**Written Assignment**

Olivia Grill

American Public University

Math320: Mathematical. Modeling

Professor Annette Burden

May 23rd, 2021

**Abstract**

For the final project I obtained my data from Census at School. It’s a database of students across the US and the variables vary from interests and test scores to age and height. When entering the website and accepting the conditions of use, you then can slightly alter your data. You can choose sample size, state, grade levels, gender, and the date collected. I didn’t choose any specifics except for the sample size which I chose to be 250. There were a lot of variables given when the data was generated, but I narrowed my data down to three; age, gender, and height. Some of the height numbers were strange so I picked out the really ridiculous numbers which left me with a sample size of 226 students. To do my calculations and graphs, I used R.

**Introduction: The Variables**

As previously stated I got my data from Census at School. This data set was randomly generated from a large database of information about students through the US. The three variables that I chose from the data set include age, height, and gender. Gender is the categorical variable while height and age are numeric values. Before doing any calculations or anything, I predicted that there would be a positive relationship between age and height, and by the end of this project I can determine if age is a good predictor of height. On average, as students get older they tend to get taller too. Age is the explanatory variable and height is my response variable. I had a few major outliers such as 1500cm or 1.4cm for height so I went ahead and deleted those ending with a sample size of 226 students.

***The Findings: Females***

The sample size for the female group was 112. These findings can be visualized in tables 1 and 2 below. The smallest height was 122cm while the tallest girl was 229cm, and the median was 160.5cm. The 25th percentile for height was 157 cm and the 75th percentile for height was 168cm. This means that ¼, or 28, of the females would have a height below 157cm. Likewise, for the 75th percentile, ¾, or 84, of the females will have a height below 168cm and 28 females will have heights above 168cm. The average height for the female group was 161.9cm. The standard deviation for the same group was 13.0148cm, the range was 72cm, and the interquartile range was 11. This means that 68% of the 112 females heights fall within 148.8852cm-174.9148cm. The range shows the difference between the tallest girl and the shortest girl, which is 72cm. Since IQR tells us the middle 50% of the data, I can conclude that the middle 50% of females (Q3-Q1) fall between 150.9-172.9cm in height. Using the formula for outliers, shown in figure 1, I determined that anything falling outside 140.5cm-184.5cm would be considered an outlier for the female group. Observing the box plots from my data, also shown in figure 2, I can see that 5 females fall below 140.5 and 4 females are above 184.5cm, so a total of 9 outliers for that set.

***The Findings: Males***

The sample size for the male group was 114 individuals. These findings can be visualized in tables 1 and 2 below. The shortest male was 141cm while the tallest was 196cm. The median was 160.5cm, the 25th percentile of males was 168cm and the 75th percentile was 180.225cm. This means that roughly 29 males fall below 160.5cm and roughly 86 males fall below 180.225cm, so 28 males are taller than 180.225cm. The average height for the 114 males was 173.4675cm with a standard deviation of 11.3990. This means that 68% of the males in this group fall between the heights of 162.0685cm-184.8665cm. The range for the male group was quite a bit lower sitting at 55. The interquartile range was 12.225, which again shows the middle 50% of data. So, the spread for the middle 50% of male heights falls between 161.2425cm-185.6925cm. Using the same equation as above, I determined anything falling outside of 149.6625cm-198.5625cm for the male group would be an outlier, shown in figure 1. Looking at my box plots I can see that 5 males of the set fall below 149.6625cm, but none are above 198.5625cm.

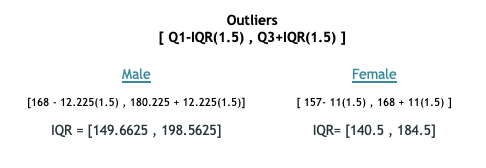
**Table 1**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Gender     n** | **Minimum** | **Q1** | **Median** | **Q3** | **Maximum** |
| Female    112 | 122 | 157 | 160.5 | 168 | 229 |
| Male       114 | 141 | 168 | 177 | 180.225 | 196 |

**Table 2**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Gender** | **Mean** | **Standard Deviation** | **Range** | **Interquartile Range (IQR)** |
| Female | 161.9134 | 13.0148 | 72 | 11 |
| Male | 173.4675 | 11.3990 | 55 | 12.225 |

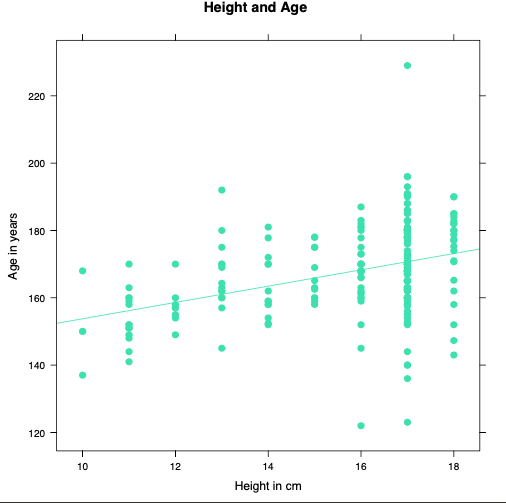
**Figure 1**



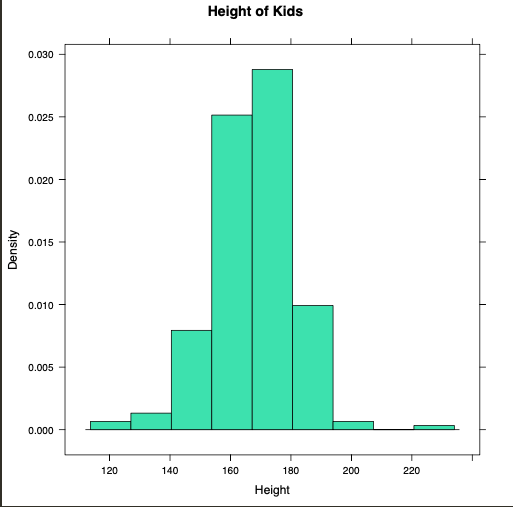
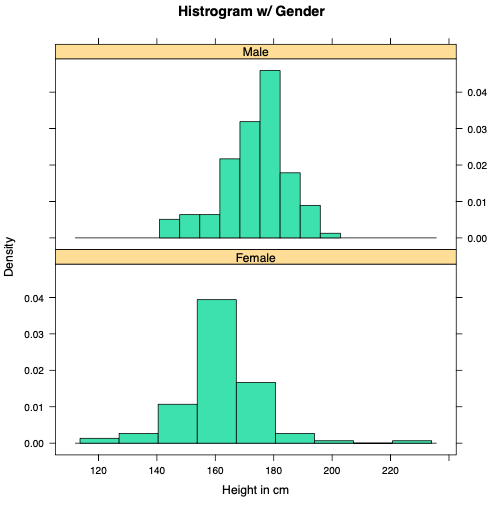
***The Graphs***

My histogram when not broken up by gender is unimodal, that is mostly bell-shaped but has a slight right skewedness. The histogram for the female set is unimodal and skewed right, while the histogram for the males looks similar to the original histogram where it’s unimodal, mostly bell-shaped but a slight right-skew. Looking at the scatter plot there is a heavy saturation around age 17 and I can notice a slight positive slope, which suggest a slight positive relationship between height and age in this data set. Throughout this project I used R to do calculations and make my graphs. The r correlation value ended up being 0.3781, which suggests a weak positive relationship.

**Figure 2 Figure 3**



**Figure 4 Figure 5**

** **

**Conclusion**

Before beginning any calculations, I made a prediction that of the 226 students in my data, there would be a positive correlation between their age and height. After computing a five-number summary, IQR, finding outliers, and drawing all the plots it showed that there was indeed a positive relationship between height and age. Based on the sample correlation r, it showed a weak positive relationship. Because of the value of this sample correlation, I didn’t compute a regression line. Aside from those numbers though, looking at the scatterplot and histogram, gives a pretty good idea of the general assumptions of the data and the slight relationship that exists between the two numeric variables.

**References**

*Welcome to Census at School - United States*. Census at School - United States. (0AD). https://ww2.amstat.org/censusatschool/.

\*I can provide my R code with comments as a reference for how I did the calculations and graph if it is needed. \*