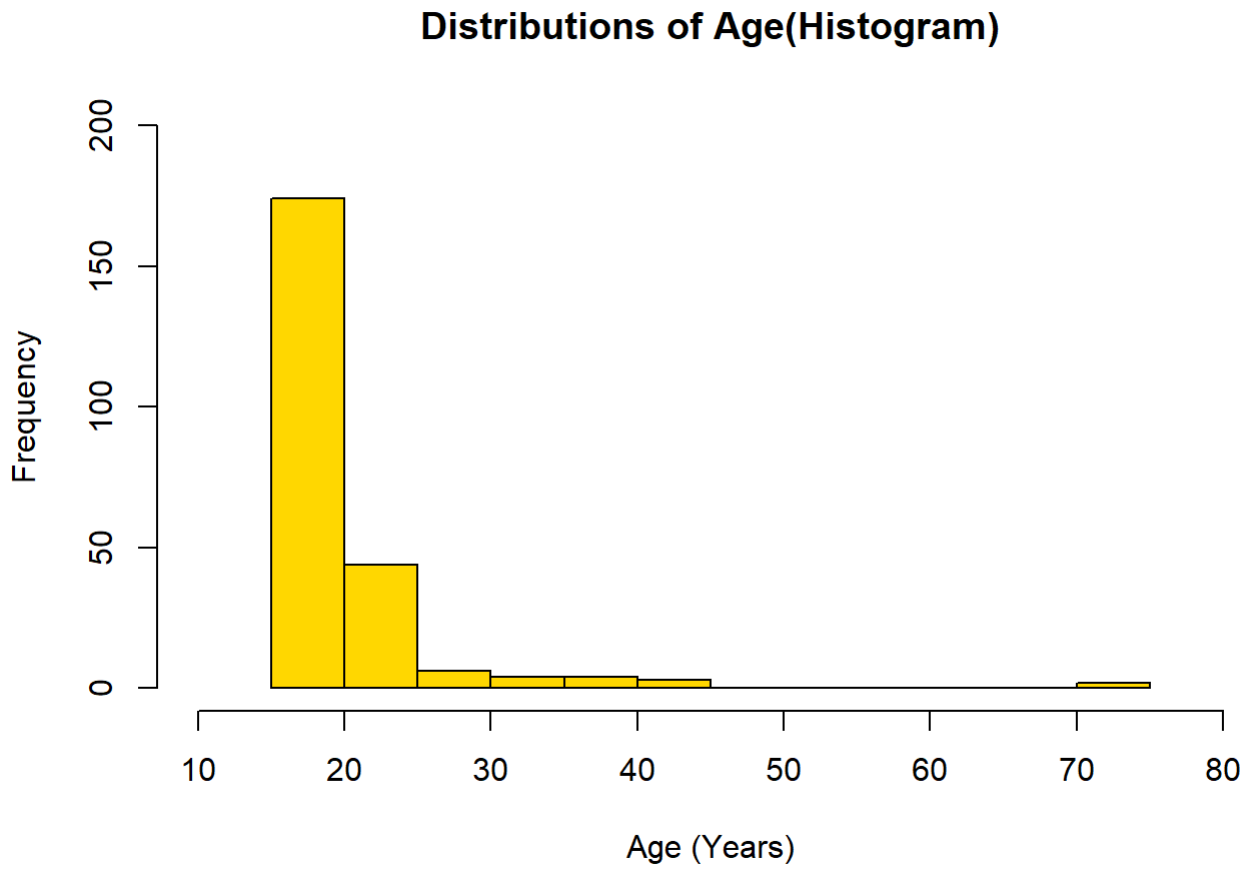
Homework 1: Summary Statistics

Ruben Pena

Due: September 29, 2019, 10:00 p.m

Data was imported renamed to “survey”. Summary statistics of the dataset shows that there are missing values. These will need to be handled either by removal, imputation, or handled within each method/function run throughout the assignment.The levels of “Sex” and “Smoke” were noted.

Student Ages



Center

Values in *AGE* range from 16.75 to 73. For this example, the **median would be a better measure of center**; this is supported by the histogram showing that a MUCH higher distribution of ages are in the 16-25 range.

Spread

The better estimate of spread, in this example, is IQR. IQR is a more robust measure of the spread in the example and will give a better estimation of the spread. The range of this sample (56.25) is heavily influenced by outliers. This can be seen in the large difference between the range and the IQR (2.5).

Skewness

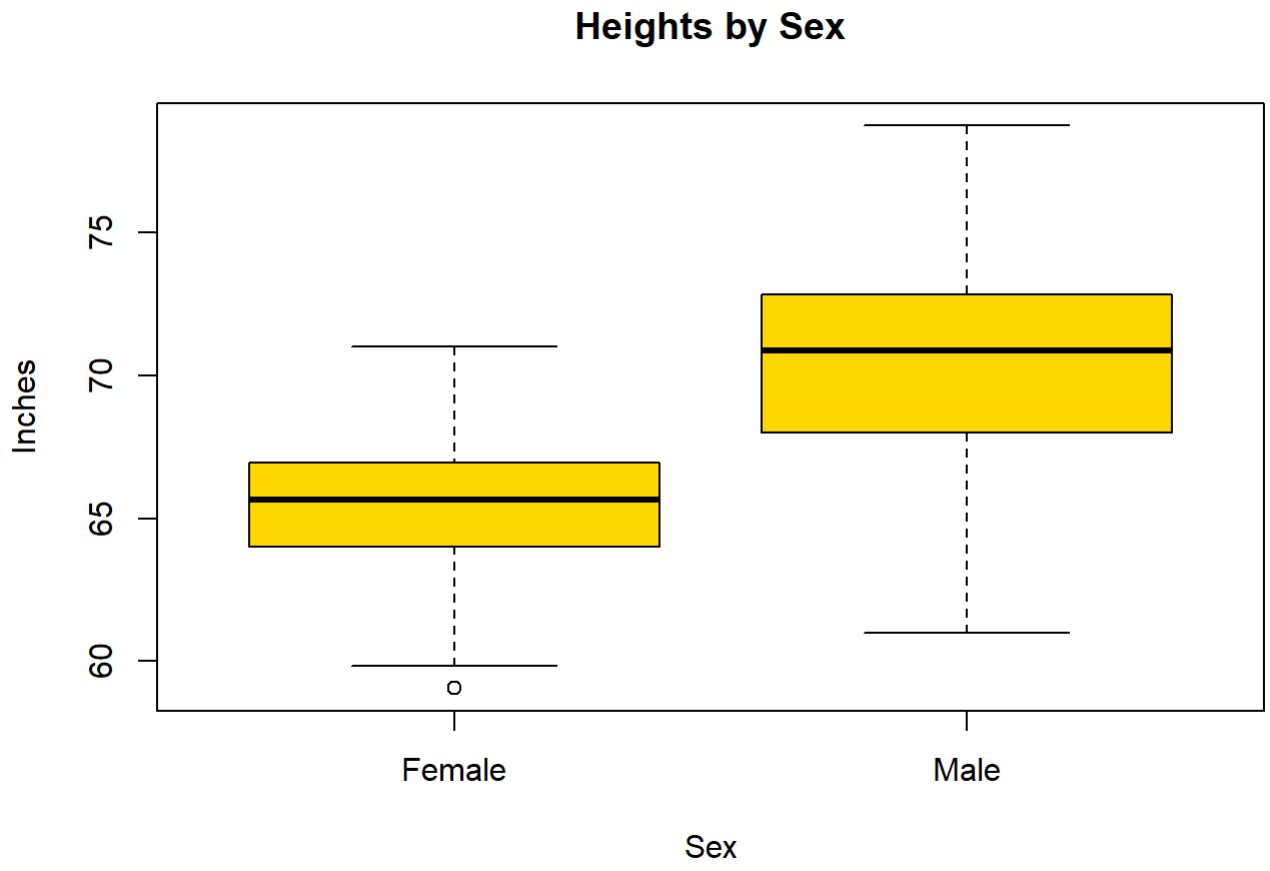
Based on these summary statistics and the histogram *AGE* can be described as being highly skewed to the right. This can also be observed in the summary statistics table (mean > median).

Outlier Analysis

Based on the distributions of *AGE*, you could justifiably label ages over 45(ish) as outliers. One could also argue that the target audience for this survey were people between 16-25, as students over these ages tend to be very rare. These outliers could be non-typical students, data entry errors, intentional input errors, or collection errors(survey sent to a faculty member or non-student).

Student Heights

a priori hypothesis of Student Heights - The mean and median heights of males will be larger than females. Males should also have a higher spread of heights compared to females as their heights have a larger range, in general.



Center

The limited range of values and lack of significant outliers for *Heights* of both sexes means that either median or mean could be used as a valid measure of center for this attribute. Looking at the boxplot it is easy to see that the *Heights* of Males is centered at a higher value than that of Female.

Spread

Either standard deviation or IQR could be used as an accurate measure of spread. As standard deviation is used in many formulae for analysis; I would choose this measurement for simplicity. As suspected, Male *Height* is more spread than Female *Height* (Male *Height* has a larger range of values than Female *Height*)

Skewness

The distribution of Female *Height* is nearly symmetrical with a negligible skew to the left. (Mean > Median) Male *Height* distribution also shows a skew to the left. (Mean > Median)
The table below shows the differences in median vs. mean. These differences seem negligible for both Males and Females.

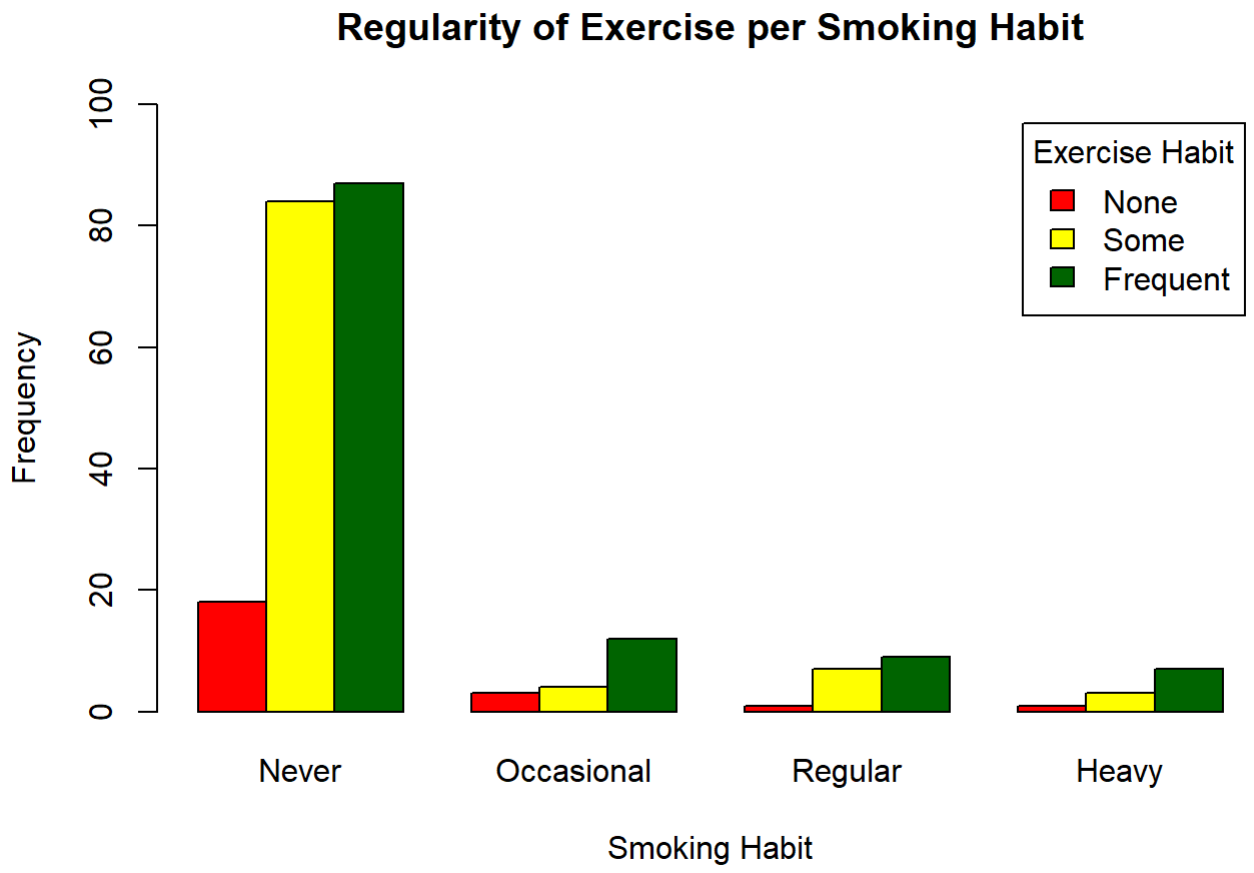
Outlier Analysis

Only one outlier was identified in the boxplot figure and this outlier is not far off from the lower fence. This is most likely not an entry or collection error. It could just be a naturally shorter student or a younger student.

```
##      Sex  Mean_Ht Median_Ht StDev_Ht
## 1 Female 65.23097  65.64961 2.421959
## 2   Male 70.40395  70.86614 3.299312
```

Exercise and Smoking Trends

a priori hypothesis of Student Exercise and Smoking Trends - Based on my experiences, I believe that Exercise and Smoking will have an inverse relationship (negative correlation; more exercise, less smoking/less exercise,more smoking, etc.)



Percentage of levels of *Exercise* (Read by row)

```
##           Smoke
## Exer            Never Occasional    Regular      Heavy
##   None     0.78260870 0.13043478 0.04347826 0.04347826
##   Some     0.85714286 0.04081633 0.07142857 0.03061224
##   Frequent 0.75652174 0.10434783 0.07826087 0.06086957
```

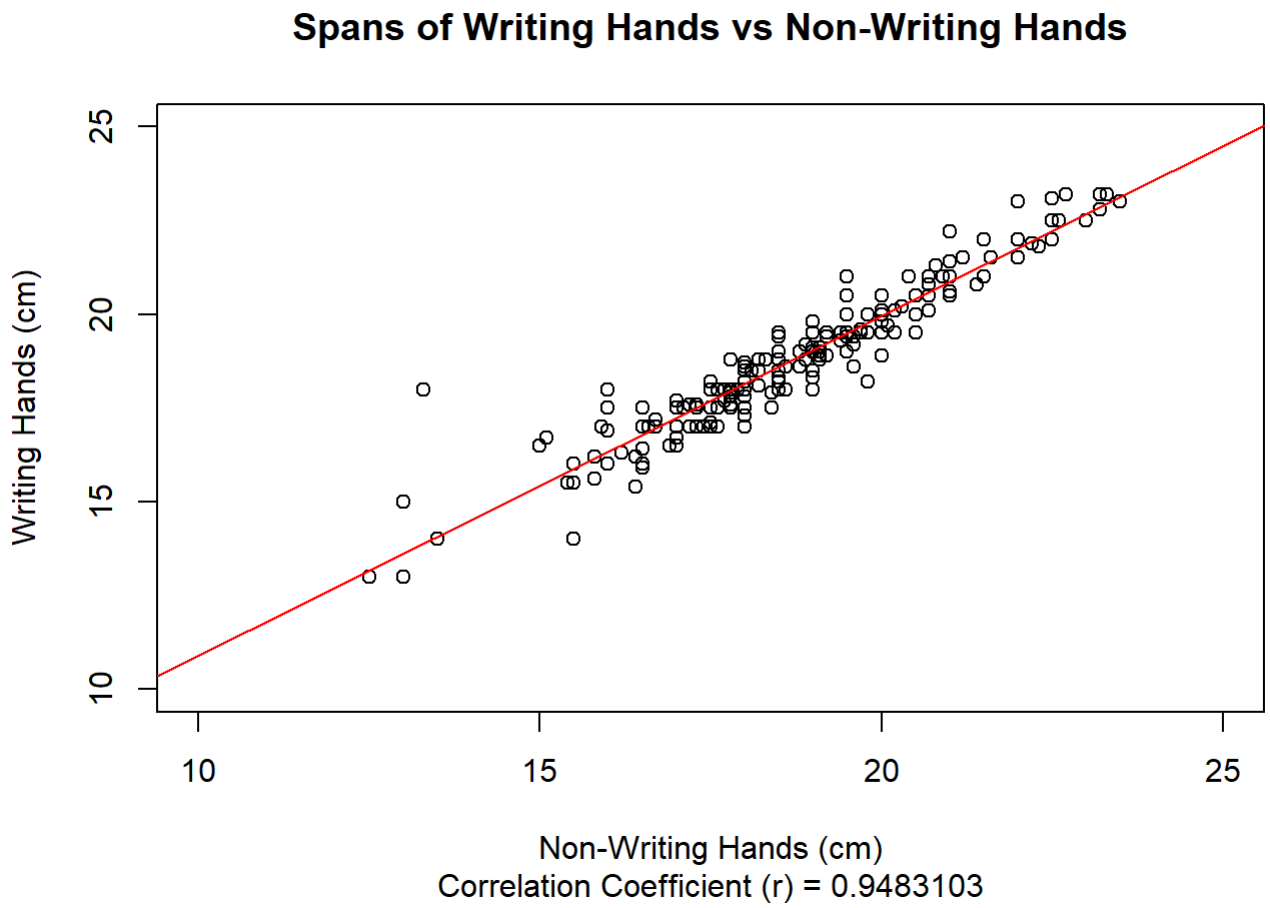
Percentage of levels of *Smoke* (Read by Column)

```
##           Smoke
## Exer            Never Occasional    Regular      Heavy
##   None     0.09523810 0.15789474 0.05882353 0.09090909
##   Some     0.44444444 0.21052632 0.41176471 0.27272727
##   Frequent 0.46031746 0.63157895 0.52941176 0.63636364
```

Relationship Analysis

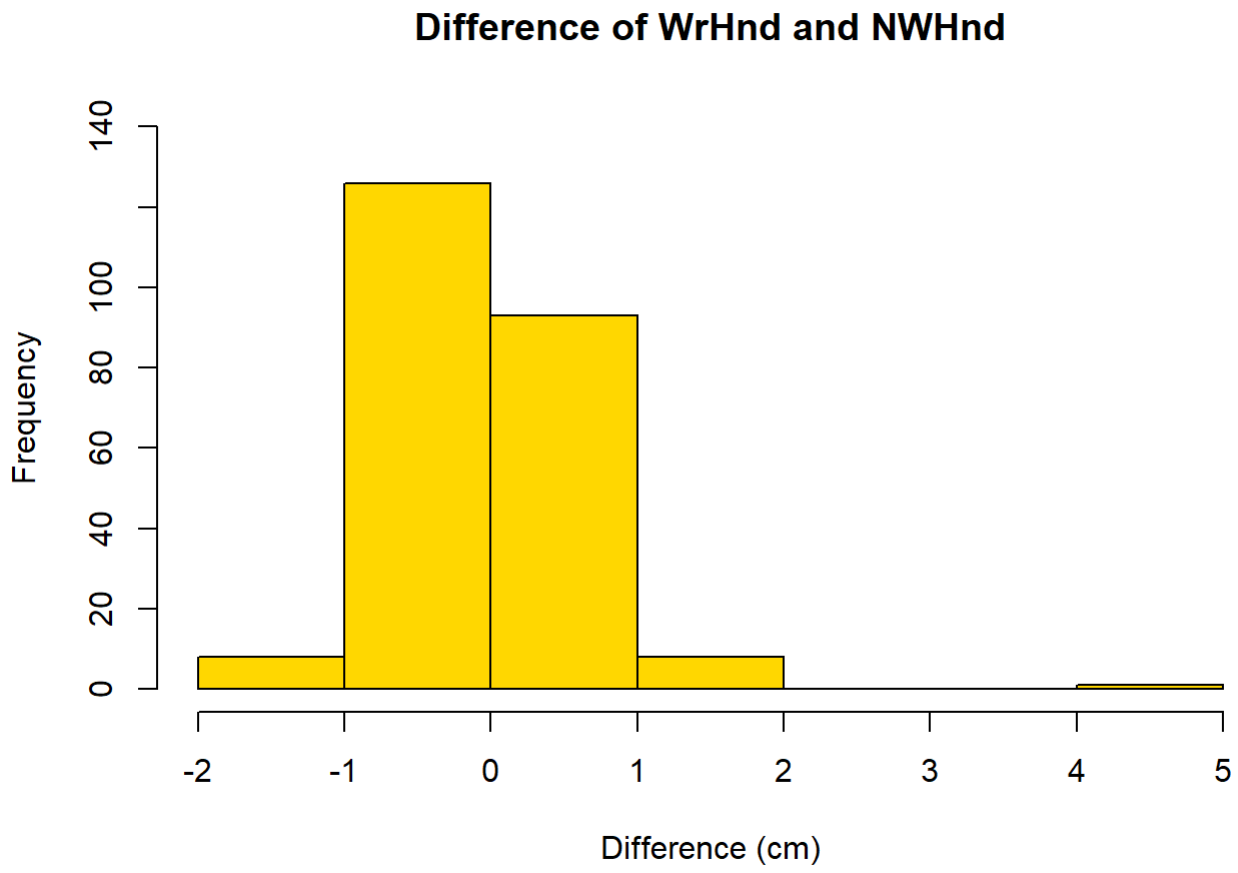
Based on both the barplot and marginal tables, there is an apparent relationship between *Exercise* and *Smoking* among students that were surveyed. Students that worked out “Frequent” or “Some” are much more likely to have answered “Never” in regards to *Smoke.* If no relationship existed between these two variables, the barplot would be much flatter and more evenly distributed. There would not be spikes of *Exercise* habits in the “Never” column. No discernible pattern would be present in the barplot.

Hand Spans



Scatterplot Analysis

The correlation coefficient of *WrHnd* and *NWHnd* is 0.9483103 and indicates a strong positive correlation. As the span of *WrHnd* increases, so does the span of *NWHnd*. This can also be seen in the abline of the linear model that I added to the scatter plot. There is strong positive linear trend.



Center of Differences

The mean of the difference between WrHnd and NWHnd is 0.0864407 and the median is 0. The mean was drawn toward a potential outlier (difference of 4.7cm). In this case, median is a more accurate estimate of center.

Spread of Differences

The minimum is -1.6, the maximum is 4.7, and the standard deviation is 0.624386. The maximum in this set throws off the calculation of the standard deviation. A more accurate estimate of spread can be found by ignoring this single outlier or by using the IQR.

Skewness and Outliers of Differences

The maximum value in this set (4.7) is most likely an outlier. The histogram shows that there is only ~1 such value and that the rest lie roughly between -2 and 2. This outlier causes the data to skew to the right(chasing the outlier).