# INST452: Health Data Analytics

## Module 1
## Introduction to R and Health Data Analytics

Dr. Nikki Sigalo

Adapted from Health Informatics course Unit 3.3 - www.measureevaluation.org

# Part 1: Health Data Analytics

# What is analytics?

The **discovery** of meaningful patterns in data

The **synthesis** of knowledge from information

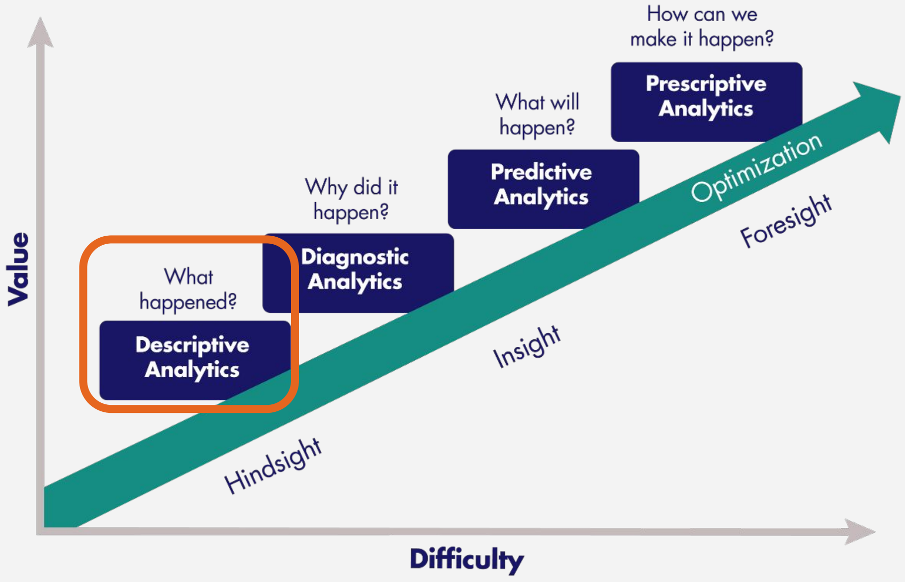
National Institute of Standards and Technology, 2013

# Types of Analytics: Overview

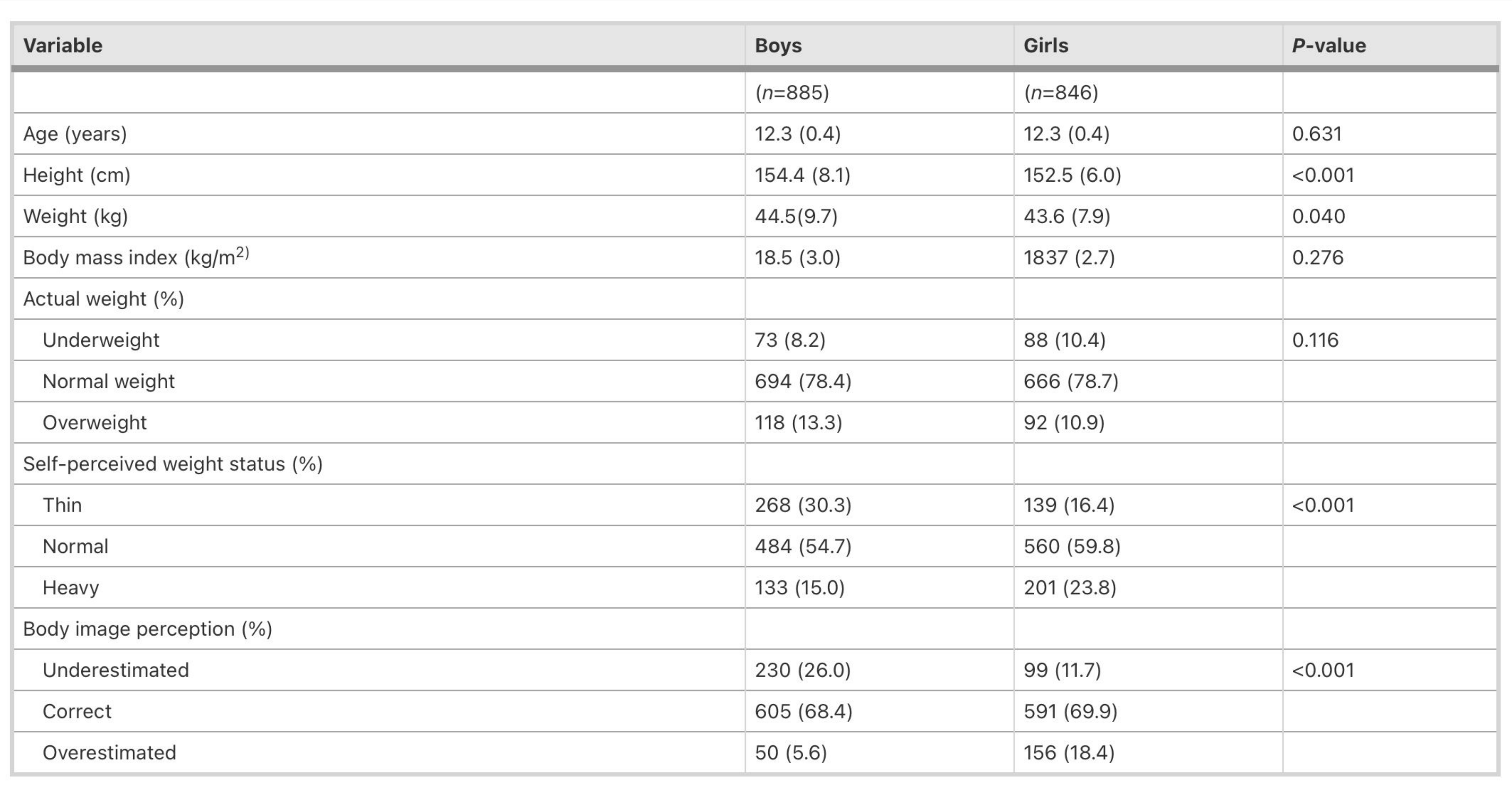
- ***Descriptive*****:** Uses business intelligence and data mining to ask: “What has happened?”
- ***Diagnostic*****:** Examines data to answer “Why did it happen?”
- ***Predictive:*** Uses statistical models and forecasts to ask: “What could happen?”
- ***Prescriptive:*** Uses optimization and simulation to ask: “What should we do?”

# Descriptive Analytics

- Describe the data
- Common statistics:
  - Measures of Central Tendency
  - Measures of Spread
  - Frequency Distributions
- Typical reporting methods:
  - Tables
  - Charts
  - Written narratives



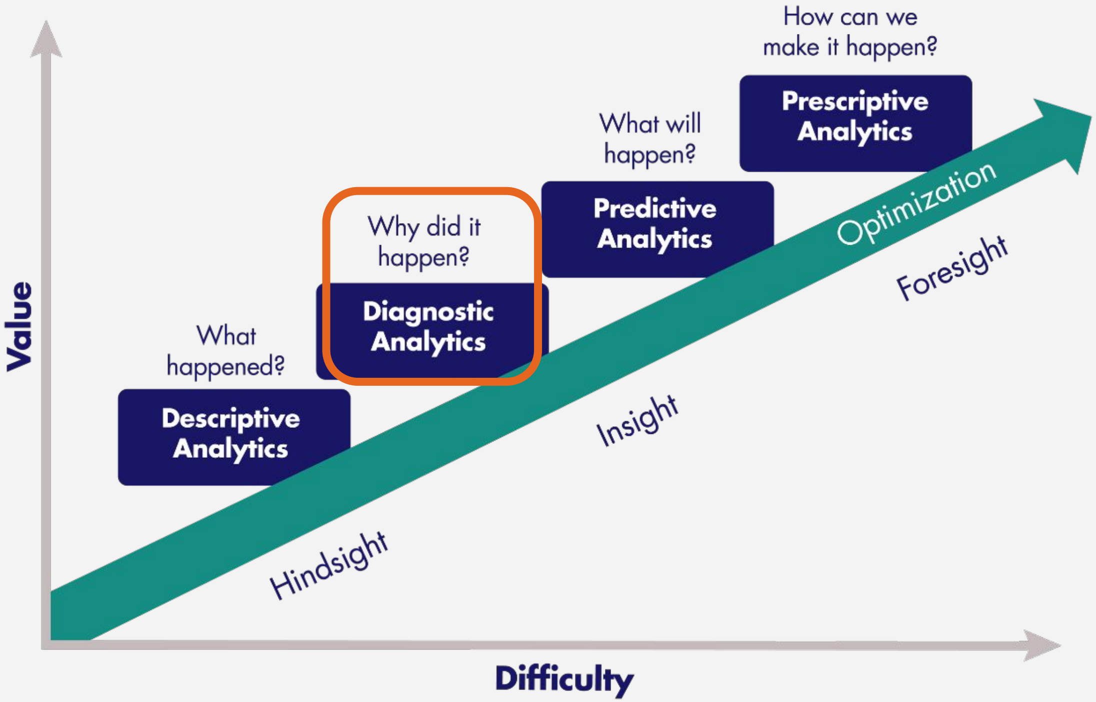
# Example: Study population characteristics from a paper on the relationship between distorted body image and lifestyle in adolescents in Japan, 2005-2009



| Variable | Boys | Girls | *P*-value |
|---|---|---|---|
| | (*n*=885) | (*n*=846) | |
| Age (years) | 12.3 (0.4) | 12.3 (0.4) | 0.631 |
| Height (cm) | 154.4 (8.1) | 152.5 (6.0) | <0.001 |
| Weight (kg) | 44.5(9.7) | 43.6 (7.9) | 0.040 |
| Body mass index (kg/m$^{2)}$ | 18.5 (3.0) | 1837 (2.7) | 0.276 |
| Actual weight (%) | | | |
| Underweight | 73 (8.2) | 88 (10.4) | 0.116 |
| Normal weight | 694 (78.4) | 666 (78.7) | |
| Overweight | 118 (13.3) | 92 (10.9) | |
| Self-perceived weight status (%) | | | |
| Thin | 268 (30.3) | 139 (16.4) | <0.001 |
| Normal | 484 (54.7) | 560 (59.8) | |
| Heavy | 133 (15.0) | 201 (23.8) | |
| Body image perception (%) | | | |
| Underestimated | 230 (26.0) | 99 (11.7) | <0.001 |
| Correct | 605 (68.4) | 591 (69.9) | |
| Overestimated | 50 (5.6) | 156 (18.4) | |

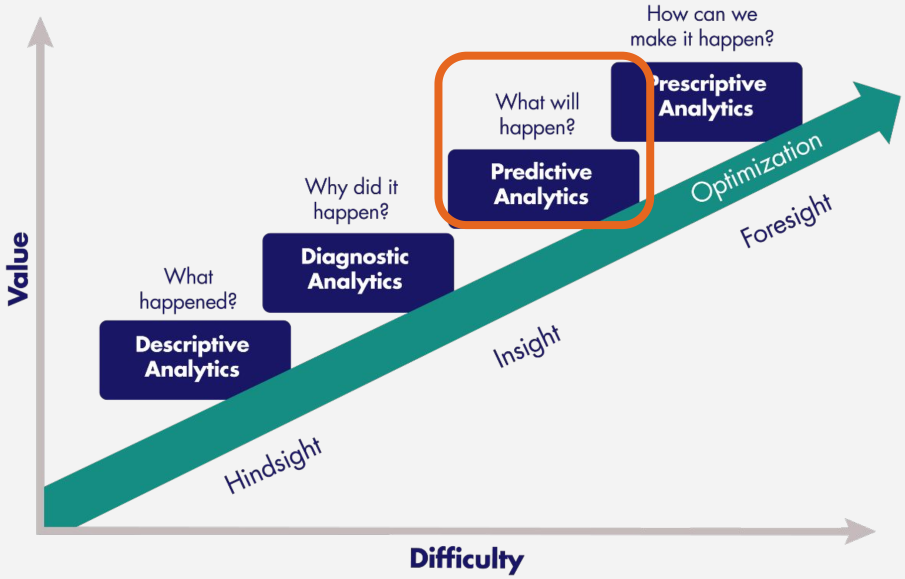
# Diagnostic Analytics

- Attempts to answer "why did it happen?"
- Drill-down techniques
- Data discovery
- Correlations



# Predictive Analytics

- Predicts instead of describing or classifying
- Rapid analysis necessary
- Relevant insights necessary
- Ease of use

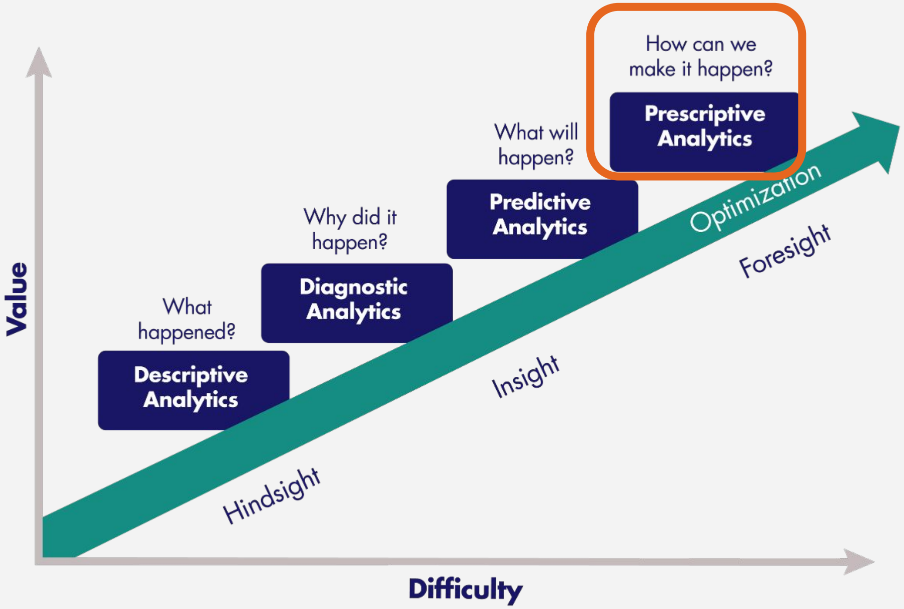


# What Predictive Analytics Cannot Do

- “The purpose of predictive analytics is NOT to tell you what will happen in the future. It cannot do that. In fact, no analytics can do that. Predictive analytics can only forecast what might happen in the future, because all predictive analytics are probabilistic in nature.”
- (Bertolucci, 2013)

# Prescriptive Analytics

- **Beyond this class, but..**
- Examines data or content to answer the question "What should be done?" or "What can we do to make _______ happen?
- Focuses on finding the best course of action in a scenario given the available data
- Related to both descriptive analytics and predictive analytics but emphasizes actionable insights instead of data monitoring
- Whereas descriptive analytics offers insights into what has happened, and predictive analytics focuses on forecasting possible outcomes, prescriptive analytics aims to find the best solution given a variety of choices
- Is characterized by techniques such as:
  - Graph analysis
  - Simulation
  - Complex event processing
  - Neural networks
  - Recommendation engines
  - Heuristics
  - Machine learning



# 10 MINUTE BREAK

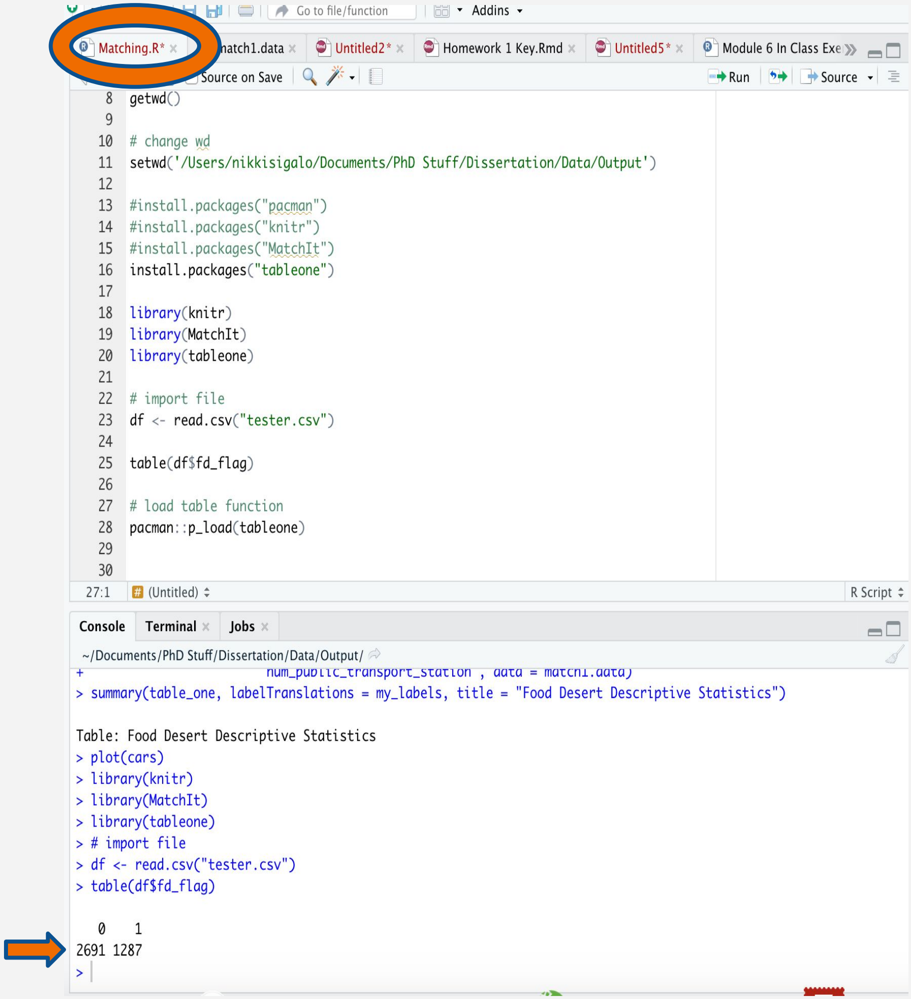
# Part 2: Introduction to R

# Why use R for analytics?

- It's free, open source, powerful and highly extensible.
- Implement many common statistical procedures
- Provide excellent graphics functionality
- A convenient starting point for many data analysis projects
  - Data transformation, cleansing, merging, aggregating
- We will start working in R next week

# R Script

- A program that contains a series of commands that you can execute one at a time (or all at once). The **script** can be saved and used later to re-execute the saved commands.
- The output from each command is displayed in the console
- Extension is .R
- Familiar with Python? This is similar to a normal Python script being executed in an IDE such as Spyder



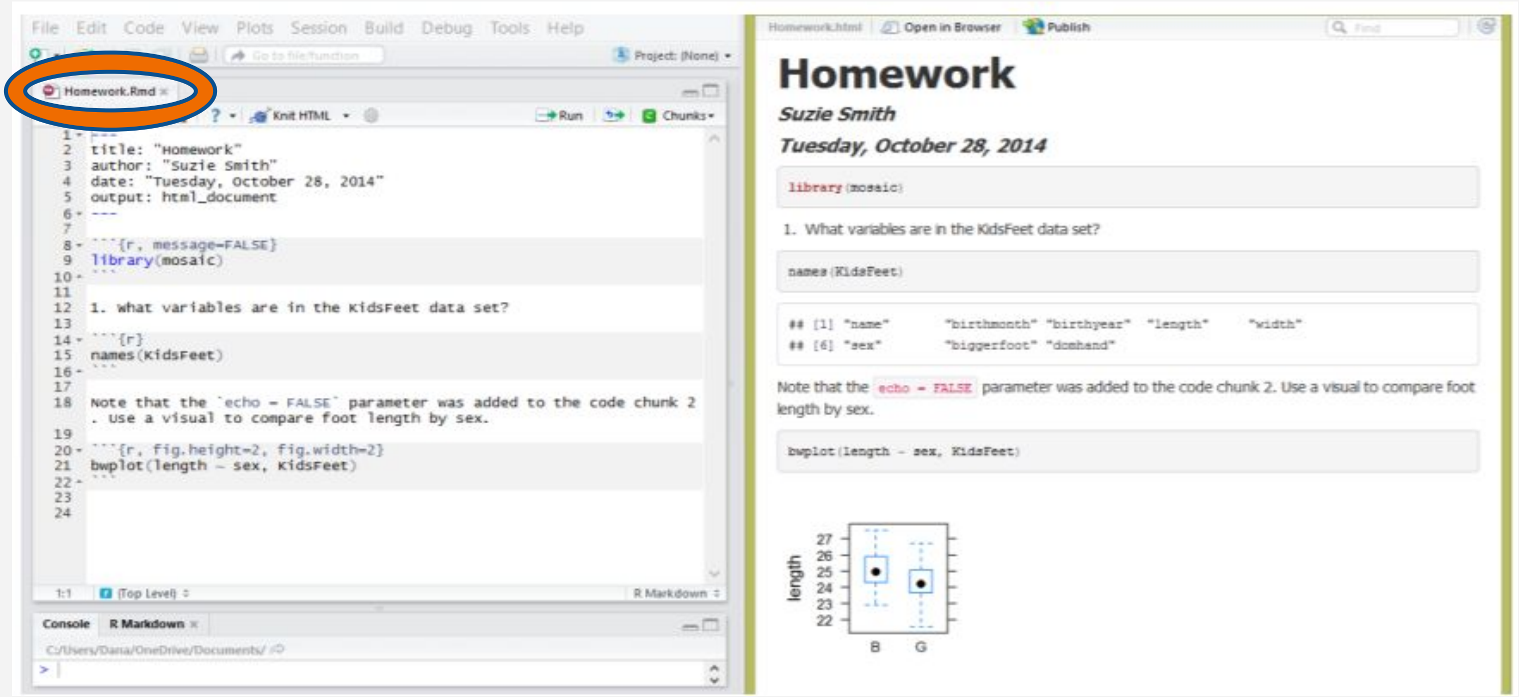
```
getwd()

# change wd
setwd('/Users/nikkisigalo/Documents/PhD Stuff/Dissertation/Data/Output')

#install.packages("pacman")
#install.packages("knitr")
#install.packages("MatchIt")
install.packages("tableone")

library(knitr)
library(MatchIt)
library(tableone)

# import file
df <- read.csv("tester.csv")

table(df$fd_flag)

# load table function
pacman::p_load(tableone)
```

~/Documents/PhD Stuff/Dissertation/Data/Output/

```
+                          num_public_transport_station , data = match1.data)
> summary(table_one, labelTranslations = my_labels, title = "Food Desert Descriptive Statistics")

Table: Food Desert Descriptive Statistics
> plot(cars)
> library(knitr)
> library(MatchIt)
> library(tableone)
> # import file
> df <- read.csv("tester.csv")
> table(df$fd_flag)

   0    1
2691 1287
>
```

# R Markdown

- Written in markdown (an easy-to-write plain text format) and contains chunks of embedded R code
- The output from each command is displayed below the code chunk
- Familiar with Python? This is similar to a Jupyter Notebook
- Can include both narrative text and code in the same document and knit the document into several different data types (i.e HTML, PDF, etc.)



~~~
---
title: "Homework"
author: "Suzie Smith"
date: "Tuesday, October 28, 2014"
output: html_document
---

```{r, message=FALSE}
library(mosaic)
~~~

1. what variables are in the KidsFeet data set?

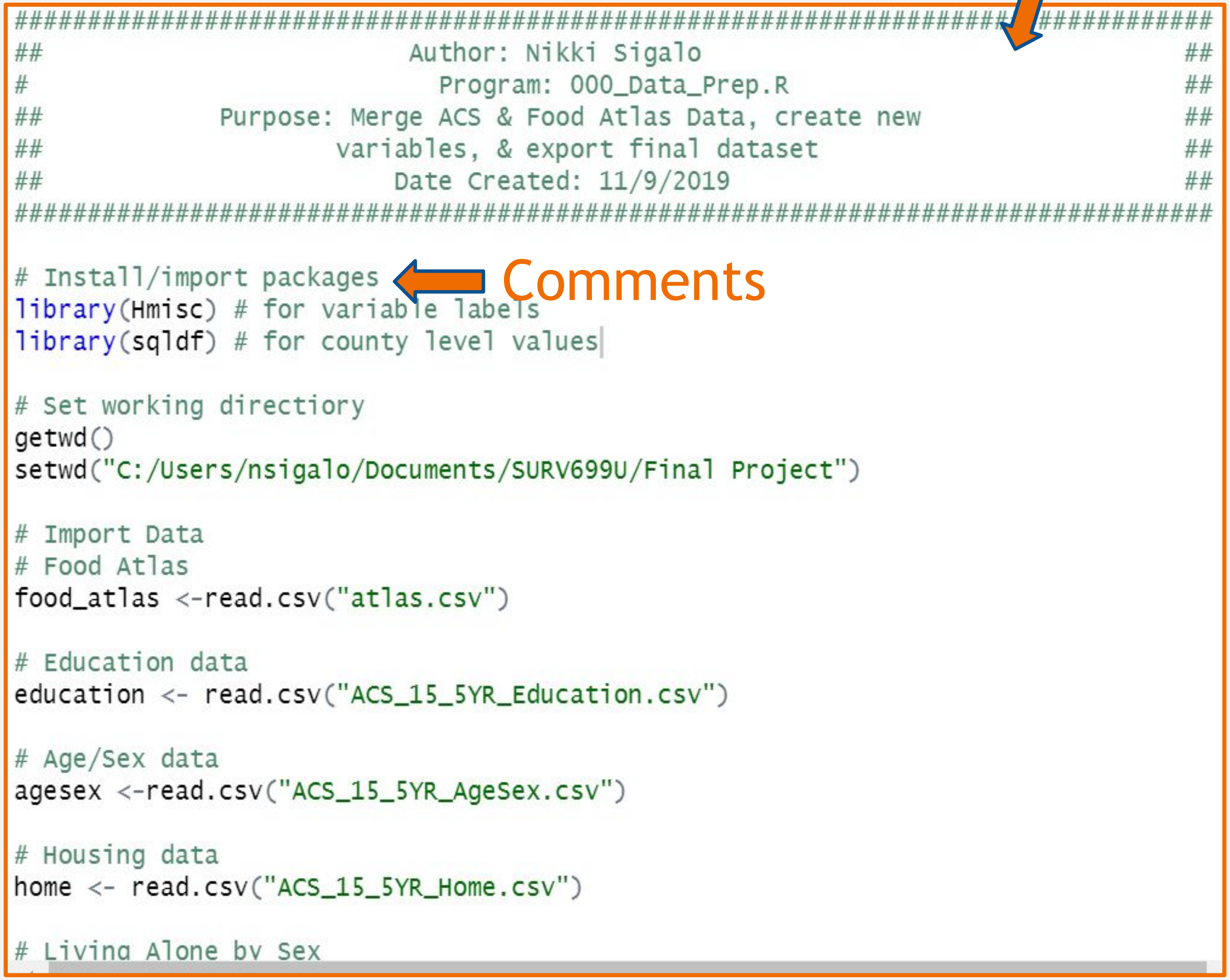
```{r}
names(KidsFeet)
```

Note that the `echo = FALSE` parameter was added to the code chunk 2
. Use a visual to compare foot length by sex.

```{r, fig.height=2, fig.width=2}
bwplot(length ~ sex, KidsFeet)
```
```

## Homework

***Suzie Smith***

***Tuesday, October 28, 2014***

```
library(mosaic)
```

1. What variables are in the KidsFeet data set?

```
names(KidsFeet)
```

```
## [1] "name"       "birthmonth" "birthyear"  "length"     "width"
## [6] "sex"        "biggerfoot" "domhand"
```

Note that the `echo = FALSE` parameter was added to the code chunk 2. Use a visual to compare foot length by sex.

```
bwplot(length ~ sex, KidsFeet)
```
```

# Style Guide - R Script

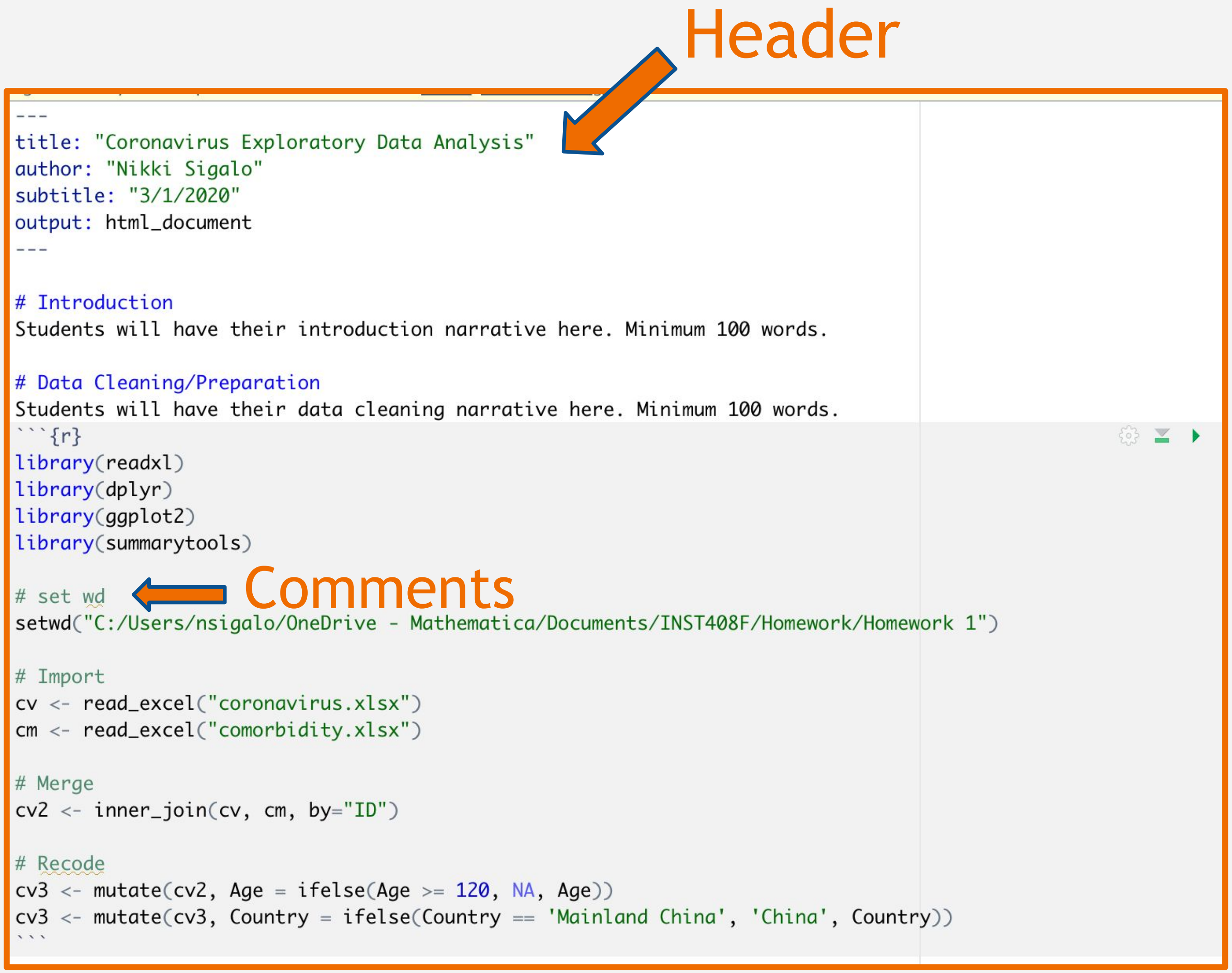
Header



```
####################################################################
##                    Author: Nikki Sigalo                        ##
#                   Program: 000_Data_Prep.R                      ##
##     Purpose: Merge ACS & Food Atlas Data, create new           ##
##           variables, & export final dataset                    ##
##                Date Created: 11/9/2019                         ##
####################################################################

# Install/import packages
library(Hmisc) # for variable labels
library(sqldf) # for county level values

# Set working directory
getwd()
setwd("C:/Users/nsigalo/Documents/SURV699U/Final Project")

# Import Data
# Food Atlas
food_atlas <-read.csv("atlas.csv")

# Education data
education <- read.csv("ACS_15_5YR_Education.csv")

# Age/Sex data
agesex <-read.csv("ACS_15_5YR_AgeSex.csv")

# Housing data
home <- read.csv("ACS_15_5YR_Home.csv")

# Living Alone by Sex
```

Comments

# Style Guide - R Markdown



Header

Comments

~~~
---
title: "Coronavirus Exploratory Data Analysis"
author: "Nikki Sigalo"
subtitle: "3/1/2020"
output: html_document
---

# Introduction
Students will have their introduction narrative here. Minimum 100 words.

# Data Cleaning/Preparation
Students will have their data cleaning narrative here. Minimum 100 words.
```{r}
library(readxl)
library(dplyr)
library(ggplot2)
library(summarytools)

# set wd
setwd("C:/Users/nsigalo/OneDrive - Mathematica/Documents/INST408F/Homework/Homework 1")

# Import
cv <- read_excel("coronavirus.xlsx")
cm <- read_excel("comorbidity.xlsx")

# Merge
cv2 <- inner_join(cv, cm, by="ID")

# Recode
cv3 <- mutate(cv2, Age = ifelse(Age >= 120, NA, Age))
cv3 <- mutate(cv3, Country = ifelse(Country == 'Mainland China', 'China', Country))
```
~~~

# R Help

- Instructor
- Google
- StackOverflow
- StackExchange
- Reddit
- Quora

# Installing R/R Studio (10 minutes)

- Available at https://posit.co/products/open-source/rstudio/