(\text{mean}(\{x\}))$$

$$= \sum_{i=1}^{N} x_i - N\Big(\frac{1}{N}\sum_{i=1}^{N} x_i\Big) = \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} x_i = 0$$

□

1.4: Show that std($\{x + c\}$) =std($\{x\}$) by substituting into the definition.

Solution:

$$\text{std}(\{x\}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\big(x_i - \text{mean}(\{x\})\big)^2} \Rightarrow \text{std}(\{x{+}c\}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\Big((x_i + c) - \text{mean}(\{x + c\})\Big)^2}$$

$$= \sqrt{\frac{1}{N}\sum_{i=1}^{N}\Big((x_i + c) - (\text{mean}(\{x\}) + c)\Big)^2}$$

$$= \sqrt{\frac{1}{N}\sum_{i=1}^{N}\Big(x_i - \text{mean}(\{x\}) + c - c\Big)^2}$$

$$= \sqrt{\frac{1}{N}\sum_{i=1}^{N}\Big(x_i - \text{mean}(\{x\})\Big)^2} = \text{std}(\{\text{x}\})$$

□

1.5: Show that $\text{std}(\{kx\}) = k \cdot \text{std}(\{x\})$ by substituting into the definition.

Solution:

$$\text{std}(\{x\}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\Big(x_i - \text{mean}(\{x\})\Big)^2} \Rightarrow \text{std}(\{kx\}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\Big(kx_i - \text{mean}(\{kx\})\Big)^2}$$

$$= \sqrt{\frac{1}{N}\sum_{i=1}^{N}\Big(kx_i - k \cdot \text{mean}(\{x\})\Big)^2}$$

$$= \sqrt{\frac{1}{N}\sum_{i=1}^{N}\Big(k\big(x_i - \text{mean}(\{x\})\big)\Big)^2}$$

$$= \sqrt{\frac{k^2}{N}\sum_{i=1}^{N}\Big(x_i - \text{mean}(\{x\})\Big)^2}$$

$$= k\sqrt{\frac{1}{N}\sum_{i=1}^{N}\Big(x_i - \text{mean}(\{x\})\Big)^2} = k \cdot \text{std}(\{x\})$$

□

1.6: Show that $\text{median}(\{x + c\}) = \text{median}(\{x\}) + c$ by substituting into the definition.

Solution: The median of a list is found by first sorting the list, so let us index each element of the sorted list as $\{x\}_s = x_1, x_2, \ldots, x_N$. $\{x\}_s$ is sorted implies that $\forall\, k, \ell \in \mathbb{N} :\ k, \ell \leq N,\ k < \ell \Rightarrow x_k < x_\ell \Rightarrow x_k + c < x_\ell + c$, so we see that $\{x+c\}_s$ is sorted with the same indices for $\{x\}_s$.

Case 1: N is odd.
Then there exists index $j = \lceil \frac{N}{2} \rceil$ that perfectly divides the data such that $\text{median}\{x\}_s = x_j$. Because $\{x + c\}_s$ is sorted with the same indices as $\{x\}_s$, the median element must occur at index j. It follows that $arg(\text{median}(\{x\}_s) = arg(\text{median}(\{x + c\}_s)) \Rightarrow \text{median}(\{x\}_s) + c = \text{median}(\{x\}) + c = \text{median}\ (\{x + c\}_s) = \text{median}(\{x + c\})$.

Case 2: N is even.
The there exists indices $j_1, j_2 = \frac{N}{2}, \frac{N}{2}+1$, respectively, such that $\text{median}(\{x\}) = \text{median}(\{x\}_s) = \frac{(x_{j_1} + x_{j_2})}{2}$. Because $\{x+c\}_s$ is sorted with the same indeices as $\{x\}_s$, to obtain the median, we must again average elements at indices j_1, j_2. It follows that $\text{median}(\{x+c\}) = \text{median}(\{x + c\}_s) = \frac{x_{j_1}+c+x_{j_2}+c}{2} = \frac{x_{j_1}+x_{j_2}+2c}{2} = \frac{x_{j_1}+x_{j_2}}{2} + c = \text{median}(\{x\}_s) + c = \text{median}(\{x\}) + c$.

□

1.10: In the case of this dataset, I do not beleive the mean to be a very useful summary because we actually see that mean number of barrels, 8582 million, produced in 1962, which is far closer to 1984 than 1880. This same phenomenon in which we suspect our data to be widely spread is confirmed when we observe that the standard deviation of the number of millions of barrels produced is 70141830.23. As Figure 1 implies, it seems that most of the data is between the 0-5000 million barrel range, but only recently has it shifted drastically upwards to around 15000+ million barrels to affect the mean in such a way.

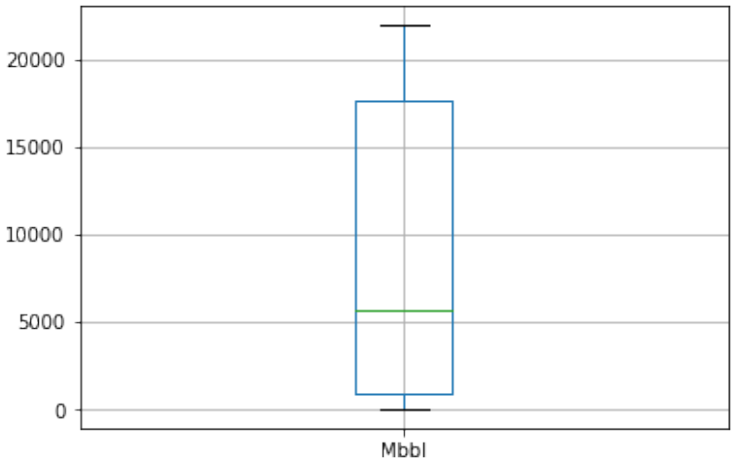


Figure 1: Number of millions of barrels of oil produced from 1880-1984

1.11:

(a) Every year between the dates 67.5 and 69.5, there is one power plant that had an 1100 Megawatt capacity, where others in that time period were mostly between 600 and 900. Also, one power plant constructed at date 71 had capacity of over 1100 Megawatts. Also, when looking at cost compared to the date, we can see that one power plant cost $90 million whereas the second highest cost was around $70 million.

(b) Mean: $46.16 million Std: $17.01 million

(c) Mean: 825.375 Megawatts Std: 189.36 Megawatts

(d) The histogram (Figure 2) isn't very skewed, probably because as technology increases, so can the capacity of the power plants, but at the same time, technology increasing generally means that the cost decreases.

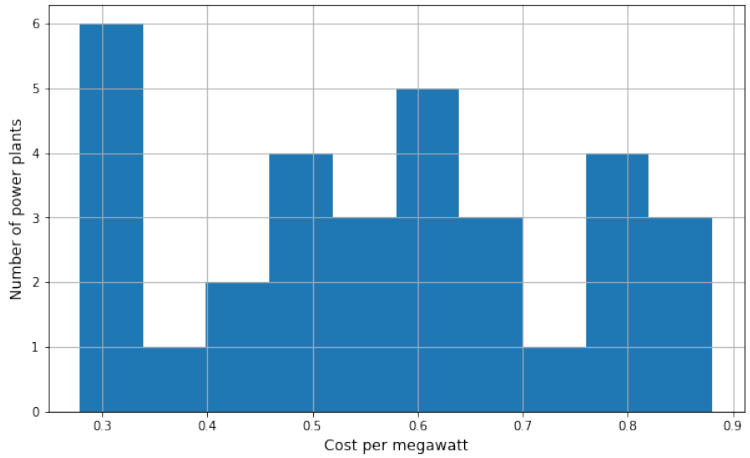


Figure 2: Histogram of Cost per Megawatt in power plants

1.12:

With respect to calories, the histograms for Beef and Poultry had a skew to the right, while the histogram for Meat had a skew to the left. The standard deviations for the calories for all three meats were very similar (∼22-25), and both Beef and Meat had very similar mean calories (∼157), but the mean calories for Poultry was far lower (∼118).

In regards to sodium, both the Beef and Poultry histograms had a left skew, while the Meat histgram had a skew to the right. Standard deviations were all in the range 84-102 and means were all within 401-418, which isn't a large percent change.

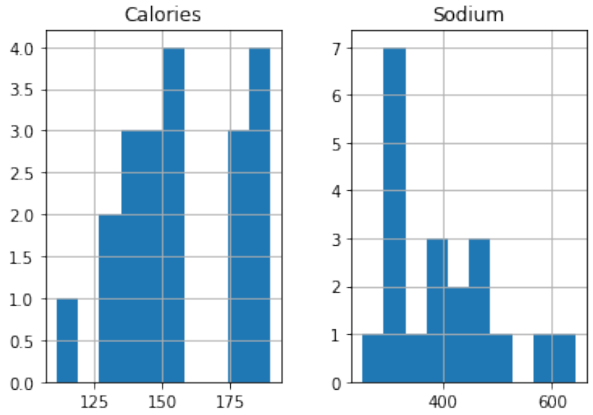


Figure 3: Histogram of calories and sodium in Beef hotdogs

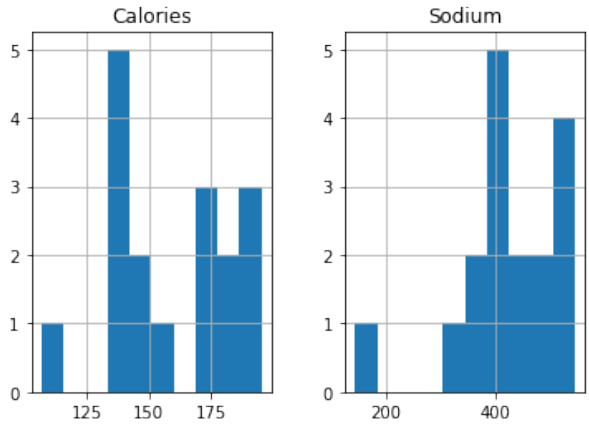


Figure 4: Histogram of calories and sodium in Meat hotdogs

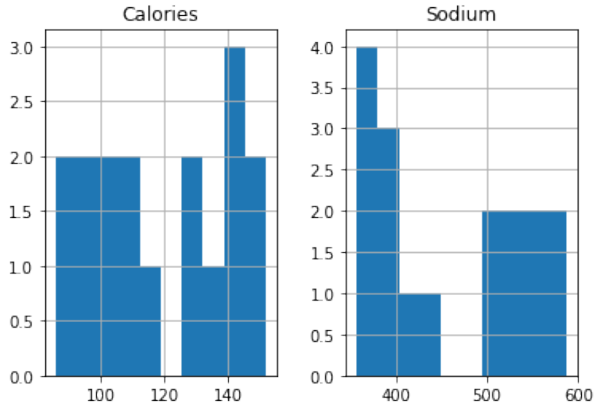
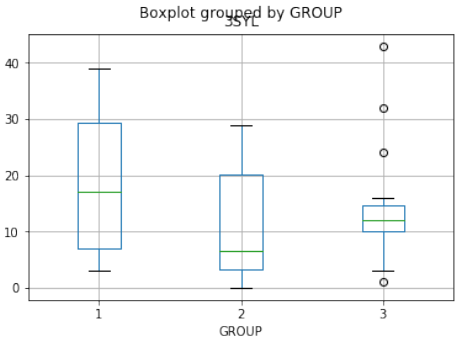


Figure 5: Histogram of calories and sodium in Poultry hotdogs

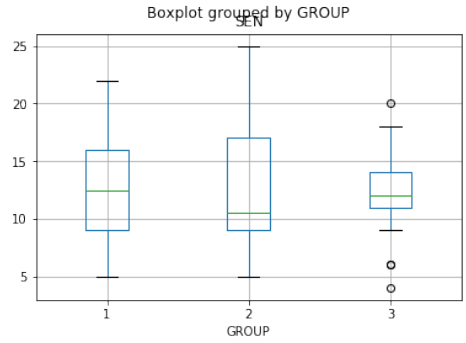
1.13:

(a) Groups 1 and 2 have a similar box shape, but the mean and total position of box 1 is greater than box 2, which follows an obvious trend that magazines of a higher education level have more 3 syllable words. Group 3 follows this trend in that it has a lower variance (most of the points fall within the 10-15 frequency range) and has a mean less than group 1. Group 3 does have a few outliers, but overall, it looks like group 3 is the group with the least amount of 3 syllable words.

(b) In examining the number of sentences, the means are approximately the same across all three groups, but the variances are what differ. The variances for group 1 and group 2 are very similar, but the variance for group 3 is smaller than the others. As with the syllables, group 3 had a few outliers in the number of sentences, this time on the lower end.



(a) 3 syllable words by magazine group



(b) Number of sentences by magazine group

1.14:

For this problem I chose not to modify the datasets to eliminate the double counting.

(a) The distributions seem almost the same, but upon examining the percentiles, the 70^{th} percentile of the Math students' drinking during the week was 1.8, in other words, 70% of the Math students' data was at a 1.8 or lower on a 1-5 scale. This is compared to the Portugese students who had a 70^{th} percentile occuring at 2.0 on a 1-5 scale. This means that slightly more Portugese students had a higher rating for drinking during the week.

```
: 0.1    1.0
  0.2    1.0
  0.3    1.0
  0.4    1.0
  0.5    1.0
  0.6    1.0
  0.7    1.8
  0.8    2.0
  0.9    3.0
```

```
: 0.1    1.0
  0.2    1.0
  0.3    1.0
  0.4    1.0
  0.5    1.0
  0.6    1.0
  0.7    2.0
  0.8    2.0
  0.9    3.0
```

Figure 7: Percentiles of levels of Math (left) and Portugese (right) students that drink during the week

(b) Upon grouping all the data together and only looking at the family size, I was able to determine that those individuals in families 'LE3' or 'Less than or equal to 3' drank on the weekends more than those individuals in the group 'GT3' or 'Greater than 3' members in the family. I was able to determine this by exmaining the histograms, but again confirmed with the percentiles of data. Around 30% of the 'LE3' data points fall within the 1.0 on the 1-5 scale for drinking while about 40% of the 'GT3' data points fall within the 1.0. Looking at the other end, about only 80% of the 'LE3' points fall within the 1.0-4.0 while 90% of the 'GT3' points fall within the 1.0-4.0 range. I conclude that more students with families less than or equal to 3 tend to drink more on the weekends than students in families of size greater than 3.

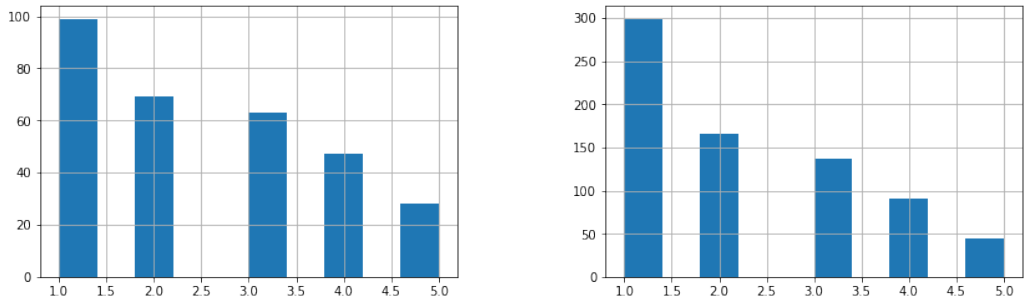


Figure 8: Histograms of levels of drinking during the week for LE3 (left) and GT3 (right) students.

0.1	1.0
0.2	1.0
0.3	1.0
0.4	2.0
0.5	2.0
0.6	3.0
0.7	3.0
0.8	4.0
0.9	4.0

0.1	1.0
0.2	1.0
0.3	1.0
0.4	1.0
0.5	2.0
0.6	2.0
0.7	3.0
0.8	3.0
0.9	4.0

Figure 9: Percentiles of levels of drinking during the week for LE3 (left) and GT3 (right students.

(c) After creating a new column with the weighted average of weekend drinking and weekday drinking, I plotted sixteen different boxplots (Figure 10) of every combination of {school, sex, famsize, romantic}. Upon examination, it was obvious that 5 data points had a higher mean than the rest. These groups had a mean total drinking score of 2.0-2.5:

- {GP, M, LE3, yes}
- {MS, M, GT3, no}
- {MS, M, GT3, yes}
- {MS, M, LE3, no}
- {MS, M, LE3, yes}

If we were to generalize these categories, we would see that students going to school at Mousinho da Silveira that were male tend to drink more than other students.

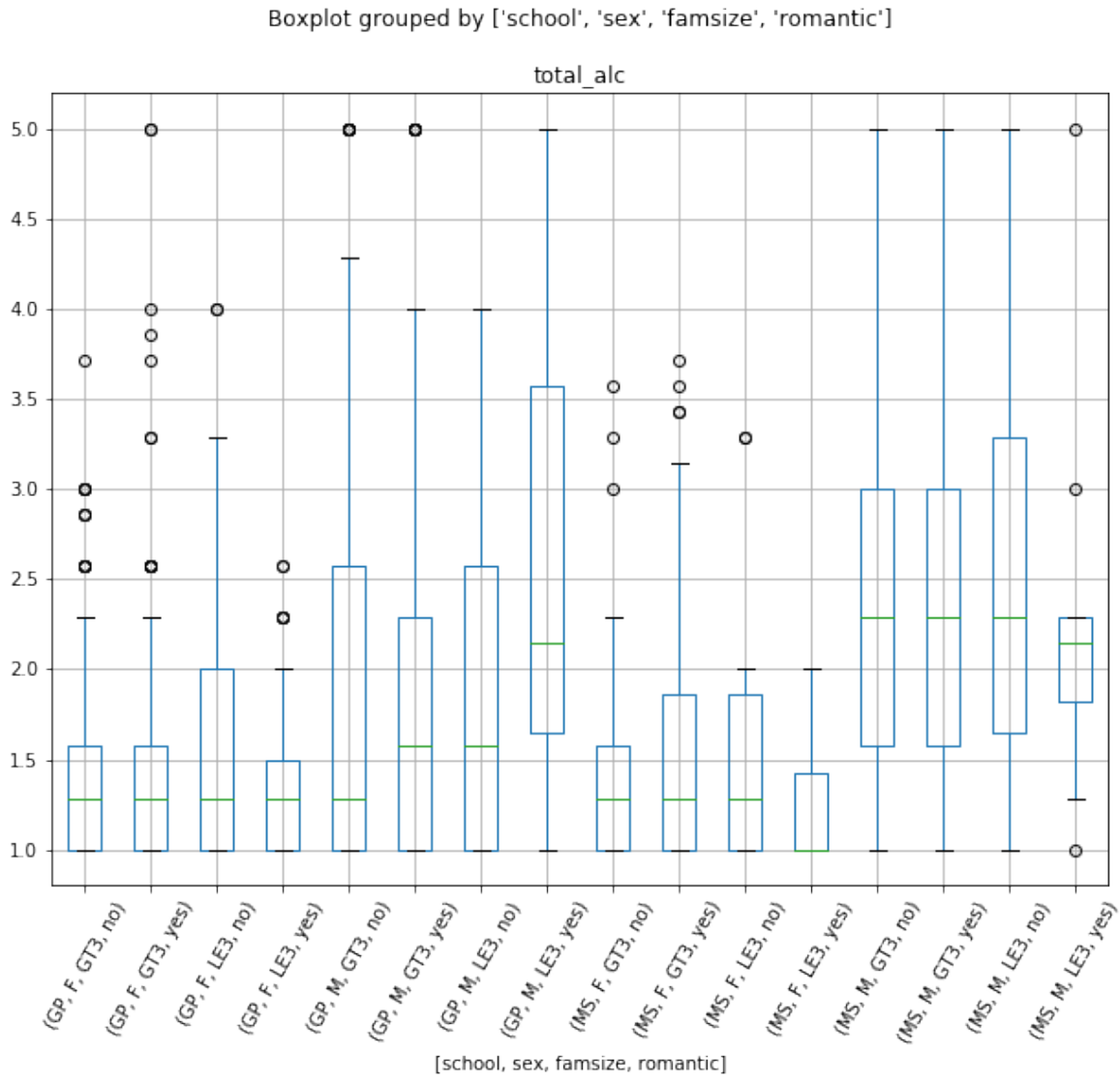

Boxplot grouped by ['school', 'sex', 'famsize', 'romantic']
total_alc
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
(GP, F, GT3, no)
(GP, F, GT3, yes)
(GP, F, LE3, no)
(GP, F, LE3, yes)
(GP, M, GT3, no)
(GP, M, GT3, yes)
(GP, M, LE3, no)
(GP, M, LE3, yes)
(MS, F, GT3, no)
(MS, F, GT3, yes)
(MS, F, LE3, no)
(MS, F, LE3, yes)
(MS, M, GT3, no)
(MS, M, GT3, yes)
(MS, M, LE3, no)
(MS, M, LE3, yes)
[school, sex, famsize, romantic]

Figure 10