An Exploratory Data Analysis: Heights of Parents and Children Classified by Gender

Amelia Kelly and Matt Wilchek

March 6, 2017

Data Set

Our Exploratory Data Analysis (EDA) focuses on an exploration of the heights of children and parents as stratified by gender. The data set PearsonLee comes from the package HistData in R [1]. The data were initially collected by Karl Pearson and Alice Lee in England during the 1890's. Pearson collected data on over 1000 families during this time.

The PearsonLee data is comprised of 746 observations with 6 variables, described below:

- Child: children's height in inches, a numeric vector
- Parent: parent's height in inches, a numeric vector
- Frequency: a numeric vector
- GP: a factor with levels fd, fs, md, and ms
- Par: a factor with levels Father and Mother
- Chl: a factor with levels Daughter and Son

Data Limitations

The following are a few data limitations that we identified initially and throughout the process of our exploratory analysis:

1. Geographically limited to England.
2. Not modern data – conditions may have changed to alter average height in the past century.
3. We do not know of the socioeconomic factors of the individuals that may impact their height
4. Parents could have multiple children; we do not know if they accounted for or potentially double counted.
5. Frequency data element is unclear; non-integer values exist.

Existing Analysis

The PearsonLee and Galton datasets of family heights data have been preeminent historical datasets in statistical analysis. Most analysts have challenged the data through regression methodologies hoping that such analysis would garner refined regression techniques for prediction analysis[2]. By taking a hard look at some of the previous statistical models done such as orthogonal regression, geometric mean regression and least squares, an analyst may be overlooking some of the most simple, but noticeable conclusions from the observations recorded. Without having to perform complex regression smoothing like other analysts have also attempted, we decided to start simple for our exploratory analysis [3].

Data Visualized

We examined multiple visualizations to gain understanding and insight into the data. We created scatterplots, histograms, and boxplots in order to help us decide how we wanted to shape our SMART question.

We first created the scatterplots below to show the heights of child and parent by different groups. We can see a clear linear relationship across each of the 4 groups. This relationship is examined in research by Wachsmuth et al. [4]

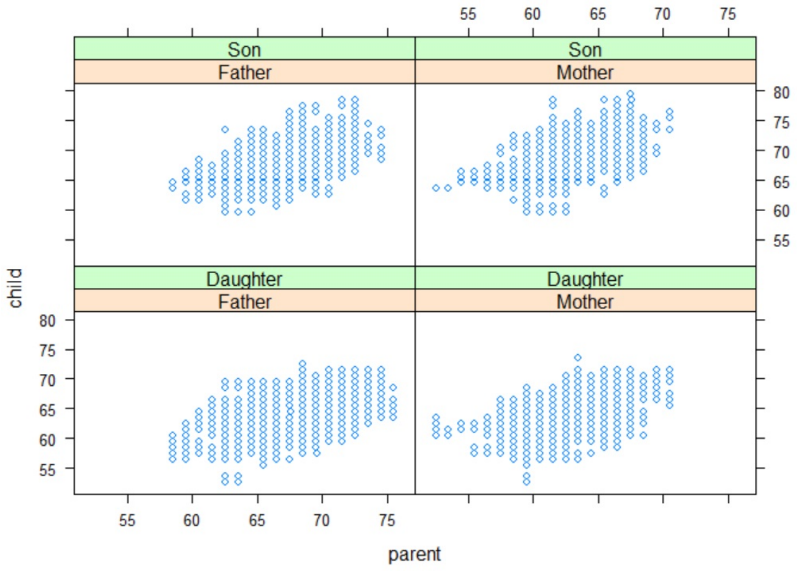

Figure 1: Scatterplots of Parent and Child Heights

We next looked at the distributions of variables of interest, child height and parent height. We created histograms of these variables (below) to visualize their spread and determine if they have approximately normal distributions. We see by the histograms that they appear to be near-normal; neither histogram has a prominent tale.

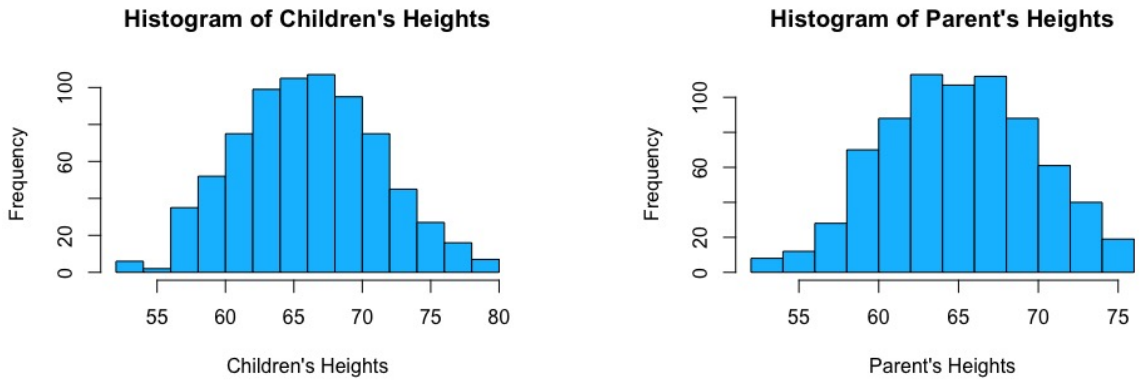

Figure 2: Histograms of Children and Parent Heights

Finally, we examined boxplots of the variables of interest. When we focused on the distributions of Child and Parent stratified by gender, we saw clear differences in the means. The means of males generally are higher than the means of the female groups. The boxplots are below.

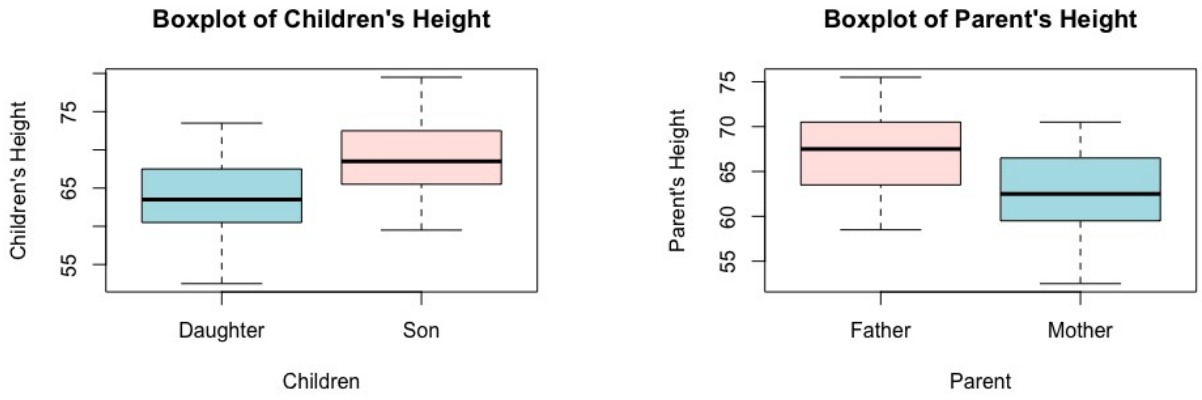

Figure 3: Boxplots of Children and Parent Heights

These data visualizations provided us with ideas of what question we wanted to consider in our EDA.

SMART Question Development

Our question aims to determine if there are differences in the means of average heights based on gender groups.

To develop our question, we began by

- Specific: Are the means of the averages of Parent/Child Heights equal?
- Measurable: We can the take means.
- Achievable: We can compare values via t test.
- Relevant: Useful for researchers looking at relationship between height and gender.
- Time-Scaled: Multiple data sets exist, so question can be answered in a timely manner

Question Modification

Initially we were interested in investigating the relationship of parent and child heights by strictly looking at the genders' summary statistics. We discovered existing research stratifying the data across four groups. This gave us the idea of looking at the groups stratified in yet another way.

When we began exploring the data through visualizations of scatterplots and boxplots, we came to the realization that we could not, as originally intended, readily compare heights considering gender alone. For one, the heights of males are generally higher than that of females; our boxplots demonstrated this, and this does not provide particularly meaningful analysis. We thus decided to create 2 groups that stratify observations into same gender verses differing gender:

- Group 1: Same gender parent-child observations (Father - Son, Mother - Daughter observations)
- Group 2: Differing gender parent-child observations (Father - Daughter, Mother - Son observations)

From here, we created another variable that calculates the average of the parent and child height for each observation. We then performed a t test to test the hypothesis that the groups have equal means.

Preliminary Conclusions

We performed a t test to test the following hypotheses:

- H_0: The means of the two groups are equal.
- H_1: The means of the two groups are <u>not</u> equal.

The output from the test in R is shown below.

```
        Welch Two Sample t-test

data:  PearsonLee$avg by PearsonLee$group
t = -0.83444, df = 719.03, p-value = 0.4043
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.8373666  0.3378614
sample estimates:
mean in group 0 mean in group 1
       65.40247        65.65223
```

Figure 4: R output for t-test

The probability of a t value of -0.834 is p = 0.4. We reject the null hypotheses at an $\alpha = 0.05$ level, and conclude that the true difference in means for the groups is not equal to 0. The sample estimate for the mean of group 0 (parent and child have same gender) is 65.4 inches while the sample estimate for the mean of group 1 (parent and child have different gender) is 65.65 inches.

We examine a boxplot of the two groups below:

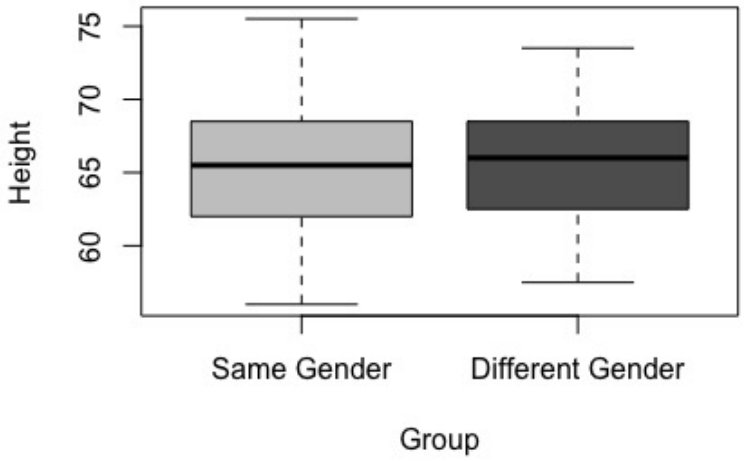

Figure 5: Boxplot of Average Heights by Group

Based on the exploratory data analysis, we begin to see that statistically, the means of

the two groups are not the same. However, from an objective point, a person would not be able to perceive the difference between a person of height 65.4 inches and another of height 65.65 inches.

Next Steps

From developing expectations, collecting data and matching our expectations to our data, performing an exploratory analysis process has been quite informative for a particular dataset. However, to explore the dataset even further after we have answered our SMART question, there are a couple of steps that would be interesting to look at.

To begin, the dataset included a column for "frequency" which we did not quite understand the purpose of. We could not find any initial background information or documentation regarding this row of information. Going forward, it may be good to take a deeper dive to understand the purpose of "frequency" and how it could have been applied to further analyze the dataset.

Additionally, in the process of answering our question, we discovered that there is more variance in the same gender group than the differing gender group. We might consider further stratifying the data to look for differences among different height groups. We could create a categorical variable defined by different ranges of heights for groups, then further analyze these for measures in variance and central tendency. Afterwards, we can further stratify using our gender categories with the height groups.

References

[1] Friendly, Michael (2017). Package 'HistData'. Retrieved from https://cran.r-project.org/web/packages/HistData/HistData.pdf

[2] Zhu, Wei., Ma, Yeming., Han, Hao (2015). Galtons Family Heights Data Revisited. Department of Applied Math and Statistics, Stony Brook University, 1-4.

[3] Wilkinson, Leland (2016). Smoothers. University of Illinois at Chicago. Retrieved from https://www.cs.uic.edu/ wilkinson/Applets/smoothers.html

[4] Wachsmuth, A.W., Wilkinson L., Dallal G.E. (2003). Galtons bend: A previously undiscovered nonlinearity in Galtons family stature regression data. *The American Statistician*, 57, 190-192.

[5] Pearson, K. and Lee, A. (1896). Mathematical contributions to the theory of evolution. On telegony in man, etc. *Proceedings of the Royal Society of London*, 60 , 273-283.

[6] Pearson, K. and Lee, A. (1903). On the laws of inheritance in man: I. Inheritance of physical characters. *Biometika*, 2(4), 357-462. (Tables XXII, p. 415; XXV, p. 417; XXVIII, p. 419 and XXXI, p. 421.)

[7] Peng, Roger D., and Matsui, Elizabeth (2017). *The Art of Data Science*. LeanPub, 2017. Retrieved from https://leanpub.com/artofdatascience