Test 1

Name: Alexander Hernandez | ID: 029531239

09/29/2022

```
library(UsingR)
```

```
## Loading required package: MASS
## Loading required package: HistData
## Loading required package: Hmisc
## Loading required package: lattice
## Loading required package: survival
## Loading required package: Formula
## Loading required package: ggplot2
##
## Attaching package: 'Hmisc'
## The following objects are masked from 'package:base':
##
##     format.pval, units
##
## Attaching package: 'UsingR'
## The following object is masked from 'package:survival':
##
##     cancer
```

```
library(VIM)
```

```
## Loading required package: colorspace
## Loading required package: grid
## VIM is ready to use.
## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues
##
## Attaching package: 'VIM'
## The following object is masked from 'package:datasets':
##
##     sleep
```

1) Short Answer Questions

a) Create the sequence: 15, 20, 20, 25, 25, 25, 30, 30, 30, 30, 35, 35, 35, 35, 35

```
rep(seq(15,35,5), times=c(1,2,3,4,5))
```

```
##  [1] 15 20 20 25 25 25 30 30 30 30 35 35 35 35 35
```

b) Generate 100 random numbers from a norm dist with mean=10, var=9. Only print first 5

```
head(rnorm(100, mean=5, sd=sqrt(9)))
```

```
## [1] 9.0611355 0.3189912 4.2948704 9.2722366 2.8733409 8.6311988
```

c) 'brightness' dataset in 'UsingR'. How many obs and print first 5

```
length(brightness)
```

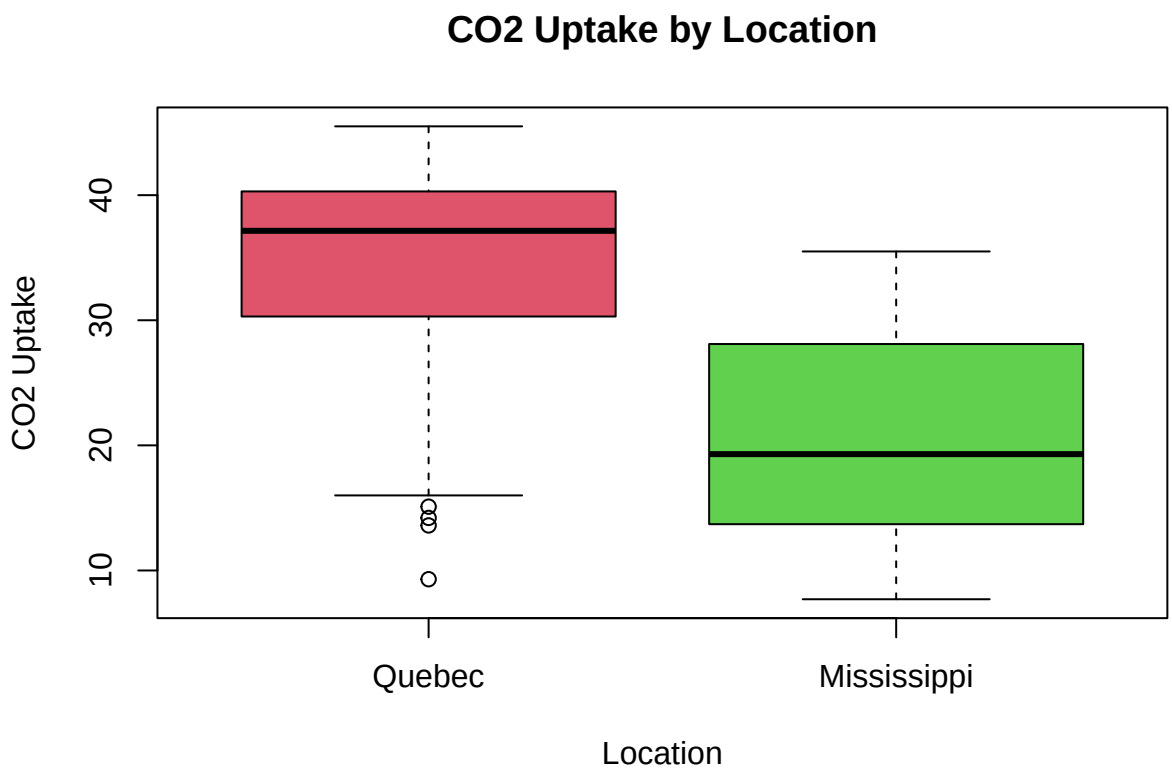
```
## [1] 966
```

```
head(brightness)
```

```
## [1]  9.10  9.27  6.61  8.06  8.55 12.31
```

d) 'CO2' dataset. Draw a side-by-side boxplot of CO2 uptake by location (quebec vs mississippi)

```
boxplot(CO2$uptake ~ CO2$Type, col=c(2,3),
        main= "CO2 Uptake by Location",
        xlab="Location", ylab="CO2 Uptake")
```



e) Import 'urine.txt' url into R and calculate average conductivity of urine

```
urine = read.table("https://www.stat.auckland.ac.nz/~wild/data/Rdatasets/txt/boot/urine.txt",
                   header=TRUE)
mean(urine$cond, na.rm=TRUE)
```

```
## [1] 20.90128
```

2) 'chickwts' dataset

a) How many variables are in the database?

```
length(names(chickwts))
```

```
## [1] 2
```

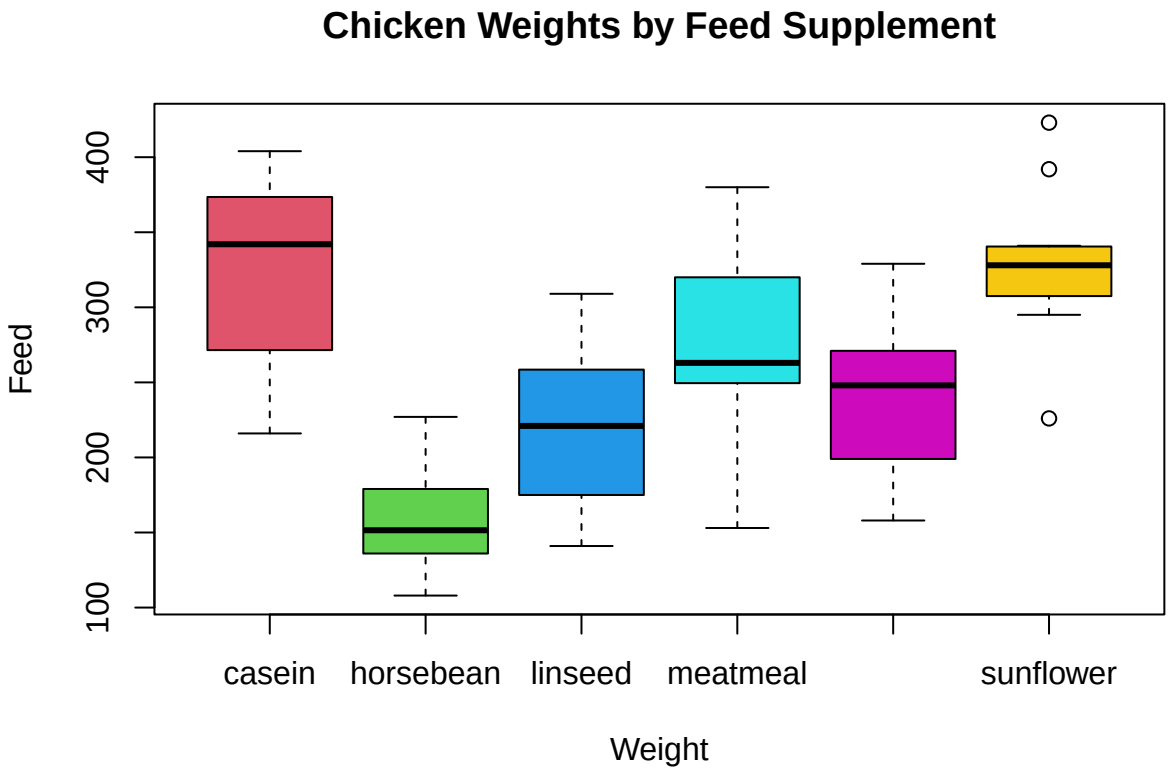
b) What is the dimension of the dataset?

```
dim(chickwts)
```

```
## [1] 71  2
```

c) Display in side-by-side boxplot using appropriate variable

```
boxplot(chickwts$weight ~ chickwts$feed, col=c(2,3,4,5,6,7),
        main= "Chicken Weights by Feed Supplement",
        xlab="Weight", ylab="Feed")
```



d) Construct a 95% confidence interval for the horsebean weight

```
t.test(chickwts$weight[chickwts$feed == "horsebean"],
       conf.level = .95)$conf.int
```

```
## [1] 132.5687 187.8313
## attr(,"conf.level")
## [1] 0.95
```

3) 'VIM' package, dataset 'tao'

a) How many variables and print their names

```
length(names(tao))
```

```
## [1] 8
```

```
names(tao)
```

```
## [1] "Year"             "Latitude"         "Longitude"        "Sea.Surface.Temp"
```

```
## [5] "Air.Temp"         "Humidity"         "UWind"            "VWind"
```

b) Remove missing values of NA's to create dataset, 'Clean'

```
Clean = na.omit(tao)
```

c) How many observations are in 'Clean'

```
nrow(Clean)
```

```
## [1] 565
```

d) Test test the hypothesis that the mean Air.Temp is greater than 25

```
# Null: u = 25
# Alt:  u > 25
t.test(Clean$Air.Temp, alt="greater", mu=25)
```

```
##
##  One Sample t-test
##
## data:  Clean$Air.Temp
## t = 4.1729, df = 564, p-value = 1.741e-05
## alternative hypothesis: true mean is greater than 25
## 95 percent confidence interval:
##  25.2068     Inf
## sample estimates:
## mean of x
##  25.34172
```

```
# As the p-value is 1.741e-05,
# there is enough evidence to reject the null hypothesis.
```

4) Insurance Quotes

a) Calculate the summary statistics for both local and online quotes

```
quotes = read.csv("C:\\repos\\STAT 50001\\Test 1\\quotes.tsv",
                  sep='\t', header=TRUE)
summary(quotes$Local)
```

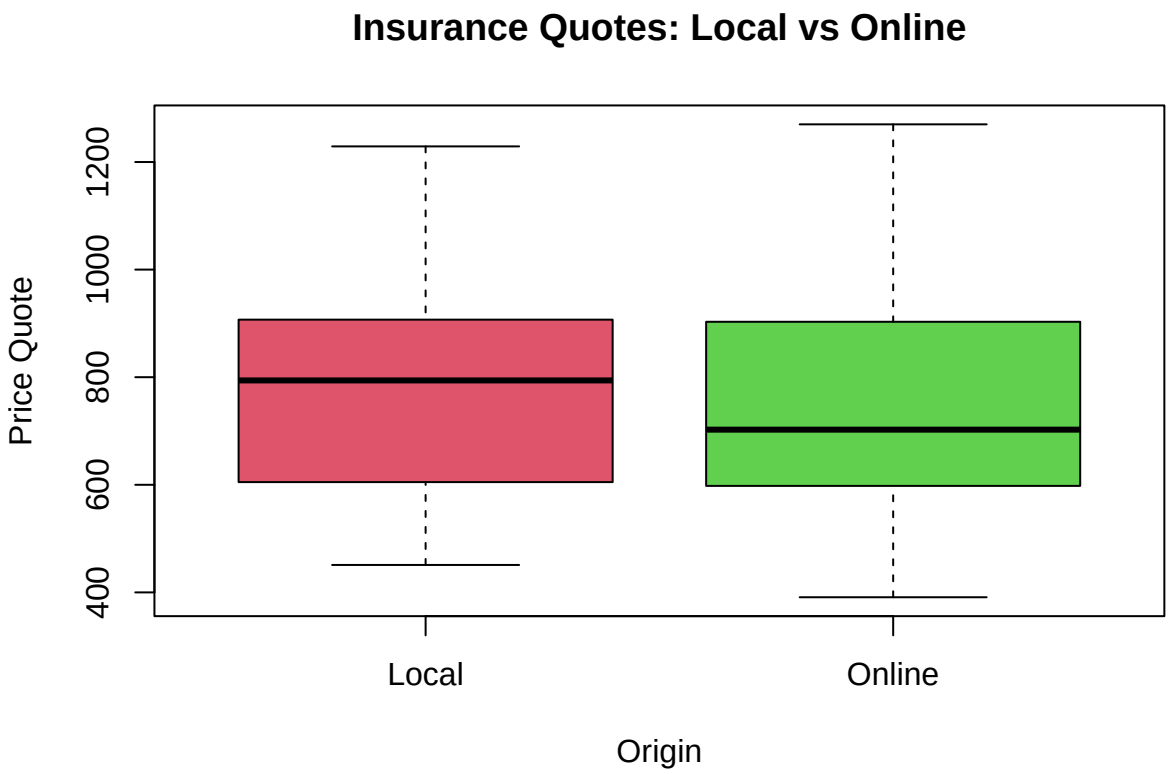
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   451.0   623.8   794.0   791.2   898.2  1229.0
```

```
summary(quotes$Online)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   391.0   599.0   702.5   753.6   881.0  1270.0
```

b) Display the information with side-by-side boxplots

```
boxplot(quotes, col=c(2,3), main="Insurance Quotes: Local vs Online",
        ylab="Price Quote", xlab="Origin")
```



c) Does the data support that the online quotes are cheaper than the one provided by a local agent?

```
# Null: u(O) - u(L) = 0
# Alt:  u(O) - u(L) < 0
t.test(quotes$Local, quotes$Online, alt="less", mu=0)
```

```
##
##  Welch Two Sample t-test
##
## data:  quotes$Local and quotes$Online
## t = 0.41497, df = 25.708, p-value = 0.6592
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 192.0612
## sample estimates:
## mean of x mean of y
##  791.2143  753.6429
```

```
# With a p-value of 0.6592,
# we do not have enough evidence to reject the null.
```

5) Chicago DOBA and Consumer Protection Taxi Data

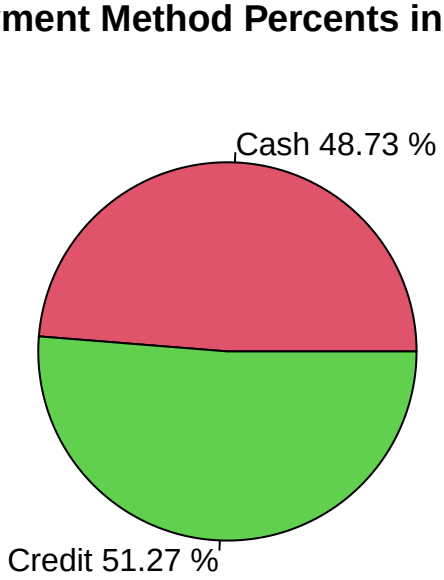
a) Import the data in R

```
taxi = read.table("C:\\repos\\STAT 50001\\Test 1\\chicago_taxi.txt",
                  sep='\t', header=TRUE)
```

b) Draw a pie chart

```
taxi_payments = table(taxi$payment_type)

total = taxi_payments[1] + taxi_payments[2]
cash_percent = round(100 * taxi_payments[1] / total,
                     digits = 2)
credit_percent = round(100 * taxi_payments[2] / total,
                     digits = 2)
taxi_labels = c(paste("Cash", cash_percent, "%"),
                paste("Credit", credit_percent, "%"))
pie(taxi_payments, labels = taxi_labels, col=c(2,3),
    main="Taxi Payment Method Percents in Chicago")
```



c) Construct a 95% confidence interval for the tip amounts based on method of payment

```
t.test(taxi$tips ~ taxi$payment_type)$conf.int
```

```
## [1] -3.949119 -3.783814
## attr(,"conf.level")
## [1] 0.95
```

d) Is there a significant difference in the tips amount by daytype?

```
# Null: u(d) - u(e)  = 0
# Alt:  u(d) - u(e) != 0
t.test(taxi$tips[taxi$daytype=="weekday"],
       taxi$tips[taxi$daytype=="weekend"], mu=0)
```

```
##
##  Welch Two Sample t-test
##
## data:  taxi$tips[taxi$daytype == "weekday"] and taxi$tips[taxi$daytype == "weekend"]
## t = 5.5765, df = 7642.2, p-value = 2.537e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2084284 0.4343954
## sample estimates:
## mean of x mean of y
##  2.075906  1.754494
```

```
# With a p-value of 2.537e-08,
# we have enough evidence to reject the null hypothesis.
```