STAT430 Project 2

Name: Ali Arshad Paracha

UID: 114689573

2. I will explore the effects of age, gender and employment type on the hours an individual spends playing video games per week. I want to see if there is a correlation between hours spent playing video games (per week) and age, gender and employment. Hence, the thing that I am trying to predict are hours and the predictors variable(s) that I will use may consist of either age, gender or employment type.

There are 5 variables in my study gender, age, employment type (student, part time, full time), preferred gaming platform and House spent playing video games per week. The data was collected by conducting surveys over multiple subreddits and through the STST430 class.

The results from project 1 show that both age and hours are skewed right and have outliers, so the best measure of center and spread for these quantitative variables was the five-number summary, which consists of the median and IQR. For the categorical data, project 1 shows that we have more males than females and more students than part/full time employees. This fact can be used to help our analysis later.

Note: Outliers were removed to get a better dataset, and to get better regression results.

3.

1. **Refer to Figure 1**:

The results show us that Age and hours spent gaming are correlated but negatively, they have a negative correlation value (-0.66) which means that as one of them increases the other decreases.

The table also tells us that, hours spent gaming and gender are correlated as well. Their correlation value is negative (-0.65) but this is because we encoded males to be the value 0 and females to be the value 1. This tells us that as the encoding increases the hours spent playing video games decreases, which in our context means that females spend lesser time playing video games than males do (which was also the conclusion we reached in project 1).

We also see that time spent playing video games is correlated to employment as well, though this one is a weaker correlation (-0.41). The correlation is negative because of our encodings. In our context and the way, we encoded the different employment types, this negative correlation tells us that students spend the most time playing video games, and the time spent playing video games decreases when individuals are part-time employees and then full-time employees.

Lastly, we see that Age and employment are heavily correlated as well (0.77). This means that as age increases the employment type goes from Student to Part time Employee to Full time employee, which is what we expect.

1. **Refer to Figure 2:**

The reason we use Adjusted R squared values is because, it takes cares of more predictors being added and increases only if adding new predictors would increase its value by that which is expected by chance otherwise it decreases. It takes the number of variables into account and this gives us a much better idea of whether we should use the variables for regression.

Looking at the results of Figure 2 we can see that we have a high adjusted r-squared value for Gender and Age (0.819), this means that these two together have an effect on the hours spent gaming, provided they are not correlated. The same can be said about Gender and Employment, since the adjusted r squared value is high (0.708), we can assume that these two independent variables together effect the hours spent gaming, again with the assumption that they are not correlated. The same can be said about gender, employment and age all three together, their adjusted r square value is 0.8199. However, we can disregard this one because we saw above that both age and employment are heavily correlated (backed up by the regression results in **Figure 3**), and we are already considering age and gender together, so we can disregard this.

1. **A) Refer to Figure 3 (Regression with Age, Gender and Employment as predictor variables):** Immediately we can see that the employment codes have a p value significantly greater than 0.05 (0.1444 and 0.9959), which tell us that employment is not a meaningful addition to the model and is not statistically significant. (This is most likely because age and employment are correlated, see results of **Figure 2**) This means that changes in its value are not related to the value of the response (Predicted) variable, so we can and should disregard it. The next regression we do (Figure 4) makes use of this fact and drops employment as one of the predictor variables.

**B) Refer to Figure 4 (Regression with Age and Gender as predictor variables):**

**ANOVA Results**: Looking at the results of Figure 4 we can see that the p - value is less than 0.05 (<0.001). This is significant because it means that our all our independent variables are most likely significant and that the y-variable (hours) and the x-variables (gender and age) are related. Furthermore, we have a small value of the MSE (8.775), which backs up the fact mentioned above and also tell us that we probably have a good model.

**Regression Results:**

1. The regression equation is:

Hours\_Spent\_Gaming = - 8.795 \* Gender - 1.306 \* Age + 44.673

2) Both gender and age have p-value less than 0.05 (< 0.001 and < 0.001 respectively), this means that both these variables are meaningful additions to the model, because changes in their values are related to changes in the response variable (variable we are predicting).

3) The residual plot for age shows that the residuals are evenly (normally) distributed, they are evenly spread out across the zero line. Furthermore, they are also independent. The residual plot for Gender shows the same, the residuals are normally distributed across the zero line and are independent. (Note, the reason all points are zero and one is because that is how we encoded the two genders Male and Female, so those two are the only values they can take). Finally, we also look at the residual plot for the Predicted value and see that all points are normally distributed, this tell us that our model is a good one and the regression results are meaningful. However, we still must compare this model to the one in which we use gender and employment as predictor variables, because since age and employment are correlated we cannot be sure which model is a better one.

C) Refer to Figure 5 (**Regression with Age and Gender as predictor variables**):

The results of figure 5 show us that the model has an adjusted r-squared value of 0.7435, which is less than the Adjusted R-squared value for our previous model (Figure 4). Also, the MSE for this model is higher (12.45) compared to our previous model, hence we can reach the conclusion that this model is not a better one than the one that contains age as a predictor variable rather than its correlated counterpart, employment.

1. **Assumptions:** 
   1. Our dependent Variable is continuous, since it deals with time (Hours spent gaming). Hence, we are trying to predict a quantitative variable so multiple linear regression can be used. Our independent variable gender is categorical, and age is quantitative.
   2. We see that our predictor terms are independent from one another. Age is independent of Gender, and also, we saw in the residual plots of Figure 4 that both age and gender are evenly distributed across the zero line, which fulfills another assumption.
   3. Lastly, the residual plots in Figure 4 showed us that age, gender and hours were normally distributed, so our model fulfills all the necessary assumptions for Multiple Linear Regression.

Since, all these assumptions are met, we can be sure that our results are meaningful as these assumptions guarantee a useful model, one that we can use to accurately predict our dependent variable.

1. **Good Sample?**

I believe the sample is good because, as mentioned above the it has a high

adjusted r squared value (0.8193). Most predictor variables are independent of one another and have normally distributed residuals. Also, the sample has a small Mean Squared Error, compared to other models (8.77), which makes our sample a good one. That being said, there is also a slight chance the sample may not be good because it contains only 60 observations and it may not be completely representative of the entire population.