INTRODUCTION TO STATISTICS FOR INDUSTRY

Dr. Jingyuan Zhao

About me (Dr. Jingyuan Zhao)

Working experience

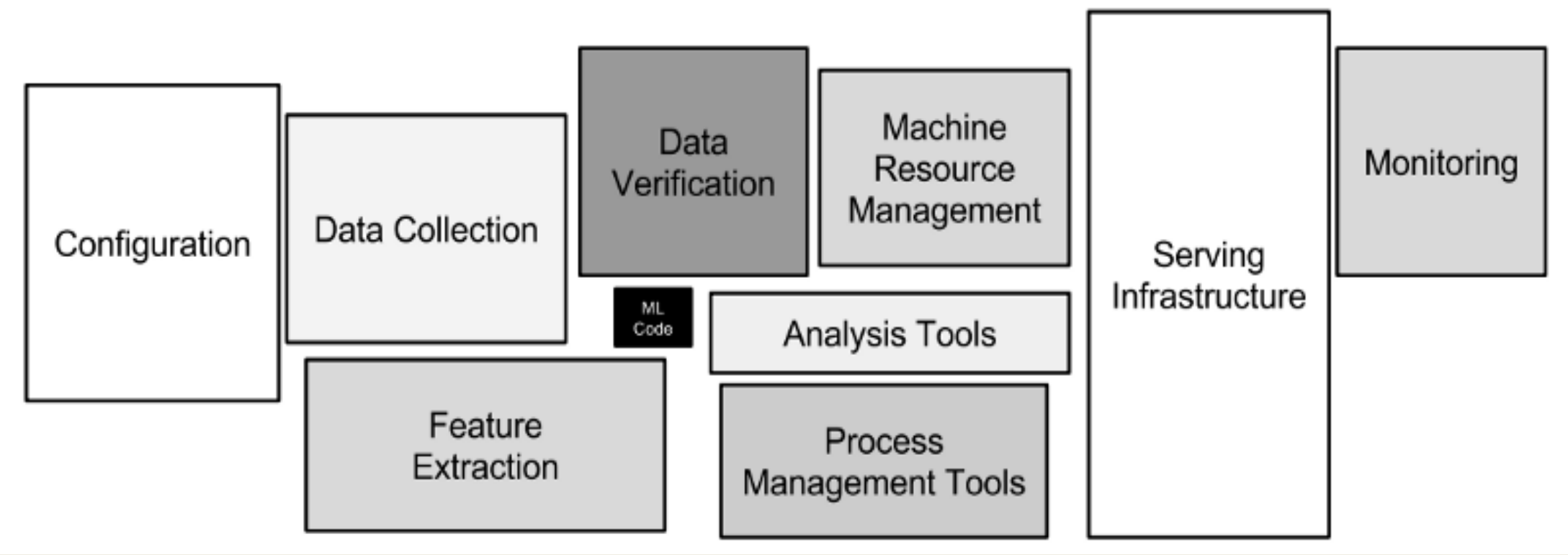
- Group Chief Data Officer, Great Eastern
- VP, AI & Analytics APAC, Capgemini
- SVP, head of Advanced Analytics Center, NTUC Enterprise
- VP, head of regional data science, Lazada, Alibaba group
- Manager of data science, global innovation center, Nielsen
- Postdoctoral scientist, Genome Institute of Singapore, A*Star

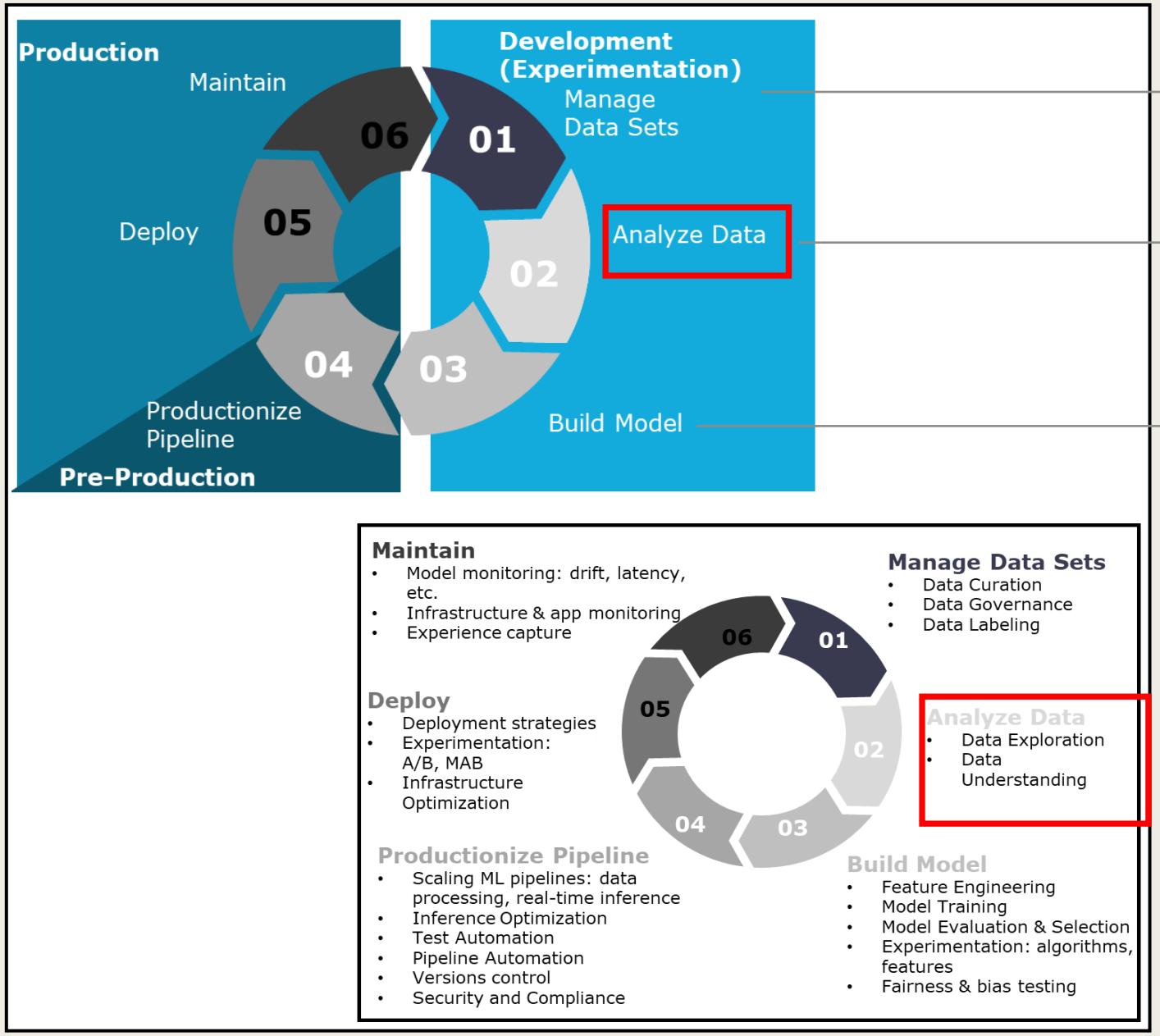
Education

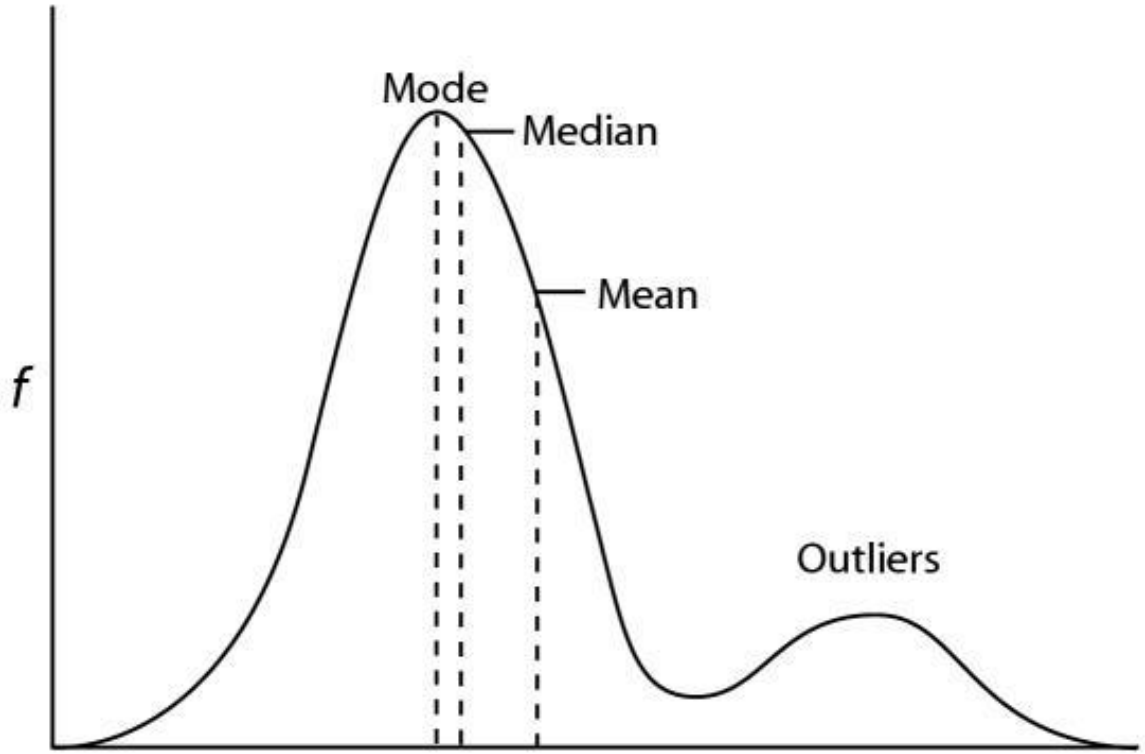
- PhD of Statistics, NUS (2004-2008)

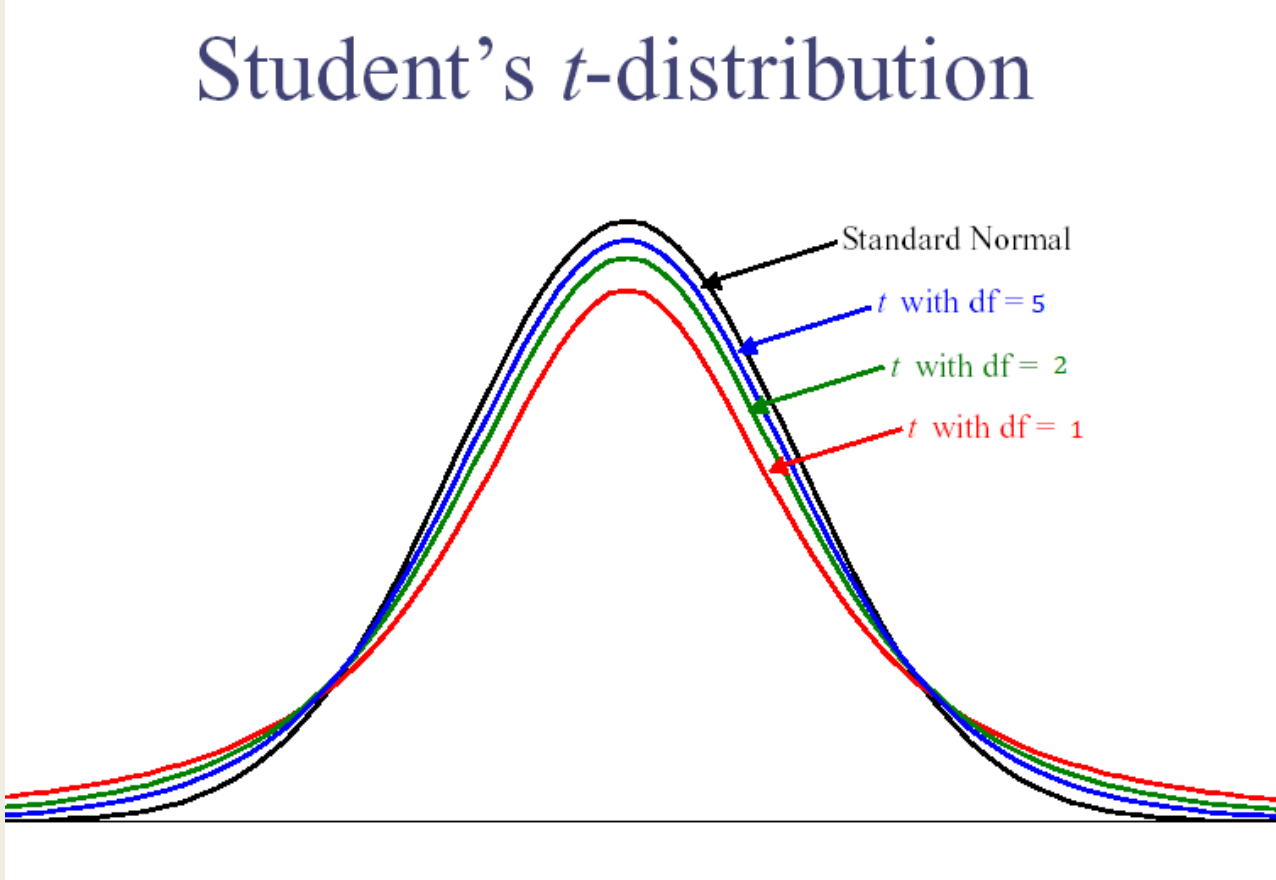
My interest End to end AI solution at scale starting from data strategy, AI roadmap, model development, deployment with big scale for immersive personalized customer experience, augmented operations and empowered employees.

Google: “Only a small fraction of a real-world ML system is composed of the ML code. The required surrounding elements are vast and complex”

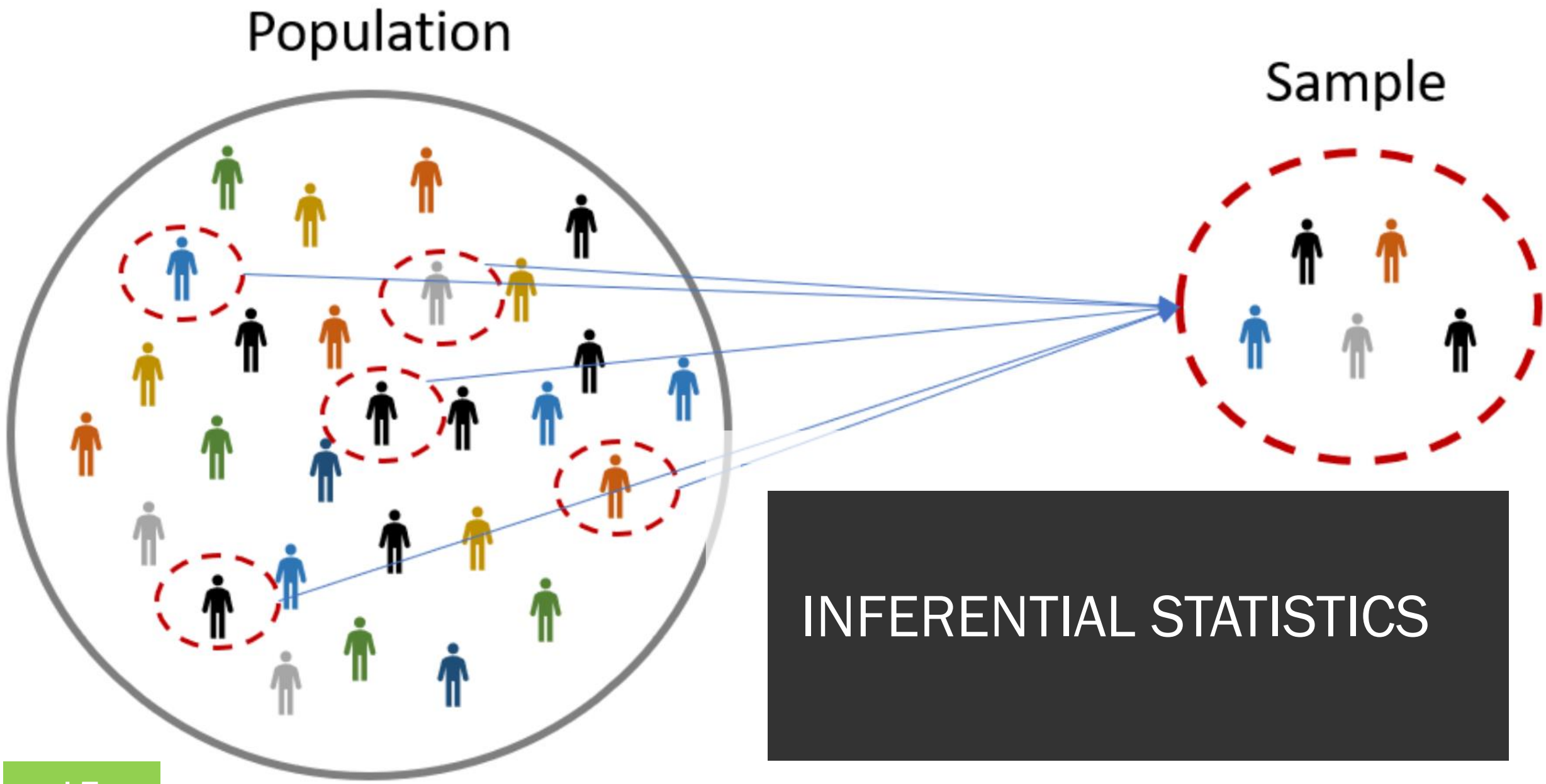



Production
Maintain
Development
(Experimentation)
Manage
Data Sets
06
01
Deploy
05
Analyze Data
02
04
03
Productionize
Pipeline
Build Model
Pre-Production
Lecture 1-2
Lecture 7-9
Lecture 3-6
Maintain
• Model monitoring: drift, latency, etc.
• Infrastructure & app monitoring
• Experience capture
Manage Data Sets
• Data Curation
• Data Governance
• Data Labeling
Deploy
• Deployment strategies
• Experimentation: A/B, MAB
• Infrastructure Optimization
Analyze Data
• Data Exploration
• Data Understanding
Productionize Pipeline
• Scaling ML pipelines: data processing, real-time inference
• Inference Optimization
• Test Automation
• Pipeline Automation
• Versions control
• Security and Compliance
Build Model
• Feature Engineering
• Model Training
• Model Evaluation & Selection
• Experimentation: algorithms, features
• Fairness & bias testing



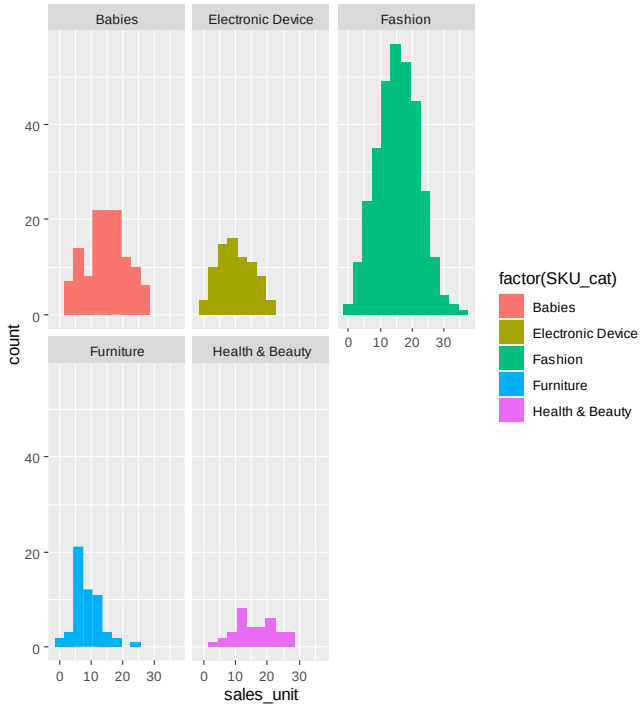


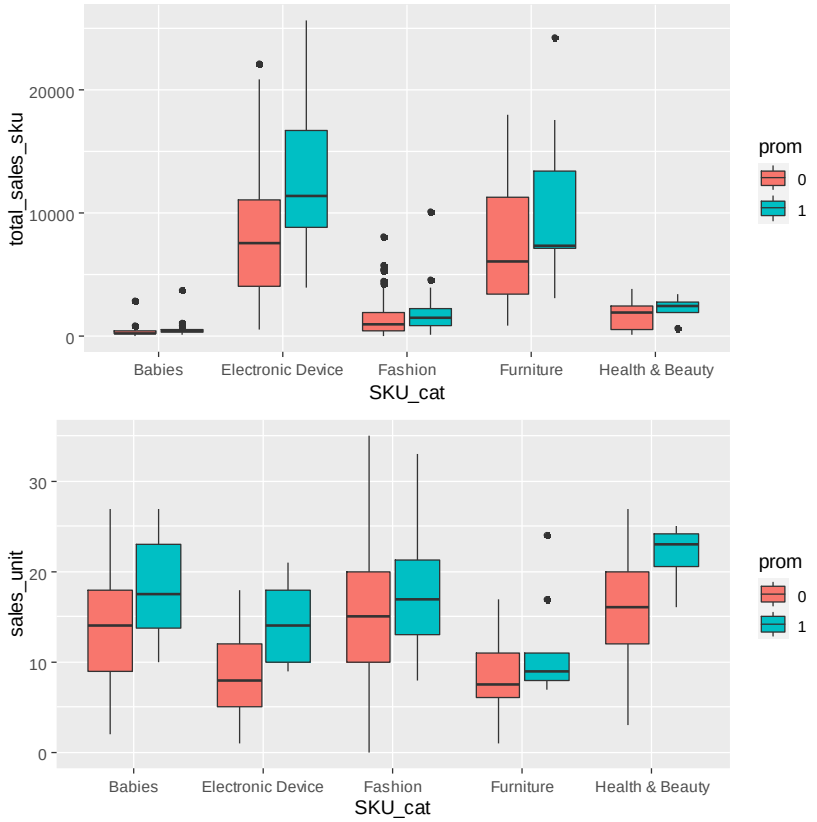
DESCRIPTIVE ANALYSIS



INFERENTIAL STATISTICS

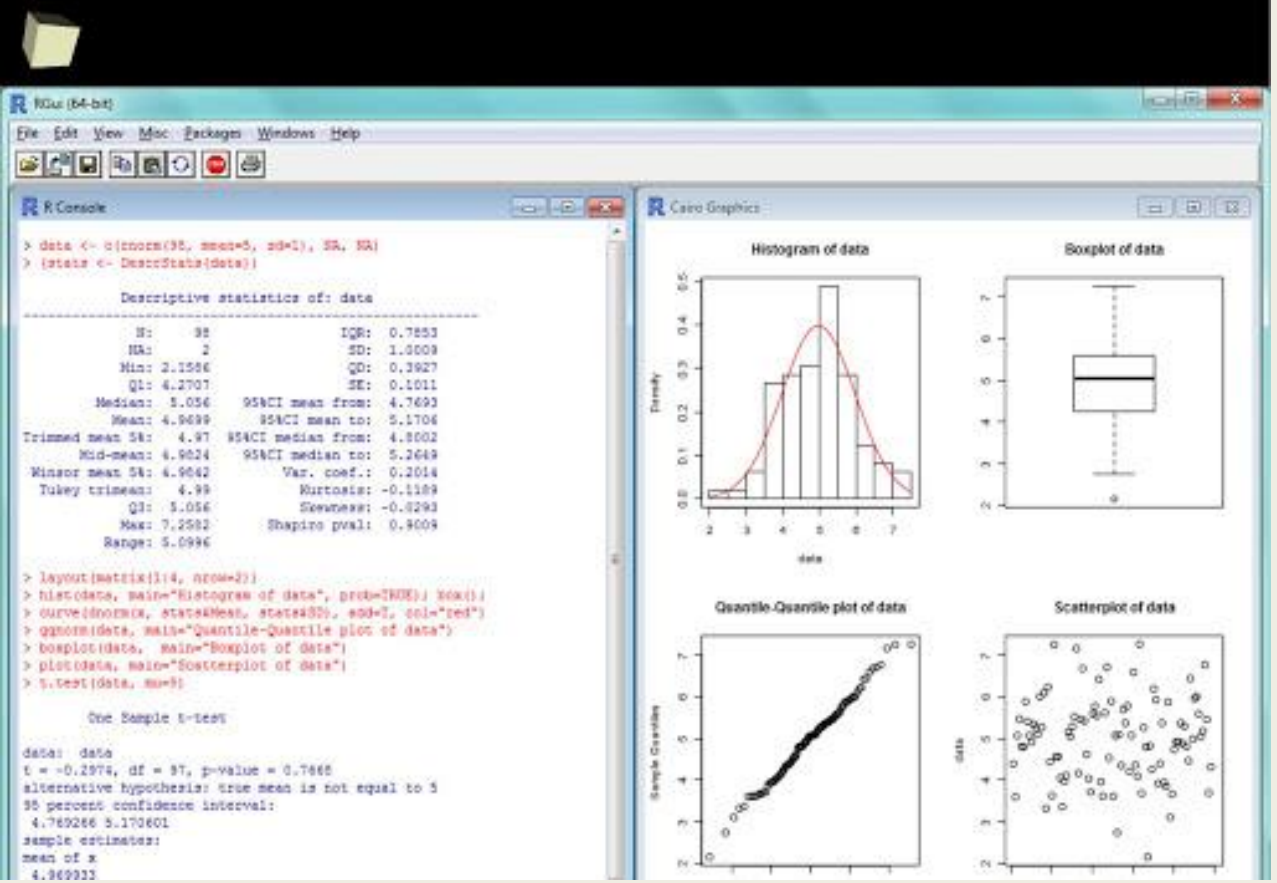
L7

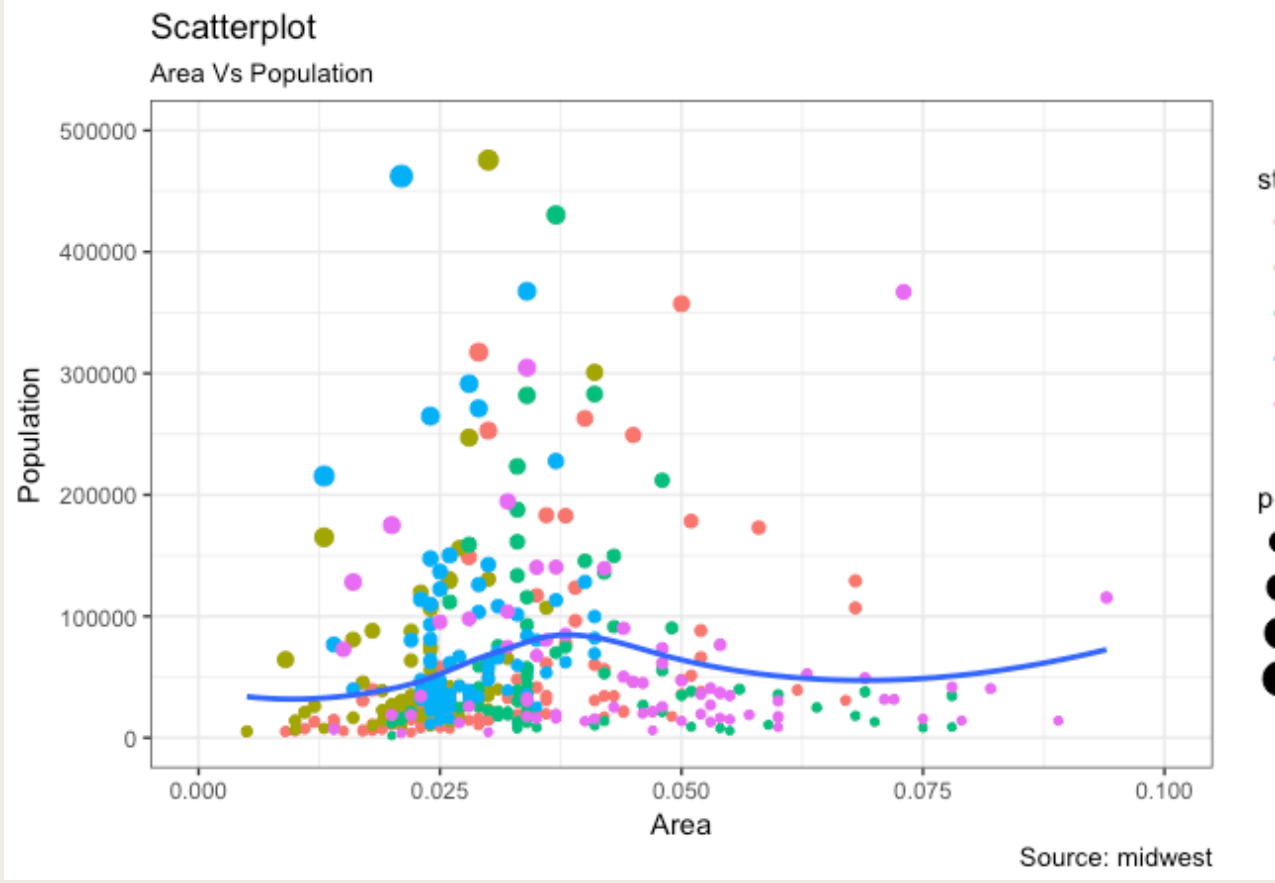




L7

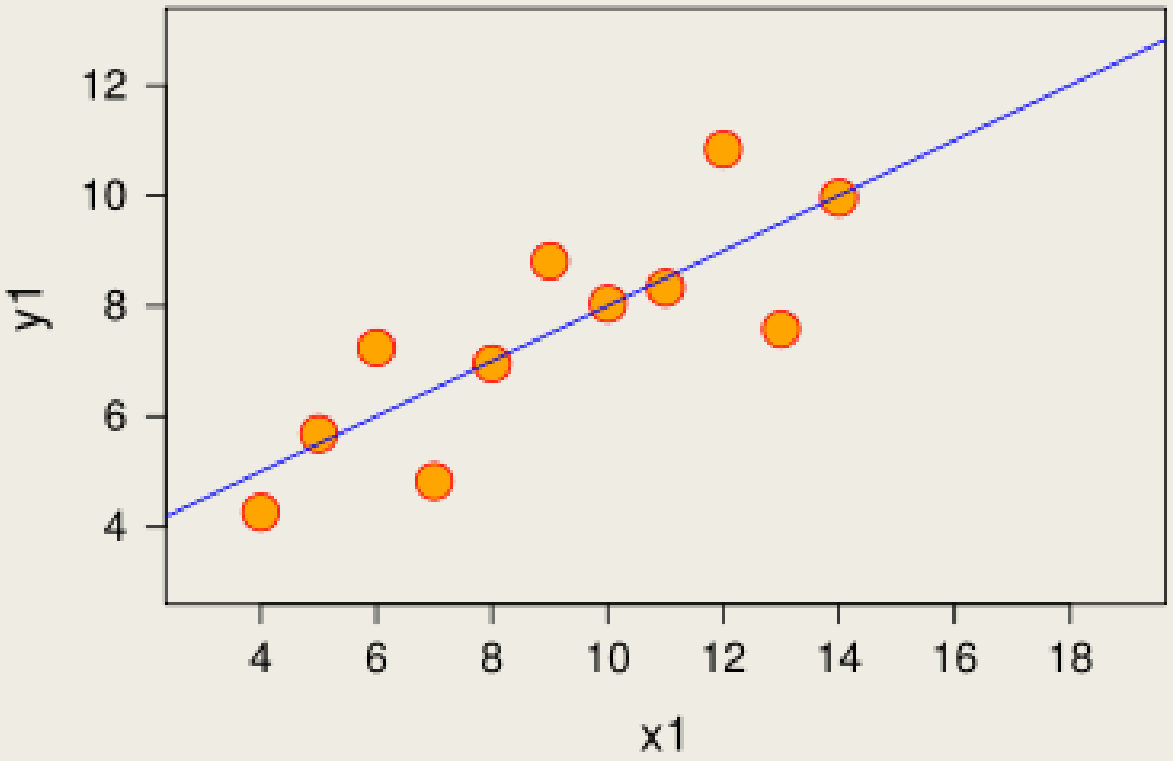
Univariate analysis

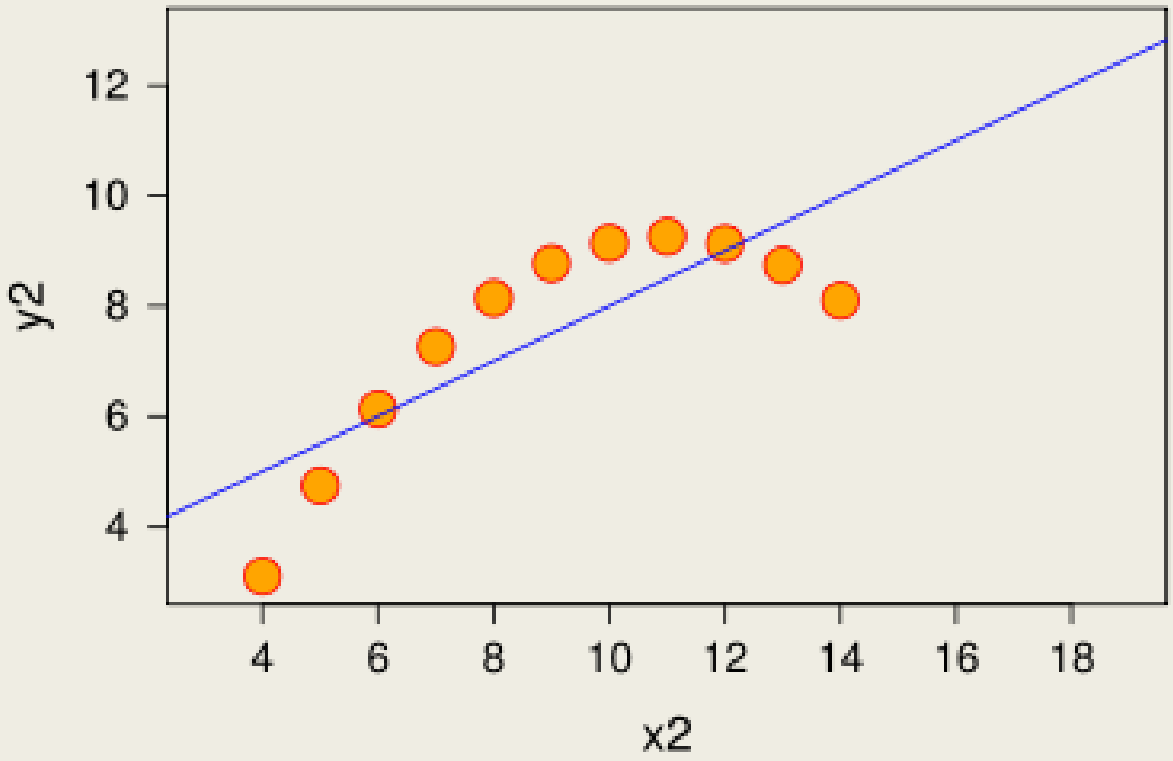


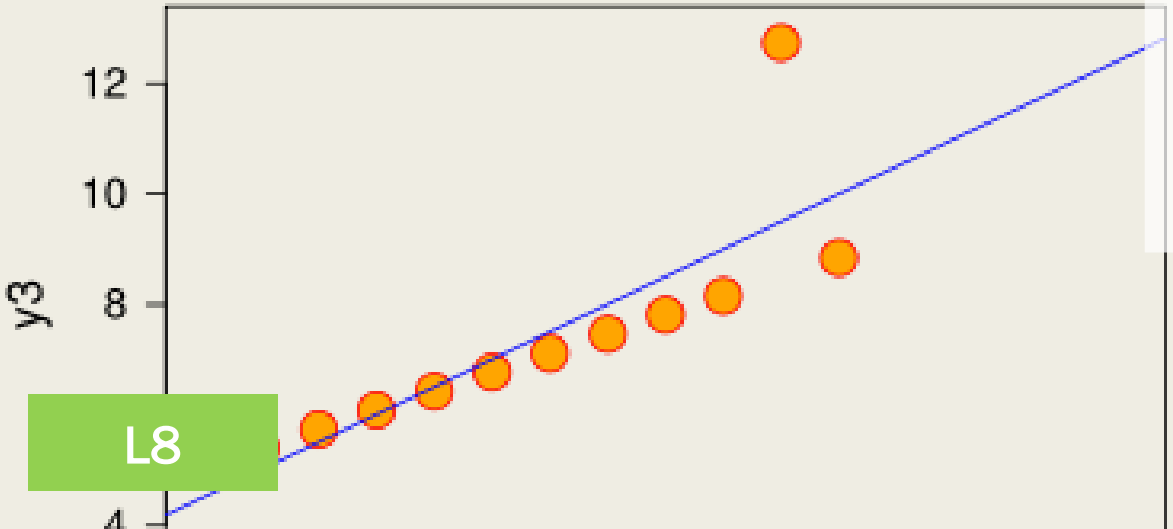


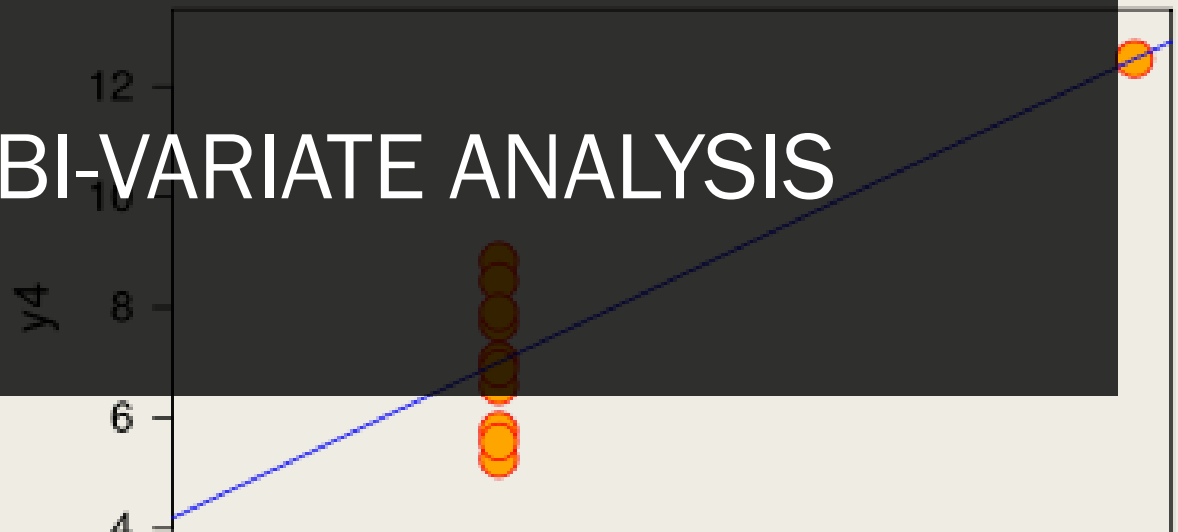
DATA ANALYSIS & VISUALIZATION IN R

L7, 8, 9





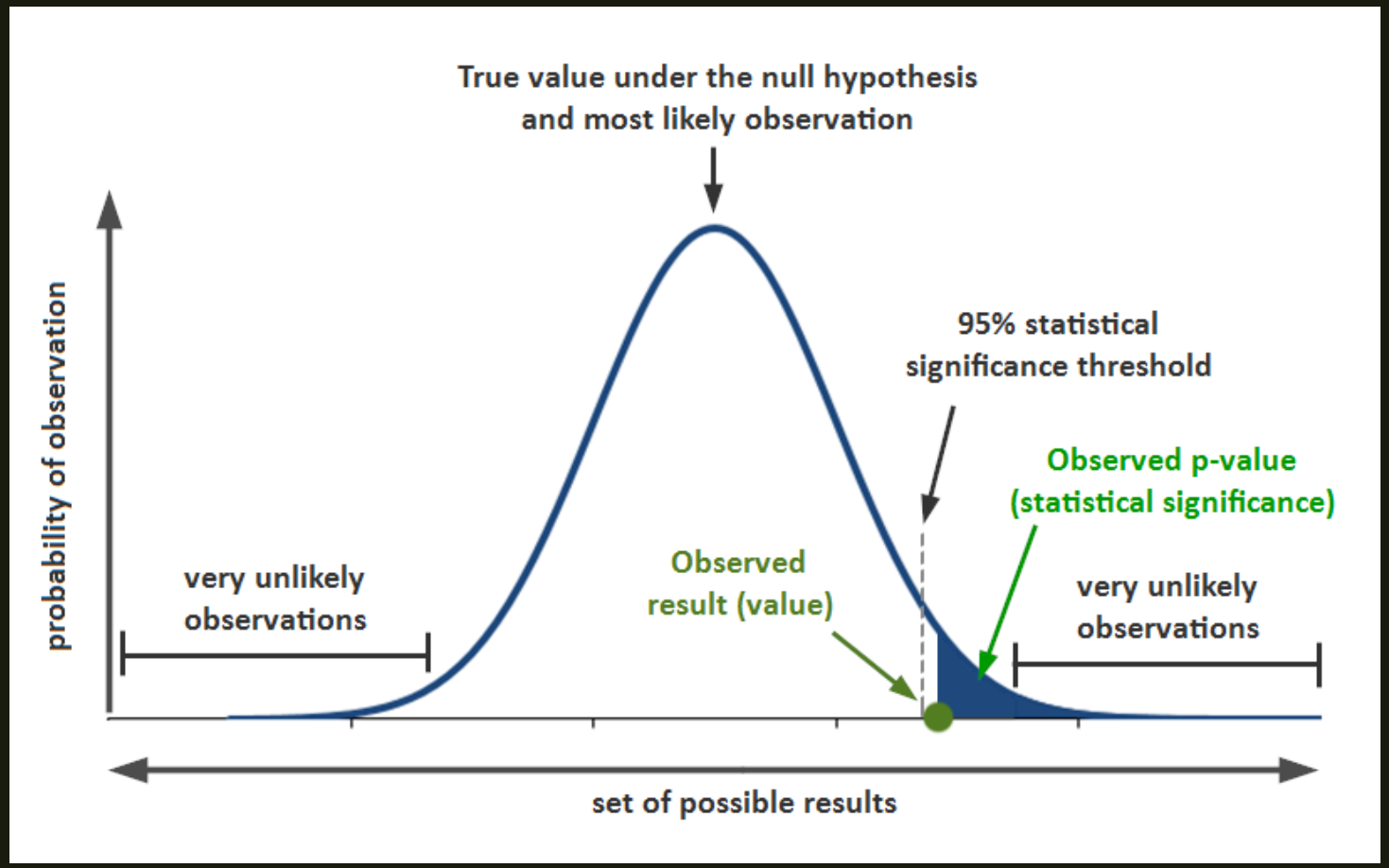




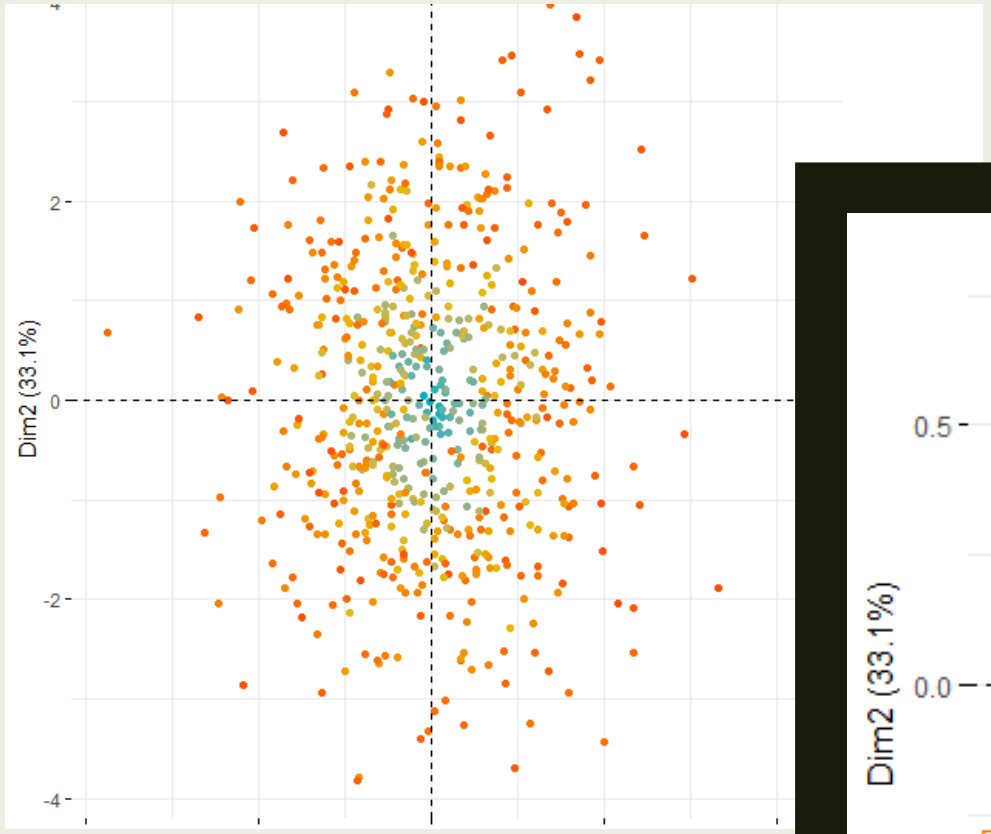
BI-VARIATE ANALYSIS

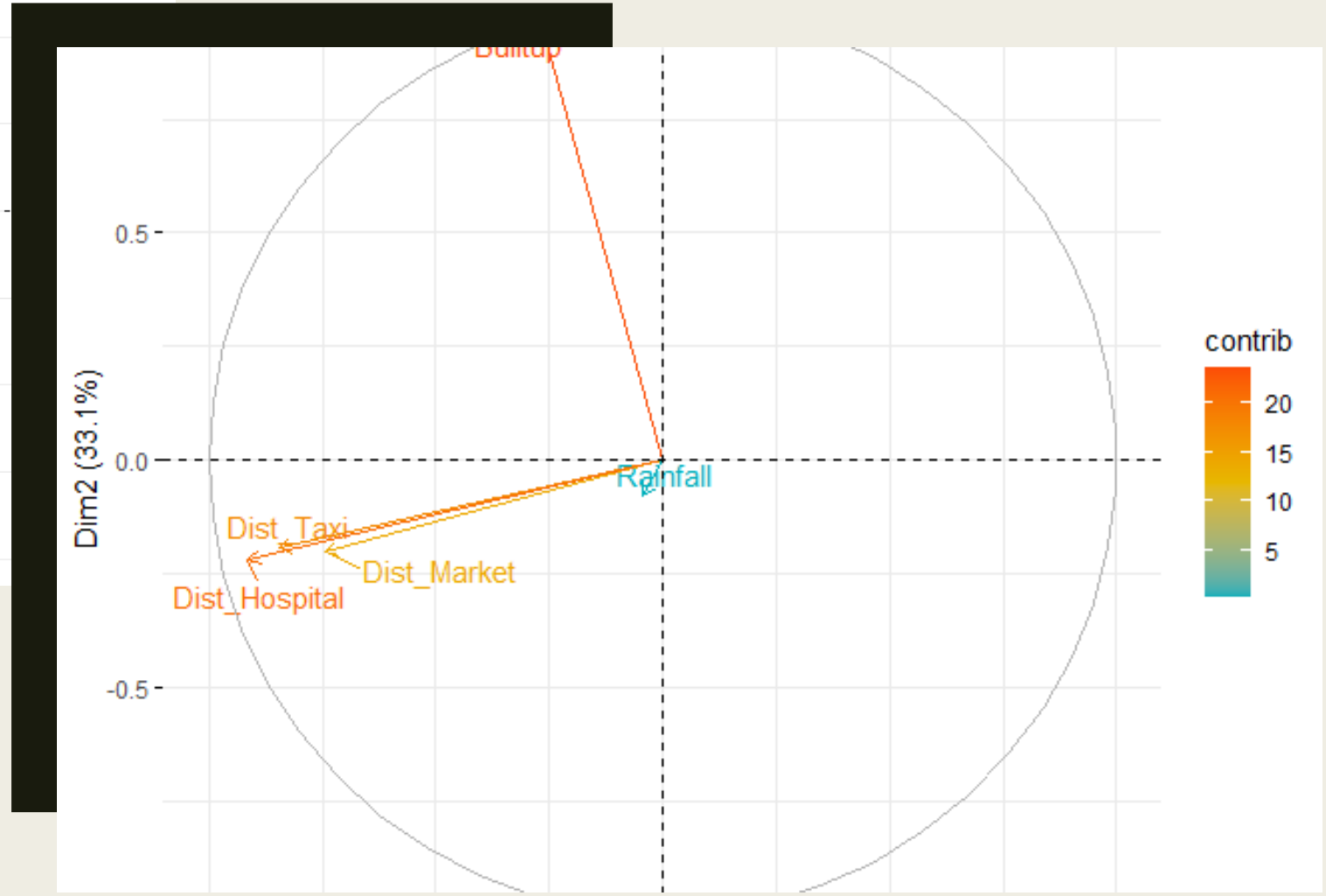
L8

HYPOTHESIS TESTING, SIGNIFICANCE INTERVAL



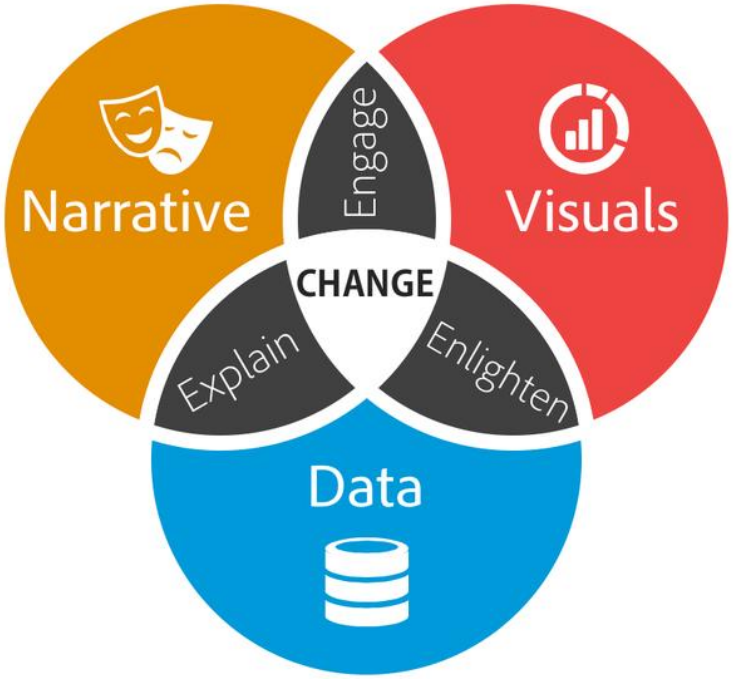
L8





L9

Multi-variate analysis



L9

Storytelling with Data: A key skill of data scientists

AGENDA

1. Two types of Statistics
2. Univariate analysis
3. Introduction to R
4. Visualization using ggplot2

Two Types of Statistics

Statistics

- ***Statistics*** is “ the science of data involving collecting, classifying, summarizing, organizing, analyzing and interpreting numerical information” –McClave, Dietrich, Sincich.

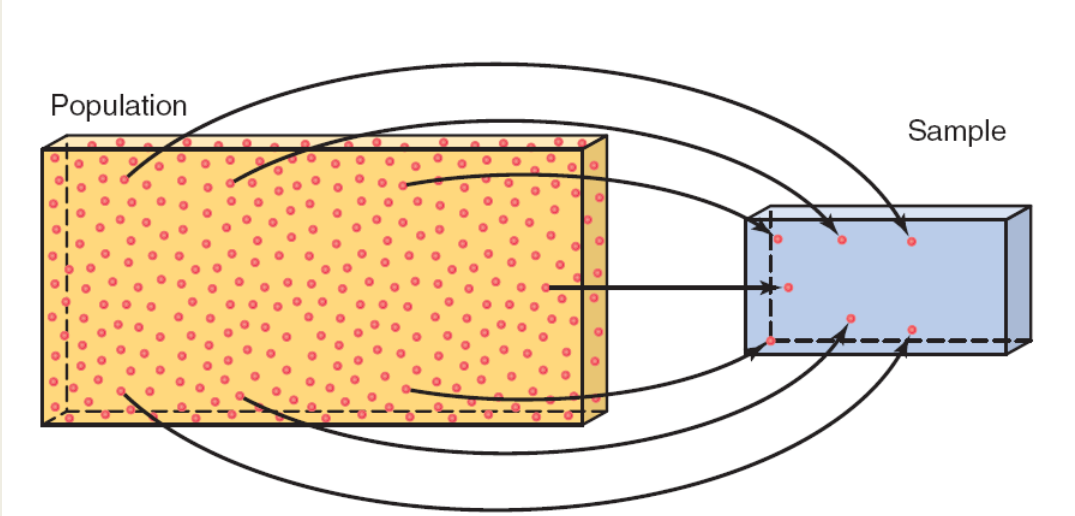
- Collecting: questions asked, sample picked, choice of geometric location
- Classifying: how is the data grouped
- Presenting: how the data is presented, type of graphs..
- Interpreting: with or without bias, predetermined.

Population

- The entire group of individuals is called the **population**.
- For example, a researcher may be interested in the relation between class size (variable 1) and academic performance (variable 2) for the population of third-grade children.

Sample

- Usually populations are so large that a researcher cannot examine the entire group. Therefore, a **sample** is selected to represent the population in a research study. The goal is to use the results obtained from the sample to help answer questions about the population.



Example of a Sample and a Population

- Population: All babies born in NC in 2004
- Sample: These six babies
- Variables: Weight, Gender, Mother Smoked?

Birth data from North Carolina 2004

Weight	Gender	Smoke
7.69	F	0
5.88	M	1
6.00	F	0
7.19	F	0
8.06	F	0
7.94	F	0

Two areas of statistics

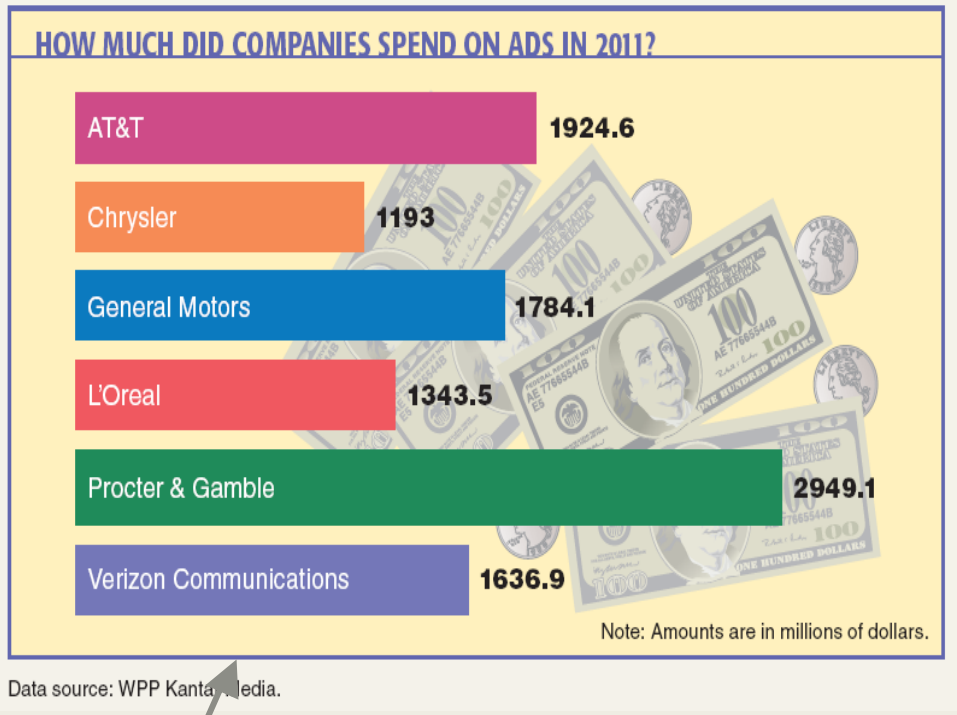
Both descriptive and inferential statistics help make sense out of row after row of data!

Descriptive Statistics: collection, presentation, and description. “Data speaks for itself”.

Inferential Statistics: making decisions and drawing conclusions about populations. “technique of interpreting”….

Descriptive Statistics

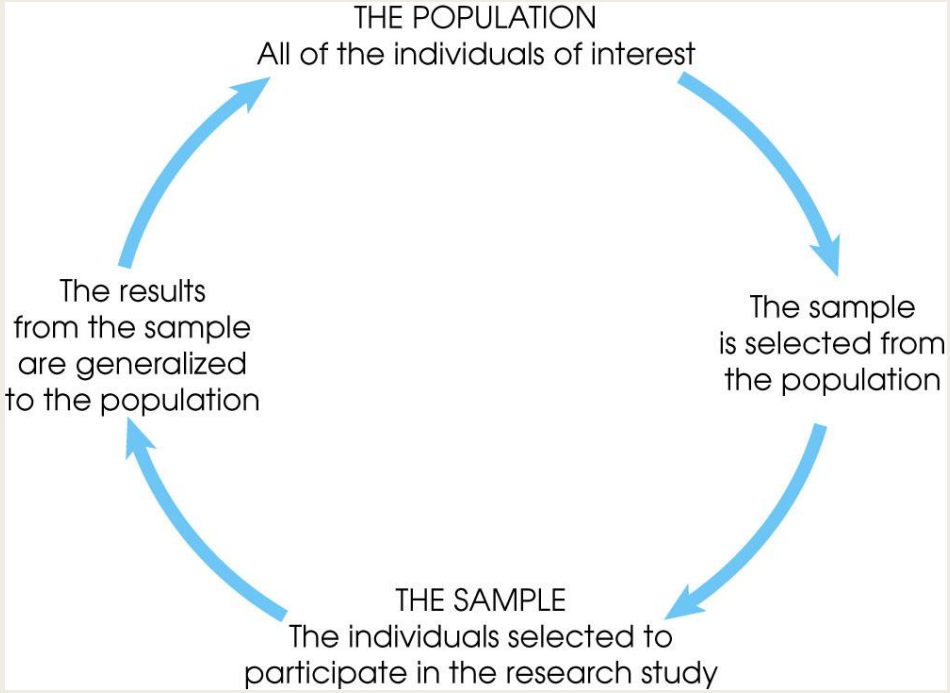
- Take a group that you're interested in, record data about the group members, and then use summary statistics and graphs to present the group properties.
- Use descriptive statistics to summarize and graph the data for a group that you choose to gain more insights and visualize data than raw data.
- With descriptive statistics, there is no uncertainty because you are describing only the people or items that you actually measure. You're not trying to infer properties about a larger population.



Description ONLY!

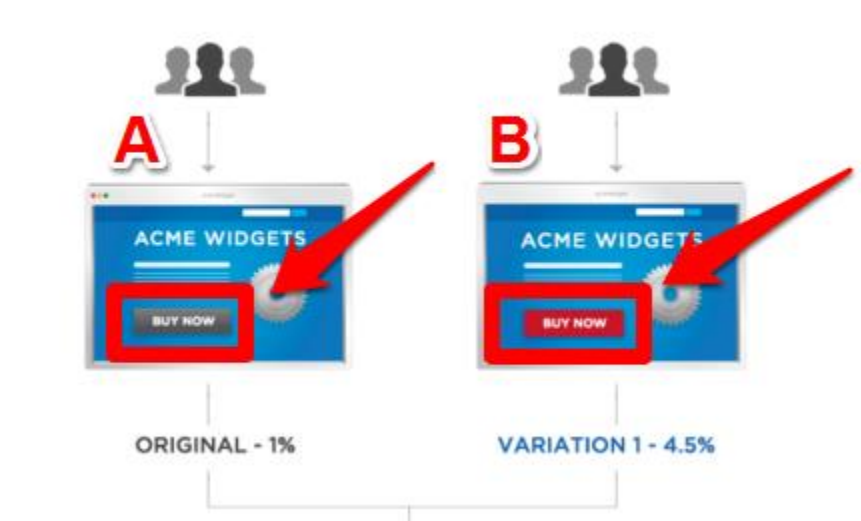
Inferential Statistics

Inferential Statistics makes inferences and predictions about a population based on a sample of data taken from the population in question.



Example: A/B testing in Ecommerce industry

How to use inferential statistics to conclude A/B testing results



Inferential statistics

https://www.onlinesalesguidetip.com/the-ultimate-split-testing-guide-for-ecommerce-sites/

Use the response of samples (selected shoppers) to infer the conversion rate & improvement of population (all website shoppers)

Key differences between inferential and descriptive statistics

Inferential Statistics	Descriptive Statistics
Using sample data to make an inference or draw a conclusion of the population	Organizing and summarizing data using numbers and graphs
The objective is to draw conclusion of the population data	Describe the characteristics of the sample or population
Drawing conclusions, performing estimations and making predictions	Collection, organizing, summarizing, presenting the data
Form of results- probability score	Charts, Graphs and Tables
Tools- Hypothesis test, ANOVA	Measure of tendency, Measure of dispersion
Use when the population data set is large	Data set is small

https://sixsigmastudyguide.com/kinds-of-statistics/

Univariate analysis

Univariate analysis

- **Central Tendency**

  Mean: adding all of the numbers together and dividing by the number of items in the set

  Median: ordering the set from lowest to highest and finding the exact middle.

  Mode: the most common number in a set

- **Disperse**

  Range, min, max, quantile, standard deviation

The Range

- The range is the distance spanned by the entire data set.
- Range = Maximum - Minimum
- The range is easy to calculate, but is subject to peculiarities of the data set and is very sensitive to outliers.
- A smaller sample size is likely to produce a smaller range. The range of a sample is a poor predictor of the range for the population.

The Formula for the Standard Deviation

- To calculate the standard deviation, use the formula:

$$s = \sqrt{\frac{\text{å}\,(x - \bar{x})^2}{n - 1}}$$

  - Σ, *read "sigma", means "add".*
  - x *represents all of the data values.*
  - n *represents the sample size.*
  - $\bar{x}$ *represents the sample mean.*

Quartiles

- The First Quartile (Q1) is the value such that 25% of the data lie at or below this value.
- Q1 is roughly the median of the lower half of the data.
- The Third Quartile (Q3) is the value such that 75% of the data lie at or below this value.
- Q3 is roughly the median of the upper half of the data.

The Interquartile Range (IQR)

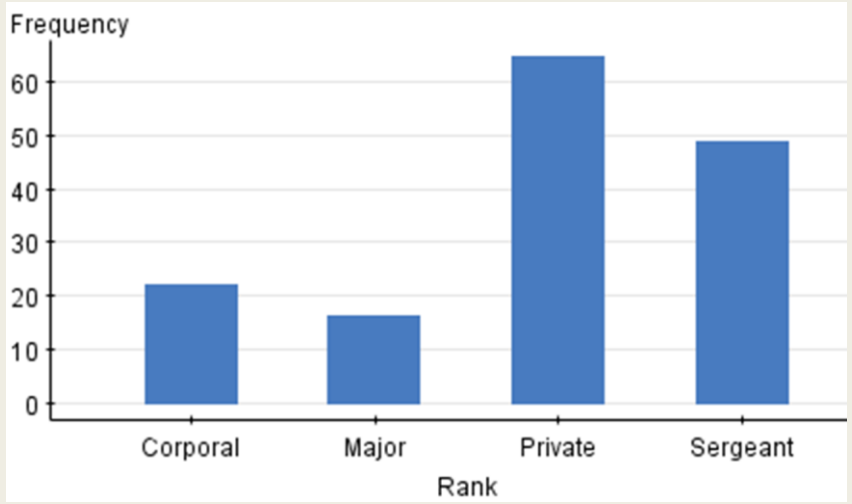
- The Interquartile Range (IQR) represents the range of the middle 50% of the data.
- Cut the ordered data into four equal parts. The distance taken up by the middle two parts is the interquartile range.
- IQR = Q3 – Q1

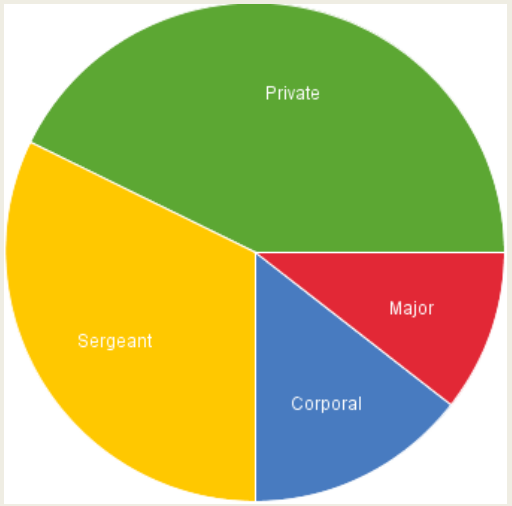
Visualizing Statistics

- Organize the data using the chart that most effectively visually summarizes the data.
- The **distribution** of the data describes the values, frequencies (counts), and "shape" of the data.
  - *Is there a data value or data values that are far from the rest of the data?*
  - *Is there symmetry?*
  - *Is there a most common value or most common range of values?*

Two Types of Charts for Categorical Data

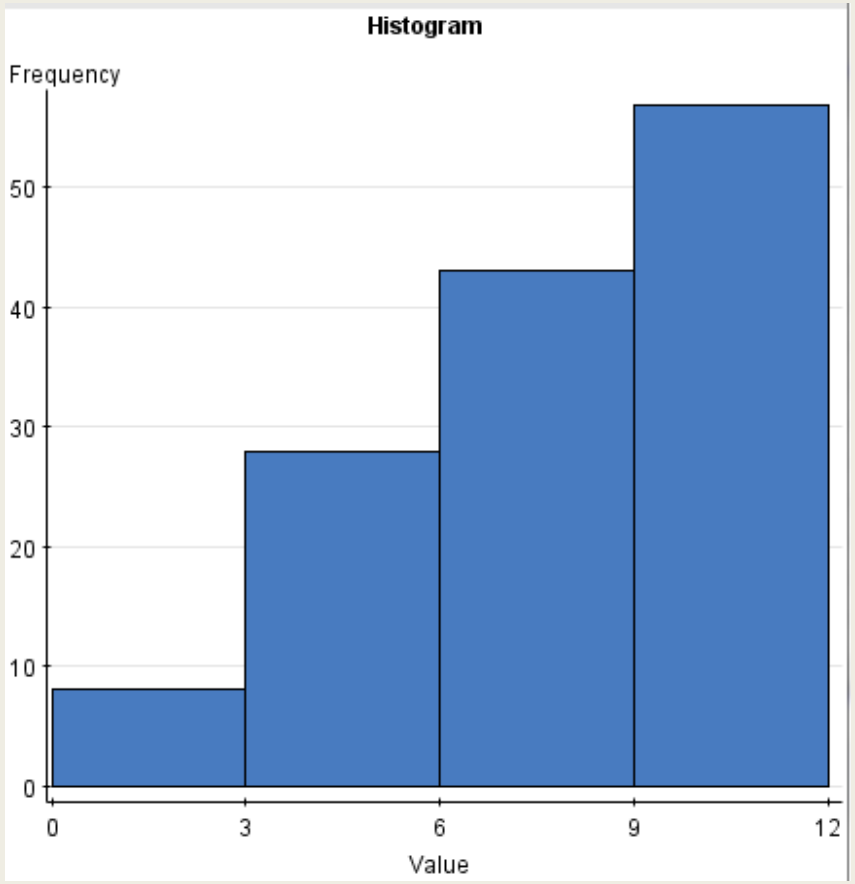
- A Bar Chart is like a histogram, but the horizontal axis can represent categorical data. A natural order may not occur.
- A Pie Chart is a circle cut into slices where the size of each slice is proportional to the frequency of the outcome that it represents.



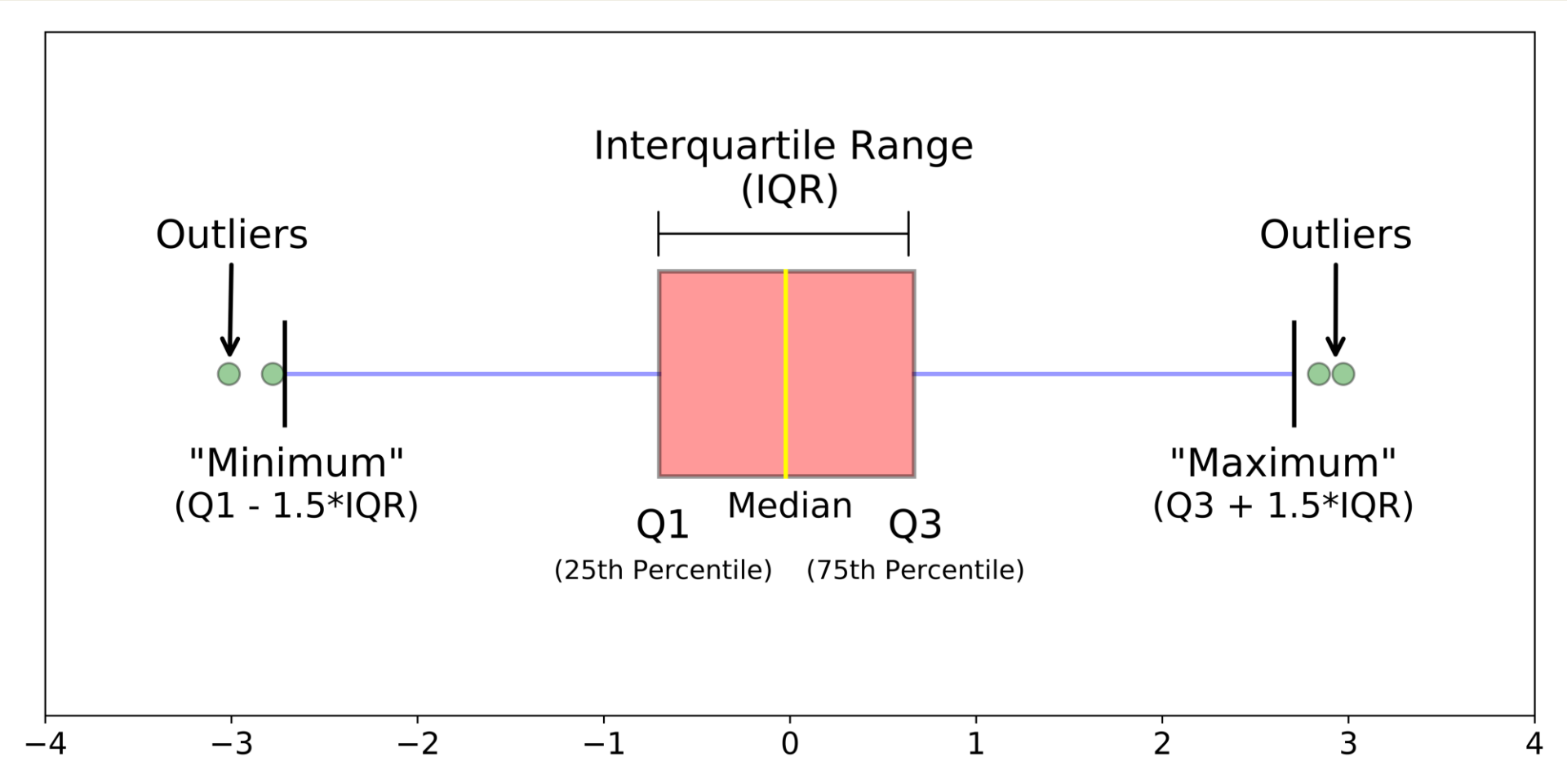


Frequency Histograms for Numerical Data

- A histogram is a type of bar graph.
- The horizontal axis is numerical.
- The vertical axis represents the frequency of the data.
- Groups the data into bins, also called intervals or classes.
- Easy to visualize the distribution.



Box Plot

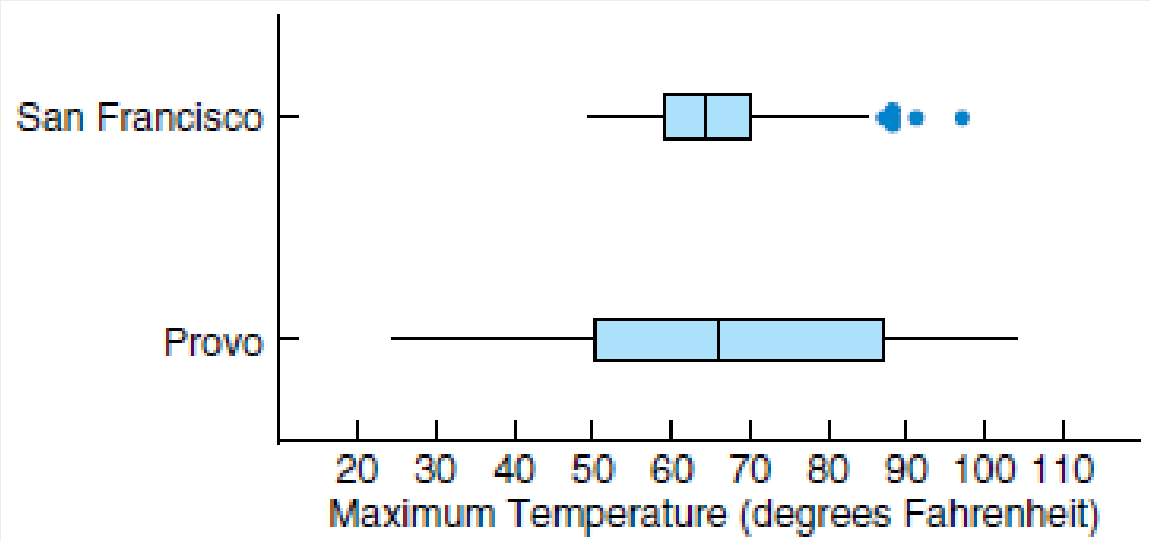


Source: https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

Boxplots

- A Boxplot is a chart that visually displays Q1, the median, Q3, and the potential outliers.
- To create a boxplot:
  1. *Plot the potential outliers*
  2. *Draw small vertical line segments at Q1, Q3, and the median.*
  3. *Draw a box with base from Q1 to Q3.*
  4. *Sketch horizontal line segments from the ends of the box to the smallest and largest values that are not potential outliers.*

Comparing Distributions with Boxplots



- Both cities have similar typical temperatures.
- Both cities have fairly symmetric distributions.
- Provo has a much greater variation in temperatures than San Francisco.

How to use Boxplot?

- Boxplots Show:
  - *Typical Range of Values*
  - *Possible Outliers*
  - *Variation*
- When to use?
  - *Useful when comparing between several groups of data sets*
  - *Used for moderate to large amount of data; when data is too small, the size of boxplot can vary significantly.*
- Boxplot vs Histogram
  - *Less detailed than histogram*
  - *but taking up less space which allows easy comparison of multiple data sets*

Outliers

- An Outlier is a data value that is either much smaller or much larger than the rest of the data.
- Some reasons for outliers
  - *Error in data collection*
  - *No error. For example, the owner's salary could be an outlier if the rest of the employees are all low wage workers*
- Need to be diligent about checking for outliers is because of all the descriptive statistics that are sensitive to outliers. The mean, standard deviation and correlation coefficient for paired data are just a few of these types of statistics.

Mean and Standard Deviation or Median and IQR?

Compare Median and Mean

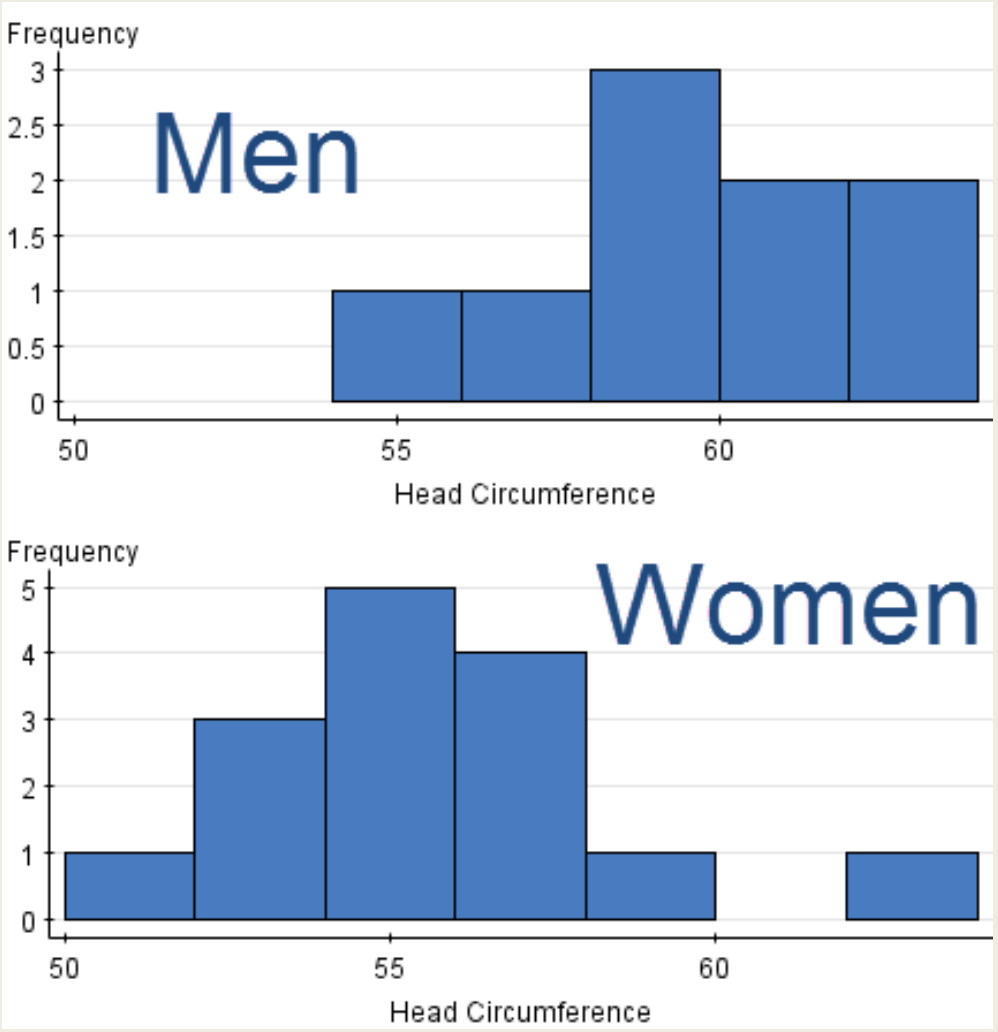
	Median	Mean
What it measures?	Balances counts.	Balances deviations.
How to calculate it?	Arrange the data from smallest to largest. If the number of data is odd, it is the middle number. If the number of data is even, it is the average of the two middle numbers.	Add up the data values and divide by the number of data values.
When to use it?	It can be used for any distribution, but is particularly useful when summarizing skewed data.	It is a useful measure of center when the distribution of data is fairly symmetric.
Is it affected by outliers?	It is not affected by outliers.	It is greatly affected by outliers.
How is it related to the histogram?	It divides the area of the histogram in half.	It is the balancing point of the histogram.
What is the related measure of spread?	The interquartile range (IQR)	The standard deviation.

https://www.onlinemathlearning.com/statistics-center-spread-hss-id2.html

Case Study: The head circumferences in centimeters for some men and women in a statistics class are given.

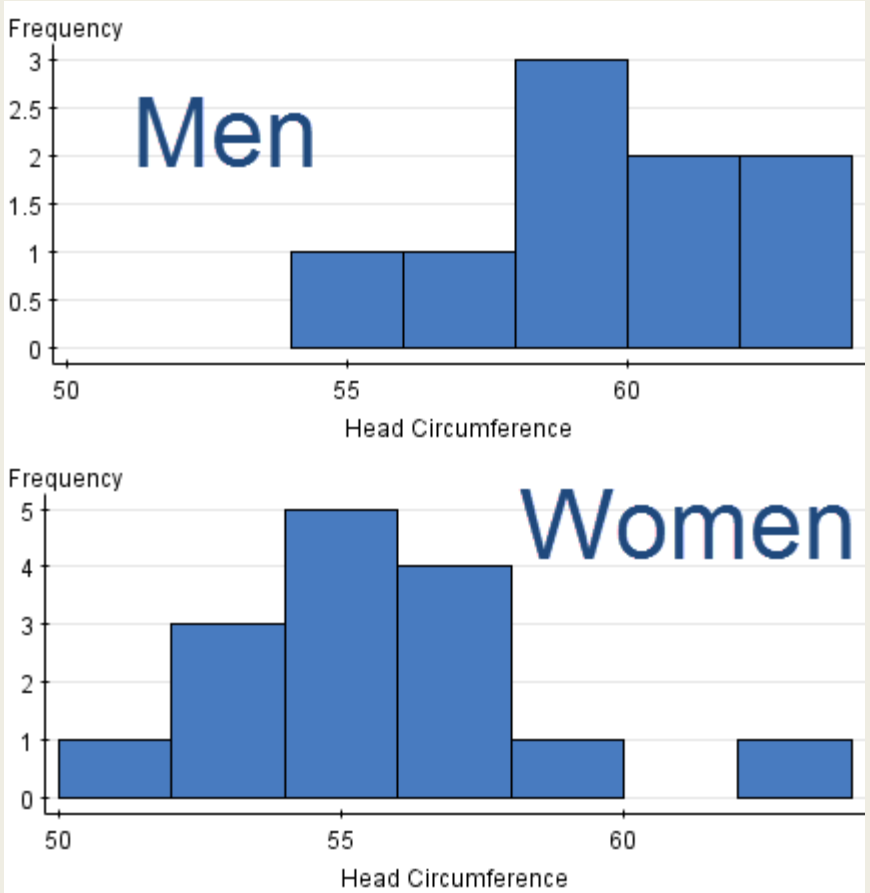
- Men: 58, 60, 62.5, 63, 59.5, 59, 60, 57, 55
- Women: 63, 55, 54.5, 53.5, 53, 58.5, 56, 54.5, 55, 56, 56, 54, 56,53, 51
- Analysis objective: To compare the circumferences of the men's and women's heads using descriptive statistics such as plots, numerical measures.

Histograms of the two sets of Data.



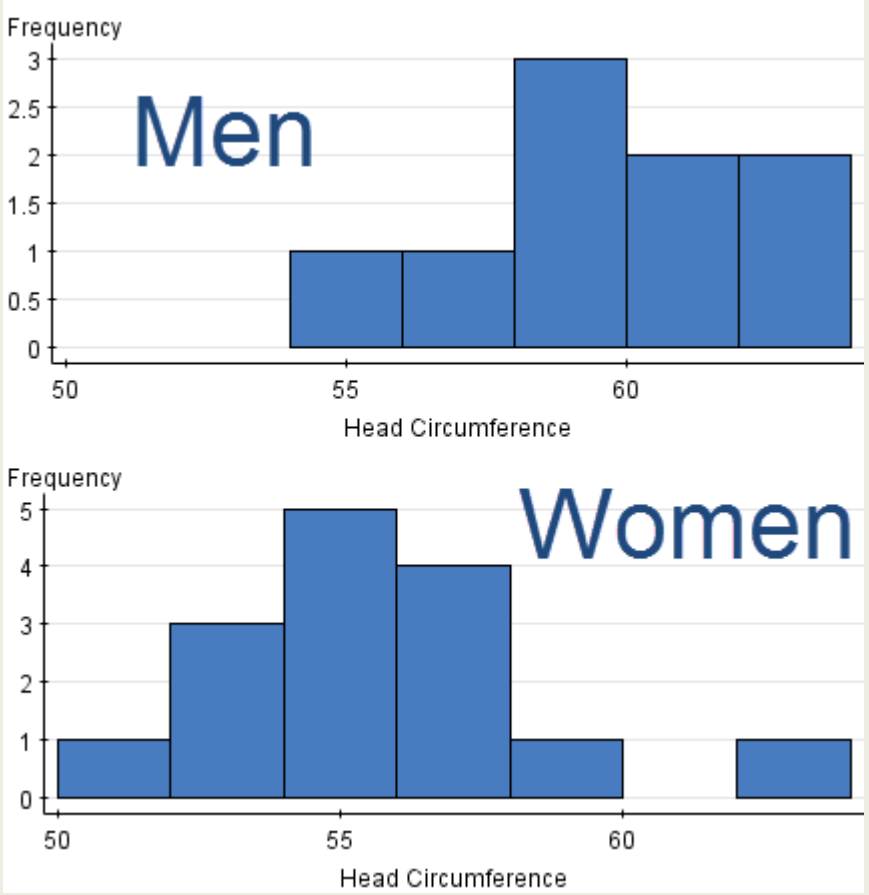
Shapes

- The distribution for men is unimodal and not too far from symmetric.
- The distribution for women is unimodal and nearly symmetric except one possible outlier.



Mean and Standard Deviation or Quartiles and IQR?

- Since the women's distribution has a possible outlier, the quartiles and IQR should be used for comparisons.



Compare Centers

Summary statistics:

Column	n	Mean	Variance	Std. Dev.	Std. Err.	Median	Range	Min	Max	Q1	Q3
Men	9	59.333332	6.3125	2.5124688	0.83748966	59.5	8	55	63	58	60
Women	15	55.266666	7.6380954	2.7637105	0.713587	55	12	51	63	53.5	56

- The median head circumference for the men was 59.5 cm, and the median head circumference for the women was 55 cm. This shows that the men tended to have larger heads.

Compare Variances

Summary statistics:

Column	n	Mean	Variance	Std. Dev.	Std. Err.	Median	Range	Min	Max	Q1	Q3
Men	9	59.333332	6.3125	2.5124688	0.83748966	59.5	8	55	63	58	60
Women	15	55.266666	7.6380954	2.7637105	0.713587	55	12	51	63	53.5	56

- The interquartile range for the head circumferences for the men was 2 cm, and the interquartile range for the women was 2.5 cm. This shows that the women tended to have more variation, as measured by the interquartile range.

Outliers

Summary statistics:

Column	n	Mean	Variance	Std. Dev.	Std. Err.	Median	Range	Min	Max	Q1	Q3
Men	9	59.333332	6.3125	2.5124688	0.83748966	59.5	8	55	63	58	60
Women	15	55.266666	7.6380954	2.7637105	0.713587	55	12	51	63	53.5	56

- Men: 58, 60, 62.5, 63, 59.5, 59, 60, 57, 55
  - *Q1 – (1.5)(IQR) = 55, Q3 + (1.5)(IQR) = 63*
  - *No Possible outliers for the men.*
- Women: 63, 55, 54.5, 53.5, 53, 58.5, 56, 54.5, 55, 56, 56, 54, 56, 53, 51
  - *Q1 – (1.5)(IQR) = 49.75, Q3 + (1.5)(IQR) = 59.75*
  - *63 is a possible outlier for the women.*

Final Comparison

- The typical head circumference for men is about 4.5 cm larger than the head circumference for women. The women's head circumference had slightly more variation than the men's.

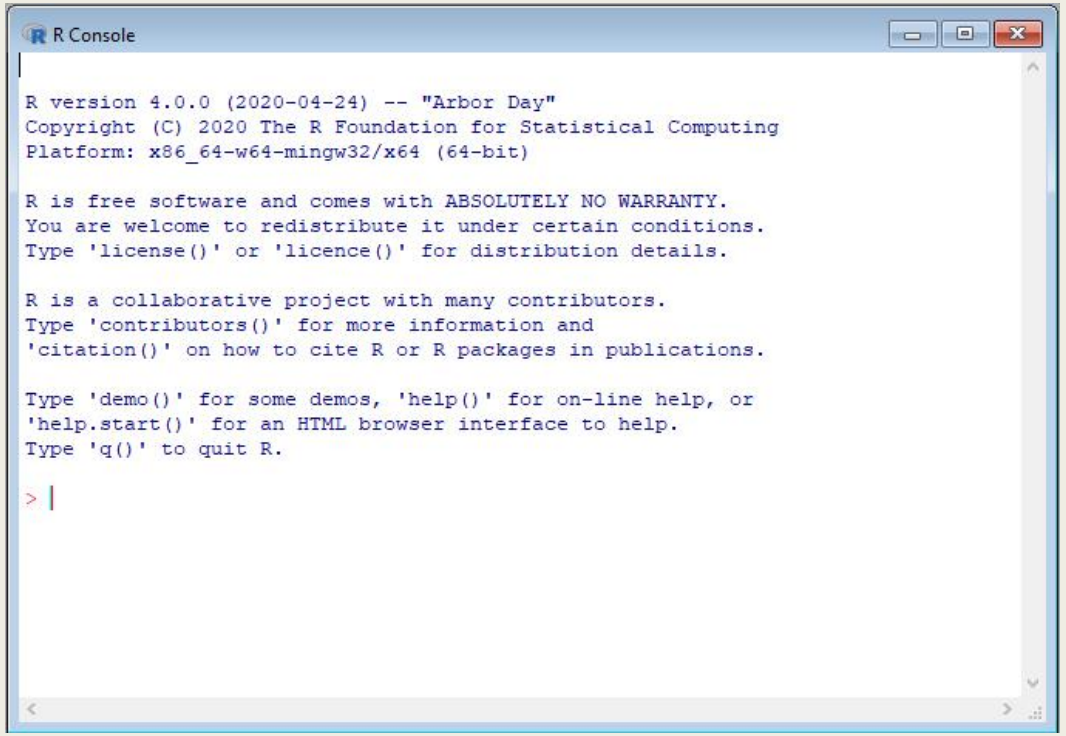
Introduction to R

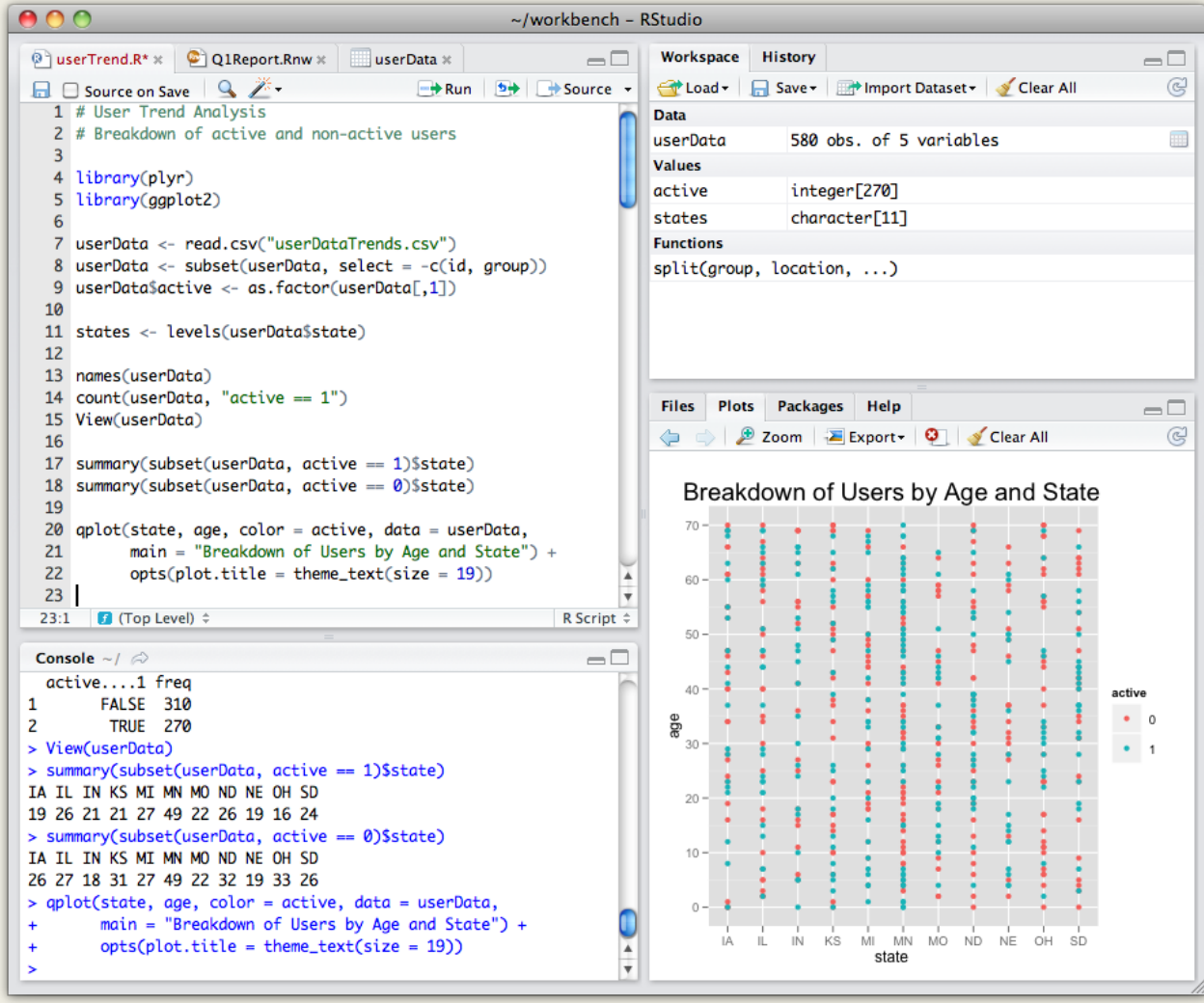
Introduction to R

- R and Python are two commonly used open source programming for data scientists
- Great for statistical analysis such as time series, survival analysis
- Beautiful visualization & Graphs like ggplot2
- Popular in research, life science, finance, media & marketing
- Easy to learn
- Large community: 12000 packages in CRAN (comprehensive R archive network); Bioconductor; GitHub
- Zoo (time series), dplyr & data.table (manipulate data), caret (machine learning), ggplot2(visualization)

Source: compare R vs. Python
https://www.guru99.com/r-vs-python.html

R & R studio, Choose one





How to install and start R

- **Installation**
- R can be downloaded from one of the mirror sites in http://cran.r-project.org/mirrors.html. You should pick your nearest location.
- Start R: Double click the R icon to activate; Alternatively, “start”->”All Programs”-> “R” -> “Rx64 4.0.2”

Basic functions

- **Variable assignment**

We assign values to variables with the assignment operator "=". Just typing the variable by itself at the prompt will print out the value. We should note that another form of assignment operator "<-" is also in use.

```
> x=1
> x
[1] 1

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

- **Functions**

R functions are invoked by its name, then followed by the parenthesis, and zero or more arguments. The following apply the function c to combine three numeric values into a vector.

```
> mean(c(1,2,3))

[1] 2
```

- **Comments**

All text after the pound sign "#" within the same line is considered a comment.

```
> mean(c(1,2,3))   #the center

[1] 2
```

- **Getting Help**

R provides extensive documentation.

For example, entering ?mean or help(mean) at the prompt gives documentation of the function c in R. Please give it a try.

R Data type

- **Numeric**

```
> x<-1
> x
[1] 1
> class(x)
[1] "numeric“
```

- **Logic**

```
y<-2
x>y
[1] FALSE
```

- Character

```
> z<-as.character(3)
> z
[1] "3“
> varname<-"number of sibling"
> varname
[1] "number of sibling"
```

Vector

- A **vector** is a sequence of data elements of the same basic type. Members in a vector are officially called **components**.

For example, # of siblings of 3 students are 1, 2, 3

> c(1,2,3)

[1] 1 2 3

For example, you are a student of Dept. Math

> c(TRUE, FALSE, TRUE)

[1] TRUE FALSE TRUE

For example, gender of student

> c("F","F","M")

[1] "F" "F" "M“

- **Vector index**

> gender<-c("F","F","M")

> gender[1]

[1] "F"

Rep()

The “rep” function replicates elements of vectors

```
> rep(1,5)
[1] 1 1 1 1 1
> rep("Sunny Day",5)
[1] "Sunny Day" "Sunny Day" "Sunny Day" "Sunny Day" "Sunny Day“
> rep(c(4,9),2)
[1] 4 9 4 9
```

Exercise:
1) use c() and rep() to generate the vector of 1,3,1,3,1,3
2) Use c() and rep() to generate the vector of 2,1,3,1,3,4

Seq()

The “seq” function creates a regular sequence of values to form a vector

```
> seq(1,10)
 [1]  1  2  3  4  5  6  7  8  9 10
> 1:10
 [1]  1  2  3  4  5  6  7  8  9 10> seq(1,10,by=2)
[1] 1 3 5 7 9
> seq(1,10,length=3)
[1]  1.0  5.5 10.0
> seq(10)
 [1]  1  2  3  4  5  6  7  8  9 10
```

Exercise:
use c(), seq() and rep() to generate the vector of 1,2,3,4,1,2,3,4,10,20

Matrix

A **matrix** is a collection of data elements arranged in a two-dimensional rectangular layout. **Elements are of the same mode**. The following is an example of a matrix with 2 rows and 3 columns.

```
> A<-matrix( c(2,4,3,1,5,7), nrow=2, ncol=3)
> A
     [,1] [,2] [,3]
[1,]    2    3    5
[2,]    4    1    7
> dim(A) # dimension of matrix A
[1] 2 3
```

$$A = \begin{bmatrix} 2 & 4 & 3 \\ 1 & 5 & 7 \end{bmatrix}$$

- An element at the m^{th} row, n^{th} column of A can be accessed by the expression A[m, n].

```
> A[2,1] # row2*col1
[1] 4
> A[,1]  #col 1
[1] 2 4
> A[2,]  #row 2
[1] 4 1 7
```

Exercise:

use seq() and matrix() to generate the matrix with 2 rows and 2 columns A=$\begin{matrix} 1 & 2 \\ 3 & 4 \end{matrix}$ and B=$\begin{matrix} 1 & 4 \\ 7 & 10 \end{matrix}$

Matrix construction

```
> B<-matrix(c(1,2,3), nrow=1, ncol=3)
> B
     [,1] [,2] [,3]
[1,]    1    2    3

> C<-rbind(A,B)
> C
     [,1] [,2] [,3]
[1,]    2    3    5
[2,]    4    1    7
[3,]    1    2    3
```

```
> B2<-matrix(c(1,1),nrow=2,ncol=1)
> D<-cbind(A,B2)
> D
     [,1] [,2] [,3] [,4]
[1,]    2    3    5    1
[2,]    4    1    7    1
```

Continuous Exercise:

Row & column combination of the 2 matrix of A=$\begin{matrix}1 & 2\\3 & 4\end{matrix}$ and B=$\begin{matrix}1 & 4\\7 & 10\end{matrix}$

Generate a new matrix C=$\begin{matrix}1 & 4\\3 & 10\end{matrix}$ using matrix A and B

Data Frame

A **data frame** is used for storing data tables. It is a list of vectors of equal length. For example, the following variable df is a data frame containing three vectors n, s, b.

Similar to matrix, but columns could have different modes

Each row represent one observation

Each column represent one variable

Example:

```
> num_sibling<-c(1,2,3)
> math_student<-c(TRUE, FALSE, TRUE)
> gender<-c("F","F","M")
> data_survey<-data.frame(num_sibling, math_student,gender)
> data_survey
  num_sibling math_student gender
1           1        TRUE      F
2           2       FALSE      F
3       3        TRUE      M
> num_sibling
[1] 1 2 3
```

Convert matrix into data frame

```
> A<-matrix(c(2,4,3,1,5,7),nrow=2,ncol=3)
> A
     [,1] [,2] [,3]
[1,]    2    3    5
[2,]    4    1    7
> dataf_test<-data.frame(A)   # convert matrix to data frame
> row.names(dataf_test)<-c("sample_1","sample_2")  #define row names
> names(dataf_test)<-c("var1","var2","var3")  # define column names
> dataf_test
         var1 var2 var3
sample_1    2    3    5
sample_2    4    1    7
```

Continuous Exercise:

Convert the new matrix C=$\begin{matrix}1 & 4\\ 3 & 10\end{matrix}$ to a data frame “data_test” with row names of “obs1”, “obs2” and col names of “col1”,”col2”

Select parts of data frame

- Selecting by columns

```
> data_survey[,1:2]
      num_sibling math_student
Anna            1        TRUE
Lucy            2       FALSE
Joshua          3        TRUE
```

- Selecting by rows

```
> data_survey[1:2,] #select the first 2 rows
     num_sibling math_student gender
Anna          1        TRUE      F
Lucy          2       FALSE      F
```

Continuous Exercise:
Select the first column of the data frame “data_test” you generated

Select parts of data frame

```
> data_survey[gender=="F",] # select observation by logic
     num_sibling math_student gender
Anna         1        TRUE      F
Lucy         2       FALSE      F

> data_survey[gender=="F"&num_sibling>=2,] # select observation by logic
     num_sibling math_student gender
Lucy         2       FALSE      F
```

Continuous Exercise:
Select the observations with col1>5 in the data frame “data_test” you generated

Combine data frames by row

- The two data frames should have the same number of variables which are in the same order
- If the variables are not the same in the two data frames, an error message will be displayed.

```
> data_survey2
        num_sibling math_student gender
Bobbi           2         TRUE      F
Sunny           3        FALSE      M
Ronnie          4         TRUE      M
> data_survey
        num_sibling math_student gender
Anna            1         TRUE      F
Lucy            2        FALSE      F
Joshua          3         TRUE      M
```

```
> data_survey_total<-rbind(data_survey, data_survey2)
> dim(data_survey_total)
[1] 6 3
> data_survey_total
       num_sibling math_student gender
Anna             1        TRUE      F
Lucy             2       FALSE      F
Joshua           3        TRUE      M
Bobbi            2        TRUE      F
Sunny            3       FALSE      M
Ronnie           4        TRUE      M
```

In some analysis, we might need to combine multiple data sets.
For example, sales table in Jul and in Aug for the monthly performance comparison
For example, clinical trial analysis using data sets from multiple health centers

Set up your working directory

> getwd() # your current working directory

[1] "C:/Users/jingzhao/Documents/R_file“

> setwd("C:/Users/jingzhao/Documents/R_file2") #set up your new working directory

> getwd()

[1] "C:/Users/jingzhao/Documents/R_file2"

Data import & export

- “read.table(…)” can be used to read data frames from free format text files
- “read.csv(…)” can be used to read data frames from files using comma to separate values
- The function “write.table” can be used to write dataframes to a file

write.table(data_survey_total,” data_survey_total.txt”) # the file will be written in your current working directory

Descriptive Statistics 1

- summary(x): a group of descriptive statistics
- max(x): maximum value of x
- min(x): minimum value of x
- range(x): min(x),max(x)
- sum(x): total of all the values in x
- mean(x): arithmetic average values in x
- median(x): median value of x
- quantile(x): Quantiles of x
- var(x): sample variance of x, with degrees of freedom=length(x)-1
- sd(x): sample standard deviation= var(x)^0.5
- skew(x): skewness of x

Descriptive Statistics (2)

- sort(x): a sorted version of x in increasing order
- rev(sort(x)): a sorted version of x in decreasing order, “rev”means “reverse”
- rank(x): vector of the ranks containing the permutation to sort x into ascending order
- length(x): number of entries in x (sample size if x is a sample of observations)
- sqrt(x): taking square-root of each entry in x
- ceiling(x): smallest integer which is larger than x
- floor(x): largest integer which is smaller than x

Visualization using ggplot2

Package ggplot2

- ggplot2 is an R implementation of Layered Grammar of Graphics which was developed by Hadley Wickham. (gg->grammar of graphics)
- It is very powerful because you are not limited to set a set of pre-speciied graphics, but you can create new graphics that are precisely tailored for your pr
- Online references if you are interested: http://ggplot2.org/

Installing a Package

install.packages("gglot2")

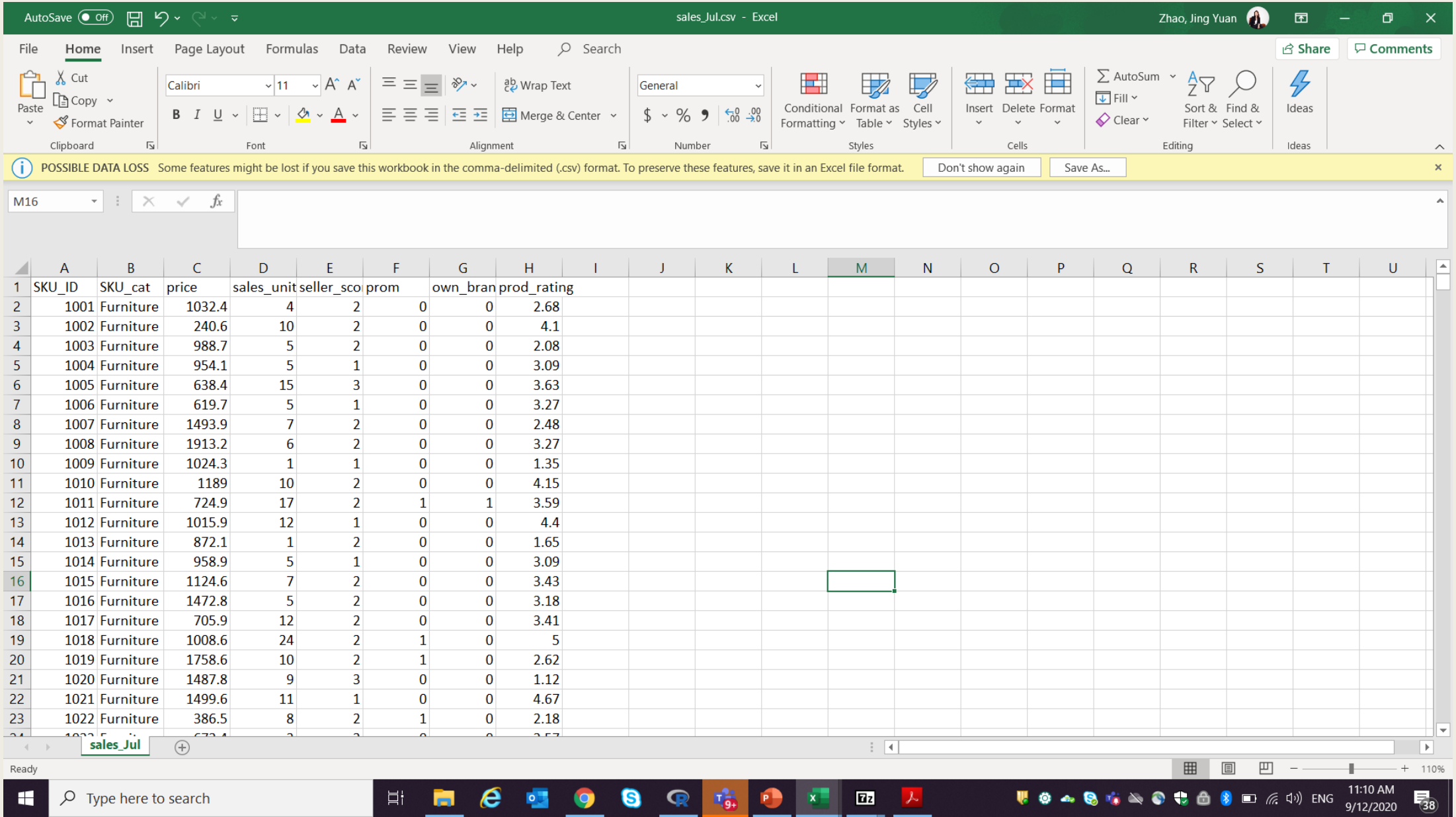
Loading a Package

> library(gglot2)

Two ways in ggplot2

- qplot()
  - *quick plot make it easy to produce basic graphs, but does not provide full capability*
  - *qplot is the simplest choice if you are dealing with input vectors*
- ggplot()
  - *grammar of graphics plot provides full implementation of The Grammar of Graphics, may have steeper learning curve but allows much more flexibility when building graphs*
  - *ggplot requires a data.frame as an input data structure*

Example: Sales data



	A	B	C	D	E	F	G	H
1	SKU_ID	SKU_cat	price	sales_unit	seller_sco	prom	own_bran	prod_rating
2	1001	Furniture	1032.4	4	2	0	0	2.68
3	1002	Furniture	240.6	10	2	0	0	4.1
4	1003	Furniture	988.7	5	2	0	0	2.08
5	1004	Furniture	954.1	5	1	0	0	3.09
6	1005	Furniture	638.4	15	3	0	0	3.63
7	1006	Furniture	619.7	5	1	0	0	3.27
8	1007	Furniture	1493.9	7	2	0	0	2.48
9	1008	Furniture	1913.2	6	2	0	0	3.27
10	1009	Furniture	1024.3	1	1	0	0	1.35
11	1010	Furniture	1189	10	2	0	0	4.15
12	1011	Furniture	724.9	17	2	1	1	3.59
13	1012	Furniture	1015.9	12	1	0	0	4.4
14	1013	Furniture	872.1	1	2	0	0	1.65
15	1014	Furniture	958.9	5	1	0	0	3.09
16	1015	Furniture	1124.6	7	2	0	0	3.43
17	1016	Furniture	1472.8	5	2	0	0	3.18
18	1017	Furniture	705.9	12	2	0	0	3.41
19	1018	Furniture	1008.6	24	2	1	0	5
20	1019	Furniture	1758.6	10	2	1	0	2.62
21	1020	Furniture	1487.8	9	3	0	0	1.12
22	1021	Furniture	1499.6	11	1	0	0	4.67
23	1022	Furniture	386.5	8	2	1	0	2.18

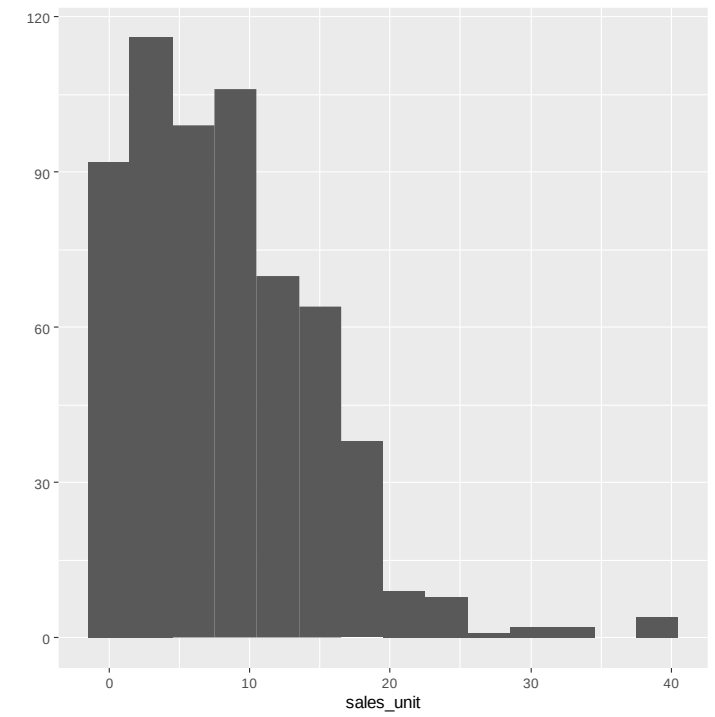
Histogram using qplot (1)

Variable

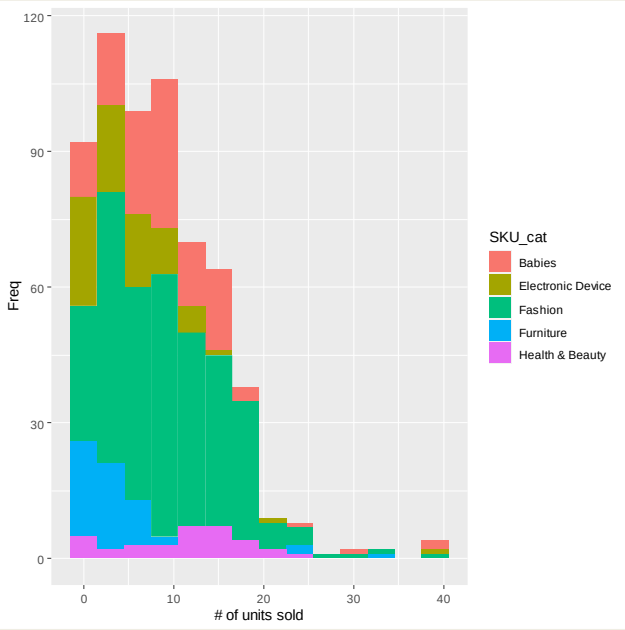
Data frame

Type of plot

- qplot(sales_unit, data=sales_jul, geom="histogram", binwidth = 3)



Histogram using qplot (2)



- To compare the distributions of multiple subgroups, just add an aesthetic mapping

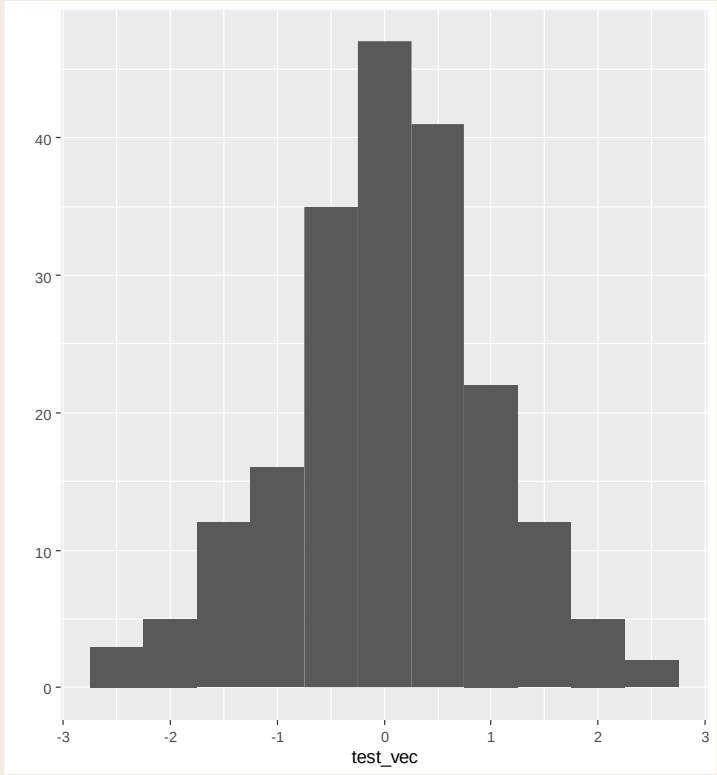
draw a histogram for each product category.

qplot(sales_unit, data=sales_jul, geom="histogram", binwidth = 3, fill=SKU_cat, xlab="# of units sold",ylab="Freq")

Different subgroups, filled by different colors

qplot works for non data frame

```
test_vec<-rnorm(200)
qplot(test_vec, geom="histogram",binwidth=0.5)
```



ggplot

1. **Data Frame**: the data set you want to analyze

   ggplot(df) → must be a data frame. If not, please convert first

2. **Aesthetics:** map your data to the visualization, such as X axis, y axis, variable used for colored (comparison)

   +aes(x=, y=, fill/color=) color variable should be categorical variable

3. **Layer:** what plots you want to see

   +geom_boxplot(), geom_bar()

4. **Faceting:** provide “drill-down” view

   +facet_grid() or facet_wrap()

5. **Label:**

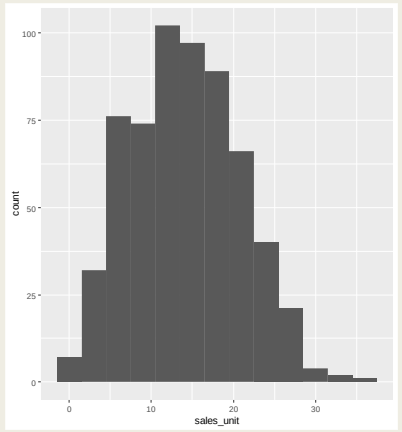
   +labs(x=“”,y=“”,title=“”)

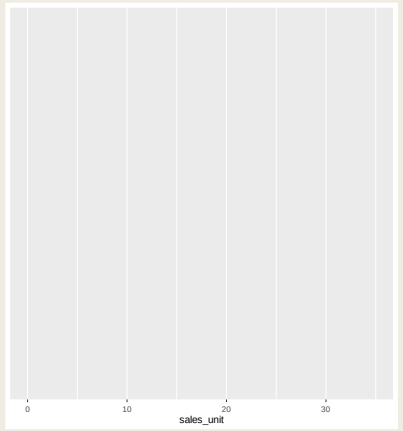
…..

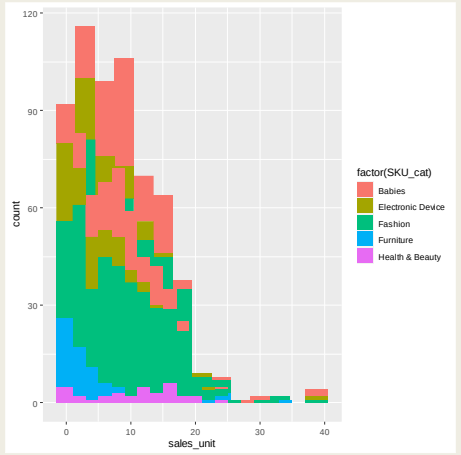
Geom

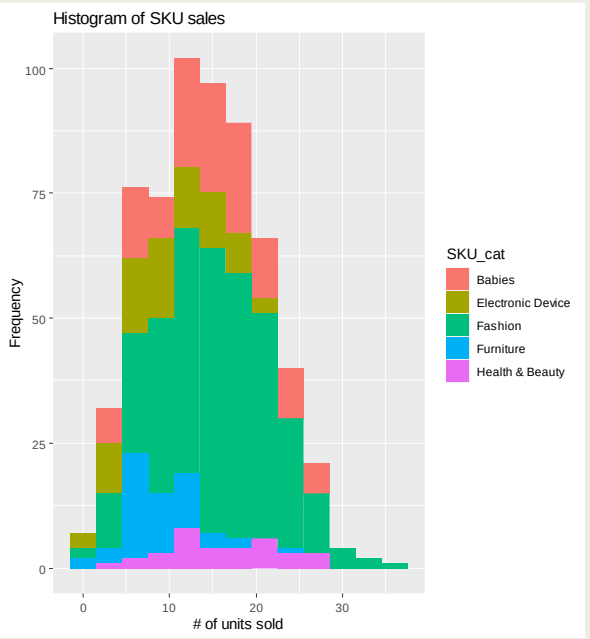
- Geometric objects (geom)
- Geometric objects are the actual marks we put on a plot.
- Examples include:
  - *geom_point produces a scatter plot*
  - *geom_text adds labels at the specied points*
  - *geom_line makes a line plot*
  - *geom_boxplot produces a boxplot*
  - *geom_jitter makes a jittered plot*
  - *geom_histogram makes a histogram*
  - *geom_density draws density curves*
  - *geom_bar makes a bar chart*

```
ggplot(sales_jul)+
  aes(x=sales_unit,fill=SKU_cat)+
  geom_histogram(binwidth=3)+
  labs(x="# of units sold", y="Frequency", title=" Histogram of SKU sales")
```

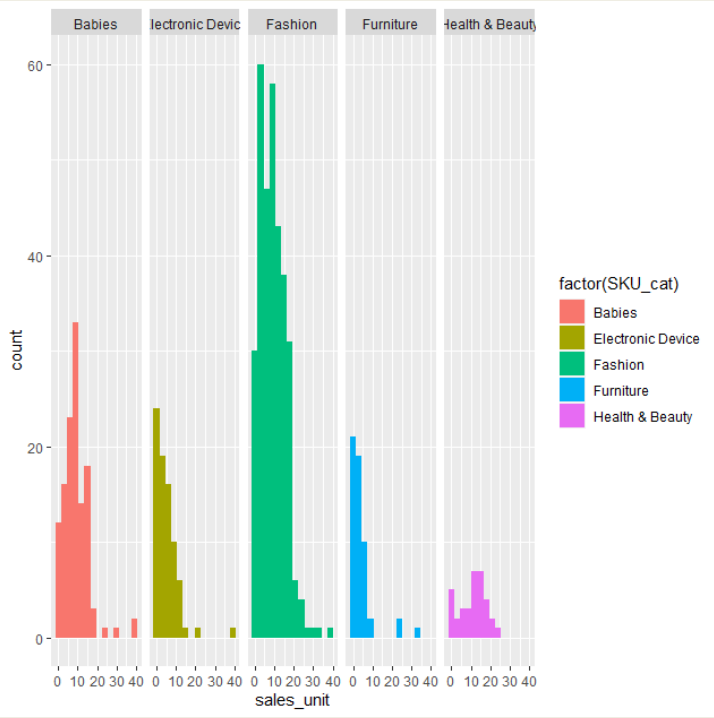




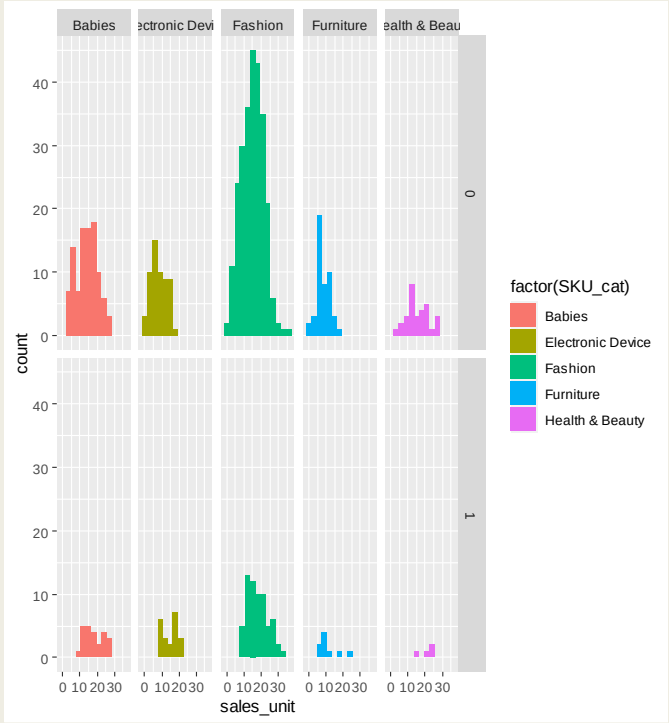


Multi-panel plots by Facet.grid()

```
ggplot(sales_jul)+
  aes(x=sales_unit,fill=factor(SKU_cat))+
  geom_histogram(binwidth=3)+
  facet_grid(.~SKU_cat)
```



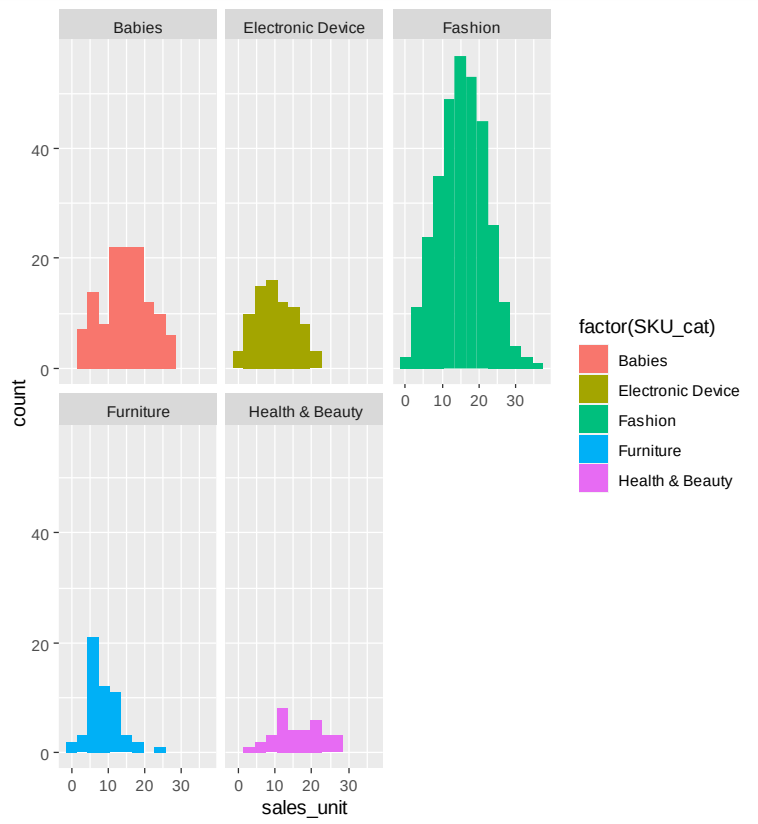

```
ggplot(sales_jul)+
aes(x=sales_unit,fill=factor(SKU_cat))+
geom_histogram(binwidth=3)+
 facet_grid(prom~SKU_cat)
```



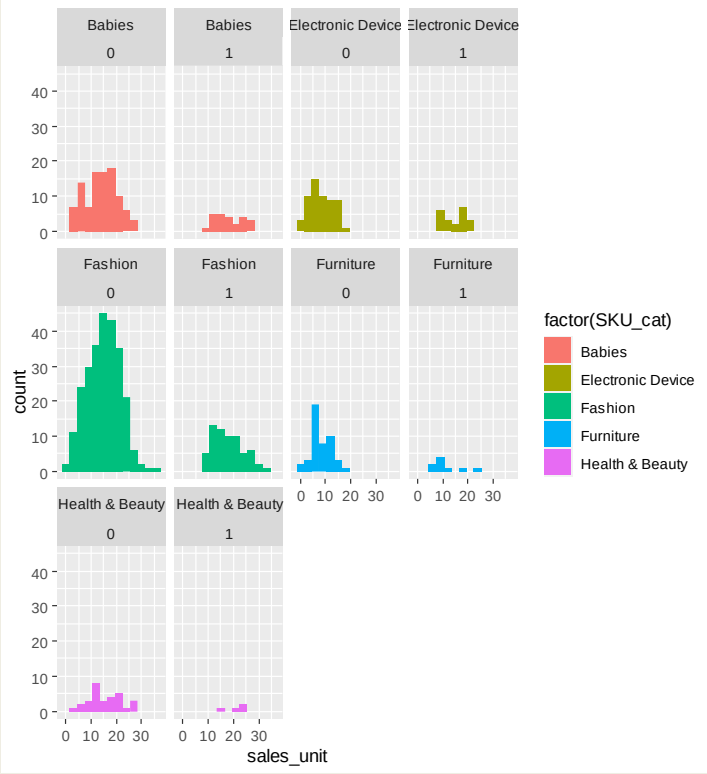
Source: http://zevross.com/blog/2019/04/02/easy-multi-panel-plots-in-r-using-facet_wrap-and-facet_grid-from-ggplot2/

Facet.wrap()

```
ggplot(sales_jul)+
  aes(x=sales_unit,fill=SKU_cat)+
  geom_histogram(binwidth=3)+
  facet_wrap(.~SKU_cat)  #facet_wrap(~ variable)
```




```
ggplot(sales_jul)+
aes(x=sales_unit,fill=factor(SKU_cat))+
geom_histogram(binwidth=3)+
  facet_wrap(.~SKU_cat+prom)
```



Mix multiple plot on the same page

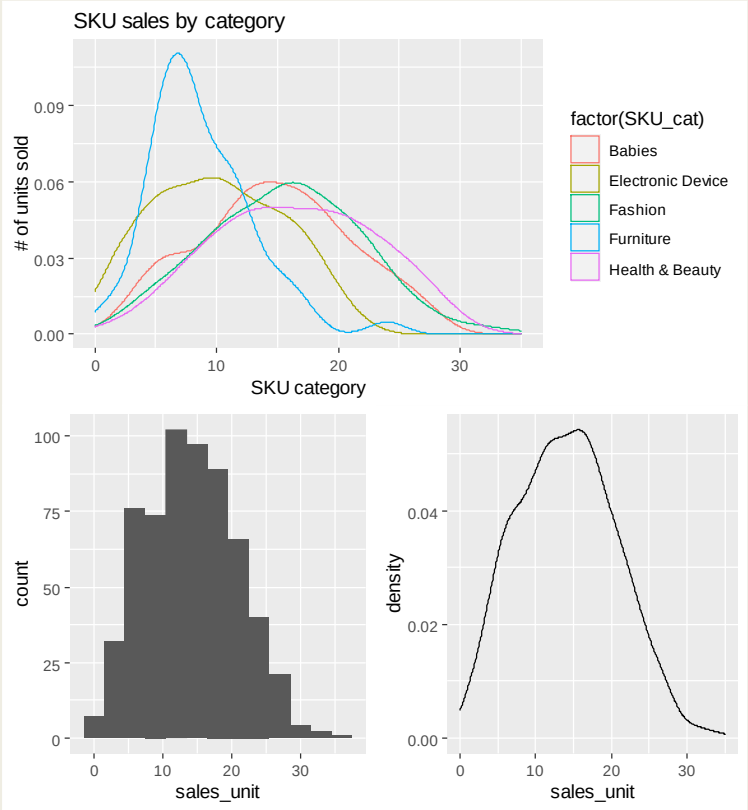
GridExtra to arrange multiple grid-based plots on a page.

library(gridExtra)

grid.arrange()

Grid.arrange()

```
grid.arrange(plot1,                                    # First row with one plot spaning over 2 columns
    arrangeGrob(plot2, plot3, ncol = 2), # Second row with 2 plots in 2 different columns
    nrow = 2)
```



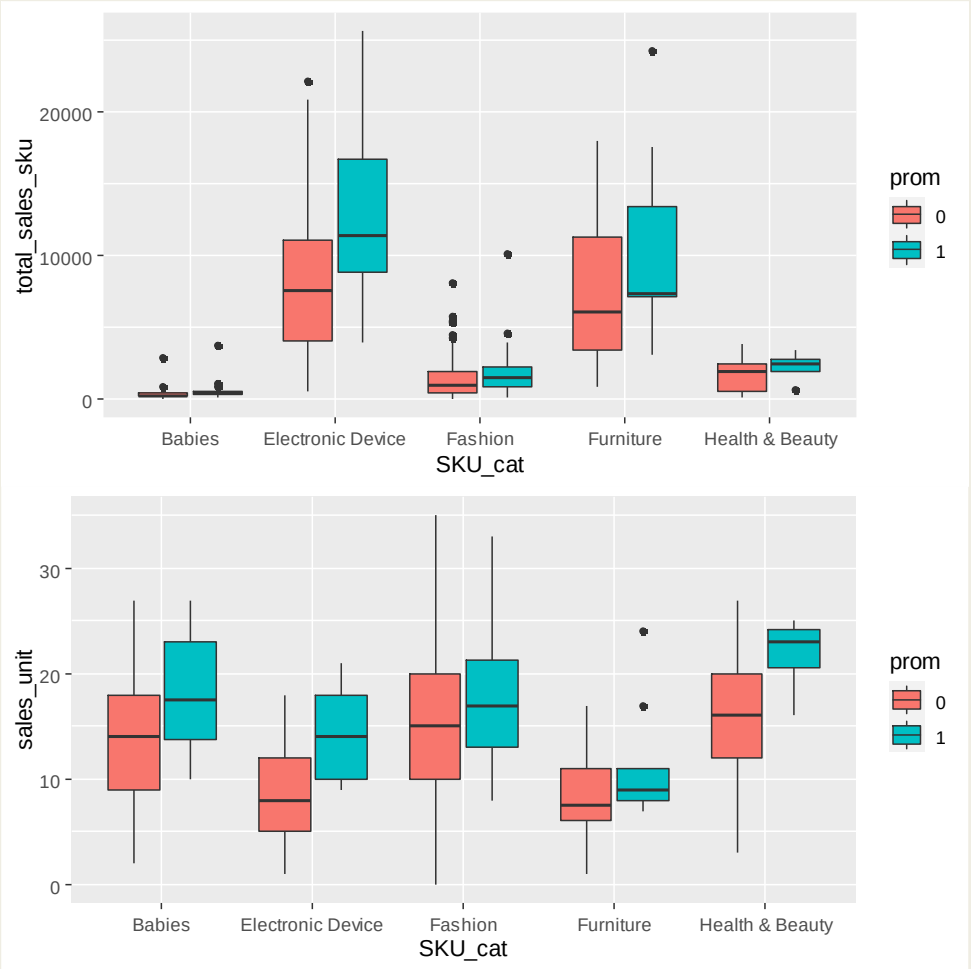
Source:
http://www.sthda.com/english/articles/24-ggpubr-publication-ready-plots/81-ggplot2-easy-way-to-mix-multiple-graphs-on-the-same-page/#:~:text=To%20arrange%20multiple%20ggplot2%20graphs,multiple%20ggplots%20on%20one%20page

Derive new features to generate additional insights

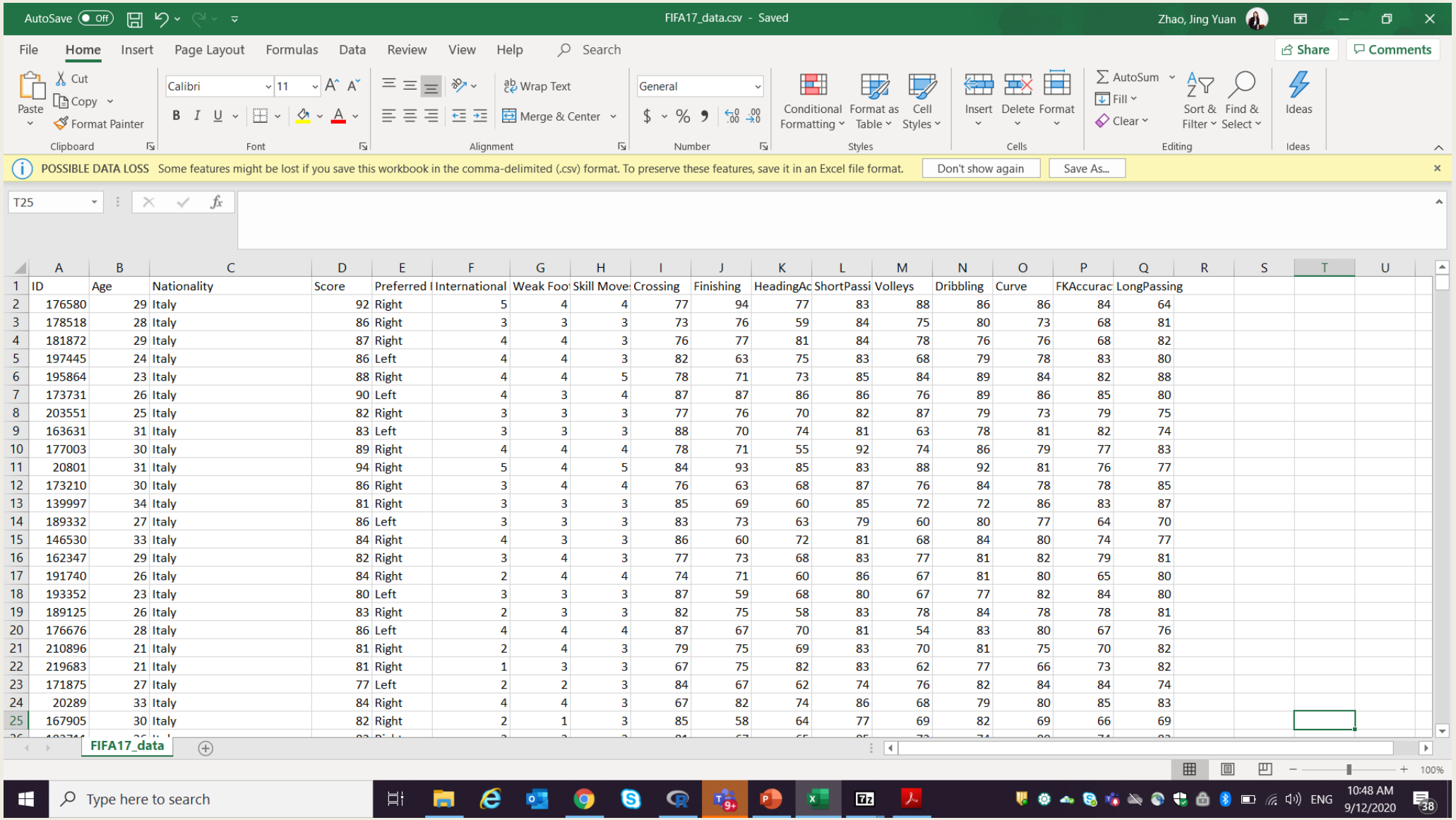
Most raw data is collected for some specific operation purpose, such as POS data, survey data, and IoT data.

However, they might not have contribution to generate insights directly.

As a data scientist, you need to derive more features based on your business understanding.



Homework: FIFA17 data



AutoSave Off | FIFA17_data.csv - Saved | Zhao, Jing Yuan

File Home Insert Page Layout Formulas Data Review View Help Search | Share | Comments

Paste | Cut | Copy | Format Painter | Clipboard | Calibri 11 | Font | Wrap Text | Merge & Center | Alignment | General | Number | Conditional Formatting | Format as Table | Cell Styles | Styles | Insert | Delete | Format | Cells | AutoSum | Fill | Clear | Sort & Filter | Find & Select | Editing | Ideas

POSSIBLE DATA LOSS Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format. Don't show again | Save As...

T25

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	ID	Age	Nationality	Score	Preferred I	International	Weak Foo	Skill Move	Crossing	Finishing	HeadingAc	ShortPassi	Volleys	Dribbling	Curve	FKAccurac	LongPassing
2	176580	29	Italy	92	Right	5	4	4	77	94	77	83	88	86	86	84	64
3	178518	28	Italy	86	Right	3	3	3	73	76	59	84	75	80	73	68	81
4	181872	29	Italy	87	Right	4	4	3	76	77	81	84	78	76	76	68	82
5	197445	24	Italy	86	Left	4	4	3	82	63	75	83	68	79	78	83	80
6	195864	23	Italy	88	Right	4	4	5	78	71	73	85	84	89	84	82	88
7	173731	26	Italy	90	Left	4	3	4	87	87	86	86	76	89	86	85	80
8	203551	25	Italy	82	Right	3	3	3	77	76	70	82	87	79	73	79	75
9	163631	31	Italy	83	Left	3	3	3	88	70	74	81	63	78	81	82	74
10	177003	30	Italy	89	Right	4	4	4	78	71	55	92	74	86	79	77	83
11	20801	31	Italy	94	Right	5	4	5	84	93	85	83	88	92	81	76	77
12	173210	30	Italy	86	Right	3	4	4	76	63	68	87	76	84	78	78	85
13	139997	34	Italy	81	Right	3	3	3	85	69	60	85	72	72	86	83	87
14	189332	27	Italy	86	Left	3	3	3	83	73	63	79	60	80	77	64	70
15	146530	33	Italy	84	Right	4	3	3	86	60	72	81	68	84	80	74	77
16	162347	29	Italy	82	Right	3	4	3	77	73	68	83	77	81	82	79	81
17	191740	26	Italy	84	Right	2	4	4	74	71	60	86	67	81	80	65	80
18	193352	23	Italy	80	Left	3	3	3	87	59	68	80	67	77	82	84	80
19	189125	26	Italy	83	Right	2	3	3	82	75	58	83	78	84	78	78	81
20	176676	28	Italy	86	Left	4	4	4	87	67	70	81	54	83	80	67	76
21	210896	21	Italy	81	Right	2	4	3	79	75	69	83	70	81	75	70	82
22	219683	21	Italy	81	Right	1	3	3	67	75	82	83	62	77	66	73	82
23	171875	27	Italy	77	Left	2	2	3	84	67	62	74	76	82	84	84	74
24	20289	33	Italy	84	Right	4	4	3	67	82	74	86	68	79	80	85	83
25	167905	30	Italy	82	Right	2	1	3	85	58	64	77	69	82	69	66	69

FIFA17_data

Type here to search | ENG | 10:48 AM 9/12/2020

Exploratory Data Analysis

1. Use R Read table
2. Understand distribution of each variable through mean, median, Quantile…
3. Draw histogram of score
4. Draw bar chat of nationality
5. Draw multi-panel histogram of score, by international reputation (need to a categorical variable)
6. Draw boxplot of score by x axis = nationality, colored by preferred foot
7. Mix plots of steps 3,4,6 with 2 rows, plot of step 6 in the first row; the rest two in the second row
8. Save the mixed plot in your working directory

Resources

- Paul Murrell, R Graphics, 2nd Ed.

  R code for all figures: *https://www.stat.auckland.ac.nz/~paul/RG2e/*

- Hadley Wickham, ggplot2: Elegant graphics for data analysis, 2nd Ed

  ggplot2 Quick Reference: *http://sape.inf.usi.ch/quick-reference/ggplot2/*

- Introduction to R

  https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf

Q&A