Statistical Brief 3

Annie Vellon

July 31, 2016

Executive Summary

On behalf of the Business Information and Analytics department at the University of Denver, I have conducted a statistical brief for the manager of Stats Dairy Ice Cream Shoppe.

Based on a Sales dataset from 2015, I have run numerous explorations in an attempt to gain a comprehensive understanding of the information, in addition to multiple staticial analyses using tests like ANOVA, correlation and t-tests to uncover what the data is telling us, and how the Stats Dairy Ice Cream Shop can use these analyses to enhnace their business model and increase profits.

In this brief you will find a section on how the data was imported for the analysis, a section on the data exploration, a section on the statistical tests/analyses, and finally, a conclusion section where final conclusion and recomendations are given based on exploration and analyses results.

Importing the Dataset

Importing the CSV File:

```
  setwd("/Users/annjosephine8/Documents")

SalesData<-read.csv("SalesData.csv")
```

Checking the Import:

```
head(SalesData, n=5)
```

```
##       Flavor Counts Order_Date Employee
## 1  Chocolate      1    12/7/15     Jose
## 2  Chocolate    104    6/26/15     Jose
## 3 Rocky Road     20     7/4/15     Jose
## 4    Vanilla     27   10/29/14     Jose
## 5    Vanilla      9    1/27/15      Sue
```

Adding and Modifying the Information:

```
Order_Date<- factor("1/15/2006 0:00:00")
as.Date(Order_Date, format = "%m/%d/%Y")
```

```
## [1] "2006-01-15"
```

```
SalesData$Day<- weekdays(as.Date(SalesData$Order_Date))

library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.2.5
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
SalesData$DayNum <- ifelse(SalesData$Day=="Sunday",1,
                           ifelse(SalesData$Day=="Monday",2,
                           ifelse(SalesData$Day=="Tuesday",3,
                           ifelse(SalesData$Day=="Wednesday",4,
                           ifelse(SalesData$Day=="Thursday",5,
                              ifelse(SalesData$Day=="Friday",6,7))))))
head(SalesData, n=5)
```

```
##       Flavor Counts Order_Date Employee    Day DayNum
## 1  Chocolate      1    12/7/15     Jose Sunday      1
## 2  Chocolate    104    6/26/15     Jose   <NA>     NA
## 3 Rocky Road     20     7/4/15     Jose Sunday      1
## 4    Vanilla     27   10/29/14     Jose   <NA>     NA
## 5    Vanilla      9    1/27/15      Sue   <NA>     NA
```

Data

Here, I will explore the SalesData dataset as I prepare to run analyses.

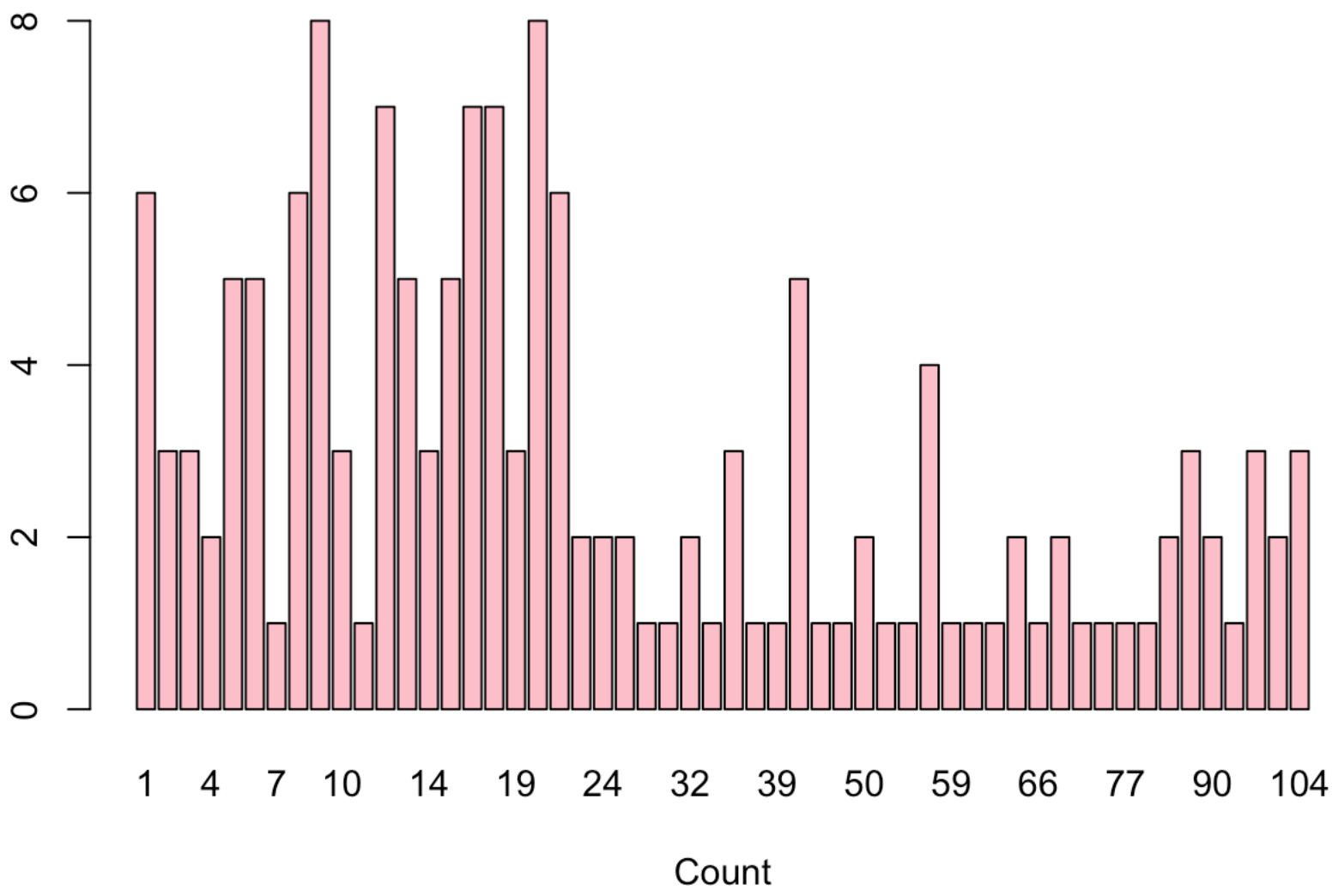
Summary Statistics of Counts:

```
summary(SalesData$Counts)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    9.00   17.00   30.36   45.00  104.00
```

Distribtuion of Counts:

```
barplot(table(SalesData$Counts), col="pink", xlab="Count")
```

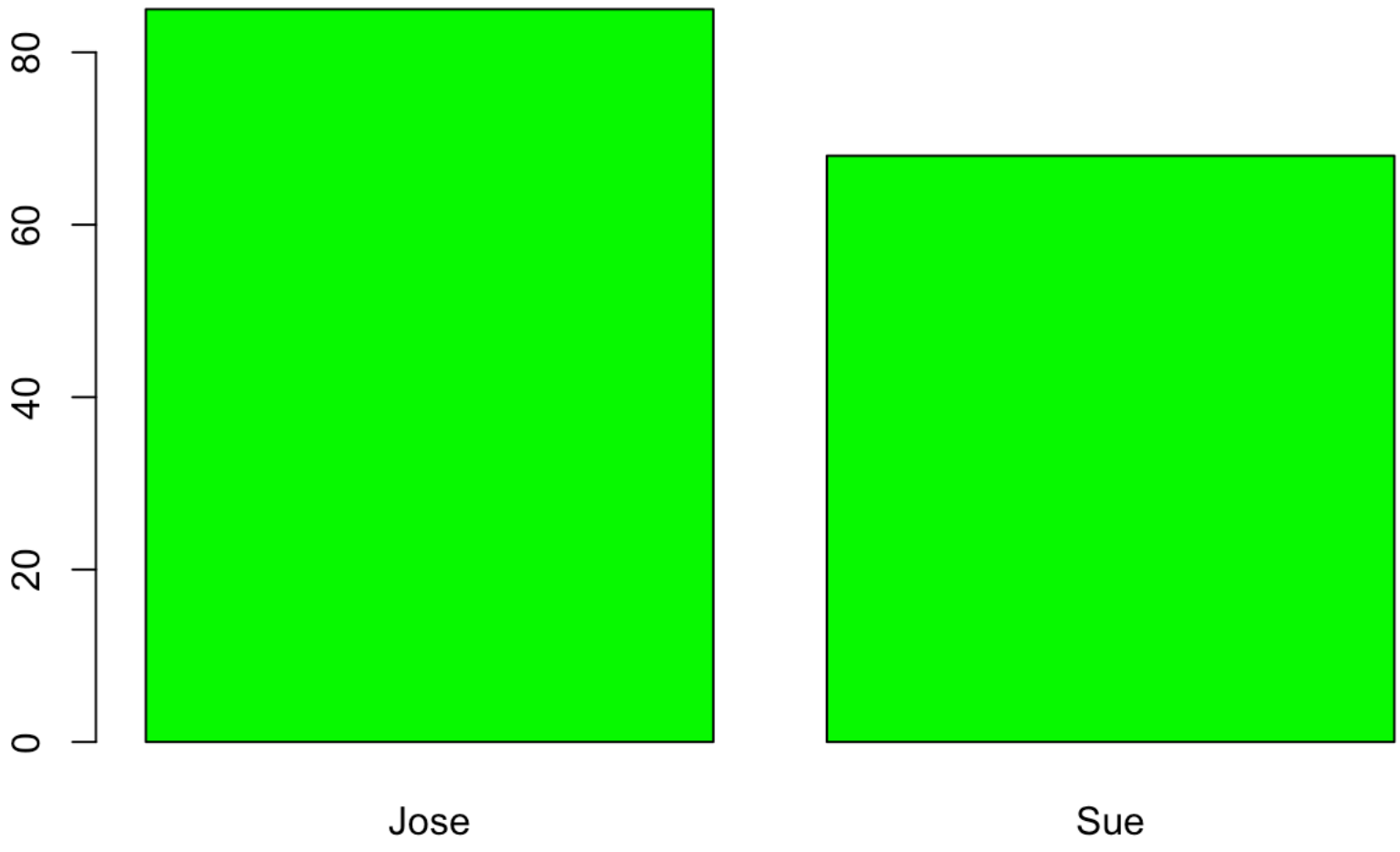


Number of Orders by ***Employee***:

```
EmployeeOrders<-table(SalesData$Employee)
EmployeeOrders
```

```
##
## Jose  Sue
##   85   68
```

```
barplot(EmployeeOrders,col="green")
```

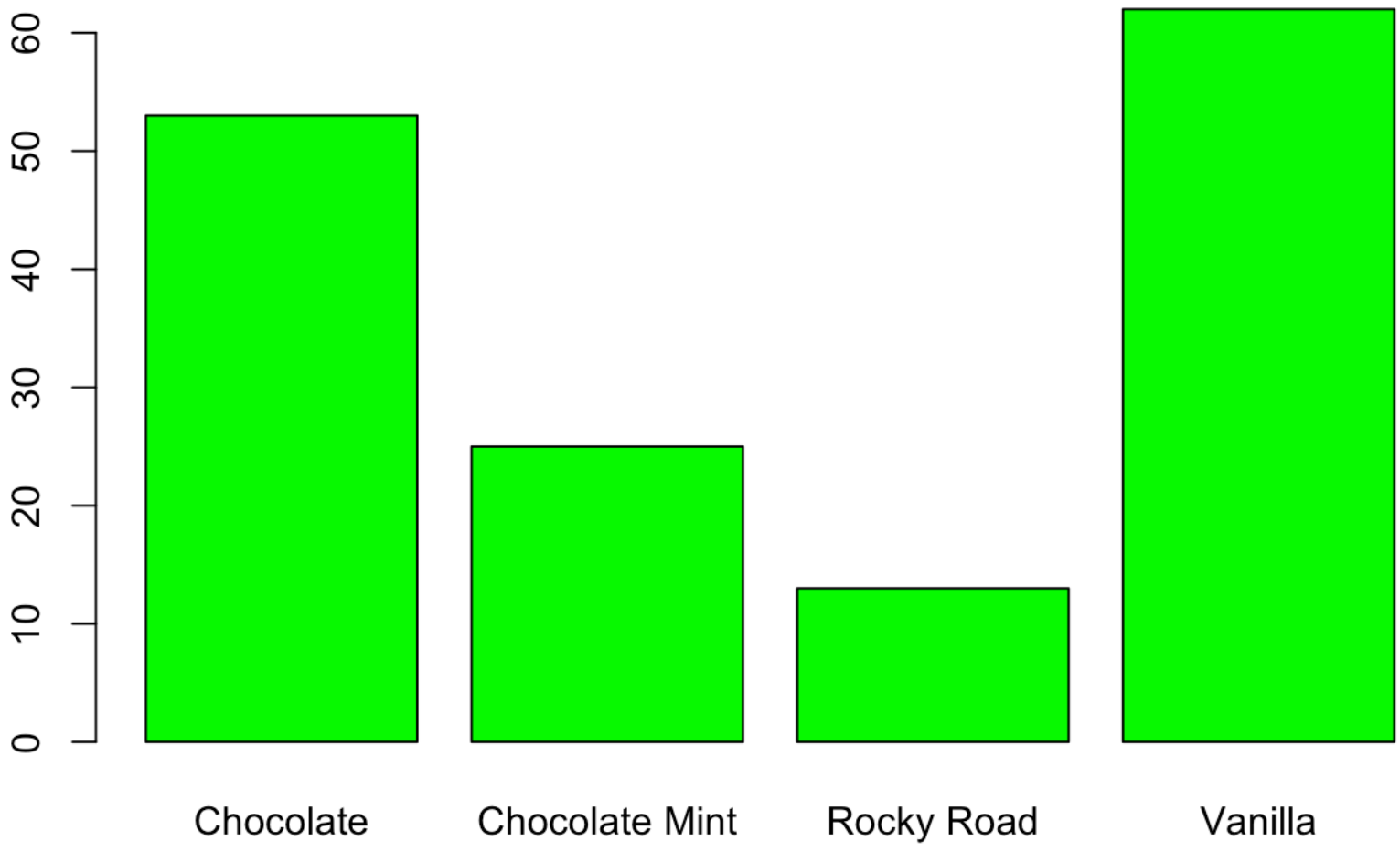


Number of Orders by ***Flavors***:

```
FlavorOrders<-table(SalesData$Flavor)
FlavorOrders
```

```
##
##      Chocolate Chocolate Mint     Rocky Road         Vanilla
##             53             25             13              62
```

```
barplot(FlavorOrders,col="green")
```



Total Quantity by ***Employee***:

```
library(dplyr)
SalesData %>%
  group_by(Employee) %>%
  summarise(Counts = sum(Counts))
```

```
## # A tibble: 2 x 2
##   Employee Counts
##     <fctr>  <int>
## 1     Jose   2803
## 2      Sue   1842
```

Total Quantity by ***Flavor***:

```
library(dplyr)
SalesData %>%
  group_by(Flavor) %>%
  summarise(Counts = sum(Counts))
```

```
## # A tibble: 4 x 2
##           Flavor Counts
##           <fctr>  <int>
## 1      Chocolate   1913
## 2 Chocolate Mint    560
## 3     Rocky Road    393
## 4        Vanilla   1779
```

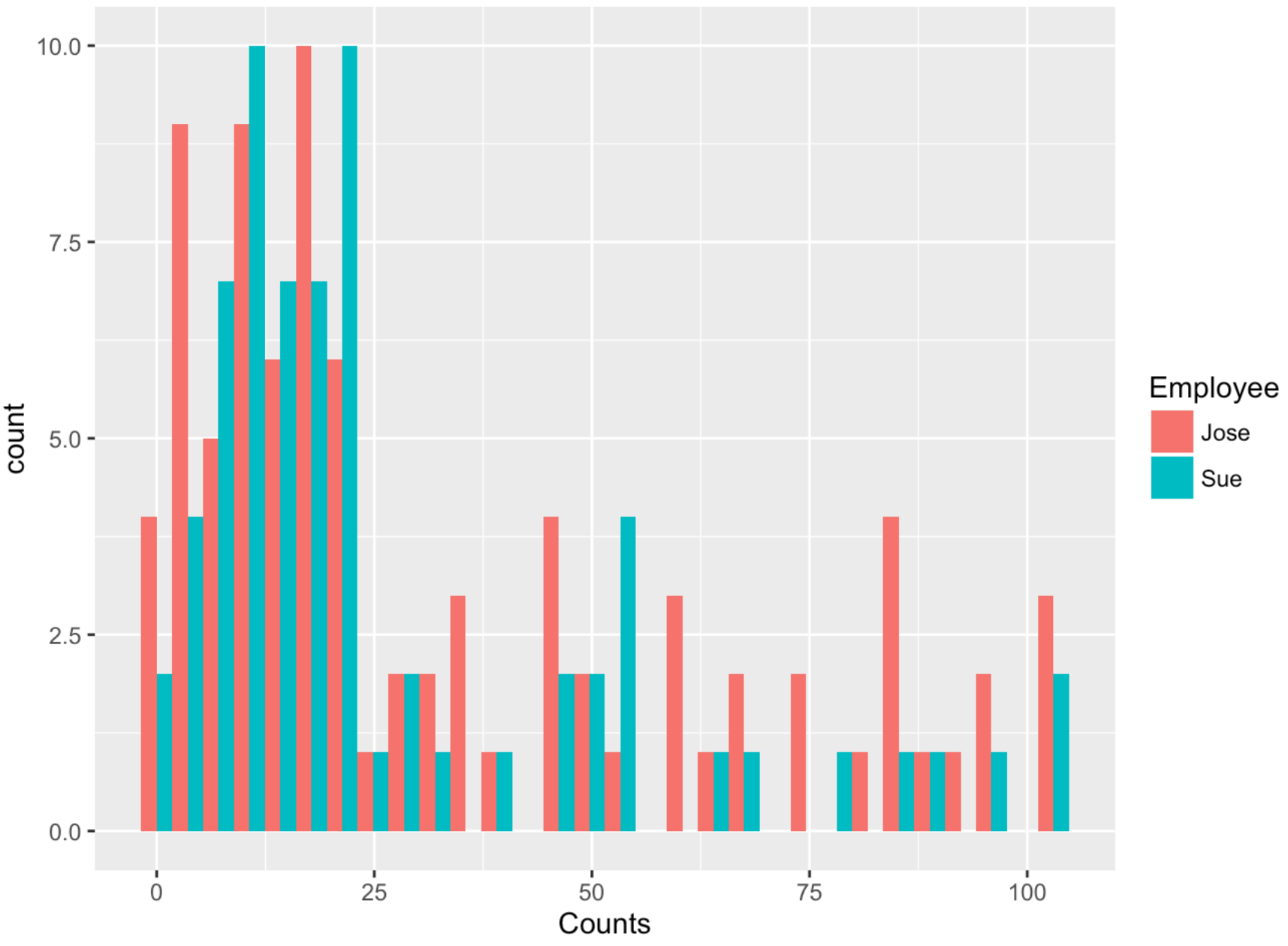
Distribution of Counts by ***Employee***:

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.2.4
```

```
ggplot(SalesData, aes(x=Counts, fill=Employee))+geom_histogram(position="dodge")
```

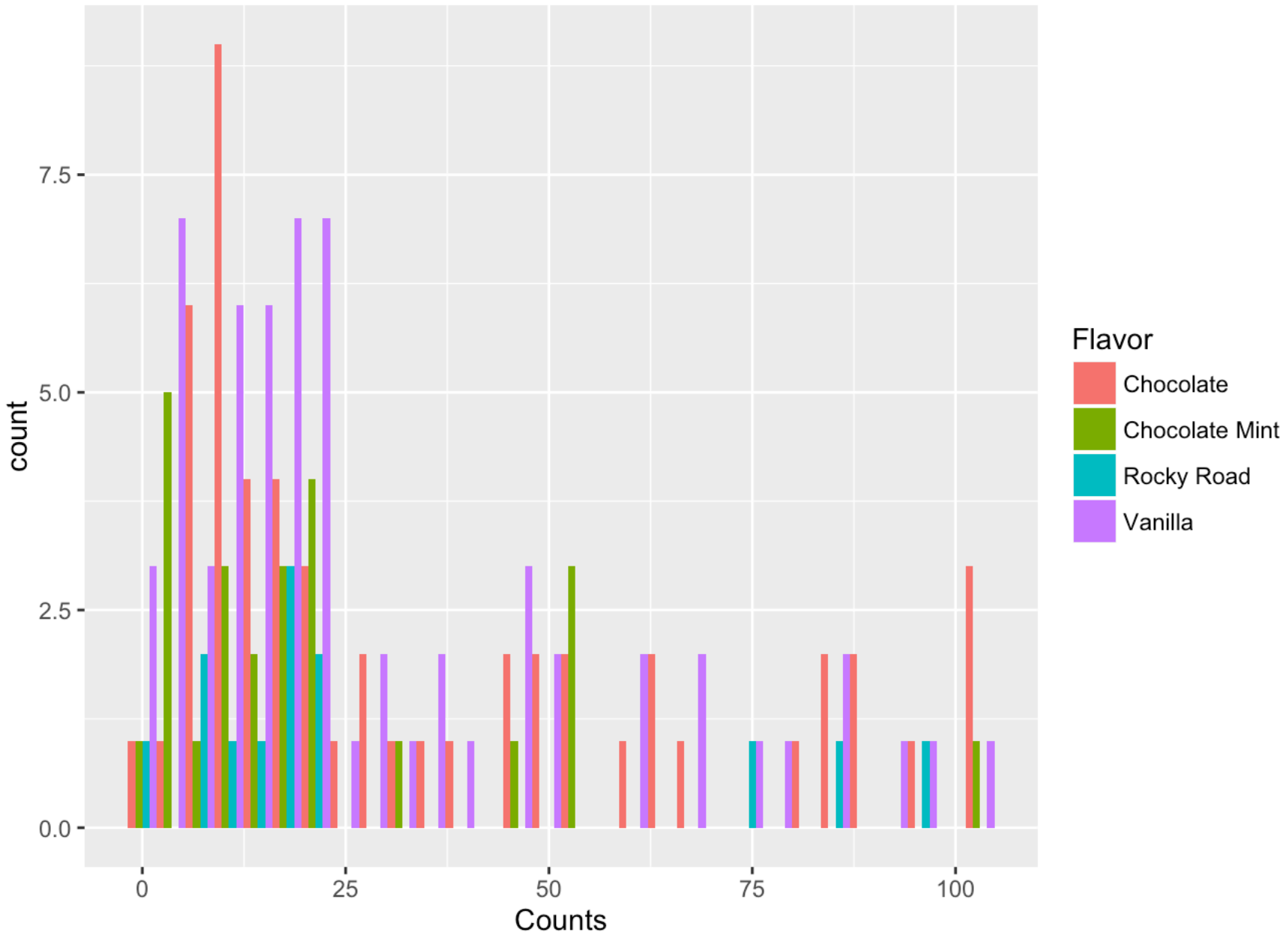
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Distribtuion of Counts by ***Flavor***:

```
ggplot(SalesData, aes(x=Counts, fill=Flavor))+geom_histogram(position="dodge")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Days of the Week Summary Statistics:

```
summarize(group_by (SalesData,Day), Mean=mean(Counts),Median=median(Counts),Min=min(C
ounts),Max=max(Counts))
```

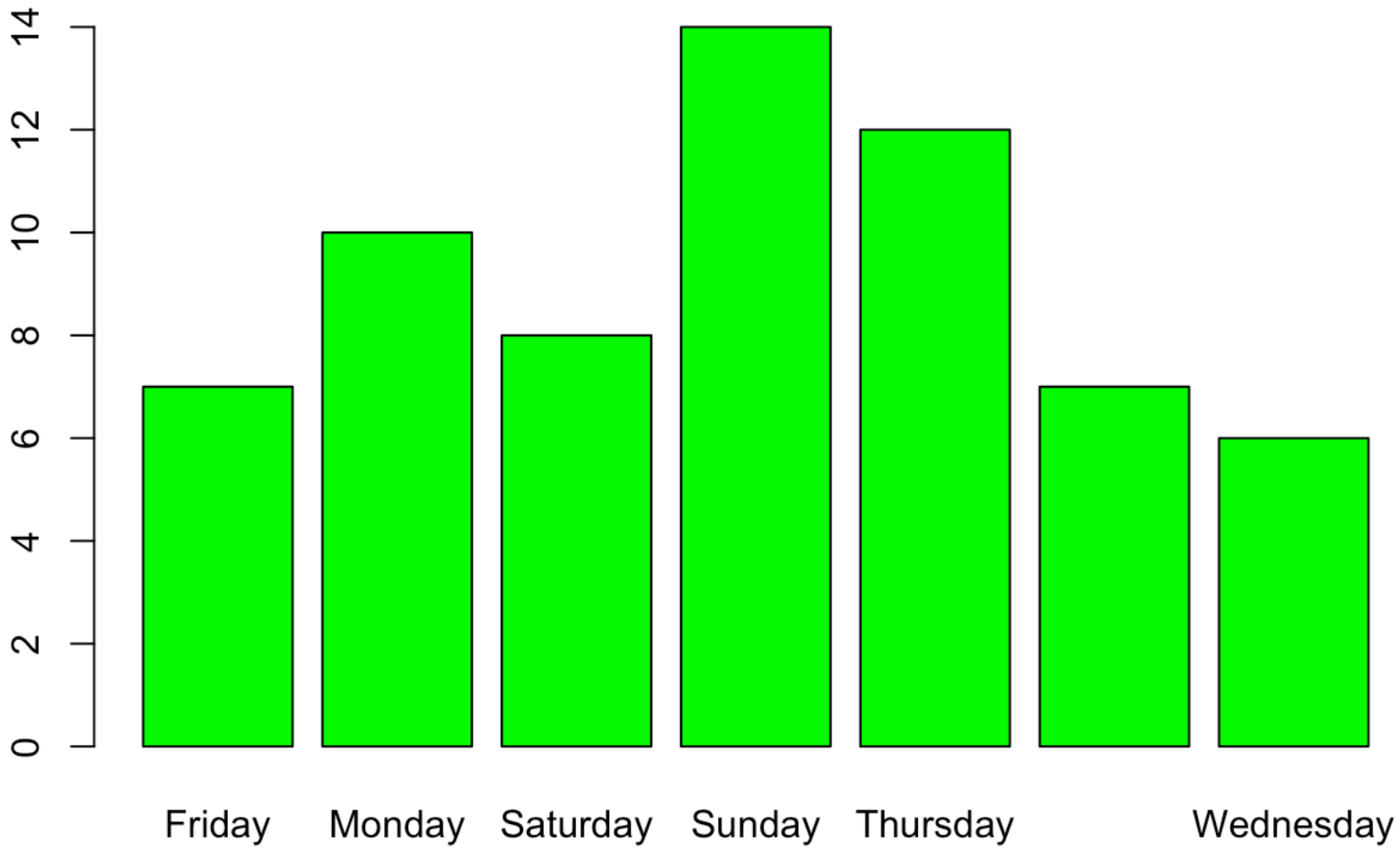
```
## # A tibble: 8 x 5
##         Day     Mean Median   Min   Max
##       <chr>    <dbl>  <dbl> <int> <int>
## 1    Friday 30.57143   20.0     3    74
## 2    Monday 23.80000   11.0     1    90
## 3  Saturday 24.25000   17.5     1    65
## 4    Sunday 29.14286   16.5     1    96
## 5  Thursday 40.66667   35.0     2    92
## 6   Tuesday 17.57143   14.0     8    45
## 7 Wednesday 24.83333   16.5     1    84
## 8      <NA> 31.80899   19.0     1   104
```

Number of Orders by ***Days of the Week***:

```
DayWeekOrders<-table(SalesData$Day)
DayWeekOrders
```

```
##
##    Friday    Monday  Saturday    Sunday  Thursday   Tuesday Wednesday
##         7        10         8        14        12         7         6
```

```
barplot(DayWeekOrders,col="green")
```



Column with a blank label is “NA”

Total Quantity by ***Days of the Week***:

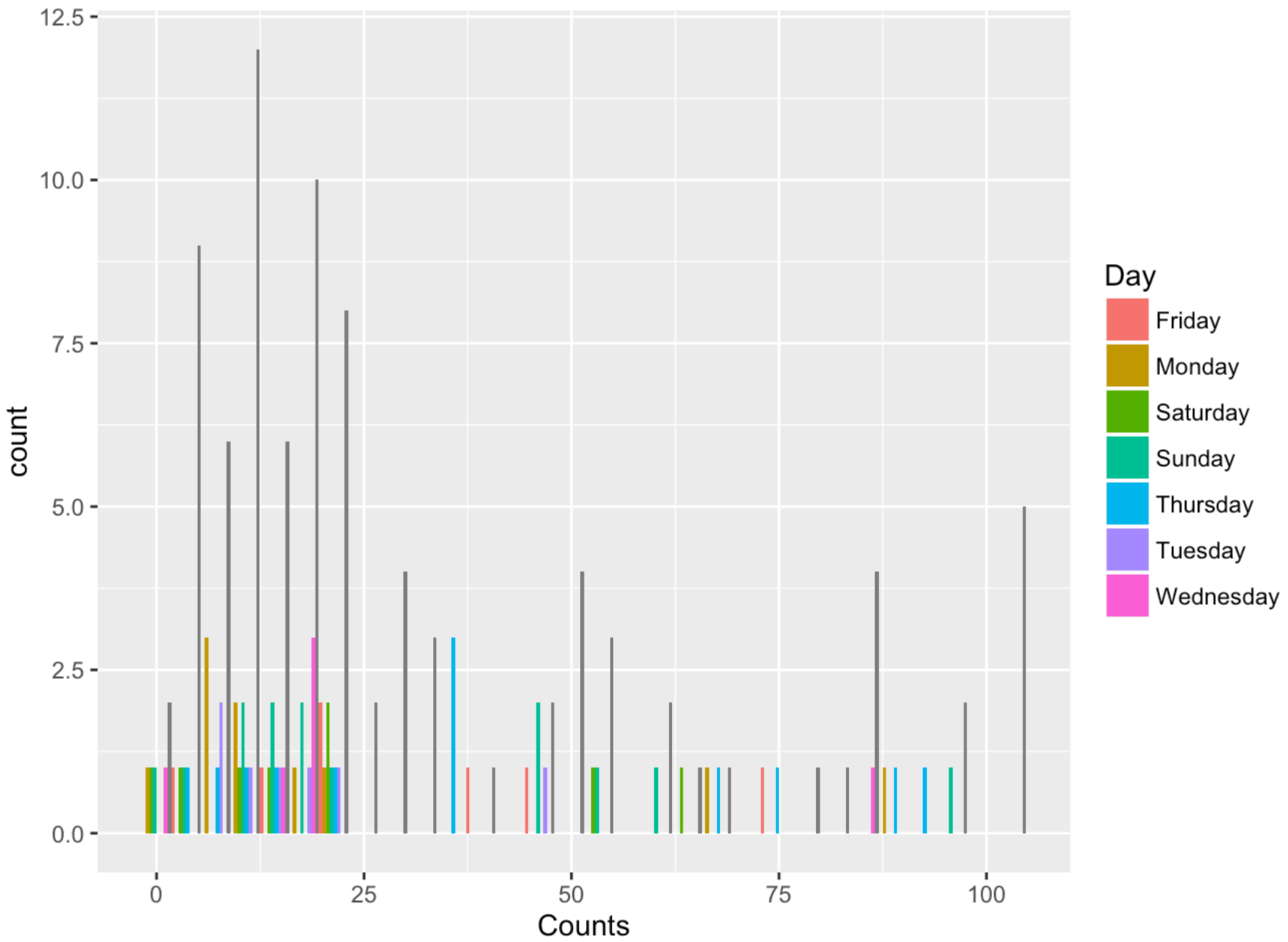
```
library(dplyr)
SalesData %>%
  group_by(Day) %>%
  summarise(Counts = sum(Counts))
```

```
## # A tibble: 8 x 2
##         Day Counts
##       <chr>  <int>
## 1    Friday    214
## 2    Monday    238
## 3  Saturday    194
## 4    Sunday    408
## 5  Thursday    488
## 6   Tuesday    123
## 7 Wednesday    149
## 8      <NA>   2831
```

Distribution of Counts by ***Days of the Week***:

```
ggplot(SalesData, aes(x=Counts, fill=Day))+geom_histogram(position="dodge")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Analyses

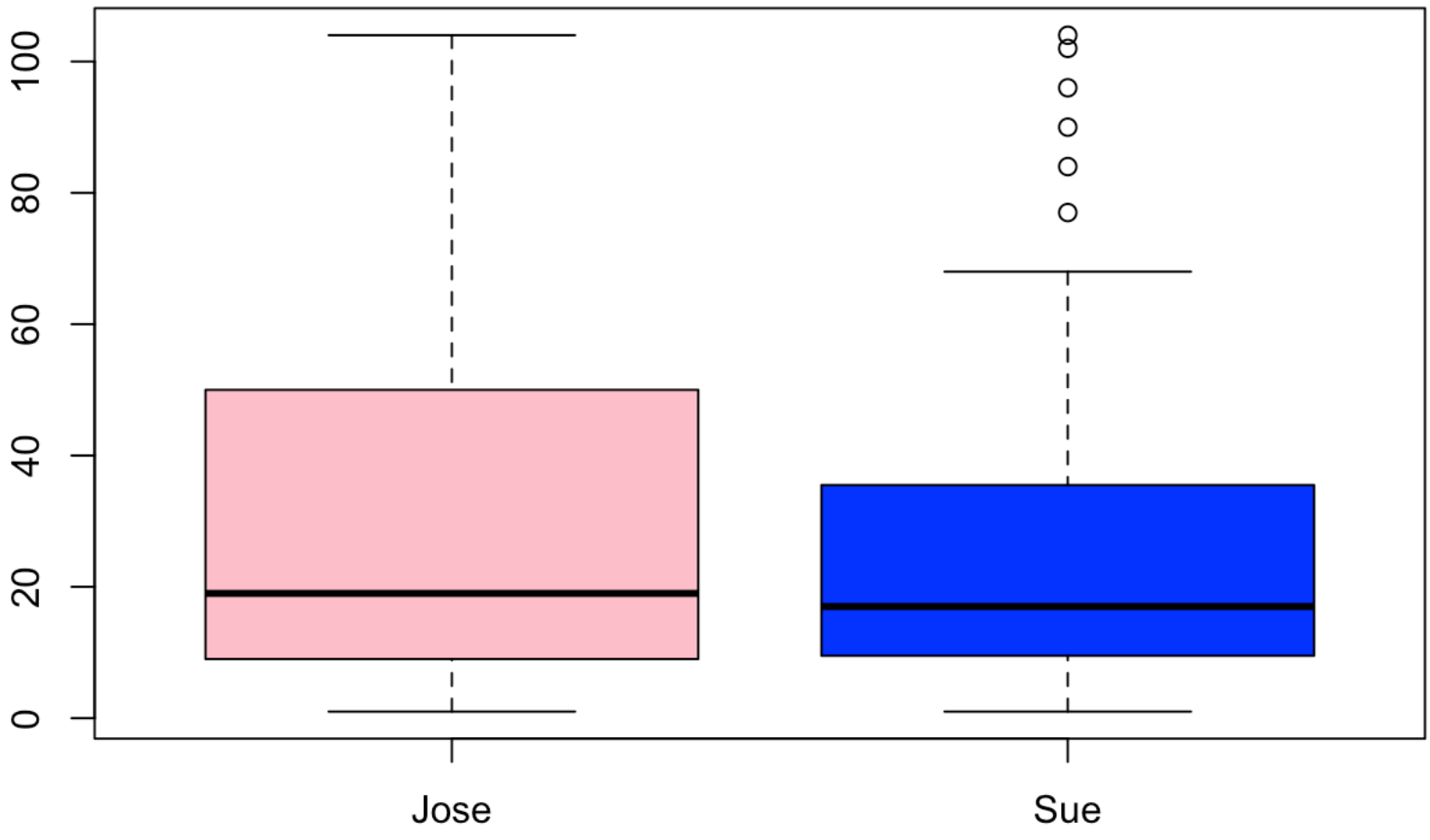
Here I have run multiple statistical tests on the SalesData data set, the tests and their results are included in this section. For these test's conclusions, please refer to the conclusions section.

Is there a difference in average Counts between Jose and Sue?

```
EmpTTest<-t.test(Counts~Employee, data=SalesData)
str(EmpTTest)
```

```
## List of 9
##  $ statistic  : Named num 1.28
##   ..- attr(*, "names")= chr "t"
##  $ parameter  : Named num 151
##   ..- attr(*, "names")= chr "df"
##  $ p.value    : num 0.202
##  $ conf.int   : atomic [1:2] -3.19 14.97
##   ..- attr(*, "conf.level")= num 0.95
##  $ estimate   : Named num [1:2] 33 27.1
##   ..- attr(*, "names")= chr [1:2] "mean in group Jose" "mean in group Sue"
##  $ null.value : Named num 0
##   ..- attr(*, "names")= chr "difference in means"
##  $ alternative: chr "two.sided"
##  $ method     : chr "Welch Two Sample t-test"
##  $ data.name  : chr "Counts by Employee"
##  - attr(*, "class")= chr "htest"
```

```
boxplot(Counts~Employee,data=SalesData,col=c("pink","blue"))
```

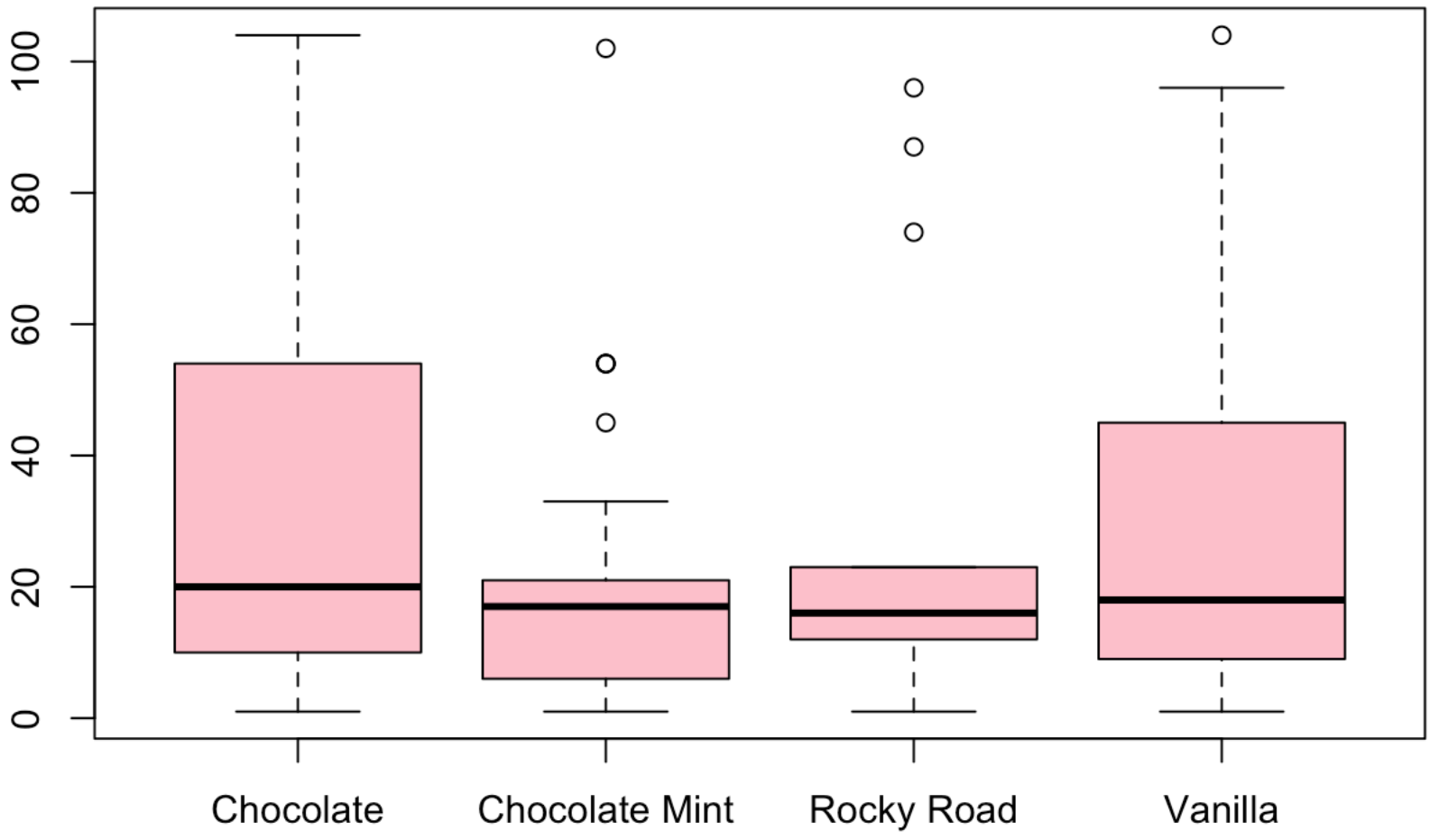


Is there a difference is average Counts between flavors?

```
FlavorANOVA<-aov(Counts~Flavor, data=SalesData)
FlavorANOVAResults<-summary(FlavorANOVA)
FlavorANOVAResults
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Flavor        3   3499  1166.4   1.415  0.241
## Residuals   149 122858   824.6
```

```
boxplot(Counts~Flavor,data=SalesData, col=c("pink"))
```



Is there a correlation between Count and DayNum?

```
SalesData_Cor<-select(SalesData,Counts,DayNum)
cor(SalesData_Cor)
```

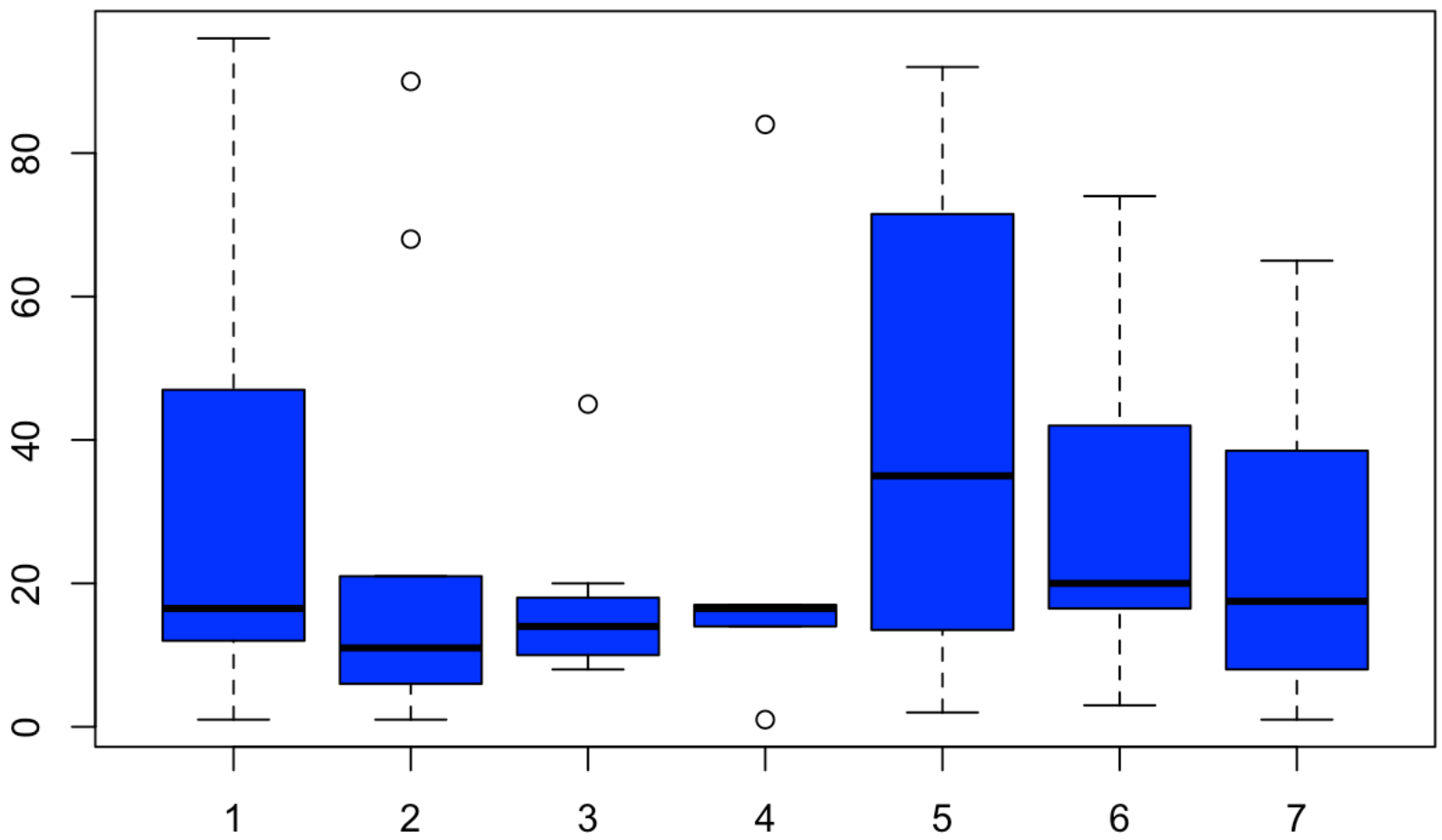
```
##         Counts DayNum
## Counts       1     NA
## DayNum      NA      1
```

Is there a difference in average Counts between DayNum?

```
DayNumANOVA<-aov(Counts~DayNum, data=SalesData)
DayNumANOVA.Results<-summary(DayNumANOVA)
DayNumANOVA.Results
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## DayNum        1    163   163.2   0.226  0.636
## Residuals    62  44709   721.1
## 89 observations deleted due to missingness
```

```
boxplot(Counts~DayNum, data=SalesData, col=c("blue"))
```



Due to the length of weekday names, I chose to use the numbers instead. 1=Sunday 7=Saturday.

Conclusions

Based on the results of the data exploration, there's a few conclusions we can jump to by looking with the naked eye. It would *appear* that the following are true based on the data exploration results…

-Joe has more orders than Sue.

-Chocolate and Vanilla are ordered more than Rocky Road and Chocolate Mint.

-Sundays are busier than the other week days.

However, we can never trust the naked eye to make a final decision. We must rely on statistical tests to help us analyze the data. When looking at the results of the tests, we are able to decipher if our origional assumptions are statistically significant, and worth investigating more. Based on the results, I have come up with the following conclusions…

-There is no difference in the number of orders between Joe and Sue.

-There is no difference in the number of orders between all of the ice cream flavors.

-There is no difference in the number of orders between all of the days of the week.

While the results show up our original assumptions may not be statistically relevent, it doens't hurt to keep those in the back of our minds. To be prepared, maybe keep an extra couple of gallons of chocolate and vanilla ice cream in the back and have 1 extra person working on Sundays to deal with the seemingly-higher influx of customers.

If you have any questions regarding any information in this statistical breif, feel free to contact me at annie.vellon@gmail.com (mailto:annie.vellon@gmail.com)