STA3000 - Statistical Computing
In-class activity #1

Class survey

- How many students filled out the survey?
  Code:
  ```
  nrow(cs)
  ```

  Answer:
  24.

- What is the percentage of students who have prior coding experience?
  Code:
  ```
  100*round(mean(cs$Do.you.have.any.prior.coding.experience.      == 'Yes'),4)
  ```

  Answer:
  70.83

- What percentage of students is from NYC?
  Code:
  ```
  100*round(mean(cs$Are.you.from.NYC. == 'Yes'),4)
  ```

  Answer:
  66.67

- Provide a table that gives the percentage of students who are from Manhattan, Brooklyn, Queens, the Bronx, and Staten Island.
  Code:
  ```
  from_mbqbs=table(cs$If.you.are.from.NYC..where.from.)
  from_mbqbs_perc=100*round(prop.table(from_mbqbs), 4)
  from_mbqbs_perc=from_mbqbs_perc[-1]
  ```

  Answer:
  ```
       Brooklyn     Manhattan        Queens Staten Island
          25.00          8.33         25.00          4.17
      The Bronx
           4.17
  ```

- Create a table that gives the percentage of students who prefer the Yankees, the Mets, or don't care by borough. Do you think that the data indicates that there is some dependence between borough and the baseball team that you prefer?
  Code:
  ```
  baseball_borough=table(cs$Which.baseball.team.do.you.prefer.,
  cs$If.you.are.from.NYC..where.from.)
  100*round(prop.table(baseball_borough), 4)
  ```

```
100*round(prop.table(baseball_borough, 1),4)
100*round(prop.table(baseball_borough, 2),4)
```

Answer:

```
> 100*round(prop.table(baseball_borough), 4)

                   Brooklyn Manhattan Queens Staten Island
I don't care 20.83    20.83      4.17  16.67          4.17
Mets          0.00     0.00      4.17   0.00          0.00
Yankees      12.50     4.17      0.00   8.33          0.00

             The Bronx
I don't care      0.00
Mets              0.00
Yankees           4.17

> 100*round(prop.table(baseball_borough, 1),4)

                    Brooklyn Manhattan Queens Staten Island
I don't care  31.25    31.25      6.25  25.00          6.25
Mets           0.00     0.00    100.00   0.00          0.00
Yankees       42.86    14.29      0.00  28.57          0.00

             The Bronx
I don't care      0.00
Mets              0.00
Yankees          14.29

> 100*round(prop.table(baseball_borough, 2),4)

                    Brooklyn Manhattan Queens Staten Island
I don't care  62.50    83.33     50.00  66.67        100.00
Mets           0.00     0.00     50.00   0.00          0.00
Yankees       37.50    16.67      0.00  33.33          0.00

             The Bronx
I don't care      0.00
Mets              0.00
Yankees         100.00
```

We can see from the tables above that there are some dependencies between borough and the baseball team the students prefer. If a student is a fan of Mets, he/she must from Manhattan according to the survey; The students who live in the Bronx have a tendency to support Yankees; More than half of students who don't live in NYC, who lives in Brooklyn and Queens tend not have any preference between Yankees and Mets; All students from Staten Island do not have a preference between Mets and Yankees.

- What is the sport that most students seem to care about? Give the percentage of students who have a preference for a football, basketball, and baseball team.

  Code:

  ```
  100*round(mean(cs$Which.basketball.team.do.you.prefer.    !=    'I don\'t care'),4)
  100*round(mean(cs$Which.baseball.team.do.you.prefer. != 'I don\'t care'),4)
  ```

```
100*round(mean(cs$Which.football.team.do.you.prefer. != 'I don\'t
care'),4)
```

Answer:

```
 > 100*round(mean(cs$Which.basketball.team.do.you.prefer. != 'I don\'t
care'),4)
 [1] 33.33

 > 100*round(mean(cs$Which.baseball.team.do.you.prefer. != 'I don\'t ca
re'),4)
 [1] 33.33

 > 100*round(mean(cs$Which.football.team.do.you.prefer. != 'I don\'t ca
re'),4)
 [1] 29.17
```

We can see that most students care more about basketball and baseball. The percentage of students who have a preference for a football, basketball, and baseball team are 29.17, 33.33, and 33.33 respectively.

- What percentage of students speak only one language?
  Code:

```
cs$How.many.languages.do.you.speak.=
tolower(cs$How.many.languages.do.you.speak.)
cs$How.many.languages.do.you.speak.<-
as.character(cs$How.many.languages.do.you.speak.)
cs$How.many.languages.do.you.speak.[cs$How.many.languages.do.you.
speak.== 'one'] <- '1'
100*round(mean(cs$How.many.languages.do.you.speak. == '1'),4)
```

  Answer:

```
8.33
```

Hair length

Are hair length and age related? The following questions use a dataset from a study that tries to answer that question.

Data: http://users.stat.ufl.edu/~winner/data/hairlength.txt
Description: http://users.stat.ufl.edu/~winner/data/hairlength.dat

Read in the data and reformat it so that the variables are of the right type and have interpretable labels. Paste the code you used below.
Code:

```
hl=read.table("http://users.stat.ufl.edu/~winner/data/hairlength.dat",
header = FALSE)
colnames(hl) = c("HairLength","Age","Count")
```

```
hl$HairLength = factor(hl$HairLength)
levels(hl$HairLength) =  c("short","medium", "long")
hl$Age = factor(hl$Age)
levels(hl$Age) = c("14-24", "25-34", "35-49", "50-60")
```

- Give the percentage of women in the sample that are in each age group. Given those percentages, do you think that the dataset is a representative sample of women in the US?
  Code:
  ```
  Age1424 = hl %>% filter(Age == "14-24")
  100*round(sum(Age1424$Count)/sum(hl$Count),4)

  Age2534 = hl %>% filter(Age == "25-34")
  100*round(sum(Age2534$Count)/sum(hl$Count),4)

  Age3549 = hl %>% filter(Age == "35-49")
  100*round(sum(Age3549$Count)/sum(hl$Count),4)

  Age5060 = hl %>% filter(Age == "50-60")
  100*round(sum(Age5060$Count)/sum(hl$Count),4)
  ```

  Answer:
  The percentage in each age group are: 21.33, 21.58, 27.67, 29.42. I think this dataset could represents the women in US because the percentage of age from 50 to 60 is a relatively large portion compared to other age groups, which also reflects the problem of aging population.

- What is the percentage of women in the sample who have short, medium and long hair?
  Code:
  ```
  short = hl %>% filter(HairLength == "short")
  100*round(sum(short$Count)/sum(hl$Count),4)

  medium = hl %>% filter(HairLength == "medium")
  100*round(sum(medium$Count)/sum(hl$Count),4)

  long = hl %>% filter(HairLength == "long")
  100*round(sum(long$Count)/sum(hl$Count),4)
  ```

  Answer:
  33.04, 43.62, and 23.33 respectively.

- Provide a table that shows the percentage of women who have short, medium, and long hair given an age group. Do you see any trends? Explain what you see.
  Code:
  ```
  install.packages("tidyr")
  library(tidyr)
  hl2 = hl %>% uncount(Count)
  100*round(prop.table(table(hl2$HairLength, hl2$Age), 1),4)
  ```

  Answer:
  We can see that as age goes up, women tend to have shorter hair.

  ```
         14-24 25-34 35-49 50-60
  short   7.19 17.53 28.50 46.78
  medium 21.97 19.77 32.95 25.31
  long   40.18 30.71 16.61 12.50
  ```

- Out of all the women in the sample who have long hair, what percentage is in the youngest group?
  Code:
  ```
  ly <- long %>% filter(Age == "14-24")
  100*round(ly$Count/sum(long$Count),4)
  ```

  Answer:
  ```
  40.18
  ```

Height, weight, and age of NBA players

In this exercise, you'll work with a sample of NBA players from the 2013-2014 season.

You can find the dataset here: http://users.stat.ufl.edu/~winner/data/nba_ht_wt.csv

- Convert the height variable from inches into meters.
  Code:
  ```
  install.packages('measurements')
  library(measurements)
  nba$Height <- conv_unit(nba$Height, "inch", "m")
  ```

  Answer:

  ```
            Player Pos Height Weight Age
   Nate Robinson   G 1.7526    180  29
   Isaiah Thomas   G 1.7526    185  24
    Phil Pressey   G 1.8034    175  22
    Shane Larkin   G 1.8034    176  20
       Ty Lawson   G 1.8034    195  25
  John Lucas III   G 1.8034    157  30
   D.J. Augustin   G 1.8288    180  25
      Kyle Lowry   G 1.8288    205  27
  ```

- Convert the weight variable from pounds into kilograms.
  Code:
  ```
  nba$Weight <- conv_unit(nba$Weight, "lbs", "kg")
  ```

  Answer:

Player	Pos	Height	Weight	Age
Nate Robinson	G	1.7526	81.64664	29
Isaiah Thomas	G	1.7526	83.91460	24
Phil Pressey	G	1.8034	79.37867	22
Shane Larkin	G	1.8034	79.83227	20
Ty Lawson	G	1.8034	88.45052	25
John Lucas III	G	1.8034	71.21401	30
D.J. Augustin	G	1.8288	81.64664	25

- Create a column that contains the body mass index (BMI) of the players.
  Code:
  ```
  nba$BMI <- nba$Weight/(nba$Height^2)
  ```

  Answer:

Player	Pos	Height	Weight	Age	BMI
Nate Robinson	G	1.7526	81.64664	29	26.58108
Isaiah Thomas	G	1.7526	83.91460	24	27.31945
Phil Pressey	G	1.8034	79.37867	22	24.40730
Shane Larkin	G	1.8034	79.83227	20	24.54677
Ty Lawson	G	1.8034	88.45052	25	27.19670
John Lucas III	G	1.8034	71.21401	30	21.89683
D.J. Augustin	G	1.8288	81.64664	25	24.41214
Kyle Lowry	G	1.8288	92.98645	27	27.80272
Sebastian Telfair	G	1.8288	79.37867	28	23.73403

- Which player has the maximum BMI in the sample?
  Code:
  ```
  nba$Player[apply(nba,2,which.max)$BMI]
  ```

  Answer:
  ```
  Glen Davis
  ```

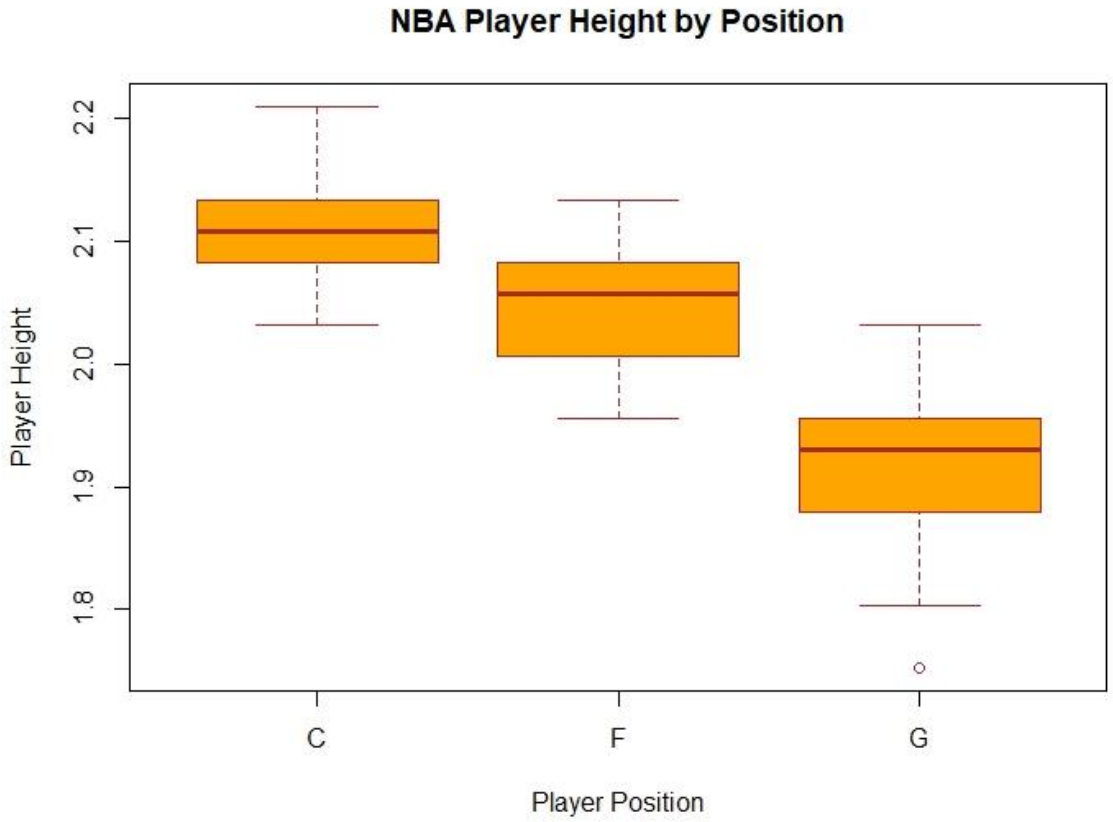
- What percentage of NBA players in the sample have a BMI over 25, which is considered "overweight"?
  Code:
  ```
  overweight <- nba %>% filter(BMI > 25)
  100*round(nrow(overweight)/nrow(nba),4)
  ```

  Answer:
  ```
  45.35
  ```

- Create a boxplot of height by position. Do you see any association between the variables?
  Code:

```
boxplot(Height~Pos,
        data=nba,
        main="NBA Player Height by Position",
        xlab="Player Position",
        ylab="Player Height",
        col="orange",
        border="brown")
```

  Answer: Yes, we can see that there's a strong correlation between the player's height and position. The players whose height is lower than 2 tend to play G position; the one with a height between 2.08 and 2 is more likely to play position F and the player whose height is larger than 2.08 most likely to play position C.



- Create a boxplot of BMI by position. Do you see any association between the variables?
  Code:

```
boxplot(BMI~Pos,
```

```
data=nba,
main="NBA Player BMI by Position",
xlab="Player Position",
ylab="Player BMI",
col="orange",
border="brown")
```

Answer: It seems like there's an association between the players' BMI and his position although it's not as strong as the one between the players' height and position. As is shown in the plot below, the players who play position C tend to have higher BMI, the players who play position F is likely to have medium BMI and the players whose position is G tend to have smaller BMI;