Exploratory Data Analysis: Analysis of Personality

Matthew Green, Grant Halver, Ling Zhu Hou, Brian Sweeney

Arizona State University

STP 420 - Introductory Applied Statistic

Neil Hatfield

Spring 2017

**Executive Summary**

The data was obtained from a study on temperament. However, many of the data fields found in the study are uninterpretable without specialized knowledge in the field of psychology. For this reason, the focus of this exploratory data analysis (EDA) will not be based entirely on the personality metrics but rather more of the demographic information. By analyzing the relationships contained within the data, we can ascertain meaning about the connections between demographics and other characteristics such as education level, income, and marital status.

Through the use of demographic characteristics, we are able to explore relationships between the data that may be the result of other personality traits. For instance, the exploration of how gender affects estimated income should show if there is a wage gap. This led to our first statistical research question: “Does the wage gap exist?” To follow up on this, we decided to explicitly explore how weight when stratified by gender can affect personality. This will be more explicit on physical characteristics and how they affect our interaction with others.

The EDA starts with exploring the data. By exploring the data first, we are able to understand more about the attributes that were measured. In turn this allows for a more meaningful analysis of the data in regards to our SRQs. Gender, age, height, weight, exercise frequency, smoking frequency, ethnicity, education level, job status, and income estimate were selected for univariate analyses. To further our understanding of the data, comparisons using more than one variable were also utilized. The bivariate relationships between gender and education level, education level and job status, and education level and smoking frequency were explored.

Since the gender divide is controversial we decided to focus our multivariate analyses there. Analyzing the effect of education level on estimated income when stratified by gender led us to conclude that although those worst off for both genders are approximately equal, the typical and higher paid men are better off than their women counterparts at every education level. The effect of exercise on weight when stratified by gender show that men have overall higher weights, and that in particular, increasing exercise for men has less of an impact on weight than for women. Our final multivariate analysis was meant to investigate whether sociability is affected by weight at different values and we also considered if gender norms played a role in this; however, this analysis showed no significant trends in the data.

Now we tackle the first SRQ: “Does the wage gap exist in the sample?” In our analysis, a Wilcoxon Rank Sum test was utilized to answer the question. The threshold of unusualness was set at 0.05. We found that there appeared to be a statistically significant difference in the estimated incomes for the two genders (S=12491641, *p*<0.0001). Thus, we rejected the null hypothesis that there was no difference in the expected income between the two genders. Approximately 62% of the time we would expect a randomly selected man to have a higher estimated income than a randomly selected woman.

The second SRQ: “Does a person's weight have an effect on sociability for either of the two genders?” was analyzed next. The threshold of unusualness was set at 0.05. However, the data does not appear to have a statistically significant relationship (H45= 36.5258, p = 0.8119). Therefore, we failed to reject the null hypothesis, that there was no difference in sociability based on a person’s weight, even when stratified by gender.

**Introduction**

Condon and Revelle originally collected this data to analyze personality constructs and how they relate to temperament. Their data used about 24,000 individuals and administered various items from personality scales, using the Synthetic Aperture Personality Assessment method (SAPA). In addition to this personality data, demographic data was also collected from the participants in the sample. This data was collected between December 8, 2013 and July 26, 2014.

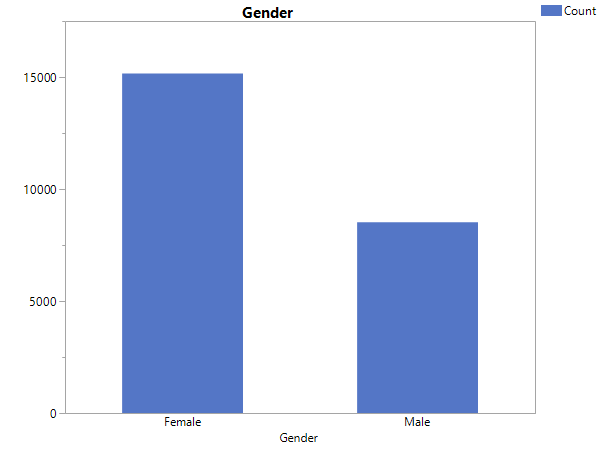
This data was selected due to a number of reasons. To begin, the personality test was an exceptionally comprehensive exam pulling questions from multiple personality assessments such as the IPIP 100 Item Big Five Factor Markers, the six-factor HEXACO-IP scales, and the Big Five Aspect Scales. In addition, the Synthetic Aperture Personality Assessment measures cognitive ability and vocational interests as well. These three categories were chosen for SAPA because they account for a substantial amount of variation without an excessive amount of repetition (Revelle, 2013). In addition, the assessment was built with questions that would contribute positively to a higher score in personality while also providing questions that would decrease your score in particular areas; however, for the purpose of this exploratory data analysis, we only utilized questions that contribute positively to an individual’s score. This means that a person with a greater value for sociability will be more social than someone with a lesser value for sociability. The survey randomly administers these questions from a pool while still maintaining a reasonable number of self-report items per person. Personality is an interesting subject for research because the literature is relatively lacking, thus performing tests on the data can show interesting trends that we might not have known otherwise. Additionally, personality is a topic that is relevant to everyone. Lastly, we chose this study due to the large sample size and the numerous variables to choose from.

The variables presented in the study helped to lead us to two statistical research questions. The first statistical research question is: Is there a difference in the income estimate due to gender in the sample, and if so, how large is the difference? This question is relevant to outside context due to the gender pay gap (Brown & Eileen, 2017). If there is a statistically and practically significant difference in the pay between men and women while controlling for other variables, this will support the idea that there is a wage gap between the two sexes. The data on income estimates for the SAPA project was calculated based on the occupation that the individual selected for the survey. This topic is extensively covered and explained, but the question is useful to see if this data shows that same trend. Secondly, we would expect that weight would lead to varying results in sociability. This could be due to people feeling insecure about their weight, thus rating themselves lower in sociability. Further, we would want to stratify by gender in order to account for how media skews what a socially acceptable body image is **(**Williams & Ricciardelli, 2014). We would expect males to be less sociable at lower than typical weights as opposed to the expectation that females who are heavier would score lower. This leads us to ask the statistical research question: Does the weight of an individual male or female have an impact on how they would score in sociability?

**Body**

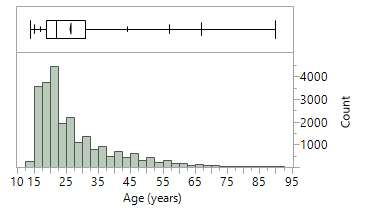
**Univariate Analyses**

In this Exploratory Data Analysis project, there were multiple variables recorded. The data was cleaned up and 10 of the variables were chosen to be looked at more closely. Those variables consist of gender, age, height, weight, exercise, smoking status, ethnicity, education-level, job status, and estimated income. For the data collection, 23,679 individuals were sampled. The first variable being observed would be gender. All of the individuals responded and there was no missing values. The data collected can be seen in Figure 1. Of the 23,679 individuals, 15,157 were female and 8522 were male. This means that 64.01% of the sample was female while 35.99% was male.



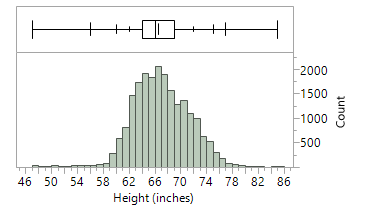
**Figure 1. Bar Chart of Gender**

Age was another demographic that was collected which is shown in Figure 2. The ages have a sample minimum of 14 years and a sample maximum of 90 years. A histogram of the ages, as seen in Figure 2, is right skewed, with a majority of ages concentrated between 15 and 30 years. The median age of the sample is 22, which means that half of the sample is 22 years old or younger. Additionally, the sample interquartile range of the ages is 12 years, meaning that the middle 50 percent of the data falls into an interval of 12 years. In this case, this interval is from 19 to 31, which are the first and third quartiles of the sample ages. Besides age, this data set also obtained measurements of physical characteristics, such as height and weight.



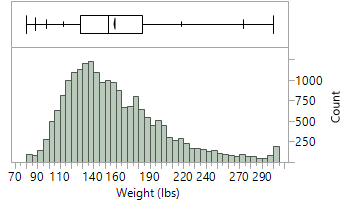
**Figure 2. Histogram and Boxplot of Age**

Unlike with gender and age, height and weight both have missing values, which are excluded in the calculations. From the sample, 20,586 individual heights were recorded. The distribution of recorded heights from the sample appear to be relatively symmetric as seen in Figure 3. The range of the heights is 38 inches, meaning that the smallest interval that all recorded heights fall within is 38 inches. Because heights are additive and the histogram appears symmetric, the sample arithmetic mean of 66.56 inches provides a good measure of adjusted group performance. This value means that the total height accumulated, in inches, by the sample is 66.56 times as large as the accumulated size of the sample. The variance of heights is approximately 17 inches squared. This is a measure of the variation as deviation of the heights, adjusted for the sample size. Another physical measurement that goes along with height is weight.



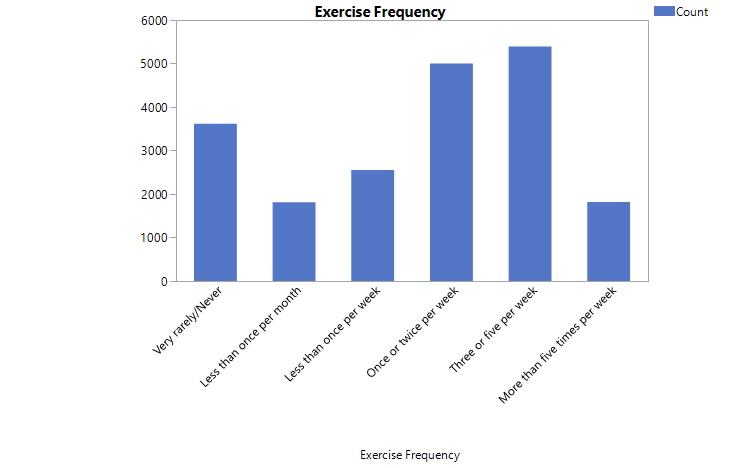
**Figure 3. Histogram and Boxplot of the Sample Heights**

There was a total of 20,227 weights recorded from the sample of individuals. The histogram, shown in Figure 4, appears somewhat right-skewed, with an unusually large number of values at the end of the right tail. The sample median weight is 153 pounds, meaning that half of the sample weighs 153 pounds or less. The range of the weights in the sample is 220 pounds. This means that the smallest interval that all weights fall within is 220 pounds wide. The sample minimum and maximum are 80 and 300 pounds respectively, so all weights in the sample are equal to or between these two values.



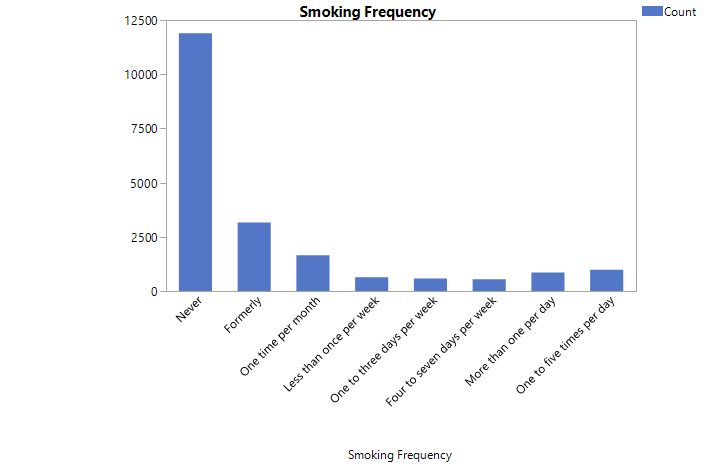
**Figure 4. Histogram and Boxplot of the Sample Weights**

Another variable recorded was from a question about how often individuals exercised. The results of this question are recorded in Figure 5. A total of 20,201 individuals gave an answer to this question while the other 3478 individuals did not provide a response. These values are not included and treated as missing in this instance. From the data, the most common response was that people exercise three to five times per week, followed by one or two times a week. These categories, along with exercising five or more times per week, represent exercising at least once a week, and contain 60.48% of responses. On the other hand, exercising very rarely or never is the third highest response with 17.90% of responses. This means that while most of the individuals in this sample tended to exercise at least once a week, many outside of that category exercised rarely or never.



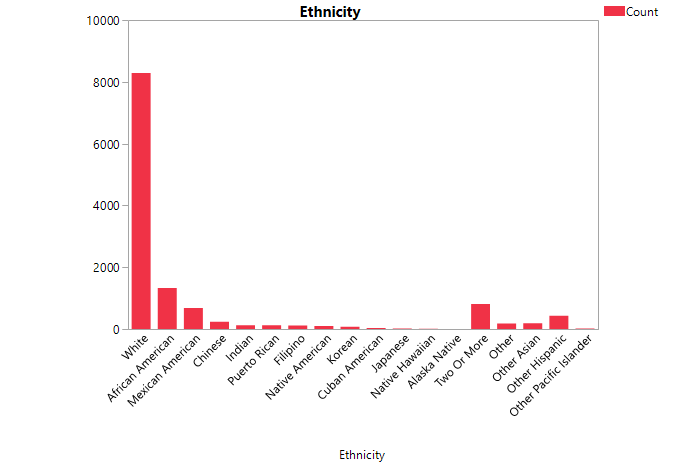
**Figure 5. Bar Chart of Exercise Frequency from the Sample**

Another question asked was did the individuals smoke and how often. Figure 6 shows the frequency of these responses. A total of 20,325 individuals gave a response to the question with the rest reporting no value. From the data collected, 11,892 of the individuals stated they never smoked. Individuals that responded as formerly smoked but stopped contained 13.40% of the recorded responses. Out of the total responses, 5262 individuals responded that they would smoke. 31.39% of those responses said they smoked once per month while individuals that stated they smoked more than one per day contains 16.32%. From the results, the majority of the responses were either not smoking or had never smoked, while multiple individuals smoke more than once per day or week. This trend shows that most people in this sample never smoked or formerly smoked and that there is variation in the smoking habits of individuals. More smokers smoke about once per month or multiple times per day than any amount in between those categories.



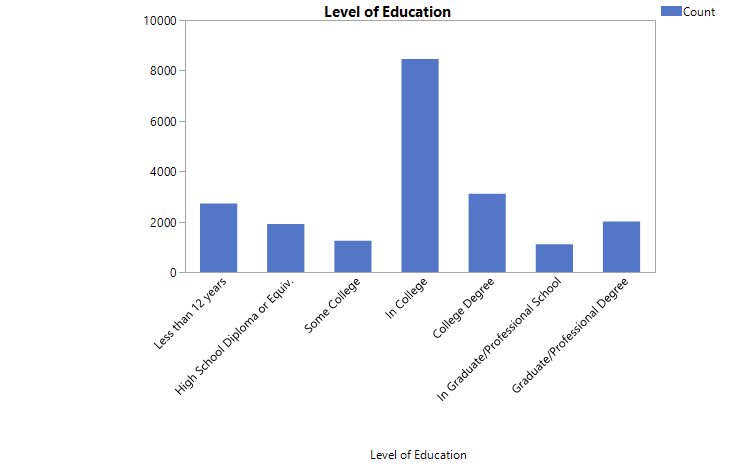
**Figure 6. Bar Chart of Smoke Frequency**

When the individuals were asked about their ethnicity, out of the total of 23,679 individuals, 10,869 individuals did not give a response thus we treat these as missing values. This leaves 12,810 responses that are displayed in Figure 7. From the results, the majority of individuals stated they were White representing 8291 individuals. The second most common was African Americans with 1329 responses. The other responses were comparatively much smaller than the first two categories. 115 individuals responded as Natives, which includes Alaskan Native, Native American, and Native Hawaiian. A total of 775 individuals identify themselves as Asian which includes categories such as Chinese, Filipino, Indian, Japanese, Korean, and other Asian. 35 individuals identify as Cuban American while 687 people are Mexican American and 434 people are identified as other Hispanics. 128 people are Puerto Rican and 22 people are other Pacific Islanders. Lastly, 809 individuals stated they had two or more ethnicities while 185 people responded as other.



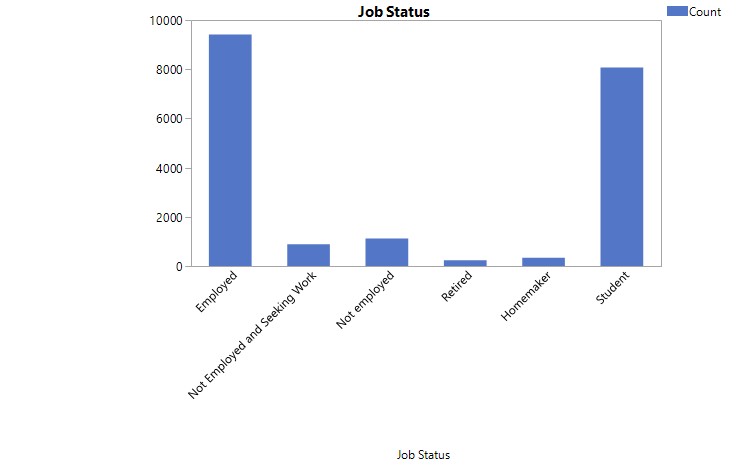
**Figure 7. Bar Chart of Ethnic Frequency**

20,571 of the individuals in the sample have a recorded level of education illustrated in Figure 8. Of these, the most common response was being in college, with 41.11% of the sample being in college. The second and third most common categories were having a college degree at 15.14% and having less than 12 years of education at 13.25%. An interesting note about this distribution though is that many of the categories build on each other, so the upper three categories would all likely have a college degree, just that some individuals are trying to obtain further education. With this in mind, 22.5% of individuals have never been to college while 30.28% of individuals in the sample have at least a college degree.



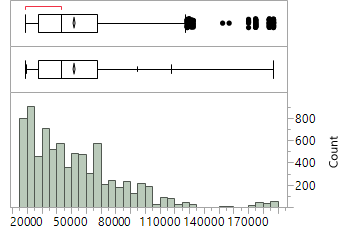
**Figure 8. Bar Chart of the Level of Education**

Another question that was asked to the individuals was about their job status. Figure 9 illustrates that a total of 20,082 individuals responded with 3597 missing responses. From the results, the majority of individuals responded as employed with 9413, and another 8073 as students. The results showed that 2019 of the individuals stated as unemployed, however only 892 are seeking for work. 341 of the individuals are homemakers and 236 are retired. From the collected data, 46.87% are currently employed and 10.05% are stated as unemployed, which includes both the individuals who are looking for work and those who are not.



**Figure 9. Bar Chart of Job Status Frequency**

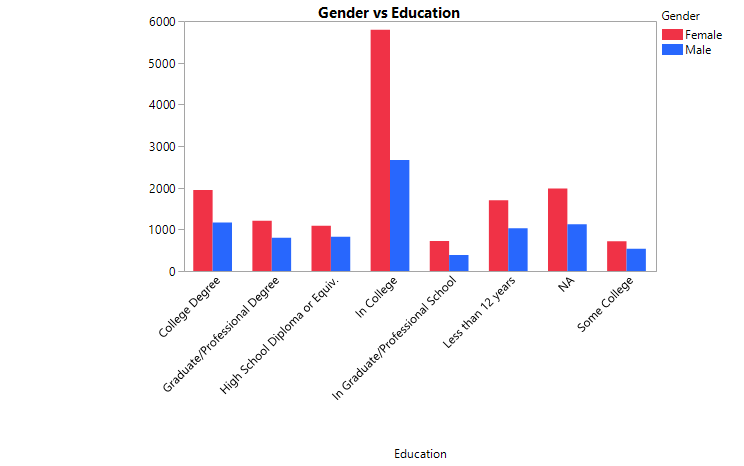
Out of the total of 23,679 individuals asked about their income estimate, only 7755 responded with 15,924 missing values. The maximum amount of income from the results were $187,200 per year while the minimum was $19,080 which can be seen in Figure 10. The sample median was $43,420 which means that 50% of the respondents made less than $43,420 per year. The first quartile was $27,850 while the third quartile was $67,490, this means 25% of the individuals made less than $27,850 and 25% made more than $67,490. The data showed an unsymmetrical data with a skewness of 1.485. From the results, most of the respondents has a high income estimate while some has a low income.



**Figure 10. Histogram and Boxplot of Income Estimate**

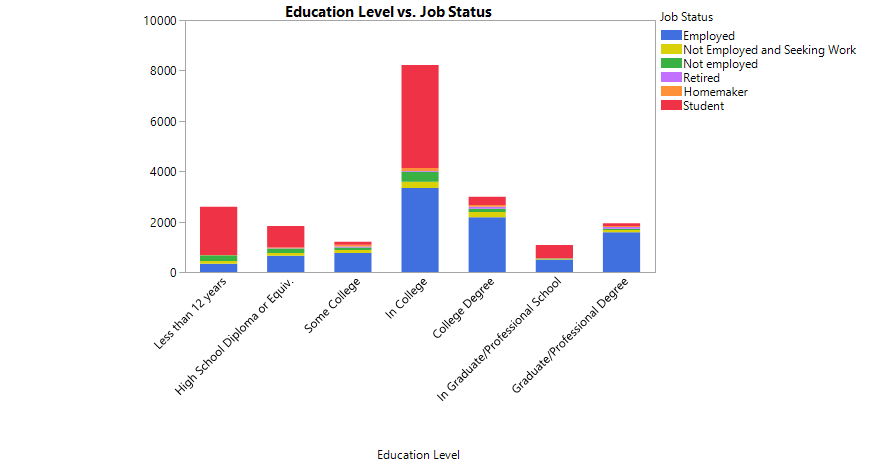
**Bivariate and Multivariate Analyses**

Now that variables have been examined separately, the variables can then put together for comparison and to see if there are any trends. Figure 11 shows the comparison of gender with the level of education. Due to the majority of the respondents being female, there are more females than males in all categories for education. The figure shows that majority of the individuals are in college with 5792 females are currently in college while only 2665 males are in college. For most of the categories, the bars for females are almost twice as big compared to the males. There are some categories where the males follow closely behind. The figure shows that majority of females and males from the sample are in college or have a college degree or higher.



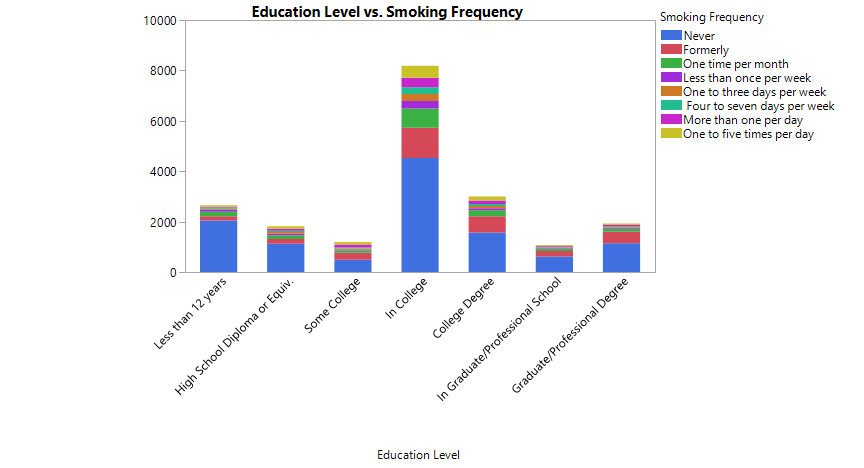
**Figure 11. Bar Chart of Gender vs. Education**

When education level is compared to job status in the sample, some trends start to emerge. Figure 12 shows a bar graph that graphs job status by the level of education. As seen in the graph, the number of students tend to decrease as the education level increases. This is likely due to people reaching a level of education they are satisfied with and devoting time and energy into finding a job instead. One interesting point about this is that some individuals are students, but are not in college nor graduate school, including 46.39% of people with a high school diploma and 11.04% of people with a college degree. This seems odd because they are not in either category of college or graduate school. Another interesting trend in this relationship is the relative amounts of employed people per education level. Individuals with graduate or professional degrees have the largest proportion employed with 81.5%. Having a college degree is second with 72.98% being employed. In the education levels where individuals are still in school, the level of employment is lower, likely because the proportion of students is larger, leaving less time for employment. Generally, the proportion of employed people increases with education level once the degree is acquired. Before that, a decent proportion of individuals in school consider themselves students and not employed.



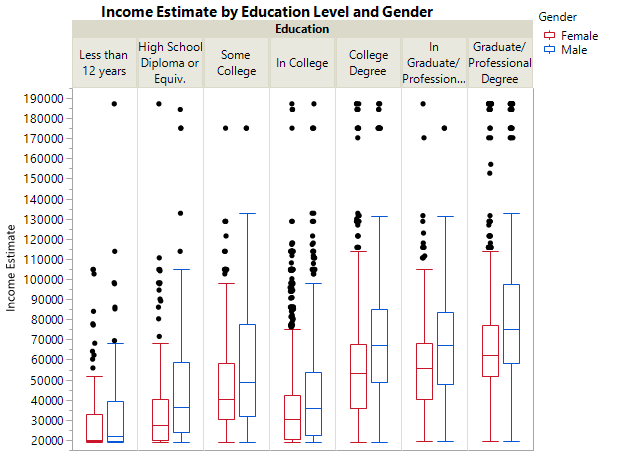
**Figure 12. Stacked-Bar Chart of Job Status by Education Level**

Education level can also be compared to smoking frequency. Figure 13 graphs smoking frequency from the sample by education level. The education level with the largest proportion of individuals who have never smoked is less than 12 years with 77.21% of the sample having never smoked. From there, the proportion of the sample that never smoked decreases with education level, peaking at the some college category, and increases slightly after that. Of the individuals in the sample who had some college, only 41.41% of them reported never smoking, while 22.88% had only formerly smoked. This means that 35.71% of the individuals in the sample who had only some college education smoke, which is the largest proportion of the education levels, followed by the in college level, where 30.06% of the sample smoked. The general relationship here is that the intermediate education levels, from having some college to being in college, tend to have slightly more smokers than the other education levels, but not by much. There is not a strong relationship between education level and smoking frequency.



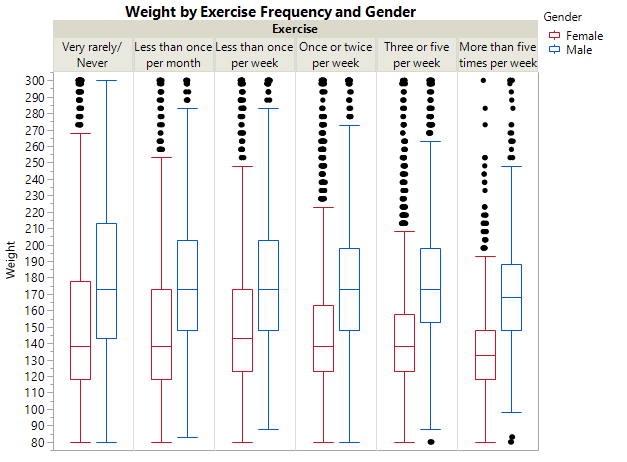
**Figure 13. Stacked-Bar Chart of Smoking Frequency by Education Level**

Moving on to more complex analyses, Figure 14 shows income estimate graphed by both education level and gender. One key point of this graph is that males have a higher income estimate than females in each education level. The center of the boxplots for the males are higher up than females at each education level. Despite this, the bottom tails are the same across both genders and all education levels. This means that the lowest incomes do not change much, regardless of gender and education level. Higher education levels lead to larger income estimates across genders as well, as seen by the increase in the middle section of the boxplot as education levels increase. There are some exceptions to this, especially with the in college level, which has a similar or even slightly lower income estimate than the high school diploma level. This is likely due to college students not focusing on work and income until after they obtain a degree. Income estimates increase with education, regardless of gender, with only slight decreases when individuals are in school. One other trend is that the income estimates are more right-skewed at the lower education levels and become more symmetric as education levels increase. This may be because at lower education levels, most income estimates are clustered at a lower level. Education gives more opportunities and widens the cluster towards a higher income estimate with education. Incomes are higher for men and for higher education levels.



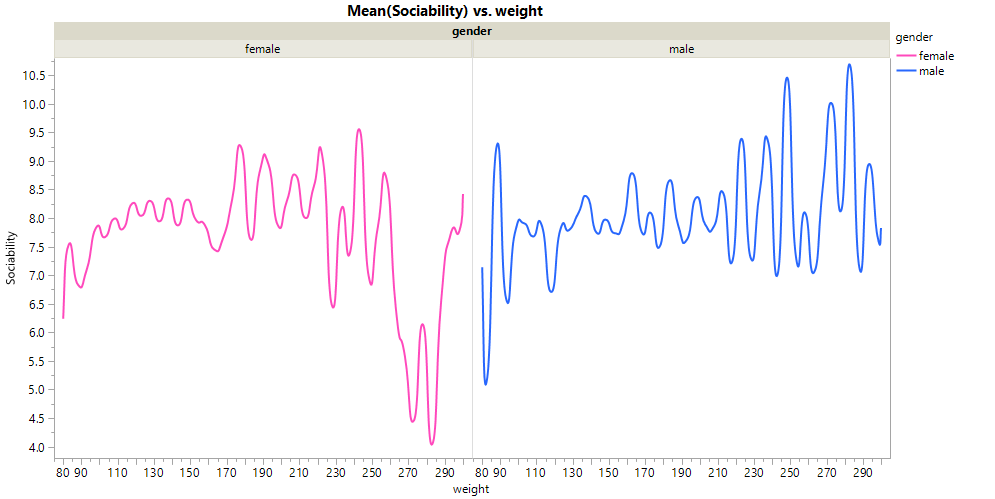
**Figure 14. Boxplots of Income Estimates by Education Level and Gender**

Another comparison that can be made is between exercise frequency and weight, stratified by gender. These variables are graphed in Figure 15. From the graph, more exercise appears to trend with lower weights in both males and females. As the amount of exercise increases, the upper half of the boxplot shifts downward. The variation in weights also decreases for both genders as exercise increases, meaning that exercise may move individuals toward a more central weight. This happens in both genders as the boxplots shrink with more exercise. The effect seems to be slightly different based on gender, with the first quartile weight increasing in males with exercise, but staying more steady in females. This means that even though variation in weight decreases for both genders, weights in females tend to decrease further downward than in males with more exercise. This effect may be due to more muscle-building exercises in males. Additionally, the comparison shows that males are generally heavier than females, even if they have similar exercise habits. The median weight and middle 50% of the weights is higher for males at all levels of exercise frequency.

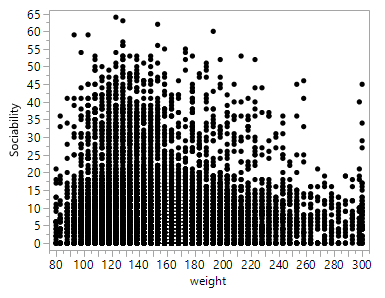
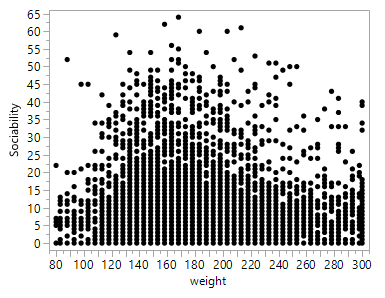


**Figure 15. Boxplots of Weight by Exercise Frequency and Gender**

One of the last trends we examined considered how weight and gender affected how someone would score in sociability. In figure 16 we have a visualization available comparing the sociability means at different weights in two separate line graphs split by gender. Figure 17 provides a scatterplot of all of the individuals and their relevant scores. The visualization appears to show that as weight increases for females there is a decrease in sociability; however, for males there is a different tendency where particularly low values in weight have a negative impact on sociability. While there is quite a range in possible values, the mean sociability score for different weights tends to reliably stay between 4 and 10 on the scale. At different weights, the mean sociability varies noticeably but the GSAM for either males or females was approximately the same with females scoring a mean of 8.01709 and males had a mean of 7.9691. In addition, the total variance for the genders were similar as well (SAVF=73.0638, SAVM=69.1759).



**Figure 16. Line Graph of the SAM of Sociability by Weight and Gender**

**Figure 17. Scatterplot of Sociability by Weight, Stratified by Gender (Male Left, Female Right)**

**Statistical Inference Tests**

The first question we wanted to investigate was whether the gender wage gap exists in our data and how large the gap really is. For this question, we started by examining the stratifying income estimates by gender. Splitting the data by gender forms two groups, making a two sample test most appropriate. Specifically, we will use a Wilcoxon Rank Sum test. For this test, the assumptions are that the data follows some continuous distribution and each observation is independent of others. Income estimates do appear to follow some continuous distribution and we assume that the observations are independent of each other just by the way the data was collected since the first question the survey asks is whether you’ve taken the survey before. Because we are using a Wilcoxon Rank Sum test, we must adjust our hypothesis to be two-tailed. The null hypothesis for this test will be that there is no statistically significant difference in income estimates due to gender. The alternative hypothesis is that there is a statistically significant difference in income estimates due to gender. For this test, we will use a threshold of unusualness of 0.05.

The results of the Wilcoxon Rank Sum test are that S=12491641, with *p*<0.001. This means that there appears to be a statistically significant difference in income estimates due to gender (S=12491641, *p*<0.0001). The median difference income estimate between genders is 11090 dollars in favor of males. Also, if we chose one male and one female, we expect the male to have a higher income estimate than the female 62.07% of the time.

This result shows that the gender gap in income is apparent and statistically significant. Males have higher income estimates as seen by the median difference which is 11090 dollars in favor of males and the probability of superiority of 0.6207.

Our second question explores what impact weight has on sociability, and if there is a difference in this effect between genders. One of the inference tests we will be doing considers whether there is any statistically significant effect that weight has at any level on sociability. In addition, we stratify the data by gender in order to account for gender norms in society. The assumptions we make for this data are independence and that scores in sociability follow some continuous distribution. In order to answer this, we will utilize the nonparametric Kruskal-Wallis test as an omnibus test. For our threshold of unusualness, we will be using an alpha of 0.05. The null hypothesis in this situation states that sociability is constant regardless of the individual's weight. On the other hand, the alternative hypothesis would imply that there is some statistically significant difference in sociability due to the weight of the individual.

After conducting some statistical analyses, we are unable to reject the null hypothesis (H45= 36.5258, p = 0.8119). According to this, weight does not appear to have an impact on sociability.

**Discussion/Conclusion**

The first statistical research question was to what extent estimated income was affected by gender. In our analysis a Wilcoxon Rank Sum test was utilized to answer the question. The threshold of unusualness was set at 0.05. We found that there appeared to be a statistically significant difference in the estimated incomes for the two genders (S=12491641, *p*<0.0001). Thus, we rejected the null hypothesis that there was no difference in the expected income between the two genders. Approximately 62% of the time we would expect a randomly selected man to have a higher estimated income than a randomly selected woman.

The implication to the answer to our statistical research question is twofold. One, the answer implies that there may be unequal pay in the workforce based upon gender. Two, the answer implies that society may reward people on the basis of our perceptions of people because of physical attributes. Further research is needed to determine whether there is a relationship between other physical attributes to determine whether the relationship we noticed is present among other physical attributes such as height, weight, and race.

The second statistical research question asks whether the weight of an individual affects their sociability when stratifying the data by gender? The threshold of unusualness was set at 0.05. However, the data does not show support for a statistically significant relationship (H45= 36.5258, p = 0.8119). Therefore, we failed to reject the null hypothesis that there was no difference in effect of a person’s weight on sociability when stratified by gender.

The answer to this statistical research question implies that there is no relationship between weight and sociability when stratified by gender. This leads us to think that there should be other factors considered for a relationship with sociability instead. In future work on sociability, factors such as age and estimated income could be considered as potential causes. In addition, the fact that data on weight was continuous was a possible confound, therefore future work on this topic could involve grouping individuals into categories of weight that range by 20lb intervals. This would allow for more accurate and useful comparisons between pairs.

Future work on this topic could involve recording more explanatory variables on the data, both personality and demographical. For instance, exercise frequency and depression could be used as the explanatory variables for weight stratified by gender. Education level could also be refined to include information about the major of study; categories such as STEM (science, technology, engineering, and mathematics), humanities, and business could be used to look at how estimated income is affected by the area studied. Other personality traits, such as the acceptance of others could also be used for statistical inference to provide information on which demographic and personality traits have strong connections. These possibilities in the future would help further account for differences and trends in personality based on demographics. Finally, utilizing a greater number of personality questions in calculating the scores for particular personality traits would result in a more robust analysis of individual’s scores.

From this exploratory data analysis, there were several major points that the group learned. The foremost is that data can be difficult to interpret directly from responses, especially with survey questions. Most of the personality assessments in this data used questions that corresponded to multiple different aspects of personality and needed to be interpreted, combined, and compared to a key to give personality scores, such as the ones we used to measure sociability. Another point that we learned is that some statistical research questions are too complicated to answer easily using the preliminary knowledge we have of statistical inference. We had originally wanted to see if the gender wage gap was still present when we accounted for education level, but our statistical inference techniques did not work well with three variables. The group explored options such as ANOVA F with blocking, but found that the different numbers of males and females in the sample made the method unstable. Further knowledge about statistical inference tests and the assumptions that need to be met are necessary to perform more complex inference tests like the ones we had originally wanted to do.

From the inference tests themselves, the group learned that the gender wage gap was apparent in the sample and that the size of this difference was large. The group also learned that weight does not appear to impact sociability, even when stratified by gender. These statistical inference tests showed us that further research and a more comprehensive, straightforward, list of variables would help more accurately pinpoint the relationships between personality and demographic traits. In conclusion, more research in this area could yield better and more accurate findings.

**References**

Brown, Anna, and Eileen Patten. "The narrowing, but persistent, gender gap in pay." *Pew Research*. N.p., 3 Apr. 2017. Web. 16 Apr. 2017.

Condon, David M.; Revelle, William, 2015, "Selected personality data from the SAPA-Project:

08Dec2013 to 26Jul2014", [doi:10.7910/DVN/SD7SVE](http://dx.doi.org/10.7910/DVN/SD7SVE), Harvard Dataverse, V3,

UNF:6:fjFuROCGJ1fFy2jQDhdogw==

Revelle, W. (2013, May 20). Personality Theory and Research. Retrieved April 17, 2017, from

<https://personality-project.org/>

Williams, R. J., & Ricciardelli, L. A. (2014). Social Media and Body Image Concerns: Further

Considerations and Broader Perspectives. *Sex Roles*, 71(11), 389-392.

Doi:10.1007/s11199-014-0429-x