Logo

Description automatically generated

**ALY 6015.80428 INTERMEDIATE ANALYTICS**

**FINAL PROJECT: INITIAL ANALYSIS REPORT**

**Submitted to: Prof. Roy, Wada**

**Submitted by: Tanusha Erpula |Kiran kasu | Kelvina Pethani| Abhinav Jain**

**Date: 05/08/2022**

**INTRODUCTION**

**Background:**

The National Longitudinal Survey of Youth in the year 1997 is a nationally representative sample of 8,984 men and women born between 1980 and 1984 who were living in the United States 1997. The participants' ages varied from 12 to 16 on December 31, 1996. Interviews were conducted annually from 1997 to 2011, and then every two years after that. The current cohort has been polled 18 times so far. Data is available from Round 1 (1997-98) until Round 19. (2019-20).

The NLSY97 collects extensive information on respondents' work habits and educational experiences. The poll also contains information about the kids' families and neighborhoods, which will help researchers figure out how schooling and other variables affect these newcomers to the country.

**Processing the data:**

The goal of the NLSY97 is to evaluate the interviews done by youth in the United States in 1979. The raw dataset has 10,402 rows and 36 fields. After cleaning the data set, there were 8130 rows and 12 fields. Year of birth is a numeric data type that indicates when a child was born. The text data types country of birth, sample race, and region are all text data types. Size of Family Number of Children is a numeric data type that refers to the number of individuals in the family, and Family Size is a numeric data type that refers to the size of the family.

**Research method:**

In this project, we will use Linear regression to build predictive models. The linear regression analysis is a regression technique that uses the least square function to model the relationship between one or more independent variables and dependent variables. This function is a linear combination of one or more model parameters (regression coefficients).

We typically add a regular term to the goal function to keep the parameters as "simple" as possible while ensuring the best fit error and making the model generalizable.

Lasso regression uses a one- and two-parameter constraint to limit the number of non-zero parameters. In the optimization process, the intersection of the function contour and the constraint space is the ideal solution, and the constraint space can be considered the canonical term.

Further develop your summary statistics tables by combing it with subset analysis (e.g. by gender or year or country)

* + Two or four of subset analysis

**ANALYSIS**

**#Descriptive Table:**

In this table it describes the descriptive values as we can see that the data is being cleaned and after cleaning the dataset, we are leftover with the Subsets like mean, sd, median, mad, min, max, range. which describes the values after cleaning the dataset.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | mean | sd | median | mad | min | max | range |
| YEAR | 2004.563 | 4.716593 | 2004.5 | 5.9304 | 1997 | 2013 | 16 |
| KEY\_SEX\* | 1.512636 | 0.499842 | 2 | 0 | 1 | 2 | 1 |
| KEY\_RACE\* | 4.262833 | 0.959299 | 5 | 0 | 1 | 5 | 4 |
| CV\_SAMPLE\_TYPE\* | 1.247669 | 0.43166 | 1 | 0 | 1 | 2 | 1 |
| KEY\_RACE\_ETHNICITY\* | 2.792542 | 1.317843 | 4 | 0 | 1 | 4 | 3 |
| KEY\_BDATE\_Y | 1982.01 | 1.3962 | 1982 | 1.4826 | 1980 | 1984 | 4 |
| CV\_MARSTAT\_\* | 6.848471 | 2.726386 | 8 | 1.4826 | 1 | 13 | 12 |
| CV\_HH\_POV\_RATIO\_ | 201.2005 | 318.1924 | 77 | 121.5732 | -5 | 3227 | 3232 |
| CV\_HH\_SIZE\_ | 2.476581 | 3.378744 | 3 | 1.4826 | -5 | 19 | 24 |
| CVC\_FIRST\_MARRY\_DATE\_Y\_XRND | 891.2906 | 999.1146 | -4 | 0 | -4 | 2014 | 2018 |
| CVC\_MARRIAGES\_TTL\_XRND | 0.497585 | 0.591577 | 0 | 0 | 0 | 5 | 5 |
| CHILD\_BIRTH\_01\_Y | 1137.014 | 995.0688 | 2000 | 16.3086 | -4 | 2014 | 2018 |
| CV\_BIO\_CHILD\_HH\_ | -2.76559 | 2.507002 | -4 | 0 | -5 | 8 | 13 |
| WAGE\_ | 9766.076 | 17719.53 | 100 | 155.673 | -5 | 180331 | 180336 |
| PARTNERS\_UID\_01\_ | 55791.32 | 89833.5 | -4 | 0 | -5 | 201306 | 201311 |
| PARTNERS\_UID\_02\_ | 1802.533 | 18958.39 | -4 | 0 | -5 | 201304 | 201309 |
| AGE\_FIRST\_CHILD | 10.91756 | 13.41126 | 17 | 16.3086 | -4 | 34 | 38 |
| AGE\_ | 21.91643 | 4.945806 | 22 | 5.9304 | 12 | 34 | 22 |

**Table 1: Descriptive Table**

**#KEY\_RACE & WAGE 2**

In this descriptive table Key race and wages describes the values after using the group by function group\_by() which helps us to form group from the data frame by multiple columns, with mean, median, sd, iqr, min and max. for key race and wages.

**Table 2: Descriptive Table of KEY\_RACE & WAGE 2**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | KEY\_RACE | mean | median | sd | iqr | min | max |
| 1 | American Indian, Eskimo, or Aleut | 8467.077 | 65 | 13498.41 | 14016 | -5 | 90000 |
| 2 | Asian or Pacific Islander | 12723.22 | -2 | 25351.05 | 15004 | -5 | 180331 |
| 3 | Black or African American | 7131.965 | -2 | 14236.07 | 8004 | -5 | 180331 |
| 4 | Something else? (SPECIFY) | 9265.261 | -2 | 16556.34 | 14004 | -5 | 180331 |
| 5 | White | 10995.04 | 600 | 18957.41 | 16004 | -5 | 180331 |

**#CV\_MARSTAT & WAGE TABLE:**In this table it describes the CV marstat and wages table using same function used for key race and wages. Where it describes the same values like min, max and other function to describe the dataframe for the CVmarstat and wages where it defines the 13 rows of the data.

**Table 3: Descriptive Table of CV\_MARSTAT & WAGE**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | CV\_MARSTAT\_ | mean | median | sd | iqr | min | max |
| 1 | -3 | 14440.38 | 8000 | 17268 | 25752 | -4 | 70712 |
| 2 | -4 | 215.7454 | -4 | 805.0332 | 124 | -4 | 40000 |
| 3 | -5 | -5 | -5 | 0 | 0 | -5 | -5 |
| 4 | Divorced, cohabiting | 19757.85 | 16000 | 22509.05 | 30644 | -4 | 180331 |
| 5 | Divorced, not cohabiting | 19478.47 | 15000 | 21762.98 | 30002 | -4 | 180331 |
| 6 | Married, spouse absent | 14487.07 | 8000 | 19629.47 | 24752 | -4 | 180331 |
| 7 | Married, spouse present | 24503.15 | 20000 | 26001.03 | 37000 | -4 | 180331 |
| 8 | Never married, cohabiting | 15753.02 | 10000 | 19264.81 | 25002 | -4 | 180331 |
| 9 | Never married, not cohabiting | 9012.167 | 1500 | 15381.31 | 13004 | -4 | 180331 |
| 10 | Separated, cohabiting | 11597.73 | 1500 | 15677.59 | 19004 | -4 | 67000 |
| 11 | Separated, not cohabiting | 15313.96 | 8000 | 18379.71 | 27002 | -4 | 121993 |
| 12 | Widowed, cohabiting | 9616.118 | 0 | 15060.85 | 15004 | -4 | 56000 |
| 13 | Widowed, not cohabiting | 11511.57 | 7000 | 13704.03 | 17002 | -4 | 50000 |

**#YEAR VS AGE TABLE:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | YEAR | mean | median | sd | iqr | min | max |
| 1 | 1997 | 14.35393 | 14 | 1.488282 | 3 | 12 | 18 |
| 2 | 1998 | 15.35393 | 15 | 1.488282 | 3 | 13 | 19 |
| 3 | 1999 | 16.35393 | 16 | 1.488282 | 3 | 14 | 20 |
| 4 | 2000 | 17.35393 | 17 | 1.488282 | 3 | 15 | 21 |
| 5 | 2001 | 18.35393 | 18 | 1.488282 | 3 | 16 | 22 |
| 6 | 2002 | 19.35393 | 19 | 1.488282 | 3 | 17 | 23 |
| 7 | 2003 | 20.35393 | 20 | 1.488282 | 3 | 18 | 24 |
| 8 | 2004 | 21.35393 | 21 | 1.488282 | 3 | 19 | 25 |
| 9 | 2005 | 22.35393 | 22 | 1.488282 | 3 | 20 | 26 |
| 10 | 2006 | 23.35393 | 23 | 1.488282 | 3 | 21 | 27 |
| 11 | 2007 | 24.35393 | 24 | 1.488282 | 3 | 22 | 28 |
| 12 | 2008 | 25.35393 | 25 | 1.488282 | 3 | 23 | 29 |
| 13 | 2009 | 26.35393 | 26 | 1.488282 | 3 | 24 | 30 |
| 14 | 2010 | 27.35393 | 27 | 1.488282 | 3 | 25 | 31 |
| 15 | 2011 | 28.35393 | 28 | 1.488282 | 3 | 26 | 32 |
| 16 | 2013 | 30.35393 | 30 | 1.488282 | 3 | 28 | 34 |

In this table years vs age table it describes same as we are using the same function group\_by() to which helps to group the dataframe by multiple columns with the same functions like min, max, mean, median, iqr. But the difference is like it is describing the data for years and age table

**Table 4: Descriptive Table of KEY\_RACE & WAGE 2**

**#Graph 1**In this graph it describes the values of race and wages where it describes where key race values like American Indian, Eskimo or Aluet, Asian or pacific Islander, Black or African American , Something else?(Specify)and white on the scale from 0 to 80000 count. where it helps to define the values as per the wages.

Chart, bar chart

Description automatically generated

**Graph 1: Bar Graph for KEY\_RACE & WAGE**

**#Graph 2**In this graph it describes the year and the age dataset on the age dataset it defines from 0 to 35 and on the y-axis it count from 0 to above 7500 where it describes the values from the year 1997 till 2013 to describe the values in the graph to define the values of the age and the year dataset.

Chart, bar chart, histogram

Description automatically generated

**Graph 2: Bar Graph of YEAR & AGE**

**#Graph 3**In this graph it describe the values like Martial Status vs Wages where it describes the values like -3, -4, -5, and many other data. where it defines the values of wages on the darker part of the graphs presented. and on the other side of the axis, it define from 0 to 60000 and above where is describes the values.

Chart, histogram

Description automatically generated

**Graph 3: Bar Graph of Marital Status vs Wage**

**#Descriptive statistics of new variables**In this Descriptive statistics of new variables where new variables where added like youth, teen and adult where added to the dataset using the transform function where we can add the predictive variables and we can increase the efficiency of the model by adding the new variable to the dataset, where it also describes the values of the functions like mean, sd, median, mad, min, max, and range to the dataset.

**Table 5: Descriptive Statistics Table of New variables**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | mean | sd | median | mad | min | max | range |
| YEAR | 2004.563 | 4.716593 | 2004.5 | 5.9304 | 1997 | 2013 | 16 |
| KEY\_SEX\* | 1.512636 | 0.499842 | 2 | 0 | 1 | 2 | 1 |
| KEY\_RACE\* | 4.262833 | 0.959299 | 5 | 0 | 1 | 5 | 4 |
| CV\_SAMPLE\_TYPE\* | 1.247669 | 0.43166 | 1 | 0 | 1 | 2 | 1 |
| KEY\_RACE\_ETHNICITY\* | 2.792542 | 1.317843 | 4 | 0 | 1 | 4 | 3 |
| KEY\_BDATE\_Y | 1982.01 | 1.3962 | 1982 | 1.4826 | 1980 | 1984 | 4 |
| CV\_MARSTAT\_\* | 6.848471 | 2.726386 | 8 | 1.4826 | 1 | 13 | 12 |
| CV\_HH\_POV\_RATIO\_ | 201.2005 | 318.1924 | 77 | 121.5732 | -5 | 3227 | 3232 |
| CV\_HH\_SIZE\_ | 2.476581 | 3.378744 | 3 | 1.4826 | -5 | 19 | 24 |
| CVC\_FIRST\_MARRY\_DATE\_Y\_XRND | 891.2906 | 999.1146 | -4 | 0 | -4 | 2014 | 2018 |
| CVC\_MARRIAGES\_TTL\_XRND | 0.497585 | 0.591577 | 0 | 0 | 0 | 5 | 5 |
| CHILD\_BIRTH\_01\_Y | 1137.014 | 995.0688 | 2000 | 16.3086 | -4 | 2014 | 2018 |
| CV\_BIO\_CHILD\_HH\_ | -2.76559 | 2.507002 | -4 | 0 | -5 | 8 | 13 |
| WAGE\_ | 9766.076 | 17719.53 | 100 | 155.673 | -5 | 180331 | 180336 |
| PARTNERS\_UID\_01\_ | 55791.32 | 89833.5 | -4 | 0 | -5 | 201306 | 201311 |
| PARTNERS\_UID\_02\_ | 1802.533 | 18958.39 | -4 | 0 | -5 | 201304 | 201309 |
| AGE\_FIRST\_CHILD | 10.91756 | 13.41126 | 17 | 16.3086 | -4 | 34 | 38 |
| AGE\_ | 21.91643 | 4.945806 | 22 | 5.9304 | 12 | 34 | 22 |
| teen | 0.282257 | 0.4501 | 0 | 0 | 0 | 1 | 1 |
| youth | 0.3125 | 0.463514 | 0 | 0 | 0 | 1 | 1 |
| adult | 0.27198 | 0.444981 | 0 | 0 | 0 | 1 | 1 |

**#Descriptive Statistics of Train dataset**In this descriptive train dataset where it defines the analysis of the predictive variables where the data is being split into the ratio from 80 to 20 where we created the training and test data to implement our model with the values defines the descriptive statistics of the train dataset.

**Table 6 : Descriptive Statistics for Train Data set**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | mean | sd | median | mad | min | max | range |
| YEAR | 2004.557 | 4.714764 | 2004 | 5.9304 | 1997 | 2013 | 16 |
| KEY\_SEX\* | 1.511645 | 0.499867 | 2 | 0 | 1 | 2 | 1 |
| KEY\_RACE\* | 4.262895 | 0.958958 | 5 | 0 | 1 | 5 | 4 |
| CV\_SAMPLE\_TYPE\* | 1.2481 | 0.431912 | 1 | 0 | 1 | 2 | 1 |
| KEY\_RACE\_ETHNICITY\* | 2.79066 | 1.317669 | 4 | 0 | 1 | 4 | 3 |
| KEY\_BDATE\_Y | 1982.012 | 1.396185 | 1982 | 1.4826 | 1980 | 1984 | 4 |
| CV\_MARSTAT\_\* | 6.844013 | 2.728529 | 8 | 1.4826 | 1 | 13 | 12 |
| CV\_HH\_POV\_RATIO\_ | 200.7971 | 318.867 | 76 | 120.0906 | -5 | 3227 | 3232 |
| CV\_HH\_SIZE\_ | 2.475324 | 3.379919 | 3 | 1.4826 | -5 | 19 | 24 |
| CVC\_FIRST\_MARRY\_DATE\_Y\_XRND | 892.7589 | 999.277 | -4 | 0 | -4 | 2014 | 2018 |
| CVC\_MARRIAGES\_TTL\_XRND | 0.498359 | 0.591832 | 0 | 0 | 0 | 5 | 5 |
| CHILD\_BIRTH\_01\_Y | 1139.44 | 994.7336 | 2000 | 16.3086 | -4 | 2014 | 2018 |
| CV\_BIO\_CHILD\_HH\_ | -2.76187 | 2.508231 | -4 | 0 | -5 | 8 | 13 |
| WAGE\_ | 9754.338 | 17721.73 | 100 | 155.673 | -5 | 180331 | 180336 |
| PARTNERS\_UID\_01\_ | 55791.17 | 89833.51 | -4 | 0 | -5 | 201306 | 201311 |
| PARTNERS\_UID\_02\_ | 1755.753 | 18713.93 | -4 | 0 | -5 | 201304 | 201309 |
| AGE\_FIRST\_CHILD | 10.94911 | 13.40802 | 17 | 16.3086 | -4 | 34 | 38 |
| AGE\_ | 21.90979 | 4.945063 | 22 | 5.9304 | 12 | 34 | 22 |
| teen | 0.28242 | 0.450178 | 0 | 0 | 0 | 1 | 1 |
| youth | 0.311904 | 0.463273 | 0 | 0 | 0 | 1 | 1 |
| adult | 0.271451 | 0.44471 | 0 | 0 | 0 | 1 | 1 |

**#Correlation Matrix**

In this Correlation Matrix where it describes the values like Wage, Year, Age, PartnersUId\_1, ParternsUID\_2 and CV\_HH\_POV\_Ratio where this values defines 91% accuracy values, where it defines on a scale of -1 to 1. Where here the P-Value is significant because it is equal to 0.000. Where on the other side it describes it is not statistically significant because the P-Value is 0.092 which is greater than the usual significant level of 0.05. Where the P-Value is the independent variable.

**Text

Description automatically generated with medium confidence**

**Graph 4: Correlation Matrix**

Chart, bubble chart

Description automatically generated

**#Linear Regression Model**

In this Linear Regression model it interprets the line of NLSY using the lm function to call the values like Year, Age, and CV\_Sample\_Type and to define it more effectively the r square values are used to describe it adj is also used to get the clear ideas of the values and using the linear graph it describes the values like on the x-axis it defines the year and on the y-axis, it defines the age till 35. In the regression output example below, the South and North predictor variables are statistically significant since their p-values are both 0.000. Due to its larger p-value, East, on the other hand, falls short of the 0.05 significant criterion (0.092). The p-values for the independent variables are displayed in the regression analysis results .Coefficient P-values are frequently used to determine if variables should be included in the final model. We would consider eliminating East based on the following findings. Model accuracy may decrease if non-statistically significant variables are retained. In Linear Relationships, Interpreting Regression Coefficients. The sign of a regression coefficient indicates whether the independent variables are connected to the dependent variable positively or negatively. The mean of the dependent variable rises as the value of the independent variable rises. As the independent variable rises, the dependent variable tends to decline, resulting in a negative coefficient. The coefficient value represents how much the mean of the dependent variable changes when the independent variable is changed by one unit while the other variables in the model remain constant. You may examine the influence of one variable on its own by keeping the other variables constant.

A screenshot of a computer

Description automatically generated with medium confidence

**A screenshot of a computer

Description automatically generated with medium confidence**

**Graph 5: Plot for linear regression model**

Chart, scatter chart

Description automatically generated

**Conclusion**

We successfully implemented the correlation, Linear Regression apart from that using the data we cleaned the dataset and after that used the groupby () function to form the group from the data set while using the dataset new variable were created and merged in the dataset like teen, youth and adult. After using the same data set it was divided into 80 to 20 ratios where we created the training and the test data to implement in our model to implement in our values to define the descriptive statistics. After that, we created the correlation where ne way for quantifying this relationship is the Pearson correlation coefficient, which is a measure of the linear association between two variables. Its value ranges from -1 to 1, with the following values: A negative linear correlation between two variables is represented by -1, whilst no linear correlation is represented by 0. 1 shows a perfectly positive linear connection between two variables. And In the regression output example below, the South and North predictor variables are statistically significant since their p-values are both 0.000. Due to its larger p-value, East, on the other hand, falls short of the 0.05 significant criterion (0.092). The p-values for the independent variables are displayed in the regression analysis results. And we are successfully able to define the Coefficient P-values are frequently used to determine if variables should be included in the final model. We would consider eliminating East based on the following findings. Model accuracy may deteriorate if variables that aren't statistically significant are kept.

**References**

Statistics review 7: Correlation and regression Retrieved on (May 8, 2022) from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC374386/>

Pasulakiransai (March 02, 2022) Guide to Data Types and How to Graph Them in Statistics Retrieved on (May 8, 2022) from  
<https://www.geeksforgeeks.org/group-by-function-in-r-using-dplyr/>

Group by one or more variables - group\_by Retrieved on (May 8, 2022) from <https://dplyr.tidyverse.org/reference/group_by.html>

(April 04, 2022) Use the group\_by Function in R Dplyr Retrieved on (May 8, 2022) from <https://www.delftstack.com/howto/r/group_by-r/>

Pearson Correlation vs Simple Linear Regression Retrieved on (May 8, 2022) from <http://www.biosci.global/customer-stories-en/pearson-correlation-vs-simple-linear-regression/>

**R code:**

library(psych)

library(tidyverse)

library(dplyr)

library(kableExtra)

library(ggplot2)

library(hrbrthemes)

library(janitor)

library(vtable)

library(jtools)

#Importing the dataset

df <- read.csv("NLYS97.csv")

df %>%

as.tibble()

describe(df)

#Data Cleaning

df$KEY\_SEX<-as.factor(df$KEY\_SEX)

df\_c <-df%>%

select(YEAR,KEY\_SEX,KEY\_RACE, CV\_SAMPLE\_TYPE,KEY\_RACE\_ETHNICITY,KEY\_BDATE\_Y,CV\_MARSTAT\_, CV\_HH\_POV\_RATIO\_,

CV\_HH\_SIZE\_,CVC\_FIRST\_MARRY\_DATE\_Y\_XRND,CVC\_MARRIAGES\_TTL\_XRND, CHILD\_BIRTH\_01\_Y

,CV\_BIO\_CHILD\_HH\_,WAGE\_,PARTNERS\_UID\_01\_,PARTNERS\_UID\_02\_,AGE\_FIRST\_CHILD,AGE\_) %>%

filter(KEY\_RACE %in% c("White","Black or African American","Asian or Pacific Islander","American Indian, Eskimo, or Aleut","Something else? (SPECIFY)")& AGE\_) %>%

na.omit()

df\_c %>%

as.tibble()

summary(df\_c)

dim(df\_c)

describe(df\_c,skew = "FALSE")

#Table1: Descriptive Analysis

write.csv(describe(df\_c),"Descriptive Table 1.csv")

#Table2- Groupby KEY\_Race and Wage\_ #Final#

df\_kr<-df\_c%>%

group\_by(KEY\_RACE)%>%

summarise(mean=mean(WAGE\_, na.rm = TRUE),

median=median(WAGE\_,na.rm = TRUE),

sd=sd(WAGE\_,na.rm = TRUE),

iqr=IQR(WAGE\_,na.rm = TRUE),

min=min(WAGE\_,na.rm = TRUE),

max=max(WAGE\_,na.rm = TRUE))

write.csv(df\_kr,"KEY\_Race and Wage\_ Table 2.csv")

# Graph1: Bar Chart

df\_c%>%

ggplot(aes(KEY\_RACE, color = "WAGE\_"))+

geom\_bar(position = position\_dodge())+

ggtitle("RACE and WAGE")

#Table 3 Groupby Year and AGE #Final#

df\_y<-df\_c%>%

group\_by(YEAR)%>%

summarise(mean=mean(AGE\_, na.rm = TRUE),

median=median(AGE\_,na.rm = TRUE),

sd=sd(AGE\_,na.rm = TRUE),

iqr=IQR(AGE\_,na.rm = TRUE),

min=min(AGE\_,na.rm = TRUE),

max=max(AGE\_,na.rm = TRUE))

write.csv(df\_y,"Year vs Age Table 3.csv")

# Graph2: Bar Chart

df\_c%>%

ggplot(aes(AGE\_, color = "YEAR"))+

geom\_bar(position = position\_dodge())+

ggtitle("YEAR and AGE")

#Table 4 Groupby CV\_MARSTAT\_ and WAGE\_ #Final#

df\_mr<-df\_c%>%

group\_by(CV\_MARSTAT\_)%>%

summarise(mean=mean(WAGE\_, na.rm = TRUE),

median=median(WAGE\_,na.rm = TRUE),

sd=sd(WAGE\_,na.rm = TRUE),

iqr=IQR(WAGE\_,na.rm = TRUE),

min=min(WAGE\_,na.rm = TRUE),

max=max(WAGE\_,na.rm = TRUE))

write.csv(df\_mr,"CV\_MARSTAT\_ and WAGE\_ Table 4.csv")

# Graph3: Bar Chart

df\_c%>%

ggplot(aes(CV\_MARSTAT\_, color = "WAGE\_"))+

geom\_bar(position = position\_dodge())+

ggtitle("Martial Status vs Wage")+

theme(axis.text.x = element\_text(angle = 90))

#Creating new variable by transform

df\_c$teen <- ifelse(df\_c$AGE\_>12 & df\_c$AGE\_<19, 1, 0)

df\_c$youth <- ifelse(df\_c$AGE\_>19 & df\_c$AGE\_<25, 1, 0)

df\_c$adult <- ifelse(df\_c$AGE\_>25 & df\_c$AGE\_<34, 1, 0)

write.csv(describe(df\_c),"Descriptive table with new variables Table 4.csv")

##Prepare train and test data

set\_train<- sort(sample(x= nrow(df\_c), size = nrow(df\_c)\*0.8))

# Descriptive table of train dataset

train\_NLSY<-df\_c[set\_train,]

write.csv(describe(train\_NLSY),"Descriptive table of train dataset Table 5.csv")

# Descriptive table of test dataset

test\_NLSY<-df\_c[-set\_train,]

write.csv(describe(test\_NLSY),"Descriptive table of test dataset Table 6.csv")

#applying correlation model

df\_cor<-df\_c[,c("WAGE\_","YEAR", "AGE\_","PARTNERS\_UID\_01\_","PARTNERS\_UID\_02\_", "CV\_HH\_POV\_RATIO\_")]

df\_mtrx<-cor(df\_cor)

df\_mtrx

write.csv(describe(df\_mtrx),"Correlation matrix Table 7.csv")

#Correctional plot

library(corrplot)

corrplot(cor(df\_cor),method="circle", title = "\nCorrelation Matrix Graph")

#Linear Regression model

lm\_model1<- lm(formula = YEAR~AGE\_+CV\_SAMPLE\_TYPE, data = train\_NLSY)

summary(lm\_model1)

summ(lm\_model1)

lm\_model1 %>%

ggplot(aes(x=YEAR, y=AGE\_))+

geom\_point()

lm\_model1 %>%

ggplot(aes(x=YEAR, y=AGE\_))+

geom\_point()+

geom\_smooth(method="lm", col="green")+

ggtitle(" Plot : Linear Regression Model of NLSY")

#++++++++++++++++++++++++++++++++++++++++++++++++++++To be continued in final Project Submission++++++