CSUS-Apr30_fruit

Sabah Ul-Hasan

4/30/2021

Install Relevant Packages and Load Libraries

```
##### Package Installation #####
# You can uncomment to re-install packages and update the versions
# Note to check your version of R + RStudio when running into installation
trouble

#  Need this package for downloading data directly from Github
## install.packages("RCurl")
## packageVersion("RCurl") # v1.98.1.3
### Helpful to know which version used for reproducibility (like lab
notebook)

#  Need this package for data clean-up
## install.packages("plyr")
## packageVersion("plyr") # v1.8.6

# Need this package for visualization of data
## install.packages("ggplot2")
## packageVersion("ggplot2") # v3.3.3
##### Package installation #####

# Can load directly, if packages already installed
##### Load libraries #####
library(RCurl)
library(plyr)
library(ggplot2)
##### Load libraries #####
```

Data Download & Organization

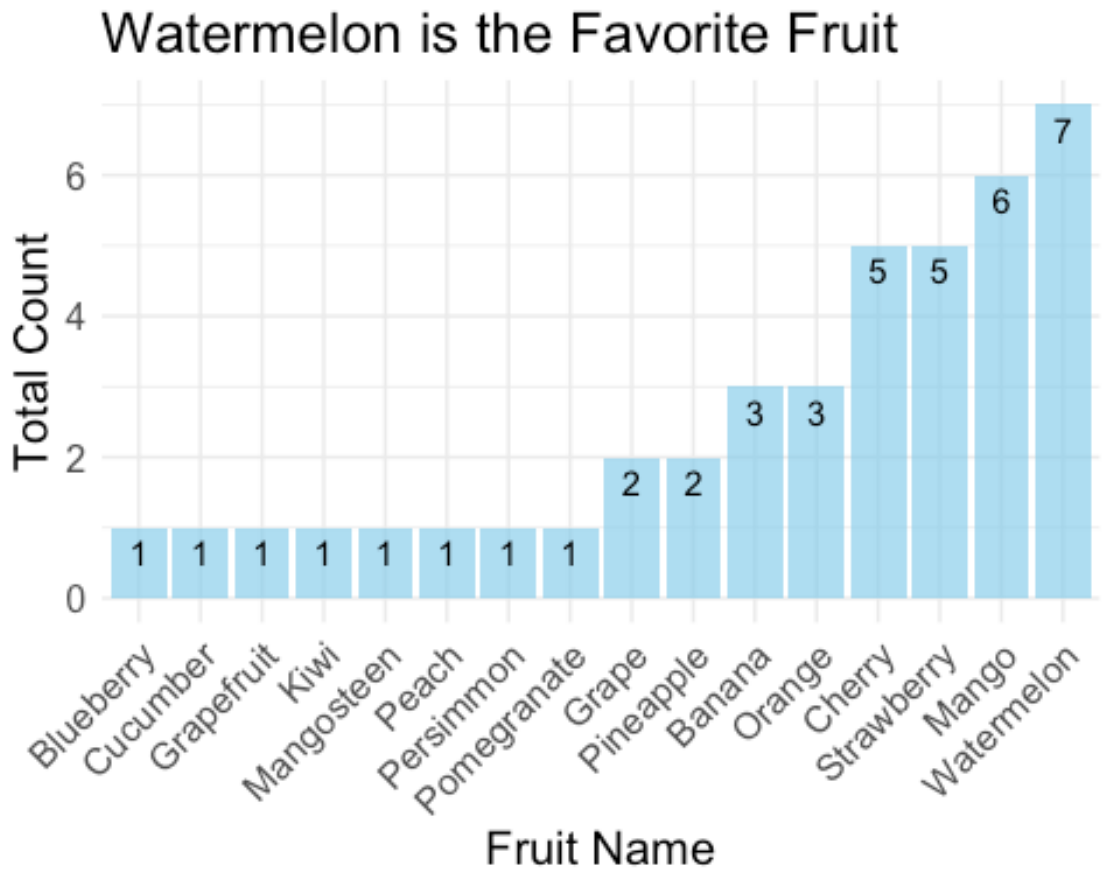
```
##### Upload Data #####
# Load data from our Github repo as 'df'
df <-
read.csv(text=getURL("https://raw.githubusercontent.com/sabahzero/dataviz/mas
ter/CSU-Stanislaus/CSUS-Apr30_fruit.csv"))

summary(df) # Quick view of our data, we can also see this in our
'Environment'

##      ï..ID        Best_Fruit            Number        Response
##  Min.   : 1.00   Length:42          Min.   :1.000   Length:42
```

```
##  1st Qu.:11.25   Class :character   1st Qu.:1.000   Class :character
##  Median :21.50   Mode  :character   Median :2.000   Mode  :character
##  Mean   :21.50                      Mean   :2.524
##  3rd Qu.:31.75                      3rd Qu.:3.750
##  Max.   :42.00                      Max.   :7.000

# About the data:
## This data is in response to the question, "What is the best fruit and
why?" at the start of the CSU Stanislaus Biology Department seminar Apr 30,
2021.
### 'ï..ID' or column 1 is a unique identifer. There were 42 responses of 69
total attendees (71 including seminar host and speaker) or ~61% of all
attendees.
### 'Best_Fruit' is the standardized fruit name (case sensitive) based on
responses attendees provided (raw response under 'Response' or column 4) in
the Zoom chat, responses are anonymized for privacy.
### 'Number' is an incremental count of how many times that fruit showed up
as a response for statistical purposes of determining mean and standard
deviation
##### Upload Data #####

##### Data Clean-up #####
# Let's rename 'ï..ID' to 'ID'
names(df)[names(df)=="ï..ID"] <- "ID"

# Let's remove the row with "Any" for focus on specific fruits
df <- df[-grep("Any", df$Best_Fruit),]

# Let's create a 5th column that categorizes the fruits
## This requires us to utilize what is known as a 'conditional' statement
df$Category <-
ifelse(grepl("Blueberry|Strawberry|Grape|Cucumber|Kiwi|Pomegranate|Banana|Per
simmon|Pineapple", df$Best_Fruit), "Berry",
                    ifelse(grepl("Orange|Grapefruit", df$Best_Fruit),
"Citrus",
                           ifelse(grepl("Mango|Mangosteen", df$Best_Fruit),
"Tropical",
                                  ifelse(grepl("Cherry|Peach",
df$Best_Fruit), "Stone",
                    "Melon"))))

# Let's create a 2nd data frame counts the number of times we see each fruit
fruit <- as.data.frame(table(df$Best_Fruit))
names(fruit)[names(fruit)=="Var1"] <- "Fruit"

# Let's create a 3rd data frame counts the number of times we see each
category
category <- as.data.frame(table(df$Category))
names(category)[names(category)=="Var1"] <- "Category"
# Let's create a 3rd column that shows number of unique fruits corresponding
```

```
to category (versus the previous doing a total count)
fruit$Category <-
ifelse(grepl("Blueberry|Strawberry|Grape|Cucumber|Kiwi|Pomegranate|Banana|Per
simmon|Pineapple", fruit$Fruit), "Berry",
                     ifelse(grepl("Orange|Grapefruit", fruit$Fruit),
"Citrus",
                            ifelse(grepl("Mango|Mangosteen", fruit$Fruit),
"Tropical",
                                   ifelse(grepl("Cherry|Peach", fruit$Fruit),
"Stone",
                     "Melon"))))

# Let's create a 4th data frame that counts how many of these unique fruits
are in a category
catunq <- as.data.frame(table(fruit$Category))
names(catunq)[names(catunq)=="Var1"] <- "CatUnq"
##### Data Clean-up #####
```

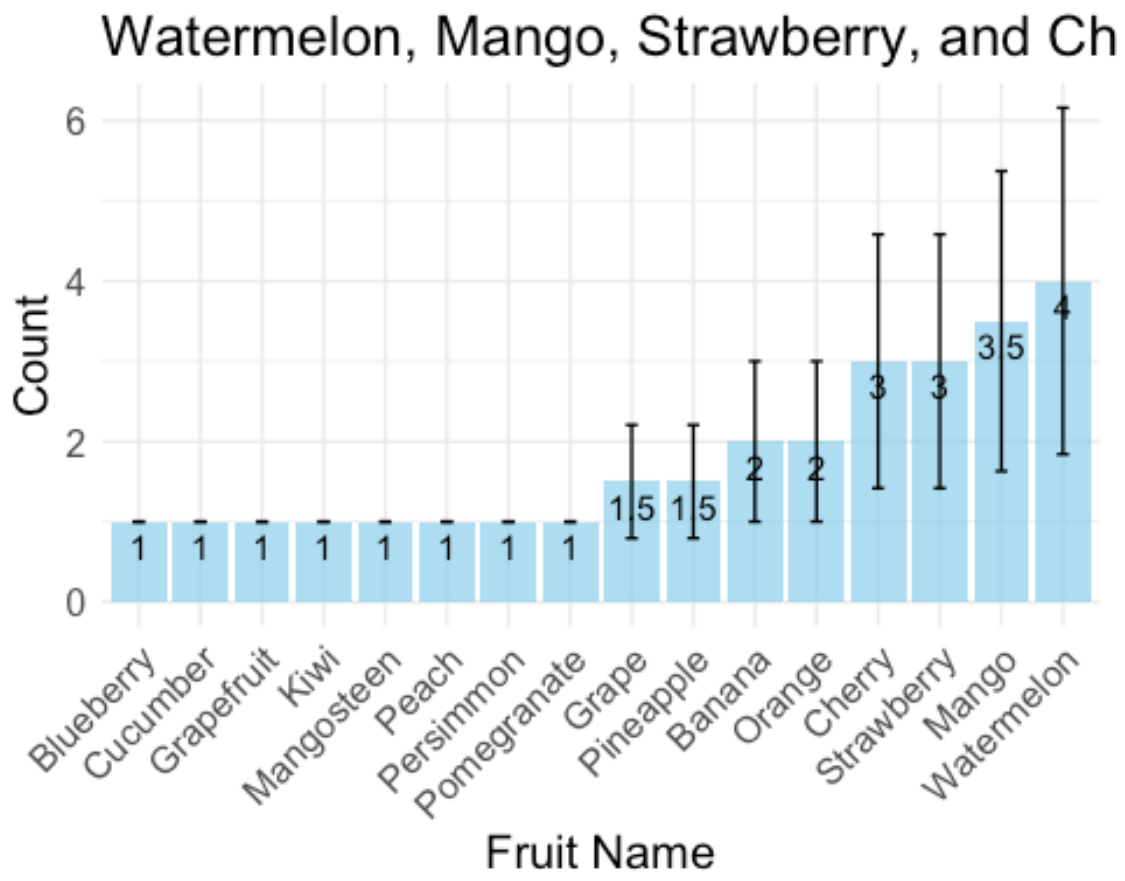
Visualizing Scientific Questions

```
# 1a. What do attendees of the Apr 30, 2021 CSUS Bio Dept Seminar consider
the 'best fruit'?
ggplot(fruit, aes(x=reorder(Fruit, Freq), y=Freq)) + # Reorder by ascending
order
  geom_col(fill="skyblue", alpha=0.7) + # bar chart with sky blue coloration
  geom_text(aes(label = Freq), vjust = 1.5, colour = "black") + # label with
number
  ggtitle("Watermelon is the Favorite Fruit") + # Title chart with answer to
1a
  xlab("Fruit Name") + # X axis title
  ylab("Total Count") + # Y axis title
  theme_minimal() + # theme selection
  theme(text = element_text(size=15),  axis.text.x = element_text(angle=45,
hjust=1)) # x-axis text label sizes
```

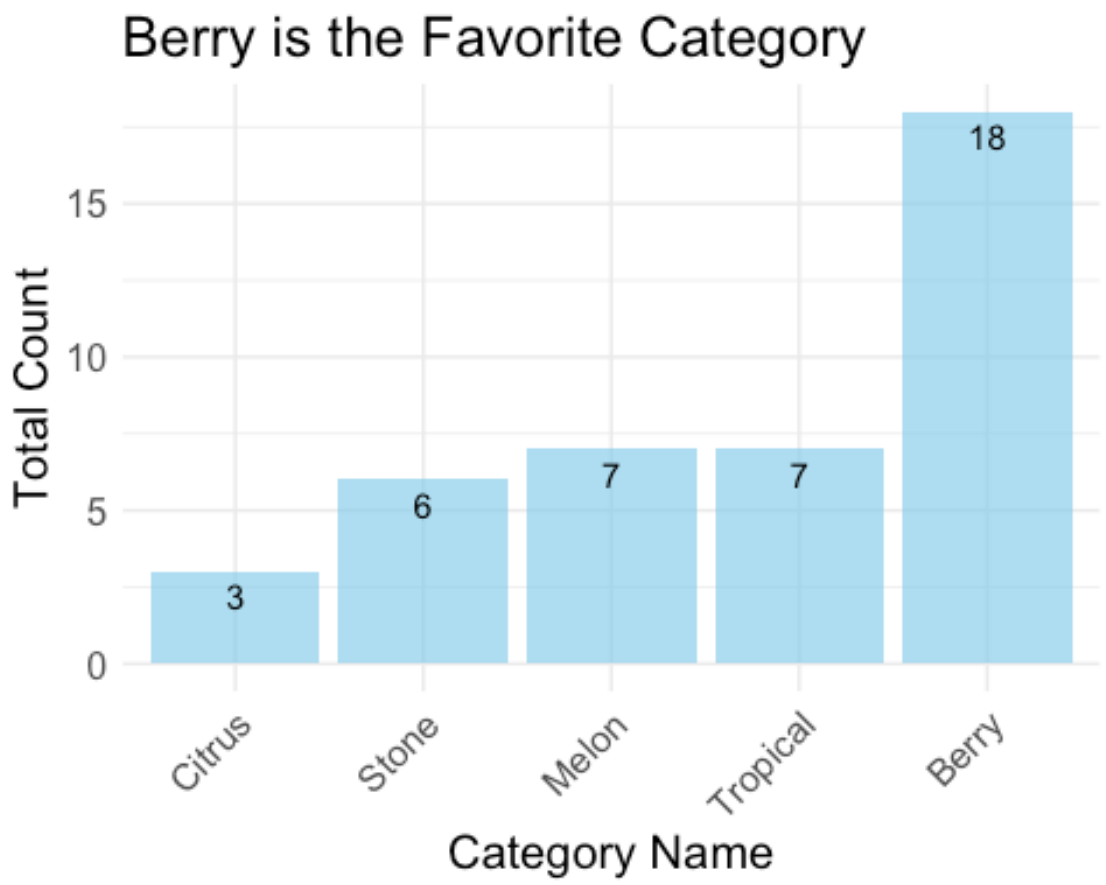



```r
# 1b. Are these results statistically significant?
## Define a function to assign mean and standard deviation
data_summary <- function(data, varname, groupnames){
  require(plyr)
  summary_func <- function(x, col){
    c(mean = mean(x[[col]], na.rm=TRUE),
      sd = sd(x[[col]], na.rm=TRUE))
  }
  data_sum<-ddply(data, groupnames, .fun=summary_func,
                  varname)
  data_sum <- rename(data_sum, c("mean" = varname))
 return(data_sum)
}
## Apply function to our data (5th data frame)
fruitsig <- data_summary(df, varname="Number", # Note that we incorporate
'Number' here
                    groupnames=c("Best_Fruit"))
fruitsig$sd[is.na(fruitsig$sd)] <- 0 # Assign NAs 0
## Plot a bar chart that includes standard deviation as error bars around
means
ggplot(fruitsig, aes(x=reorder(Best_Fruit, Number), y=Number)) +
  geom_col(fill="skyblue", alpha=0.7) +
  geom_text(aes(label=round((Number), digits = 2)), vjust = 1.5, colour =
```

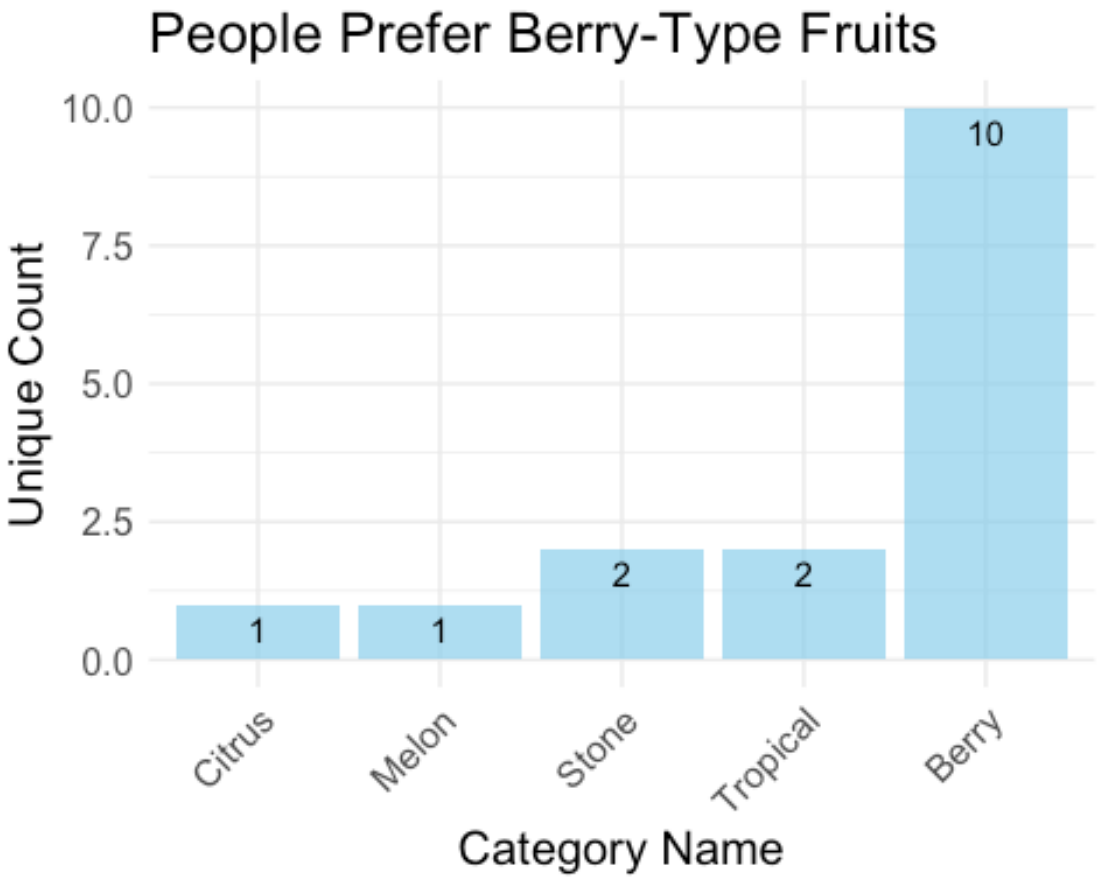
```
"black") + # Note the 'round()' function
  geom_errorbar(aes(ymin=Number-sd, ymax=Number+sd), width=.2,
position=position_dodge(.9)) + # Error bars
  ggtitle("Watermelon, Mango, Strawberry, and Cherry (1 sd)") +
  xlab("Fruit Name") +
  ylab("Count") +
  theme_minimal() +
  theme(text = element_text(size=15),  axis.text.x = element_text(angle=45,
hjust=1))
```




```
# 2. What is the top fruit category?
ggplot(category, aes(x=reorder(Category, Freq), y=Freq)) +
  geom_col(fill="skyblue", alpha=0.7) +
  geom_text(aes(label = Freq), vjust = 1.5, colour = "black") +
  ggtitle("Berry is the Favorite Category") +
  xlab("Category Name") +
  ylab("Total Count") +
  theme_minimal() +
  theme(text = element_text(size=15),  axis.text.x = element_text(angle=45,
hjust=1))
```




```
# 3. Which category has the highest number of unique fruits?
ggplot(catunq, aes(x=reorder(CatUnq, Freq), y=Freq)) +
  geom_col(fill="skyblue", alpha=0.7) +
  geom_text(aes(label = Freq), vjust = 1.5, colour = "black") +
  ggtitle("People Prefer Berry-Type Fruits") +
  xlab("Category Name") +
  ylab("Unique Count") +
  theme_minimal() +
  theme(text = element_text(size=15),  axis.text.x = element_text(angle=45,
hjust=1))
```



What Next?

```
# i. How would you figure out if the results for 2 or 3 were statistically
significant?
# ii. How would you plot all 4 bar charts into one graph?
# iii. What's a question you would ask based on the data available, and how
would you use code to answer it?
# iv. How would you clean-up or edit this code?
```