STA130 Fall 2019 – T0107

Week 2: Data Wrangling and R help

(Materials used in this presentation are provided by the U of T Statistical Sciences Department.

This presentation was prepared by Vivian Ngo.)

Github.com/viviannngo97/STA130-Fall-2019

viv.ngo@mail.utoronto.ca

Reminders

- Tutorials start 10 minutes after the hour

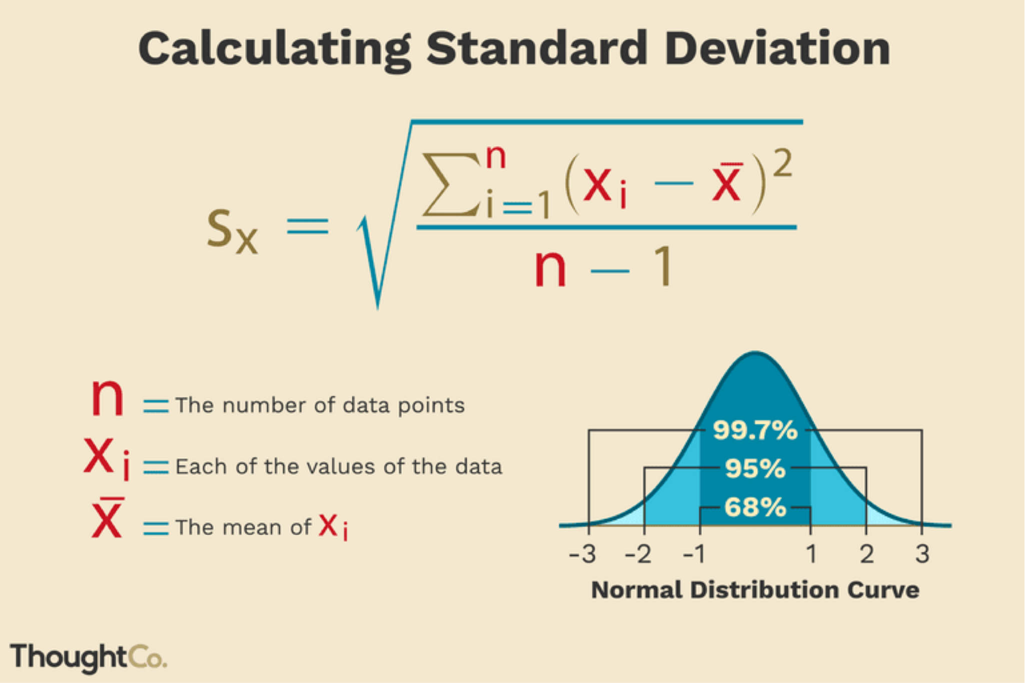
Agenda

- Vocabulary
- Group Discussion: Homework Question 1
- Writing Activity
- Mentorship Program

Vocabulary for this week's material

- Mean, average
- Median
- Standard deviation
- Variance
- Boxplot
- Interquartile range
- Quartile
- Proportion
- Outlier
- R object
- Vector
- Types of variables: e.g. character, numeric, logical
- Data frame
- Summary table, summary statistics

Vocabulary



Calculating Standard Deviation

$$S_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

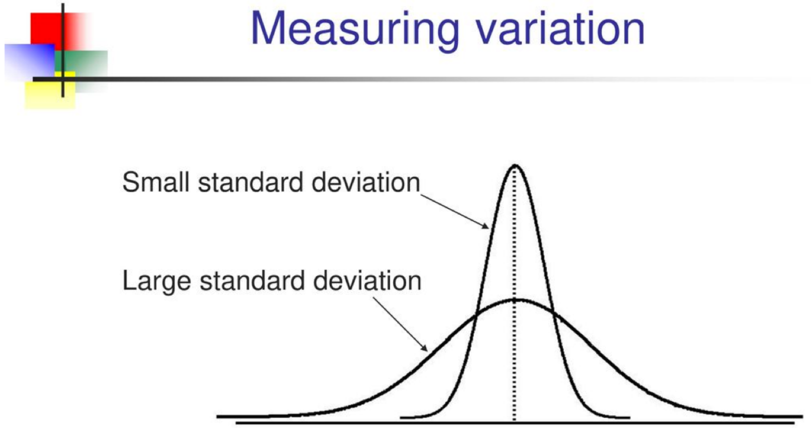
n = The number of data points

x_i = Each of the values of the data

$\bar{x}$ = The mean of x_i

Normal Distribution Curve

ThoughtCo.

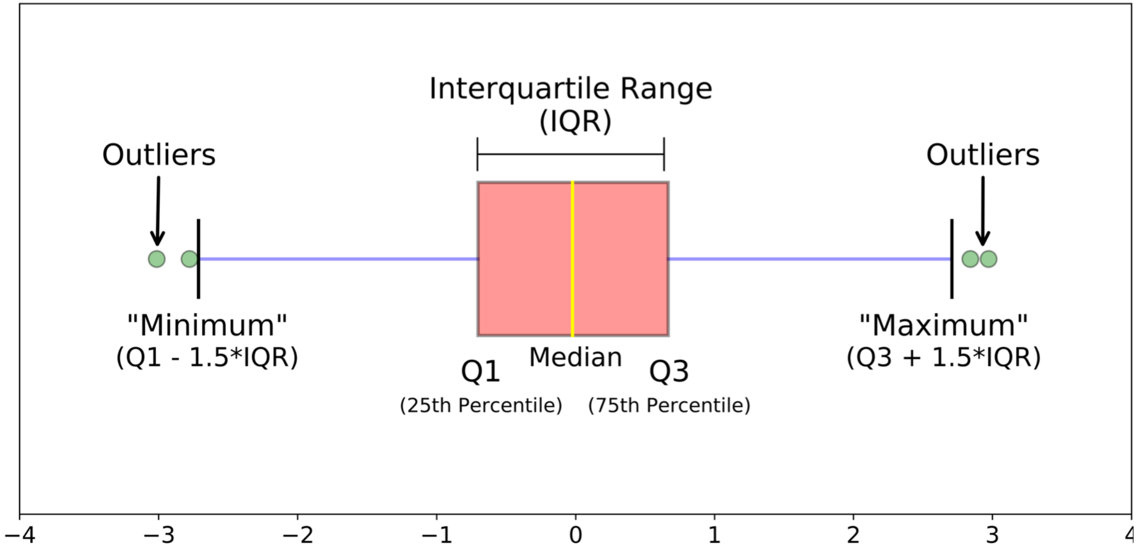


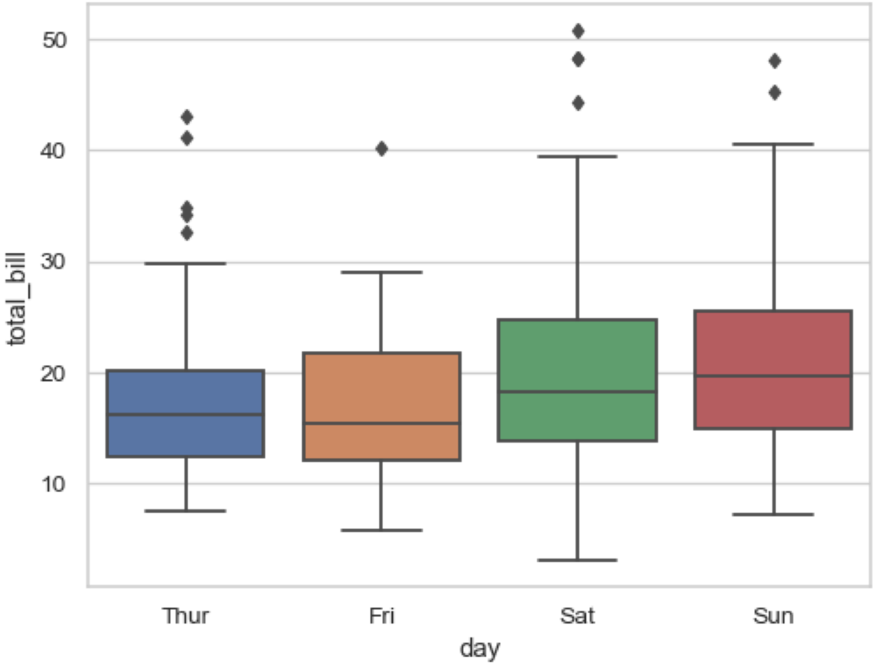
Measuring variation

- How far, on average, does our data deviate from the mean?
- Tells us about the spread
- Smaller standard deviation -> less spread

- Mean, average
- Median
- **Standard deviation**
- Variance
- Boxplot
- Interquartile range
- Quartile
- Proportion
- Outlier
- R object
- Vector
- Types of variables: e.g. character, numeric, logical
- Data frame
- Summary table, summary statistics

Vocabulary



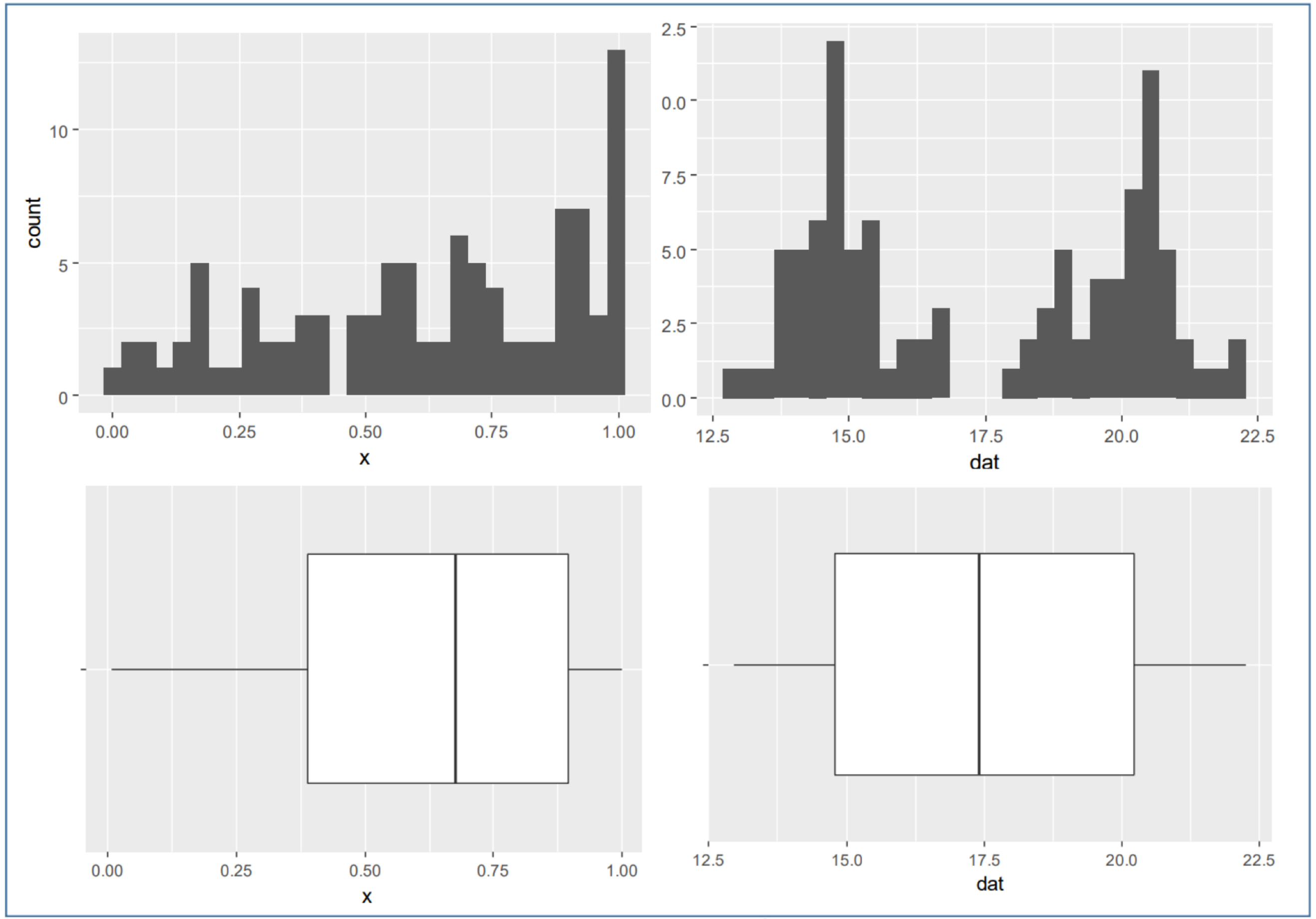


- Mean, average
- Median
- Standard deviation
- Variance
- **Boxplot**
- Interquartile range
- Quartile
- Proportion
- Outlier
- R object
- Vector
- Types of variables: e.g. character, numeric, logical
- Data frame
- Summary table, summary statistics

- When to use boxplots?
  - To summarize the distribution of a quantitative variable
  - To compare distributions or summarize based on a categorical variable of interest
- Examples?

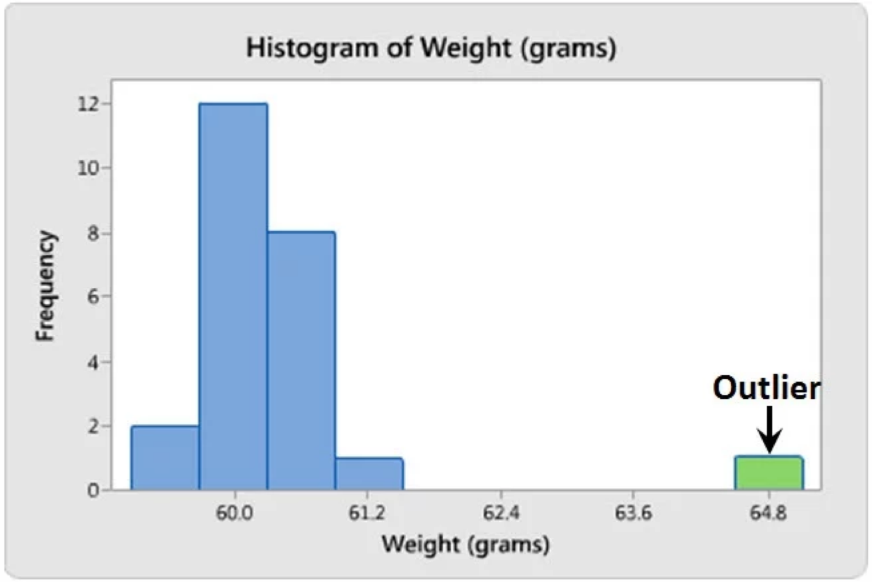
Vocabulary

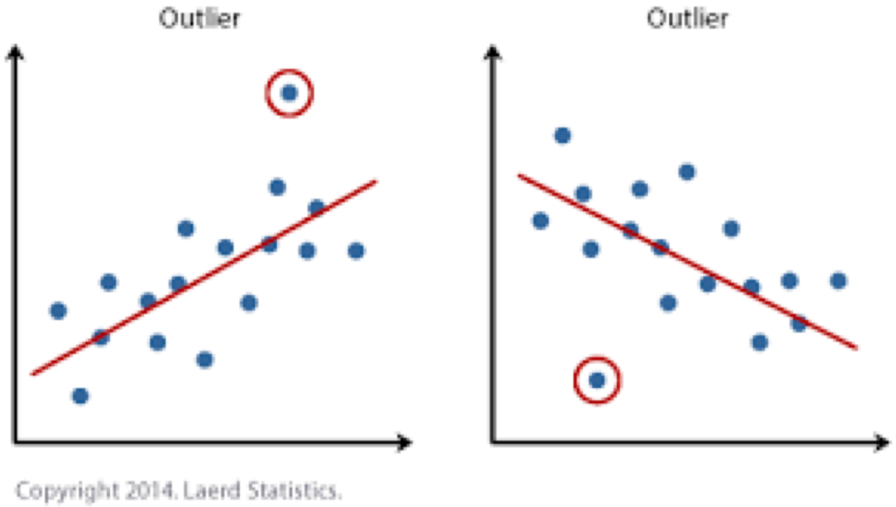
Boxplots vs Histograms



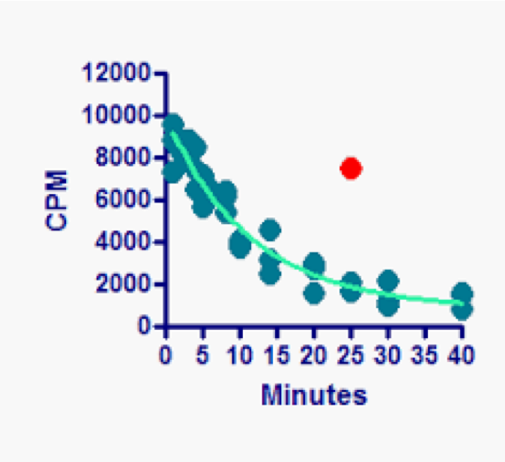
- Mean, average
- Median
- Standard deviation
- Variance
- **Boxplot**
- Interquartile range
- Quartile
- Proportion
- Outlier
- R object
- Vector
- Types of variables: e.g. character, numeric, logical
- Data frame
- Summary table, summary statistics

Vocabulary



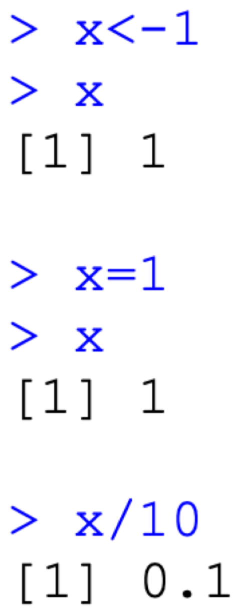


Copyright 2014. Laerd Statistics.



- Mean, average
- Median
- Standard deviation
- Variance
- Boxplot
- Interquartile range
- Quartile
- Proportion
- **Outlier**
- R object
- Vector
- Types of variables: e.g. character, numeric, logical
- Data frame
- Summary table, summary statistics

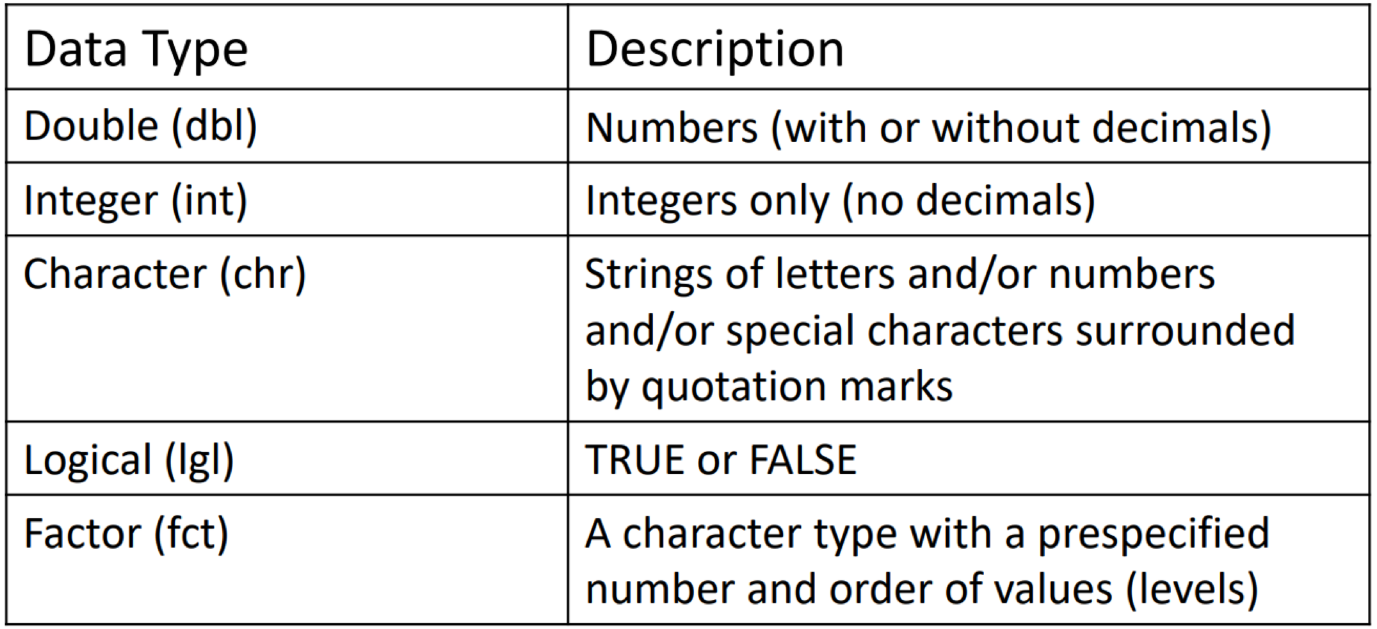
Vocabulary



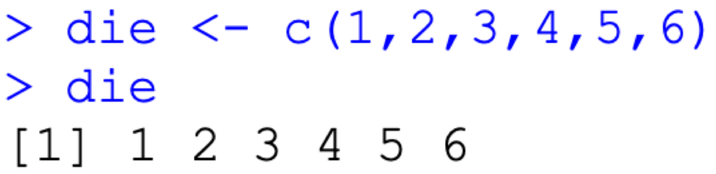
```
> x<-1
> x
[1] 1

> x=1
> x
[1] 1

> x/10
[1] 0.1
```



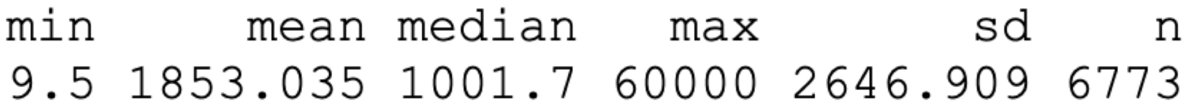
Data Type	Description
Double (dbl)	Numbers (with or without decimals)
Integer (int)	Integers only (no decimals)
Character (chr)	Strings of letters and/or numbers and/or special characters surrounded by quotation marks
Logical (lgl)	TRUE or FALSE
Factor (fct)	A character type with a prespecified number and order of values (levels)



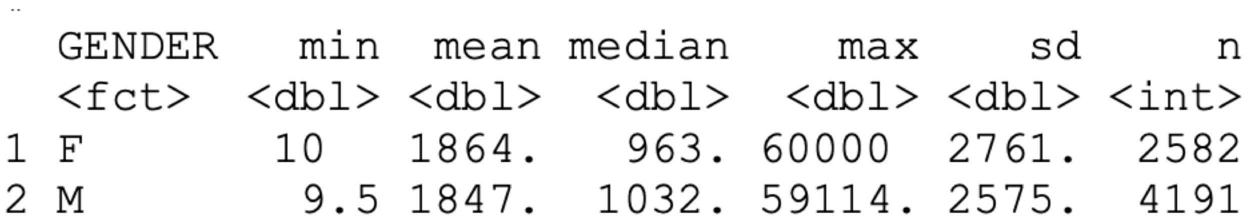
```
> die <- c(1,2,3,4,5,6)
> die
[1] 1 2 3 4 5 6
```

- Mean, average
- Median
- Standard deviation
- Variance
- Boxplot
- Interquartile range
- Quartile
- Proportion
- Outlier
- **R object**
- Vector
- Types of variables: e.g. character, numeric, logical
- Data frame
- Summary table, summary statistics

Vocabulary



```
 min     mean median   max       sd    n
 9.5 1853.035 1001.7 60000 2646.909 6773
```

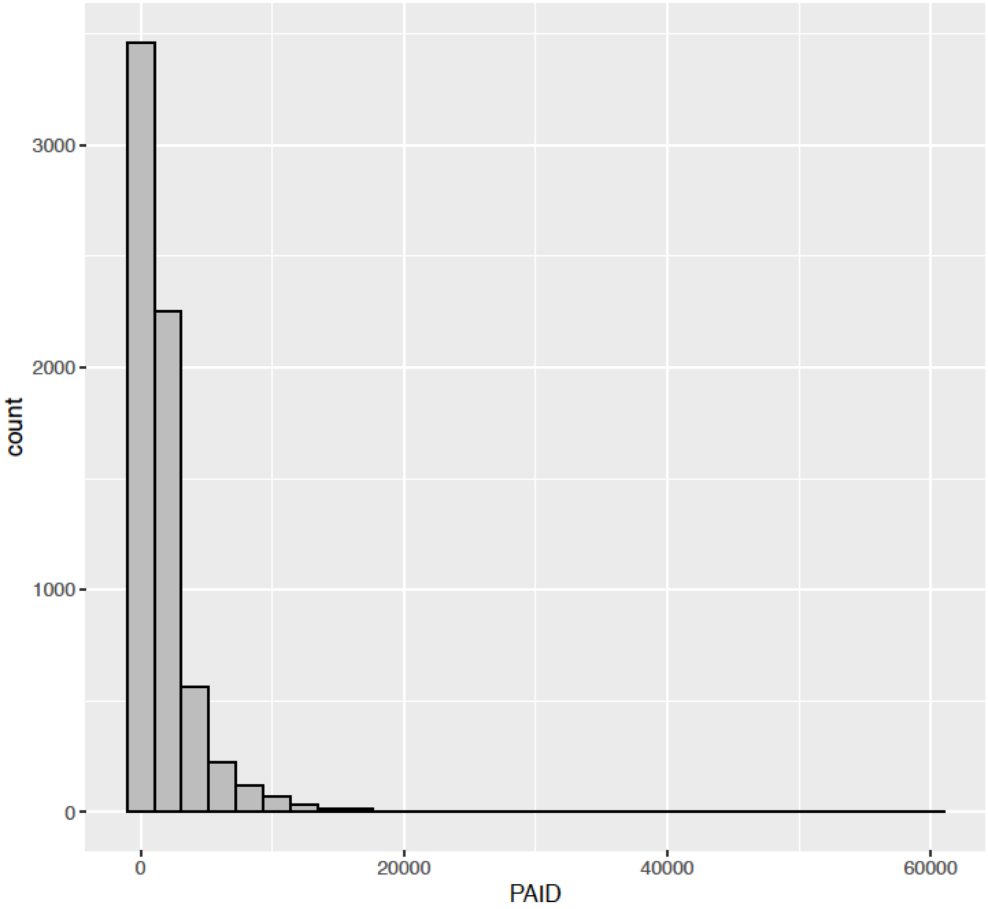


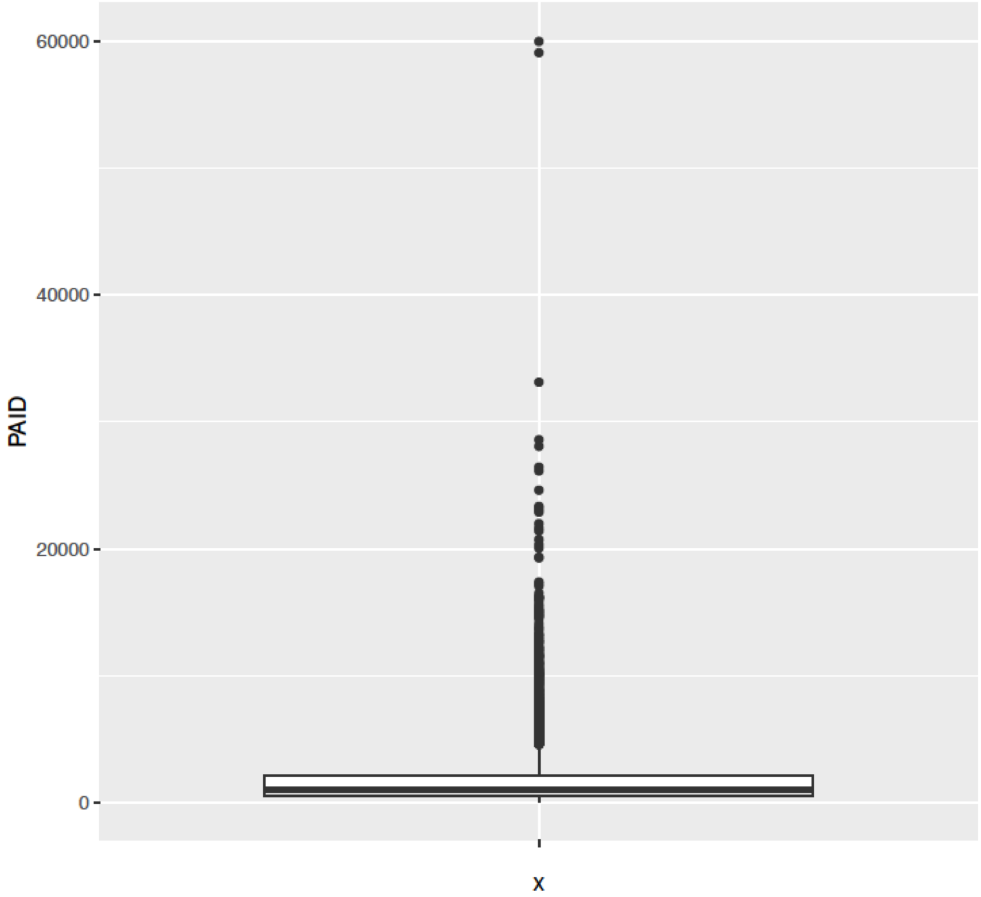
```
  GENDER   min  mean median   max    sd     n
  <fct>  <dbl> <dbl>  <dbl> <dbl> <dbl> <int>
1 F         10 1864.   963. 60000 2761.  2582
2 M        9.5 1847.  1032. 59114. 2575.  4191
```

- Mean, average
- Median
- Standard deviation
- Variance
- Boxplot
- Interquartile range
- Quartile
- Proportion
- Outlier
- R object
- Vector
- Types of variables: e.g. character, numeric, logical
- Data frame
- **Summary table, summary statistics**

Group Discussion: Homework Question 1

- *For Question 1b, you used both histograms and boxplots to visualize your data. Which features were easier/harder to observe from each of the visualizations? In what situations may you want to choose a boxplot over a histogram, or vice versa? Explain.*





Writing Activity

Self-Reflection:

- *What questions, if any, do you have so far regarding the course materials?*
- *What is one of your favorite things about tutorial? Least favorite?*

Mentorship Program