```
/* NOTE: Whenever I mention hours spent gaming below, I am referring to hours spent
playing video games PER WEEK, not mentioning it again and again to remove redundancy  */

/* Reading the dataset */
Data survey;
    INFILE '/home/aliparacha19960/EPG194/Survey.csv' delimiter=',';
    Length Platform $20;
    Length Employment $20;
    Input Gender $ Age Platform $ Hours_Spent_Gaming Employment $;
run;

/* Creating a new dataset that has encodings for all the catrgorical variables
(Gender, Platform and Employment type)  */
data survey2;
    set survey;

    if Gender='Male' then
        do;
            gender_code=0;
        end;
    else
        do;
            gender_code=1;
        end;

    if Platform='PC' then
        do;
            platform_code=0;
        end;
    else if Platform="XBOX" then
        do;
            platform_code=1;
        end;
    else
        do;
            platform_code=2;
        end;

    if Employment='Student' then
        do;
            employment_code1=0;
            employment_code2=0;
        end;
    else if Employment="Part - time Employee" then
        do;
            employment_code1=0;
            employment_code2=1;
        end;
    else
        do;
            employment_code1=1;
            employment_code2=0;
        end;
run;

/* Part 1: Find Correlation Cooefficient  */
/* Finding Pearson's coorelation coefficient, this will help us see which variables might
be correlated with one another  */
proc corr data=survey2;
Title "Figure 1";
run;
/* The results show us that Age and hours spent gaming are correlated but negatively,
they have a negative coorelation 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 coorelated as well. Their
 coorelation value is negative as well (-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 coorelation (-0.41). The coorelation is negative because of our
 encodings. In our context and the way we encoded the different emplloyment types, this
 negative coorelation tell 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 amployment are heavil coorelated 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.
 */

/* Part 2: Perform Regression  */

/* I will now plot the Adjusted R-squared values to find the best model*/
proc reg data=survey2;
    model hours_spent_gaming = gender_code employment_code1 employment_code2 age / selection = adjrsq;
    Title "Figure 2";
run;
/* The reason we use Adjusted R squared values is because, it takes care 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. 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 really high adjusted r square
value for Gender and Age (0.819), this means that most likely these two together have an effect
on the hours spent gaming. 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. The same can be said about gender, employment and age all three
together, their adjusted r square value is 0.8199 which means all three together might also effect the
time spent playing video games provided there isn't a correlation between the variables. */


/* Performing Regression by using gender, age and employment_type as predictor variables.
Hours_Spent_Gaming is the predicted variable.  */
proc reg data=survey2;
     model hours_spent_gaming = gender_code employment_code1 employment_code2 age;
     Title "Figure 3";
run;
/* Immediately we can see that the employment codes have a p value siginificantly 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 means that changes in its value are not related to
the value of the response (Predicted) vairable, so we can and should disregard it. */


/* Performing Regression by using gender and age as the predictor variables.
Hours_spent_gaming is the predicted variable */
proc reg data=survey2;
     model hours_spent_gaming = gender_code age;
     Title "Figure 4";
     plot hours_spent_gaming*age;
run;
/* Looking at the ANOVA 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 indpendent variables are 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.
```

```
Looking at the REGRESSION results of Figure 4 we see that

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 reisudal 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 distrbiuted 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 can not
be sure which model is a better one.
*/

/* Performing Regression that uses gender and employment as predictor variabes */
proc reg data=survey2;
     model hours_spent_gaming = gender_code employment_code1 employment_code2;
     Title "Figure 5";
run;
/* The results of figure 5 show us that the model had a 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. */


/* Part 3: Assumptions

i) 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.
ii) 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.
iii) 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 assunptions are met, we can be sure that our results are meaningful as these
assumptions gurantee a useful model, one that we can use to accurately predict our
dependent variable.

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 sampple 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.
 */
```