Statistics for Engineering, 4C3/6C3

Assignment 2

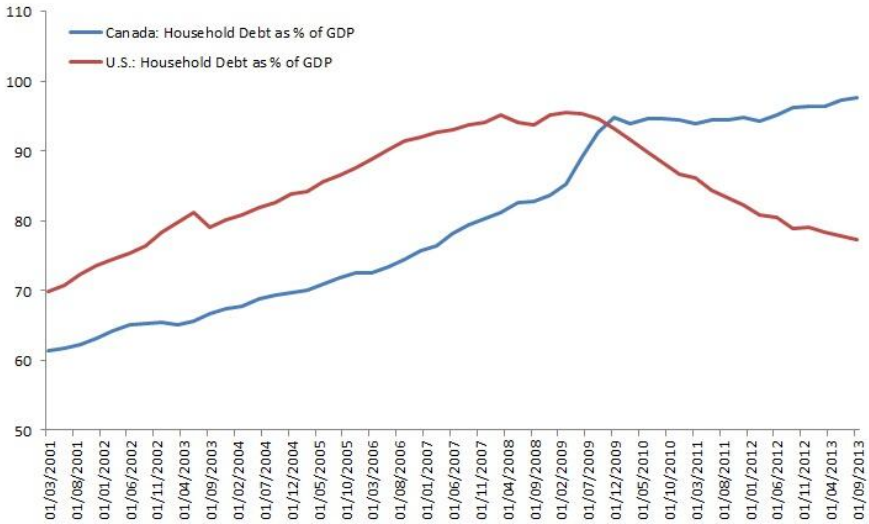
Kevin Dunn, kevin.dunn@mcmaster.ca **Due date: 23 January 2014**

Assignment objectives: interpreting data visualizations; univariate data analysis

Question 1 [5]

Similar to a question on the final exam, 2013:

The following visualization appeared in the Twitter feed of a Globe and Mail columnist, Scott Barlow.



1. The article itself is behind a paywall, so it is not accessible. What might have been the message Mr Barlow was wishing to convey with this plot?
2. Referring the principles of data visualization we learned about, what might he have improved on the graph?

Solution

1. It would appear that his article is related to the increasing amounts of debt being taken on by Canadians, as compared to their US counterparts. From 2001 to 2008 the Canadian and US debt levels were highly correlated. The brief recession/depression in 2009 seemed to reverse the amount of debt taken on by US citizens, declining at roughly the same rate it climbed by prior to 2009. Canadians have decreased their rate of debt increase since 2009, but the trend is still upward. There was interesting jump in Canadian debt levels at the end of 2009. Likely the rest of Mr Barlow's article would have been reasons for these features in the data.
2. Some brief enhancements might include:
    - clear labelling of the vertical axis
    - better spacing and choice of text for the time axis: it is harder to read vertically, the spacing is 5 months apart (which is unusual for financial data which are most often presented in multiples of 3 months; quarters, halves or full years)
    - consider using only years for the time axis, as that minimizes data ink, and one can see more clearly when the events occurred within the year
    - rather use dark colours for the plot's two series, and not rely on colour to differentiate the two

- clip the y-axis to range from 60% to 100% rather
- it might be interesting to see a bit more context of time, prior to 2001.

Question 2 [3]

1. Why are robust statistics, such as the median or MAD, important in the analysis of modern data sets? Explain, using an example, if necessary.
2. What is meant by the break-down point of a robust statistic? Give an example to explain your answer.

Solution

1. Data sets you will have to deal with in the workplace are getting larger and larger (lengthwise), and processing them by trimming outliers (see Question 5 later) manually is almost impossible. Robust statistics are a way to summarize such data sets without point-by-point investigation.

   This is especially true for automatic systems that you will build that need to (a) acquire and (b) process the data to then (c) produce meaningful output. These systems have to be capable of dealing with outliers and missing values.

2. The breakdown point is the number of contaminating data points required before a statistic (estimator) becomes unbounded, i.e. useless. For example, the mean requires only 1 contaminating value, while the median requires 50% + 1 data points before it becomes useless.

   Consider the sequence $[2, 6, 1, 91511, -4, 2]$. The mean is 15253, while the median is 2, which is a far more useful estimate of the central tendency in the data.

Question 3 [8]

A food production facility fills bags with potato chips with an advertised bag weight of 50.0 grams.

1. The government's *Weights and Measures Act* requires that at most 1.5% of customers may receive a bag containing less than the advertised weight. At what setting should you put the target fill weight to meet this requirement exactly? The check-weigher on the bagging system shows the long-term standard deviation for weight is about 2.8 grams.
2. Out of 100 customers, how many are lucky enough to get 55.0 grams or more of potato chips in their bags?

Solution

1. Given that it is a long-term standard deviation, we have $\sigma = 2.8\,\text{grams}$. Calculate the z-value and find which fraction of z falls at or below 1.5% of the probability area. From the tables this is $z = -2.17$ = `qnorm(0.015)`.

   Then solve for μ:

   $$z = \frac{50 - \mu}{2.8} = -2.17$$
   $$\mu = 56.08 \text{ grams}$$

   The check weigher should be set at 56.08 grams.

2. From the prior answer, we can see many customers will receive a bag with 55.0 grams or more. Probability of 55.0 grams of more is the area area the corresponding z-value:

   $$z > \frac{55 - 56.08}{2.8}$$
   $$z > -0.385$$

The exact answer is `(1 - pnorm(-0.385))*100 = 64.98`, so around 65 customers out of every 100 (you might have a slightly different number if you used tables to answer your question - make sure you can use the statistical tables to answer this problem too.)

Question 4 [20]

The following confidence interval is reported by our company for the amount of sulphur dioxide measured in parts per billion (ppb) that we send into the atmosphere.

$$123.6\,\text{ppb} \leq \mu \leq 240.2\,\text{ppb}$$

Only $n = 21$ raw data points (one data point measured per day) were used to calculate that 90% confidence interval. A z-value would have been calculated as an intermediate step to get the final confidence interval, where $z = \dfrac{\overline{x} - \mu}{s/\sqrt{n}}$.

1. What assumptions were made about those 21 raw data points to compute the above confidence interval?
2. Which lower and upper critical values would have been used for z? That is, which critical values are used before unpacking the final confidence interval as shown above.
3. What is the standard deviation, s, of the raw data?
4. Today's sulphur dioxide reading is 460 ppb and your manager wants to know what's going on; you can quickly calculate the probability of seeing a value of 460 ppb, or greater, to help judge the severity of the pollution. How many days in a 365 calendar-day year are expected to show a sulphur dioxide value of 460 ppb or higher?
5. Explain clearly why a wide confidence interval is not desirable, from an environmental perspective.

Solution

1. The 21 data points are independent and come from *any distribution* of finite variance.
2. From the t-distribution at 20 degrees of freedom, with 5% in each tail: $c_t = 1.72$ = `qt(0.95, df=20)`. The t-distribution is used because the standard deviation is estimated, rather than being a population deviation.
3. The standard deviation may be calculated from:

$$\begin{aligned} UB - LB = 240.2 - 123.6 = 2 \times c_t \frac{s}{\sqrt{n}} &= (2)(1.72)\frac{s}{\sqrt{n}} \\ s &= \frac{(116)(\sqrt{n})}{(2)(1.72)} \\ s &= 154.5\,\text{ppb} \end{aligned}$$

   Note the very large standard deviation relative to the confidence interval range. This is the reason why so many data points were taken (21), to calculate the average, because the raw data comes from a distribution with such a large variation.

   An important note here is that many of you picked up on this large estimate for the standard deviation and realized it was so wide, that it would imply the distribution produced values with negative sulphur dioxide concentration (which is physically impossible). However, note that when dealing with large samples (21 in this case), the distinction between the normal and the t-distribution is minimal. Further, the raw data are not necessarily assumed to be from the normal distribution, they could be from any distribution, including one that is heavy-tailed, such as the F-distribution (see the yellow and green lines in particular).

4. The probability calculation requires a mean value. Our best guess for the mean is the midpoint of the confidence interval, which is always symmetric about the estimated process mean, $\overline{x} = \dfrac{240.2 - 123.6}{2} + 123.6 = 181.9$.

Note that this is not the value for μ, since μ is unknown.

$$z = \frac{460 - 181.9}{154.5} = 1.80$$

Probability is `1 - pt(1.8, df=20)` $= 1 - 0.9565176 = 0.0434824$, or about $0.0434824 \times 365 = 15.9$, or about 16 days in the year (some variation is expected, if you have used a statistical table)

5. A wide confidence interval implies that our sulphur dioxide emissions are extremely variable (the confidence interval bounds are a strong function of the process standard deviation). Some days we are putting more pollution up into the air and balancing it out with lower pollution on other days. Those days with high pollution are more environmentally detrimental.

Question 5 [10]

Many students in the course expressed an interest in analyzing a large data set. Here's a baby-step: your manager has asked you to describe the flow rate characteristics of the overhead stream leaving the top of the distillation column at your plant. You download one month of data, available on the course website.

The data are from 1 March to 31 March, taken at one minute intervals.

Solution

Please see one of the students solutions to this question at the end; thanks to Ghassan Marjaba for allowing us to use his solution for this question.

Question 6 [600-level students only: 12]

In the course notes on the section on comparing differences between two groups we used, without proof, the fact that:

$$\mathcal{V}\{\bar{x}_B - \bar{x}_A\} = \mathcal{V}\{\bar{x}_B\} + \mathcal{V}\{\bar{x}_A\}$$

Using the fact that $\mathcal{V}\{cx\} = c^2\mathcal{V}\{x\}$, you can show that:

$$\mathcal{V}\{\bar{x}_B + \bar{x}_A\} = \mathcal{V}\{\bar{x}_B\} + \mathcal{V}\{\bar{x}_A\}$$

1. The first equation is only correct when an important assumption is true; what is that assumption?
2. *Based on an actual industrial problem*: A filling machine doses a drug to a canister. The patient will inhale the drug (imagine an asthma pump). The weight of the drug in the canister must be added as precisely and accurately as possible, to avoid patient over- or under-dosing.

   The weight filled will fluctuate with temperature in the building and is theoretically calculated as having a standard deviation of 32mg due to typical temperature variations. The filling line has 6 machines that fill the canisters and the variability from machine-to-machine is 40mg. The operators calibrate the machines at the start of each shift, and their estimated calibration accuracy is estimated at 15mg. The wear and tear on the machine parts over the year is estimated to only add an extra 10mg of variation.

   What is the expected long-term standard deviation of fill weights recorded from this process? What assumption(s) do you have to make to calculate this?

Solution

1. Assume the operation in system A is independent of system B's operation. This implies $\overline{x}_A$ is independent of $\overline{x}_B$.
2. Using a similar concept to the above, the variation from multiple sources can be added up, as long the variation from each source is independent of the other. In this case, it would require the variance due to temperature is unrelated to the machine-to-machine variance. We'd have to check every pair of combinations to ensure they are independent.

   If this assumption is true, then the total variance in the process is $\sigma^2_{\text{total}} = 32^2 + 40^2 + 15^2 + 10^2 = 54.3^2$, or $\sigma_{\text{total}} = 54.3$ mg.

Question 7 [600-level: 0 points]

The solution appears in Process Improvement using Data.

The paper by PJ Rousseeuw, "Tutorial to Robust Statistics", *Journal of Chemometrics*, **5**, 1-20, 1991 discusses the breakdown point of a statistic.

1. Describe what the breakdown point is, and give two examples: one with a low breakdown point, and one with a high breakdown point. Use a vector of numbers to help illustrate your answer.
2. What is an advantage of using robust methods over their "classical" counterparts?

END

Thanks for Ghassan Marjaba to allow us to use his solution for this question. The point of this solution is to show the intensive, iterative nature of dealing with real data sets.

Question 5

I have downloaded the data, the looked for the: summary to see the mean, median, quartiles, etc. Then I used the box plots and histograms.

The box plots were used to determine any outliers, corrupt data, missing data, etc. The histogram was used to determine if there is a distribution.

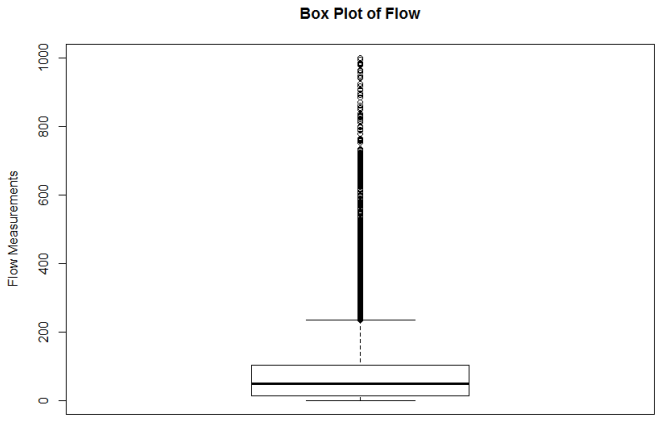
The data altogether was first analyzed. The summary provided by R was:

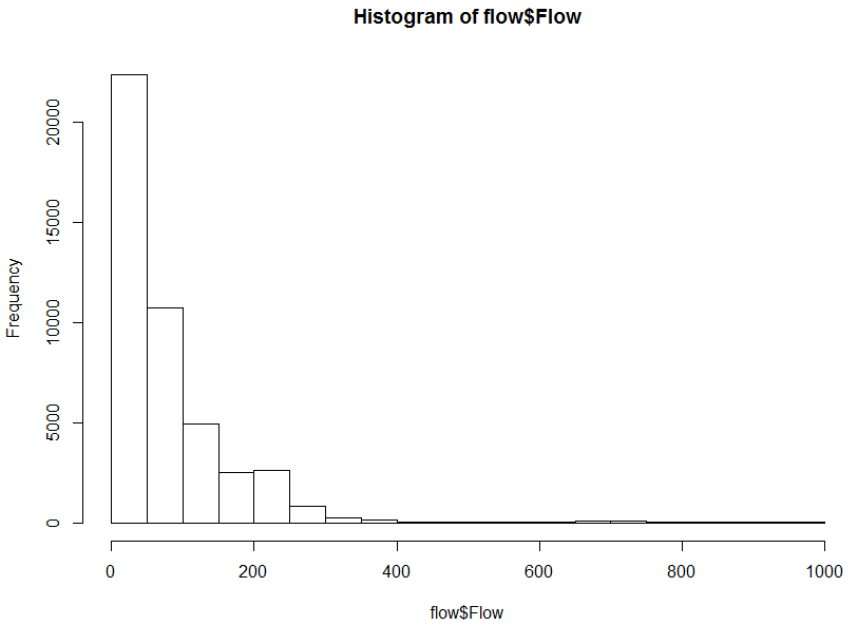
```
> flow <-
read.csv('http://datasets.connectmv.com/file/distillate-flow.csv')
> summary(flow)
      Flow
 Min.   :   0.00
 1st Qu.:  12.94
 Median :  49.62
 Mean   :  75.71
 3rd Qu.: 102.05
 Max.   :1000.00
```

Of note was the minimum and maximum values, as well as the difference in mean and median. Also, the IQR was very interesting. Before I discuss these, I used the following commands to plot in R:

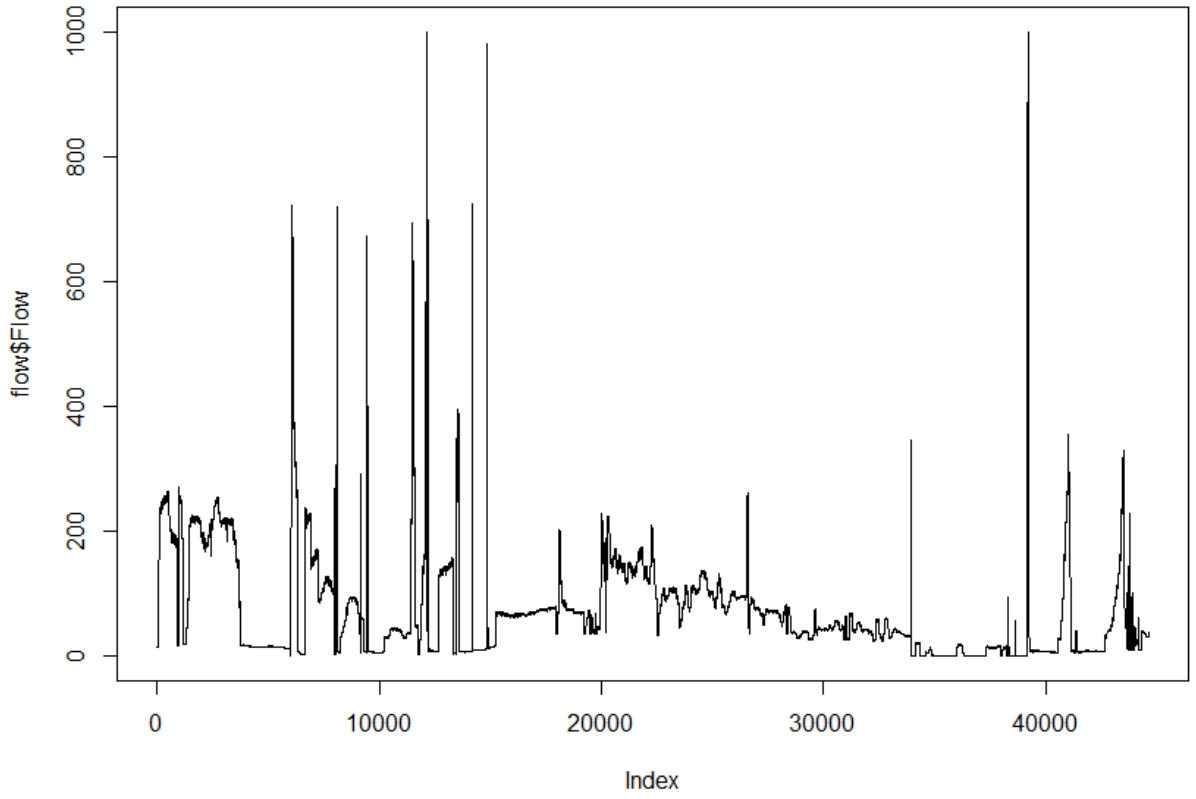
```
boxplot(flow, ylab = "Flow Measurements", main ='Box Plot of Flow')
hist(flow$Flow)
plot(flow$Flow, type="l")
```

The resulting 3 plots are shown here:





The three plots are really of no use. The data is too noisy. The plots just confirm that. The IQR was of interest in the box plot showing that the values over 200 seem to corrupt the data. Also, there seemed to be a lot of zeros which refer to missing data.



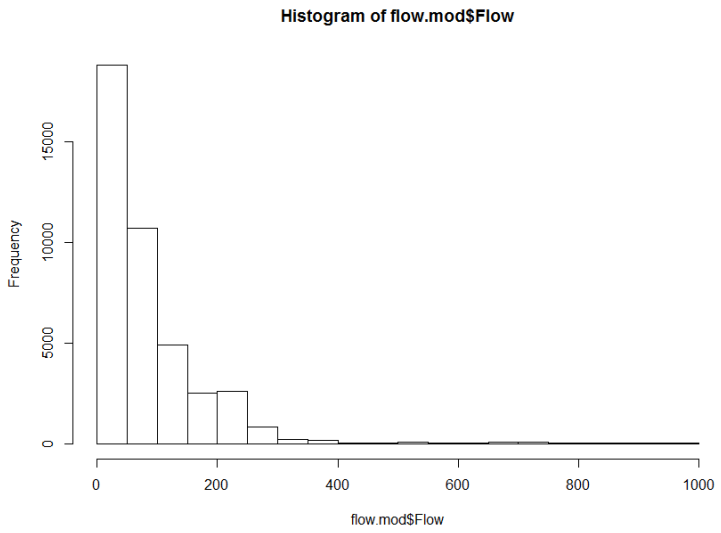
Although this can be done through R, I first removed all the zeros from the data set manually and re-did the work (without the line plot which was not of much help. The summary was:

```
> flow.mod <- read.csv('distillate-flow_mod.csv')
> summary(flow.mod)
      Flow
 Min.   :   0.0174
 1st Qu.:  17.5010
 Median :  64.3960
 Mean   :  82.1858
 3rd Qu.: 106.8800
 Max.   :1000.0000
```

This time, the summary provided insight into the fact that the values larger than 200

and below 15 are also corrupting the data and seem to be errors in measurements. To confirm, the box plot and histogram were plotted as shown below.

```
> boxplot(flow.mod$Flow)
> hist(flow.mod$Flow)
```



The above graphs confirmed that the data small data and large data are corrupting the data.

The values above 200 were removed and the exercise was repeated.

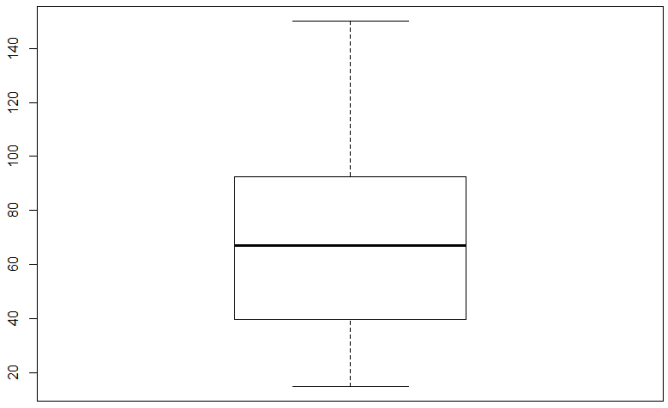
```
> summary(flow.small)
   X0.017365
 Min.   :  0.08682
 1st Qu.: 14.37600
 Median : 47.50300
 Mean   : 60.12994
 3rd Qu.: 90.27100
 Max.   :199.99000
```

The plots were now:

Once again, to get any useful data, we need to remove data above 150 and below 15.

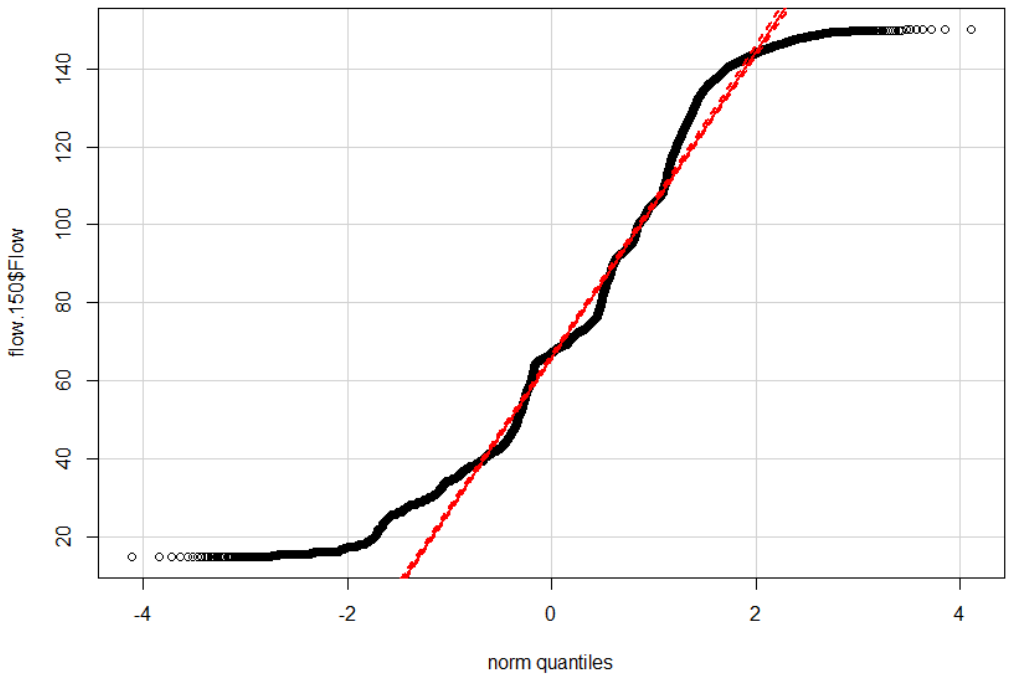
```
> summary(flow.150)
      Flow
 Min.   : 15.00
 1st Qu.: 39.55
 Median : 67.19
 Mean   : 69.03
 3rd Qu.: 92.50
 Max.   :150.00
```

Now we can analyze a distribution of some sort. The mean and the median are more closely matched, indicating fewer outliers (at least in the context of this portion of the dataset). To confirm, we plot.




```
> boxplot(flow.150$Flow)
```

I want to check if these measurements are normally distributed:



The data is clearly not distributed, especially around the tails. I would be inclined to further shrink the data set to check if any portion is normally distributed.

As a summary to my manager:

- we have a lot of noise in our data.
- The majority of the flow values ranges between 15 and 150.
- Our median is around 67 after we filter out most of the noise
- Our IQR is around 53 after we filter out most of the noise
- The data could be analyzed in flow ranges.
- I would like to spend more time analyzing the data, since from this quick analysis I was not able to be confident about my solution. (same in the assignment context).