Analysis of Data on Household Income and Household Arrangements by Race

Introduction

To create my final dataset for this project I used a total of five datatsets all sourced from the U.S. Census Bureau. The datasets "wc_LA", "bc_LA", and "ac_LA" all contain data on the living arrangements of white children, black children, and children of all races under 18 years old from 1968 through 2020, respectively. All three data sets state the total amount of children under 18 in two parent, one parent, mother-only or father-only households, and the amount of children living with relatives or nonrelatives. The fourth dataset, "household_income" contains data on the household income of some racial/ethinic groups from the years 1967 through 2019. The variables include the median income, mean income, and the percernt distribution of households that fell under 15k all the way up to households with over 200k income for that year. Although this dataset also contains information of hispanic and asian households, their data does not go back all the way to 1967 and there was not enough data points in order to include those categories. There was a lot of tidying that had to be done such as removing unnecessary rows with NAs and renaming the columns.

The "household_income" dataset was separated into three datasets for the three categories- white, black, and all races. I then joined the living arrangement information and the household income information for each respective group by using the inner_join() function. Finally, I created an empty data frame in which I could enter the rows and columns in a more coherent manner. My final dataset "Project_2" contains 53 observations for each racial group from the years 1968 to 2019 and contains all of the variables from the living arrangement and household income datasets (for a total of 20 variables).

Families and the way they are structured are important components of society. I chose to look into these datasets because I wanted to know more about how the family unit has evolved throughout the years and if this differs between groups. It would also be interesting to see how living arrangements are correlated with household income. I suspect that

household income would be negatively impacted with an increase of single parent households while also significantly differing between racial groups.

```
#wc_LA <- read_excel("C:/Users/anaes/Desktop/SDS/ch2.xls")
#bc_LA <- read_excel("C:/Users/anaes/Desktop/SDS/ch3.xls")
#ac_LA <- read_excel("C:/Users/anaes/Desktop/SDS/ch1.xls")
#household_income <- read_excel("C:/Users/anaes/Desktop/SDS/tableA2.xlsx")



#getting rid of unneccesary rows
#bc_LA <- bc_LA %>%
#slice(9:65)

#renaming columns
#bc_LA <- bc_LA %>%
#rename("year" = `Table with row headers in column A, and column headers in rows 10 t

#omitting NAs and extra columns
#bc_LA <-  bc_LA%>%
#select(-`...4`, -`...8`)%>%
#slice(4:57)

#getting rid of unneccesary rows
#ac_LA <- ac_LA %>%
#slice(9:65)

#renaming columns
#ac_LA <- ac_LA %>%
#rename("year" = `Table with row headers in column A, and column headers in rows 10 t

#omitting NAs and extra columns
#ac_LA <-  ac_LA%>%
#select(-`...4`, -`...8`)%>%
#slice(4:57)


#getting rid of unneccesary rows
#wc_LA <- wc_LA %>%
#slice(9:65)

#renaming columns
#wc_LA <- wc_LA %>%
#rename("year" = `Table with row headers in column A, and column headers in rows 10 t

#omitting NAs and extra columns
#wc_LA <-  wc_LA%>%
#select(-`...4`, -`...8`)%>%
#slice(4:57)
```

```
#renaming columns
#household_income <- household_income %>%
#rename("year" = `Table with row headers in column A and column headers in rows 4 thr

#getting rid of columns
#all_races_hi <- household_income %>%
#select(-`...2`,-`...3`, -`...14`,-`...16`)%>%
#slice(5:58)

#ommitting NAs
#all_races_hi <- na.omit(all_races_hi)

#getting rid of columns
#white_alone_hi <- household_income %>%
#select(-`...2`,-`...3`, -`...14`,-`...16`)%>%
#slice(59:113)

#ommitting NAs
#white_alone_hi <- na.omit(white_alone_hi)

#getting rid of columns
#black_hi <- household_income %>%
#select(-`...2`,-`...3`, -`...14`,-`...16`)%>%
#slice(187:241)

#ommitting NAs
#black_hi <- na.omit(black_hi)




#joining household incomes and living arrangements for each race by year
#all <- ac_LA %>%
#inner_join(all_races_hi,by = "year")

#white <- wc_LA %>%
  #inner_join(white_alone_hi, by = "year")

#black <- bc_LA %>%
#  inner_join(black_hi, by = "year")


#temp <- all %>%
#  inner_join(white, by = "year")

#temp2 <- temp %>%
#  inner_join(black, by = "year")
```

```
#creating empty dataset to combine all three datasets
#final_set <- setNames(data.frame(matrix(ncol = 20, nrow = 156)), c("year", "children

#final_set$year <- c(temp2$year)
#final_set$children_under_18 <- c(temp2$children_under_18.x, temp2$children_under_18.
#final_set$two_parents<- c(temp2$two_parents.x, temp2$two_parents.y, temp2$per_two_pa
#final_set$one_parent <- c(temp2$one_parent.x, temp2$one_parent.y, temp2$per_one_pare
#final_set$mother_only <- c(temp2$mother_only.x, temp2$mother_only.y, temp2$mother_on
#final_set$father_only <- c(temp2$father_only.x, temp2$father_only.y, temp2$father_on
#final_set$other_relatives <- c( temp2$other_relatives.x, temp2$other_relatives.y, te
#final_set$non_relatives <- c(temp2$non_relatives.x, temp2$non_relatives.y, temp2$non


#final_set$under_15k <- c(temp2$under_15k.x, temp2$under_15k.y, temp2$under_15k)
#final_set$`15_25k` <- c(temp2$`15_to_25k.x`, temp2$`15_to_25k.y`, temp2$`15_to_25k`)
#final_set$`25_35k` <- c(temp2$`25_to_35k.x`, temp2$`25_to_35k.y`, temp2$`25_to_35k`)
#final_set$`35_50k` <- c(temp2$`35_to_50k.x`, temp2$`35_to_50k.y`, temp2$`35_to_50k`)
#final_set$`50_75k` <- c(temp2$`50_to_75k.x`, temp2$`50_to_75k.y`, temp2$`50_to_75k`)
#final_set$`75_100k` <- c(temp2$`75_to_100k.x`, temp2$`75_to_100k.y`, temp2$`75_to_10
#final_set$`100_150k` <- c(temp2$`100_to_150k.x`, temp2$`100_to_150k.y`, temp2$`100_t
#final_set$`150_200k` <- c(temp2$`150_to_200k.x`, temp2$`150_to_200k.y`, temp2$`100_t
#final_set$`200_over`<- c(temp2$`200_over.x`, temp2$`200_over.y`, temp2$`200_over`)
#final_set$median <- c(temp2$mean_income.x, temp2$mean_income.y, temp2$mean_income)
#final_set$mean_income <- c(temp2$mean_income.x, temp2$mean_income.y, temp2$mean_inco
#final_set$race <- c(rep("all", times = 52),rep("white", times = 52), rep("black", ti



#write_csv(final_set,"Project2_data.csv")
```

Exploratory Data Analysis

To explore my data I first created two new variables, “per_one_parent” and “per_two_parents”, to get the percent of children growing up in one-parent and two-parent households instead of the raw number. I felt the percentage would be more representative of the real trend since population numbers are going up in general. After finding the basic information about my dataset, like number of rows and columns, I proceeded to calculate the mean, standard deviation, quantile values, min and max for several of the numeric variables. I then produced a correlation matrix between the numeric variables mean_income, per_one_parent, per_two_parents, under_15k, and 200k_over. To my surprise, there was a negative correlation between two-parent households and mean income, even more so than for one-parent households. Realizing this could be the effect of

other variables at play, I produced various scatterplots to visualize how mean income and household arrangements changed through time and how these were different between groups as well as visualize the general trends between one-parent and two-parent households.Both one parent households and mean income increased since 1968 while two parent households decreased since 1968.

```
#importing dataset
Project_2 <- read.csv("C:/Users/anaes/Desktop/SDS/Project2_data.csv")

Project_2 <- Project_2 %>%
   mutate( per_one_parent = (one_parent/ children_under_18) *100,
           per_two_parents = (two_parents/ children_under_18)*100)


Project_2 %>%
  summarize('number of rows' = n(),
            'number of colums' = ncol(Project_2),
            'number of categories' = n_distinct(race))%>%
kbl() %>%
  kable_paper("hover", full_width = F) #used to make tables
```

number of rows	number of colums	number of categories
156	22	3

```
Project_2 %>%
  summarize('Mean Income($)'= quantile(mean_income),
       'Median Income ($)'= quantile(median),
          'Two Parents'=  quantile(per_two_parents),
    'One Parent %'=  quantile(per_one_parent),
    'Two Parents %' = quantile(per_two_parents),
   'Under 15k'= quantile(under_15k),
   'Over 200k' = quantile(`X200_over`)) %>%
  data.frame(Stats = c("min", "Q1", "median", "Q3", "max"))%>%
kbl() %>%
  kable_paper("hover", full_width = F)%>%
  add_header_above(c(" ", "Quantile Values" = 7))
```

	Quantile Values			
Mean.Income…	Median.Income….	Two.Parents	One.Parent..	Two.Parent
37392.00	37392.00	5.275304	4.339789	5.2753
55650.50	55650.50	6.380755	8.176975	6.3807

64576.50	64576.50	7.790775	10.688177	7.7907

Quantile Values

Mean.Income...	Median.Income...	Two.Parents	One.Parent...	Two.Parent
81605.25	81605.25	36.564783	45.836106	36.5647
101732.00	101732.00	58.828316	57.537467	58.8283

```
Project_2 %>%
  summarize('Mean Income($)'= mean(mean_income),
       'Median Income ($)'= mean(median),
          'Two Parents'=  mean(per_two_parents),
    'One Parent %'=  mean(per_one_parent),
    'Two Parents %' = mean(per_two_parents),
   'Under 15k'= mean(under_15k),
   'Over 200k' = mean(`X200_over`)) %>%
  kbl() %>%
  kable_classic_2(full_width = F)%>%
  add_header_above(c(" ", "Mean Values" = 6))
```

	Mean Values					
Mean Income($) </th> <th style="text-align:right;"> Median Income ($)	Two Parents	One Parent %	Two Parents %	Under 15k	Over 200k	
66598.29	66598.29	18.36874	22.60215	18.36874	14.42436	3.473

```
Project_2 %>%
  summarize('Mean Income($)'= sd(mean_income),
```

```
        'Median Income ($)'= sd(median),
           'Two Parents'=  sd(per_two_parents),
     'One Parent %'=  sd(per_one_parent),
    'Two Parents %' = sd(per_two_parents),
    'Under 15k'= sd(under_15k),
    'Over 200k' = sd(`X200_over`)) %>%

  kbl() %>%
  kable_classic_2(full_width = F)%>%
  add_header_above(c(" ", "Standard Deviation" = 6))
```

	Standard Deviation					
Mean Income($) </th> <th style="text-align:right;"> Median Income ($)	Two Parents	One Parent %	Two Parents %	Under 15k	Over 200k	
15705.83	15705.83	16.60639	20.21519	16.60639	5.948189	2.537

```
Project_2 %>%
  summarize(min_mean_inc=min(mean_income),
            max_mean_inc=max(mean_income),
    min_median_inc=min(median),
     max_median_inc=max(median),
    min_under_15k=min(under_15k),
          max_under_15k =max(under_15k),
    min_X200k_over= min(`X200_over`),
          max_edu=max(`X200_over`))%>%
kbl() %>%
  kable_classic_2(full_width = F)%>%
  add_header_above(c("Median Income (%)" = 2, "Mean Income" = 2,"Under 15k"= 2, "200k
```

Median Income (%)		Mean Income		
min_mean_inc	max_mean_inc	min_median_inc	max_median_inc	m
37392	101732	37392	101732	

```
# mean of mean income
Project_2 %>%
  group_by(race)%>%
  summarise(mean(mean_income))
```

```
## # A tibble: 3 x 2
##   race  `mean(mean_income)`
## * <chr>               <dbl>
## 1 all                73708.
## 2 black              49400.
## 3 white              76687.
```

```
#sd of mean income
Project_2 %>%
  group_by(race)%>%
  summarise(sd(mean_income))
```

```
## # A tibble: 3 x 2
##   race  `sd(mean_income)`
## * <chr>             <dbl>
## 1 all              10612.
## 2 black             7586.
## 3 white            11075.
```

```
#mean under 15k
Project_2 %>%
  group_by(race)%>%
  summarise(mean(under_15k))
```

```
## # A tibble: 3 x 2
##   race  `mean(under_15k)`
## * <chr>             <dbl>
## 1 all               11.2
## 2 black             22.4
## 3 white              9.69
```

```
#sd of under 15l
Project_2 %>%
  group_by(race)%>%
  summarise(sd(under_15k))
```

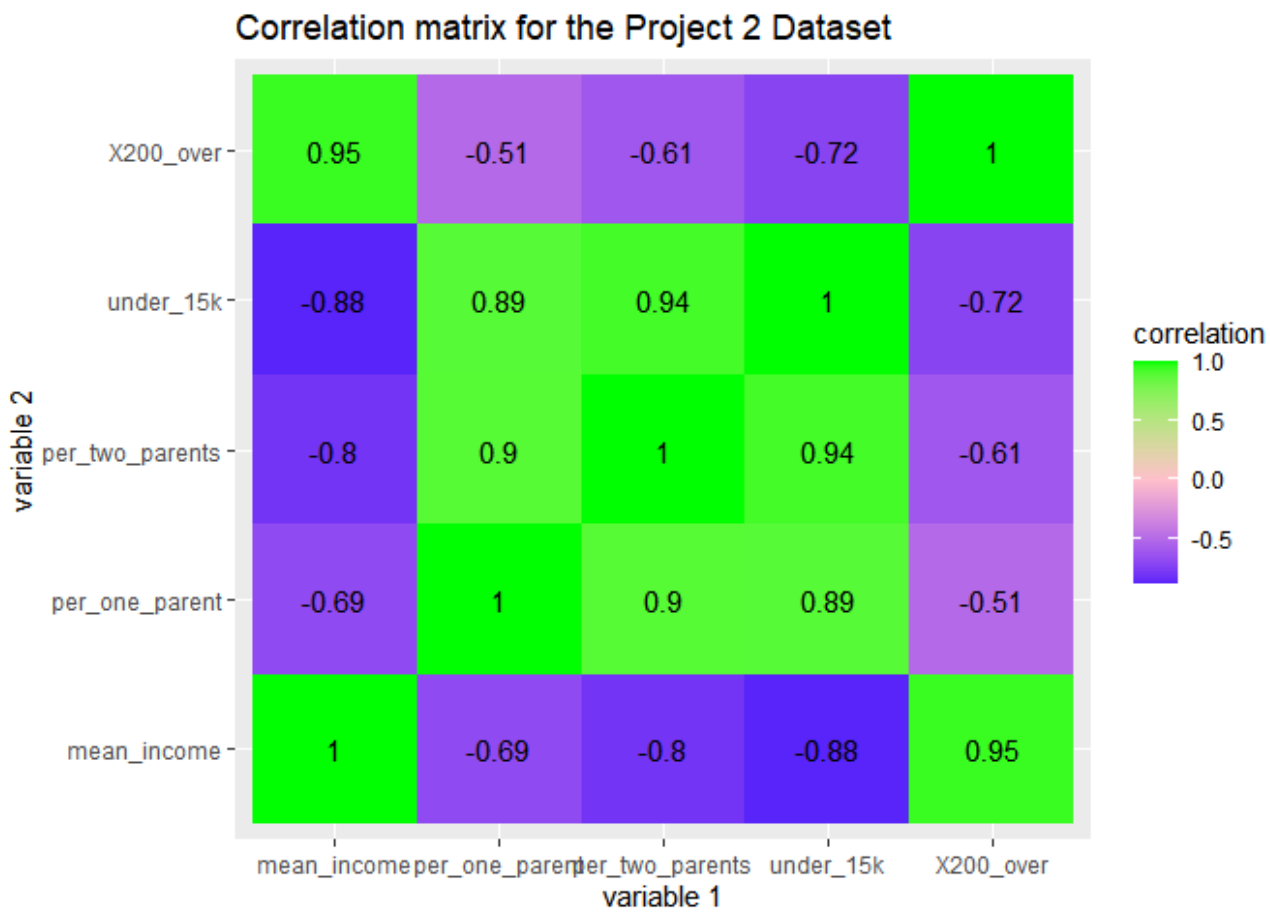
```
## # A tibble: 3 x 2
##   race  `sd(under_15k)`
## * <chr>           <dbl>
## 1 all              1.14
## 2 black            2.72
## 3 white            1.17
```

```
final_num <- Project_2 %>%  #mean income
  select(per_one_parent, per_two_parents, mean_income, X200_over, under_15k)
```

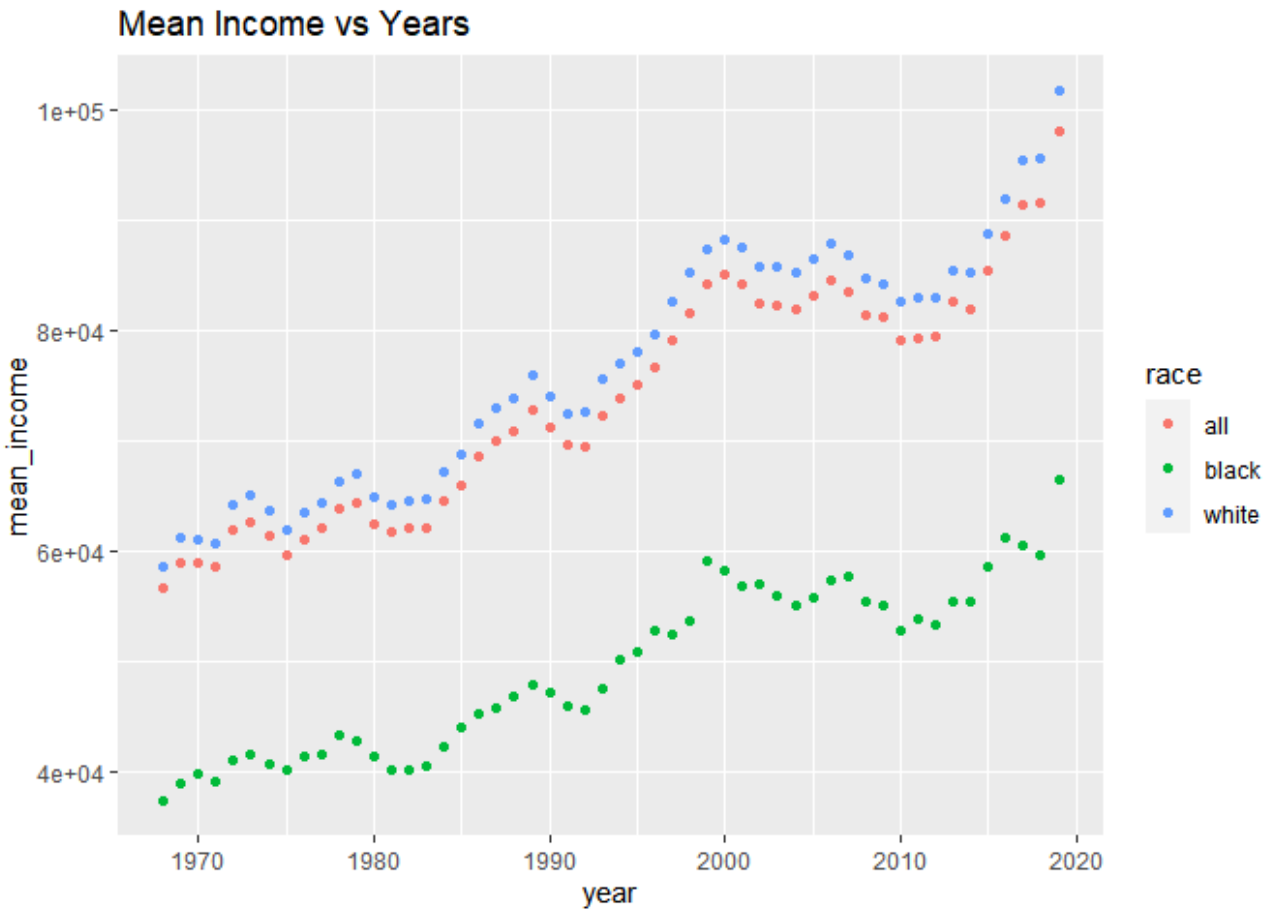
```
cor(final_num, use = "pairwise.complete.obs")
```

```
##                 per_one_parent per_two_parents mean_income  X200_over
## per_one_parent       1.0000000       0.8969573  -0.6866667 -0.5060940
## per_two_parents      0.8969573       1.0000000  -0.8036588 -0.6121766
## mean_income         -0.6866667      -0.8036588   1.0000000  0.9533679
## X200_over           -0.5060940      -0.6121766   0.9533679  1.0000000
## under_15k            0.8936236       0.9371151  -0.8842257 -0.7158536
##                  under_15k
## per_one_parent   0.8936236
## per_two_parents  0.9371151
## mean_income     -0.8842257
## X200_over       -0.7158536
## under_15k        1.0000000
```

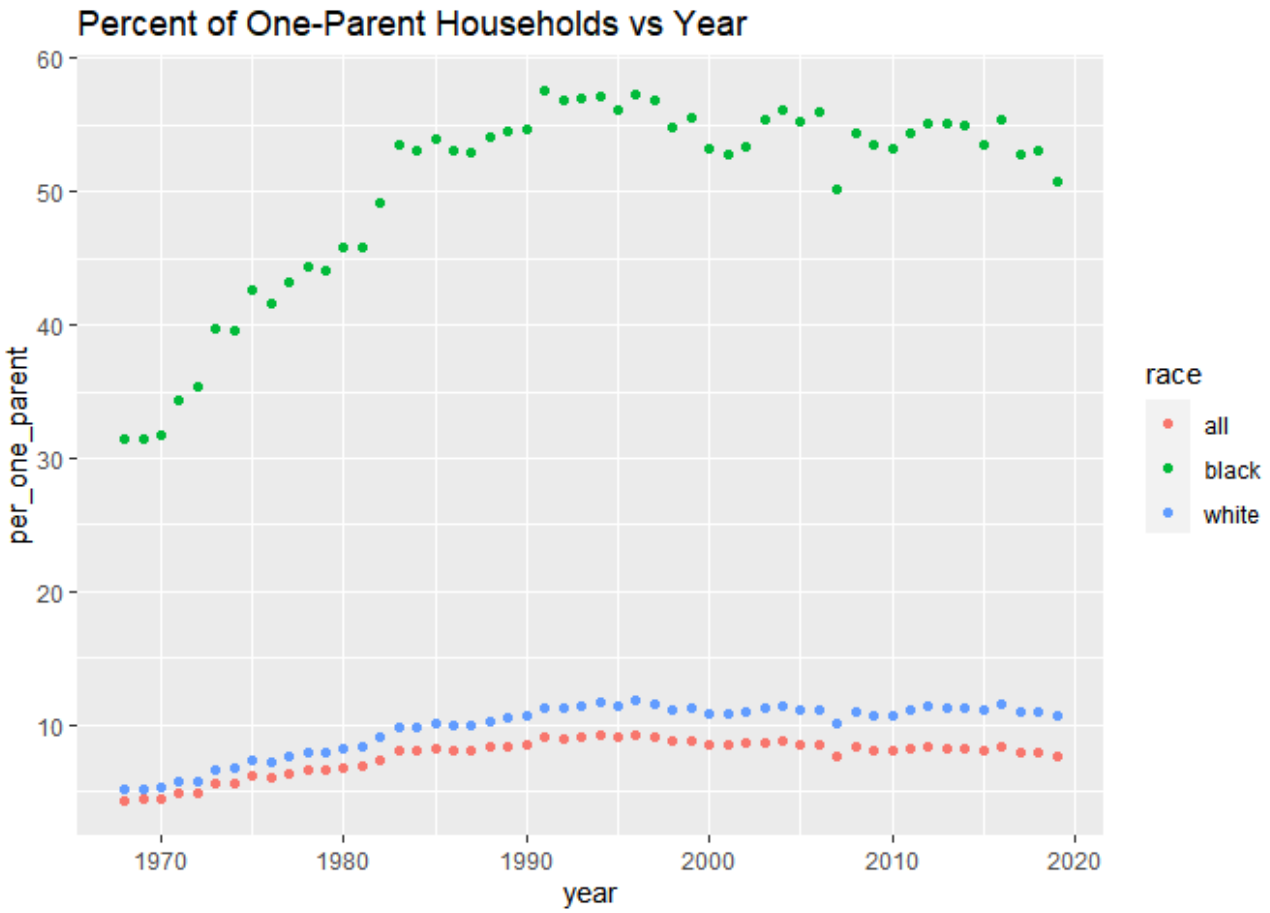
```
cor(final_num, use = "pairwise.complete.obs") %>%
  # Save as a data frame
  as.data.frame %>%
  # Convert row names to an explicit variable
  rownames_to_column %>%
  # Pivot so that all correlations appear in the same column
  pivot_longer(-1, names_to = "other_var", values_to = "correlation") %>%
  ggplot(aes(rowname, other_var, fill=correlation)) +
  # Heatmap with geom_tile
  geom_tile() +
  # Change the scale to make the middle appear neutral
  scale_fill_gradient2(low="blue",mid="pink",high="green") +
  # Overlay values
  geom_text(aes(label = round(correlation,2)), color = "black", size = 4) +
  # Give title and labels
  labs(title = "Correlation matrix for the Project 2 Dataset", x = "variable 1", y =
```



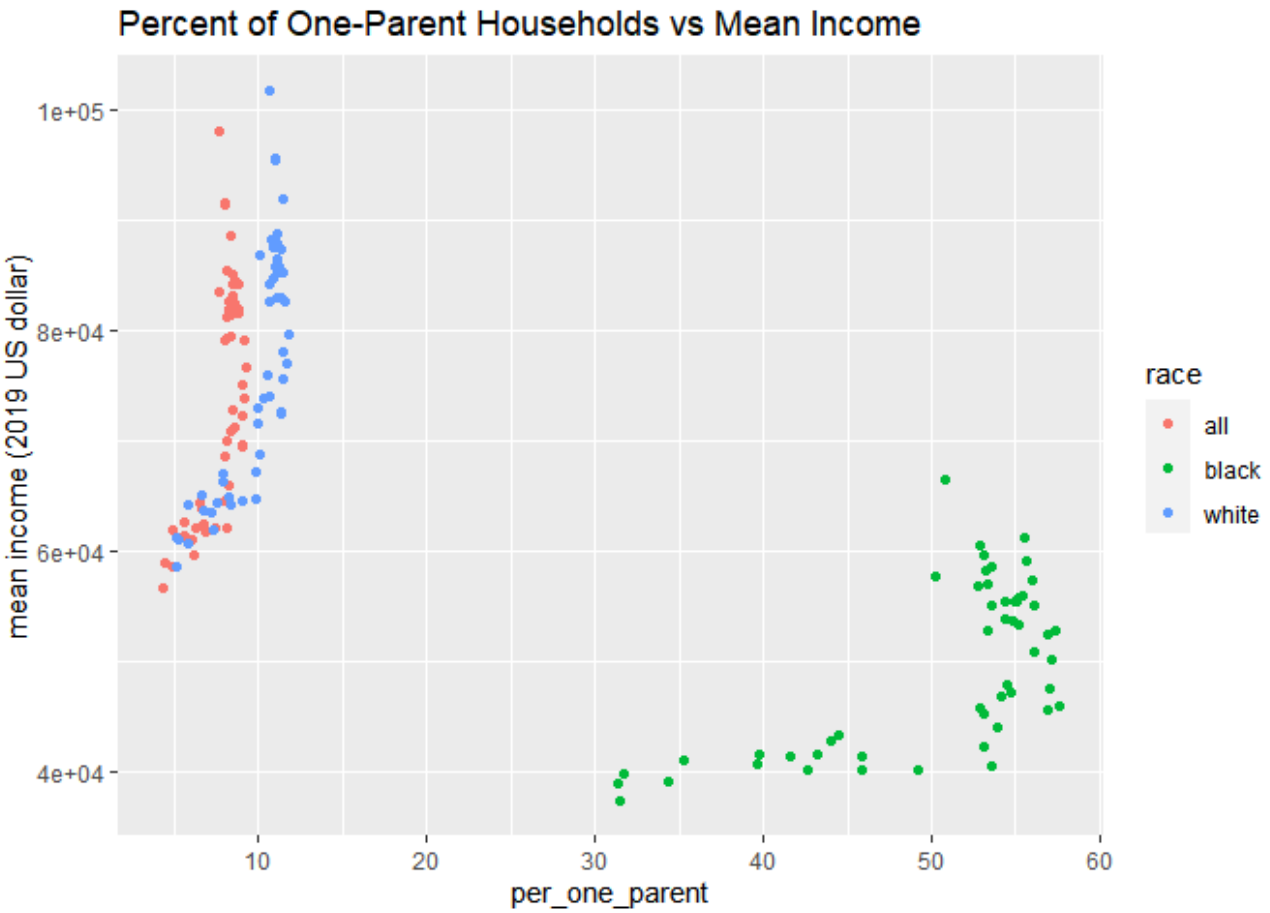

```
ggplot(Project_2, aes(x = year, y = mean_income, color = race)) + geom_point() + ggti
```




```
ggplot(Project_2, aes(x = year, y = per_one_parent, color = race)) + geom_point() + g
```




```
ggplot(Project_2, aes(x = per_one_parent, y= mean_income, color = race)) + geom_point
```



MANOVA

The next step in this analysis was to conduct a multivariate analysis of variances (MANOVA. This compares the numeric responses "mean income" and "percent of mother only households" between our different groups - white and black racial categories. The null hypothesis is that the mean of each of these two response variables does not differ between groups. The alternative hypothesis is that for at least one response variable, at least one of the group mean will differ. First the assumptions of a MANOVA, which are numerous, were visualized. The assumptions of homogeneity of within group variance, multivariate normality, random sample, and no colliniearity were violated

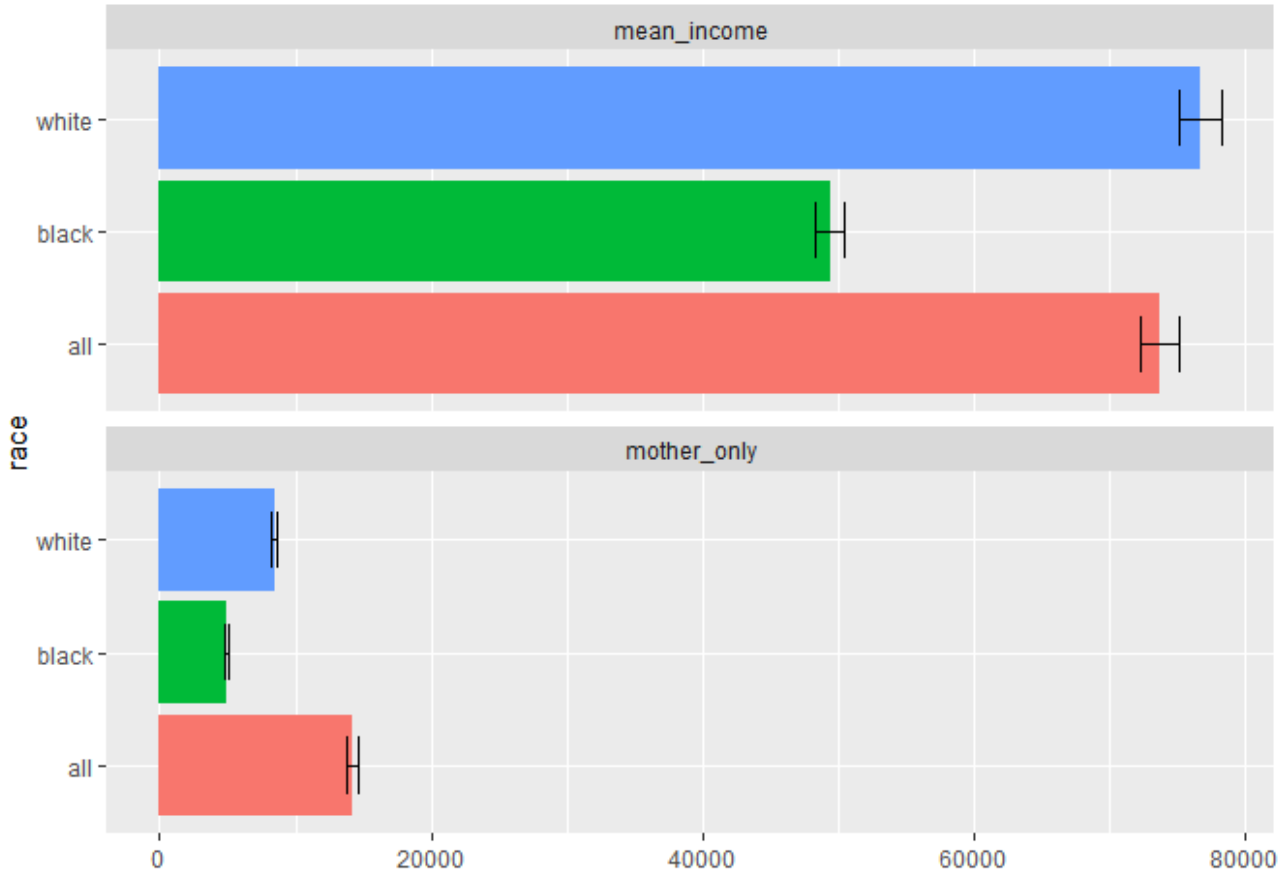
A one-way MANOVA was conducted to determine the effect of the three categories on two dependent variables (mean income and percentage of mother only households). Significant differences were found among the three groups for at least one of the dependent variables (Pillai's trace = 0.96, pseudo F= 101, p< .0001).

Univariate ANOVAs for each dependent variable were conducted as follow-up tests to the MANOVA, were also significant for mean income (F = 214.85, p< .0001) and percent

mother only (F= 922.73 , p< 0.0001).Post hoc analysis was performed conducting pairwise comparisons. All three groups were significantly different for the "percent mother only" variable and "mean_income" variable as expected from the ANOVA tests with p values < 0.0001. Adjusting for multiple comparisons (Bonferroni a = 0.0166), did not change the results. Given that there wre 9 total tests, the probability of a type 1 error, a false positive, is 0.142625.

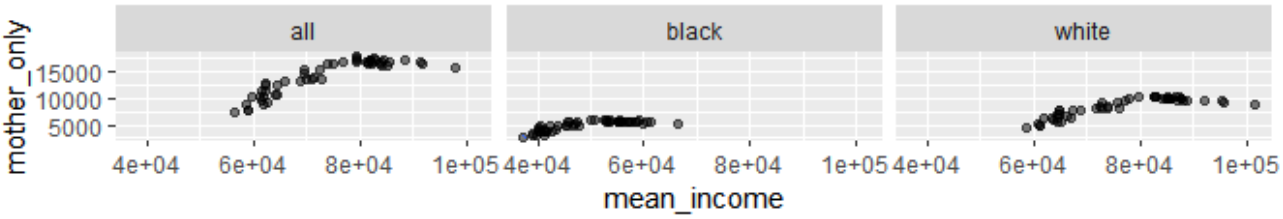
####Contains revisions based on comments for project 2 (only comparing two groups instead of three now)

```
Project_2 <- Project_2 %>%
  mutate( per_mother_only = (mother_only/children_under_18)*100)

#Project_2 %>%
  #group_by(race) %>%
 # summarize(mean(mean_income), mean(mother_only))

# Represent the means per race
Project_2 %>%
  select(race,mean_income,mother_only) %>%
  pivot_longer(-1,names_to='DV', values_to='measure') %>%
  ggplot(aes(race,measure,fill=race)) +
  geom_bar(stat="summary", fun = "mean") +
  geom_errorbar(stat="summary", fun.data = "mean_se", width=.5) +
  facet_wrap(~DV, nrow=2) +
  coord_flip() +
  ylab("") +
  theme(legend.position = "none")
```




```
# Inspect multivariate plots of response variable for each race
ggplot(Project_2, aes(x = mean_income, y = mother_only)) +
  geom_point(alpha = .5) +
  geom_density_2d(h=2) +
  coord_fixed() +
  facet_wrap(~race)
```

```
## Warning: stat_contour(): Zero contours were generated

## Warning in min(x): no non-missing arguments to min; returning Inf

## Warning in max(x): no non-missing arguments to max; returning -Inf

## Warning: stat_contour(): Zero contours were generated

## Warning in min(x): no non-missing arguments to min; returning Inf

## Warning in max(x): no non-missing arguments to max; returning -Inf
```




```
p2_for_man <- Project_2 %>%
  slice(53:156)


#performing
manova_p2 <- manova(cbind(mean_income, per_mother_only) ~ race, data = p2_for_man)


#get a summary
summary(manova_p2)
```

```
##            Df  Pillai approx F num Df den Df    Pr(>F)
## race        1 0.96224     1287      2    101 < 2.2e-16 ***
## Residuals 102
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#perform ANOVA on each numeric response
summary.aov(manova_p2)
```

```
##  Response mean_income :
##              Df     Sum Sq    Mean Sq F value    Pr(>F)
## race          1 1.9358e+10 1.9358e+10  214.85 < 2.2e-16 ***
## Residuals   102 9.1903e+09 9.0101e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response per_mother_only :
##              Df Sum Sq Mean Sq F value    Pr(>F)
## race          1  25819   25819  922.73 < 2.2e-16 ***
## Residuals   102   2854      28
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
 #perform t-tests between groups
  pairwise.t.test(p2_for_man$mean_income, p2_for_man$race, p.adj ="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  p2_for_man$mean_income and p2_for_man$race
##
##       black
## white <2e-16
##
## P value adjustment method: none
```

```
  pairwise.t.test(p2_for_man$per_mother_only, p2_for_man$race, p.adj="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  p2_for_man$per_mother_only and p2_for_man$race
##
##       black
## white <2e-16
##
## P value adjustment method: none
```

```
# Type 1 error 1- 0.95^(number of tests)
1-0.95^3
```

```
## [1] 0.142625
```

```
#bonferroni
0.05/3
```

```
## [1] 0.01666667
```

Randomization Test

*Given that the data did not satisfy the necessary MANOVA assumptions, I conducted a PERMANOVA which would lessen the restrictions by scrambling the relationship between variables and creating a null distribution. The null hypothesis would be that “mean income” and “percent of mother only households” does not differ between the “black”, “white”, or “all [races]” groups. The alternative hypothesis is that the mean for at least one of the response variables differs between groups. *

The actual observed F-statistic from the PERMANOVA was 119.36 with a p value of 0.001 which means we can reject the null hypothesis. This p-value is greater than the p value with the regular MANOVA. A histogram of the null dustribution was produced but the observed F statistic was too large to appear on the plot.

####contains revisions based on comments on project 2 (I rejected null hypothesis)

```
library(vegan)
```

```
## Warning: package 'vegan' was built under R version 4.0.5
```

```
## Loading required package: permute
```

```
## Warning: package 'permute' was built under R version 4.0.5
```

```
## Loading required package: lattice
```

```
## This is vegan 2.5-7
```

```
dists <- Project_2 %>%
  select(mean_income, per_mother_only) %>%
```

```
  dist

# Perform PERMANOVA on the distance matrix
adonis(dists ~ race, data = Project_2)
```

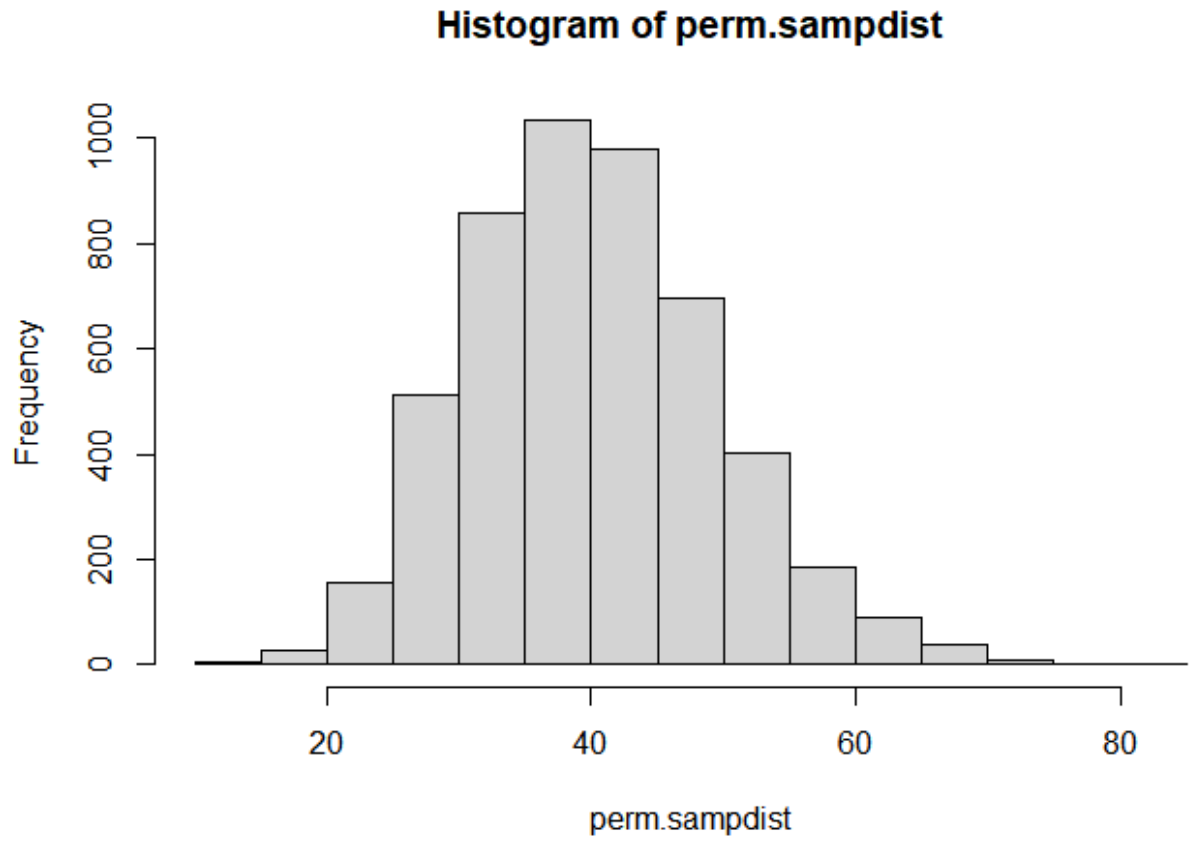
```
##
## Call:
## adonis(formula = dists ~ race, data = Project_2)
##
## Permutation: free
## Number of permutations: 999
##
## Terms added sequentially (first to last)
##
##            Df  SumsOfSqs    MeanSqs F.Model      R2 Pr(>F)
## race        2 2.3301e+10 1.1650e+10  119.36 0.60942  0.001 ***
## Residuals 153 1.4933e+10 9.7604e+07         0.39058
## Total     155 3.8234e+10                    1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Fstat = 119.36
```

```
set.seed(348)
perm.sampdist<-replicate(5000,{

randp2<-Project_2
randp2$race<-sample(Project_2$race)

euc_dist<-dist(randp2[,c("mean_income","per_mother_only")],method="euclid")
euc_dist_white<-dist(randp2[randp2$race=="white",c("mean_income","per_mother_only")],
euc_dist_black<-dist(randp2[randp2$race=="black",c("mean_income","per_mother_only")],
euc_dist_all<-dist(Project_2[Project_2$race=="all",c("mean_income","per_mother_only")

SSR<-sum(euc_dist_white^2)/53+sum(euc_dist_black^2)/53 + sum(euc_dist_all^2)/53
SST<-sum(euc_dist^2)/156

(SST-SSR)/(SSR/153)
}
)

mean(perm.sampdist>Fstat)
```

```
## [1] 0
```

```
hist(perm.sampdist, breaks=20); abline(v = Fstat,col='red')
```



Multi-factor Linear Regression

The next part of this experiment was to perform a multi-factor linear regression that would predict the variable of "mean income" from "race", "per_one_parent", and "year" and the interaction between "race" and "per_one_parent". I mean centered and scaled my variables to get more descriptive results. The assumption of independent values and equal variance were violated but the assumptions of normality of residuals and linearity were met.

The null hypothesis for this test would be that all of the estimates are equal to zero and thus have no impact on the response variable. The alternative hypothesis is that the estimates are not equal to zero. Each coefficient represents the change in the mean of the response variable "mean income" per unit increase in the associated variable when all the other predictors are at their average. When controlling year and household arrangement, the

mean income of the category "black" is -23955 less than for "all" races. For the category "white", the mean income is 3416 more than for "all" races.

*When controlling for year and race, the mean income for one-parent households is 8776 more than for two parent household. The interaction for race and "per_one_parent" was The interaction between race:black and one parent household was -13754, which means that the variables negatively alter each other's slope. The intercept, 80189, is the mean income of a person from the "all races" category when per_one_parent_c is at its average and the year is 1993. According to the adjusted r-squared value, my model accounts for 95% of the variation in response. The "race:black" group and the "year_c" groups were significant predictors with p values less than 0.001. *

An interaction tests if the slope of one variable affects the slope of another variable. A graph was made to represent the interaction between race and percent one-parent households on mean income. One can see by the graph that the positive relationship between "percent one-parent" households and "mean income" is weakest when it comes to the "black" category and strongest when it comes to the "all" category (the steepest slope). A second graph was made to visualize the interaction between years and "percent one-parent" households on mean income by race. One can see that the slope of each regression line is almost the same with the starting y-intercepts being different, suggesting that the mean income has increased at rought the same rate for each category from 1968 to 2019. As mentioned above, the signficant predictor variables were racial category "black" and "year". When calculating robust standard errors both those variables remained significant with the additional significant variable of the interaction term between "race:black" and "per_one_parent_c". I also calculated the robust standard errors which were quite different from the original fit.

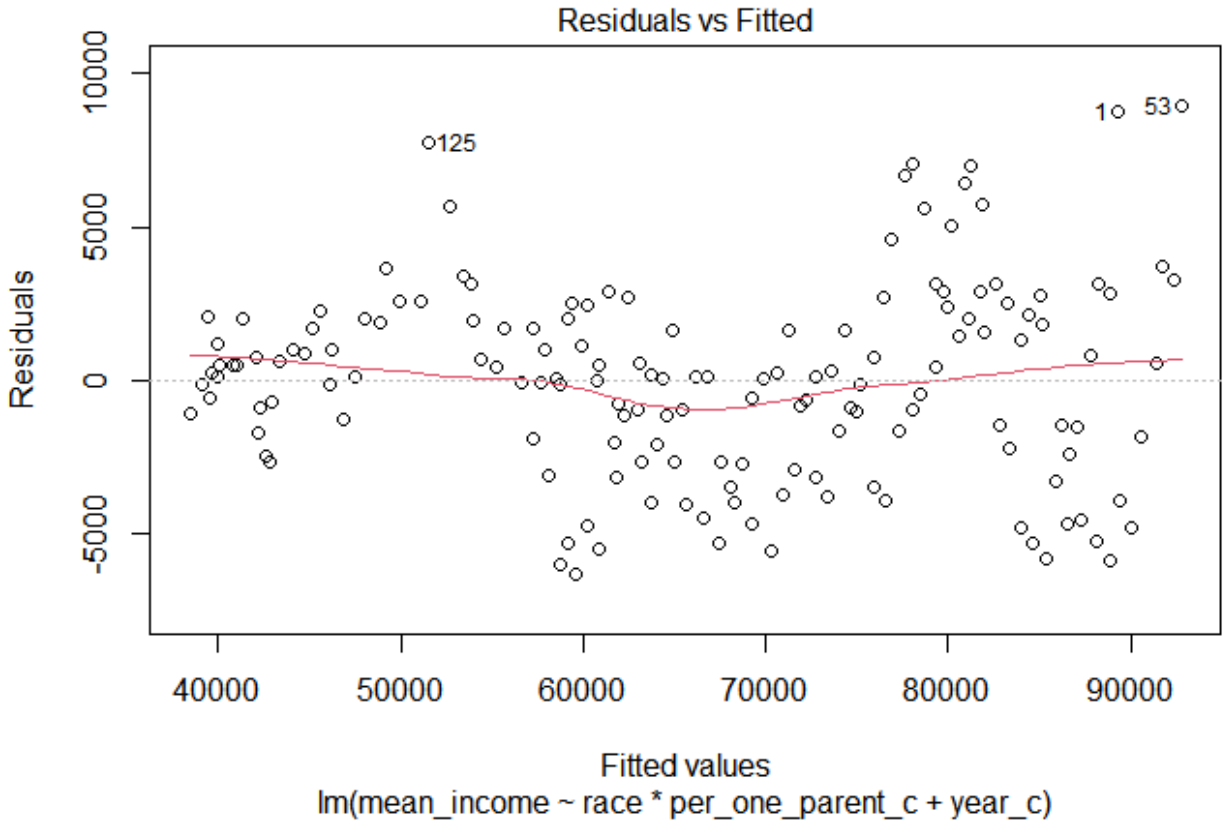
####contains revisions based on comments from project 2 (I attempted to correct my interpretations)

```
#mean centering and scaling data
p2_for_lm <- Project_2

p2_for_lm$per_one_parent_c <- scale(p2_for_lm$per_one_parent - mean(p2_for_lm$per_one

p2_for_lm$per_two_parents_c <- scale(p2_for_lm$per_two_parents - mean(p2_for_lm$per_t

p2_for_lm$per_mother_only_c <- scale(p2_for_lm$per_mother_only - mean(p2_for_lm$per_m

p2_for_lm$under_15k_c <- scale(p2_for_lm$under_15k - mean(p2_for_lm$under_15k))

p2_for_lm$X200_over_c <- scale(p2_for_lm$X200_over - mean(p2_for_lm$X200_over))
```

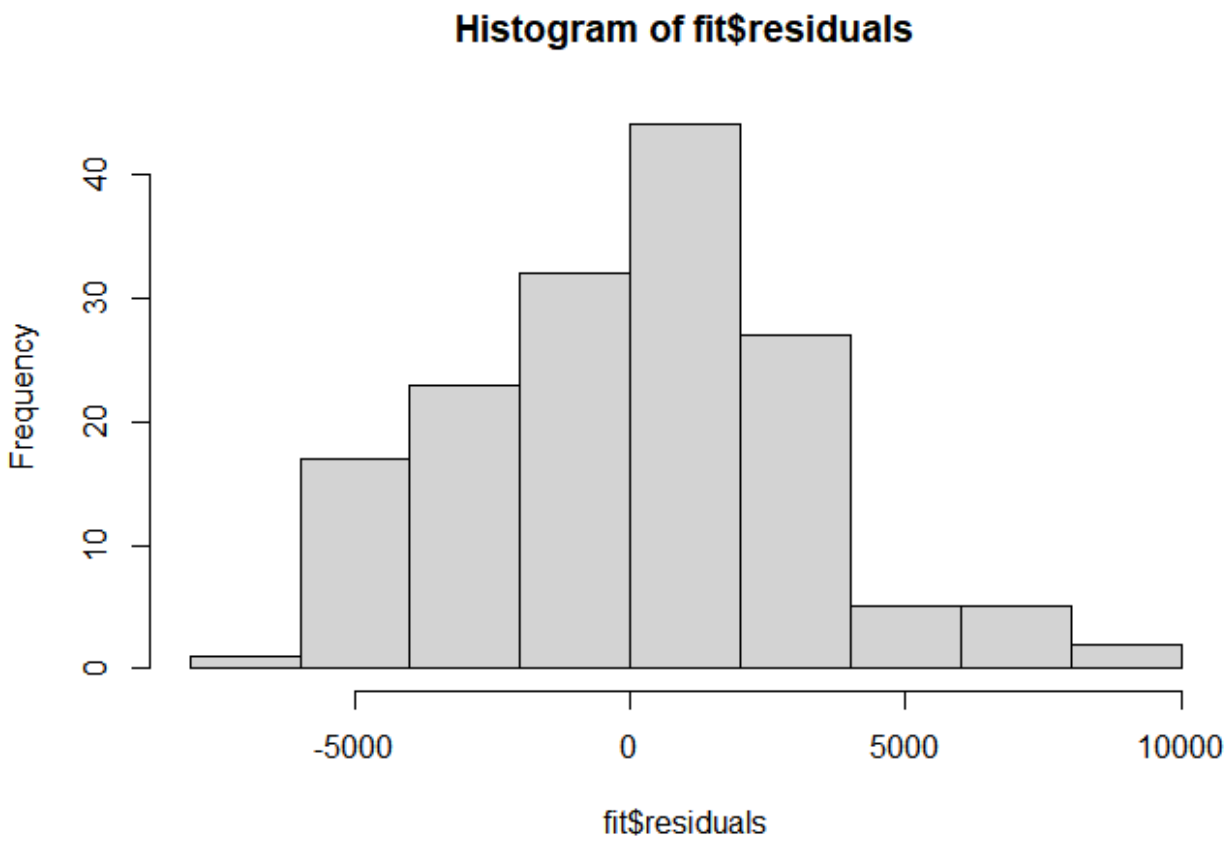
```r
p2_for_lm$year_c <- scale(p2_for_lm$year - mean(p2_for_lm$year))



#conduction multi linear regression
fit <- lm(mean_income ~ race * per_one_parent_c + year_c ,data= p2_for_lm)

summary(fit)
```

```
##
## Call:
## lm(formula = mean_income ~ race * per_one_parent_c + year_c,
##     data = p2_for_lm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6295.3 -2128.9   114.1  1996.7  8958.7
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    80189       5808  13.806  < 2e-16 ***
## raceblack                     -23955       6670  -3.591 0.000446 ***
## racewhite                       3416       5826   0.586 0.558502
## per_one_parent_c                8776       7841   1.119 0.264865
## year_c                          9223        394  23.411  < 2e-16 ***
## raceblack:per_one_parent_c    -13754       7557  -1.820 0.070774 .
## racewhite:per_one_parent_c      2130       8123   0.262 0.793553
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3216 on 149 degrees of freedom
## Multiple R-squared:  0.9597, Adjusted R-squared:  0.9581
## F-statistic: 591.2 on 6 and 149 DF,  p-value: < 2.2e-16
```

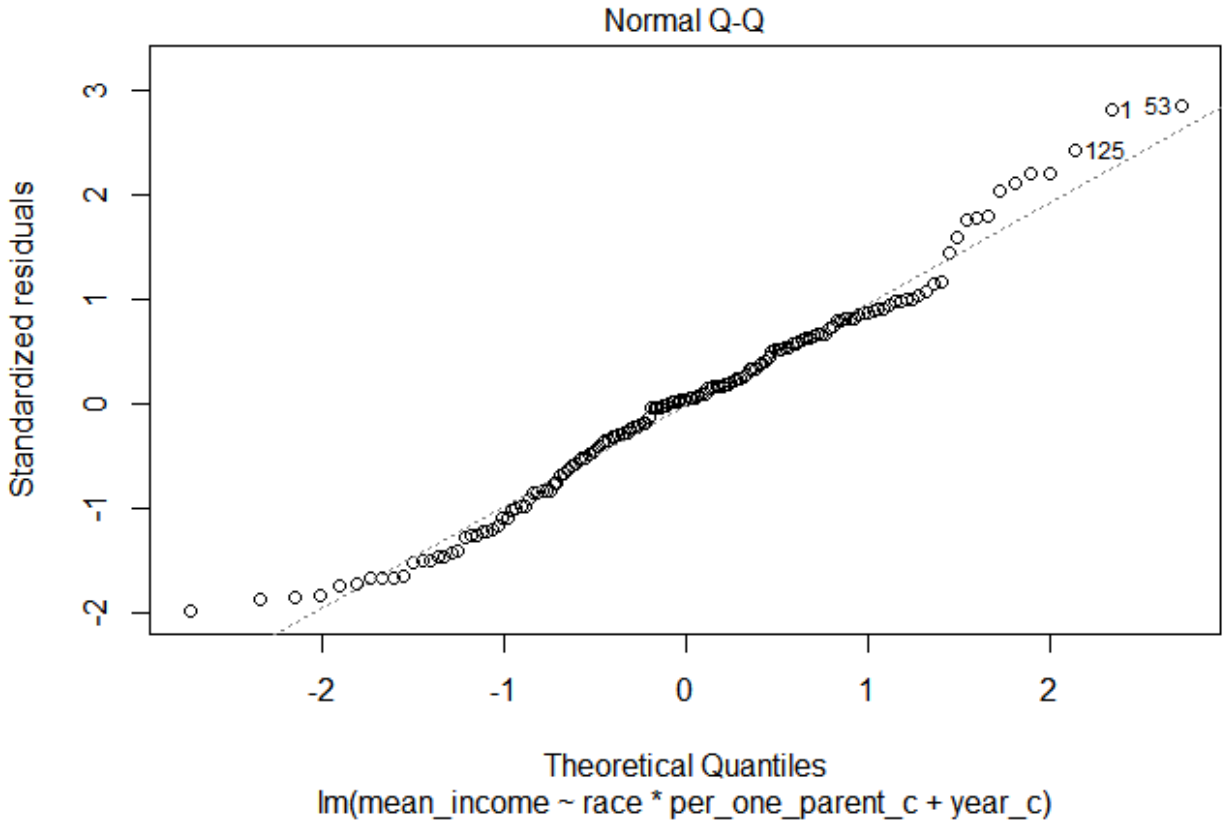
```r
# Residuals vs Fitted values plot
plot(fit, which = 1)
```




```
# Histogram of residuals
hist(fit$residuals)
```

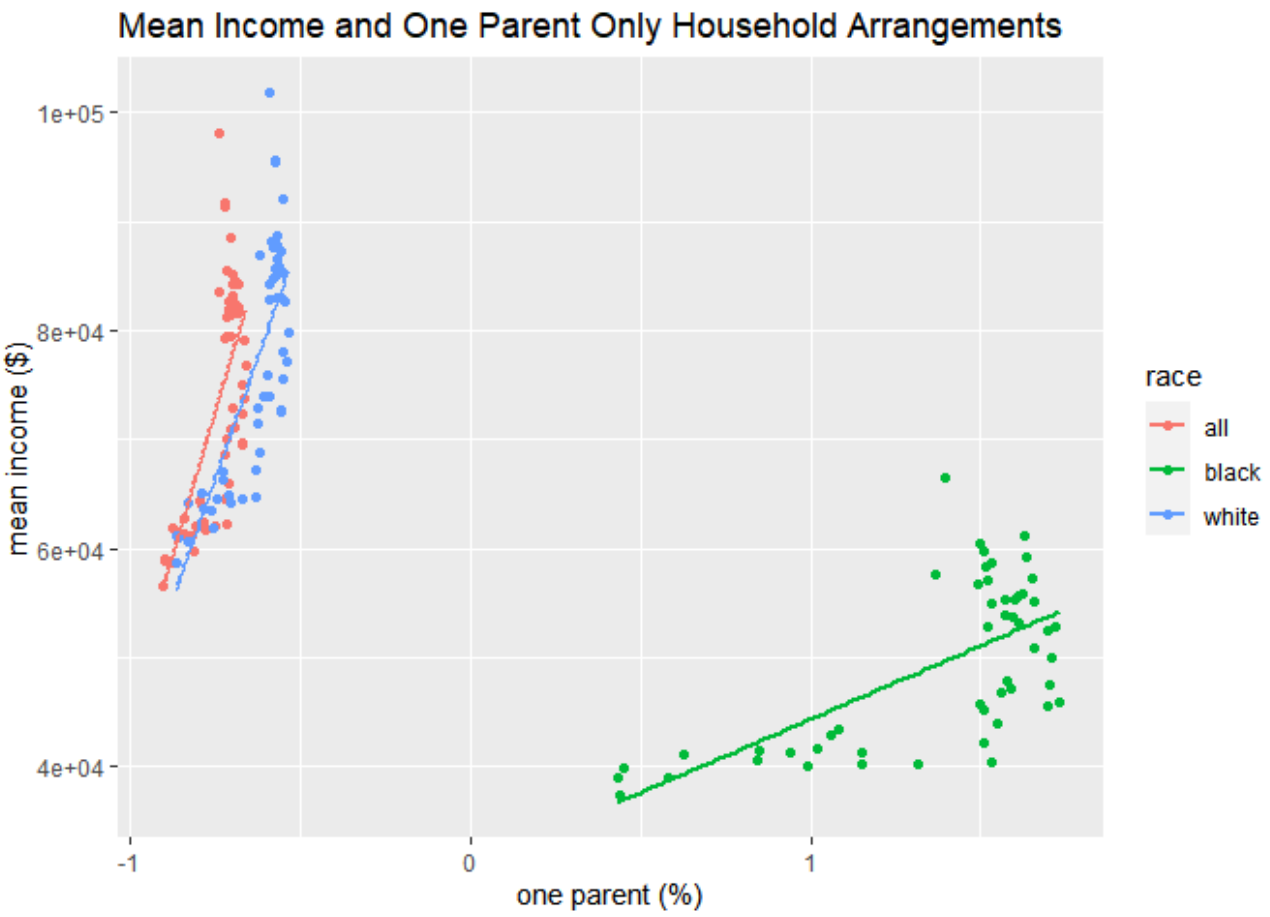



```
# Q-Q plot for the residuals
plot(fit, which = 2)
```



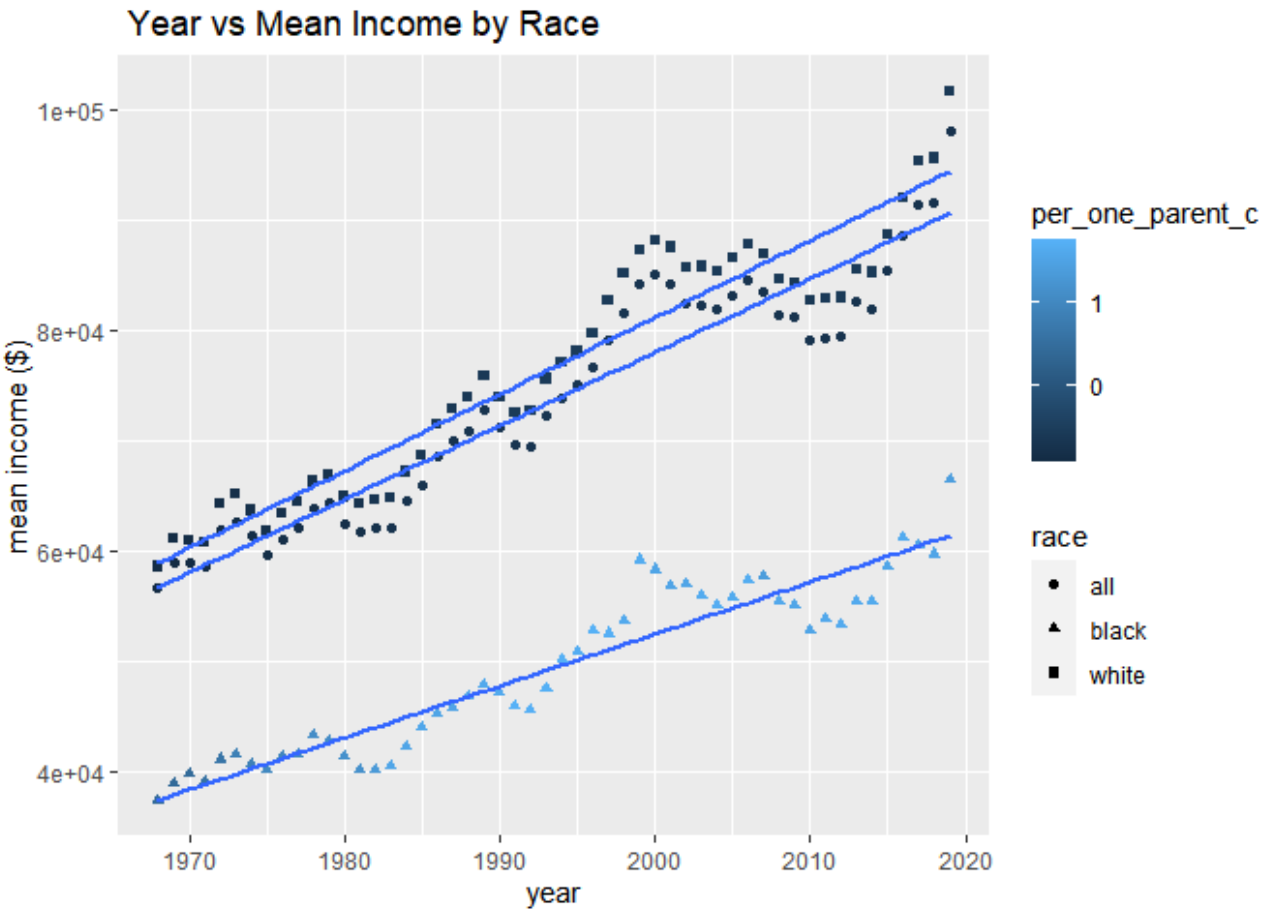

```
#plot visualizing linear regressions
ggplot(p2_for_lm, aes(x = per_one_parent_c, y = mean_income, color = race), se=FALSE)
  geom_point() +
  geom_smooth(method = lm, se = FALSE) +
  labs(title ="Mean Income and One Parent Only Household Arrangements ",
       x = "one parent (%)", y = "mean income ($)")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
#plot visualizing linear regressions
ggplot(p2_for_lm, aes(x = year, y = mean_income, color = per_one_parent_c, shape = ra
  geom_point() +
  geom_smooth(method = lm, se = FALSE) +
  labs(title =" Year vs Mean Income by Race ",
       x = "year", y = "mean income ($)")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
#robust SEs
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 4.0.5
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 4.0.4
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(sandwich)
```

```
## Warning: package 'sandwich' was built under R version 4.0.5
```

```
coeftest(fit, vcov = vcovHC(fit))
```

```
##
## t test of coefficients:
##
##                             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                  80189.2     5059.7 15.8487 < 2.2e-16 ***
## raceblack                   -23954.5     6010.6 -3.9854 0.0001051 ***
## racewhite                     3416.5     4684.5  0.7293 0.4669532
## per_one_parent_c              8775.9     6742.4  1.3016 0.1950660
## year_c                        9223.2      480.6 19.1909 < 2.2e-16 ***
## raceblack:per_one_parent_c -13753.7     6215.2 -2.2129 0.0284238 *
## racewhite:per_one_parent_c   2129.7     6047.5  0.3522 0.7252182
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Repeat bootstrapping 5000 times, saving the coefficients each time
samp_SEs <- replicate(5000, {
  # Bootstrap your data (resample observations)
  boot_data <- sample_frac(p2_for_lm, replace = TRUE)
  # Fit regression model
  fitboot <- lm(mean_income ~ race + year * per_one_parent_c, data = boot_data)
  # Save the coefficients
  coef(fitboot)
})
```

```
# Estimated SEs
samp_SEs %>%
  # Transpose the obtained matrices
  t %>%
  # Consider the matrix as a data frame
  as.data.frame %>%
  # Compute the standard error (standard deviation of the sampling distribution)
  summarize_all(sd)
```

```
##   (Intercept) raceblack racewhite     year per_one_parent_c
## 1    43405.12  2100.029  649.9348 21.55429         33765.01
##   year:per_one_parent_c
## 1              16.79519
```

Logistic Regression

The final part of the project was to perform a logistic regression on a binary variable. I made a new categorical variable "dominant arrangement" that states whether or not the majority of children were living in two parent or one parent households. If the majority was two parents that would be equal to 1, if the majority was one parent it would show a 0. To interpret the coefficients, controlling for income, being in the category "black" decreases the log odds of having two parents by 8.489. The odds of having two parents is 2.058 times smallet for "black" than for "all". Being in the category "white" increases the log odds of having two parents by 1.065. The odds of having two parents is 2.902 times greater for "white" than for "all". Every one-unit increase in the percentage of people under a $15k income increases the log odds of having two parents by 4.152 and multiplies the odds by 6.36. The confusion matrix for this logistic regression predicts a two-parent household with a probability greater than 0.5. It calculated 119 true negatives, 7 false positives, 27 false negatives, and 3 true positives. The accuracy was 0.78. The sensitivity was 0.1, the specificity was 0.94, and the precision was 0.3. The darker area in the density graph of the log odds displays the clumps that have been misclassfied. Under the AUC rules of thumb, the model can be said to be "fair" givent that randomly selected household that has two parents has a higher predicted probability than a randomly selected household with one parent 79.4% of the time.

####Has revised interpretation of coefficients and includes interpretions of odds ratios. (i also added density graph and ROC curve)

```
#creating binary variable

p2_for_lm <- p2_for_lm %>%
  mutate( dominant_arrangement = case_when(per_one_parent > per_two_parents ~ "single
                                   per_two_parents > per_one_parent ~ "both_parents"))
p2_for_log <- p2_for_lm %>%
  mutate( y = case_when( dominant_arrangement == "both_parents" ~ 1 ,
                         dominant_arrangement == "single_parent" ~ 0 ))


#logistic regression
fit_log <- glm(y ~ race + under_15k_c, data = p2_for_log, family = "binomial")
summary(fit_log)
```

```
##
## Call:
## glm(formula = y ~ race + under_15k_c, family = "binomial", data = p2_for_log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

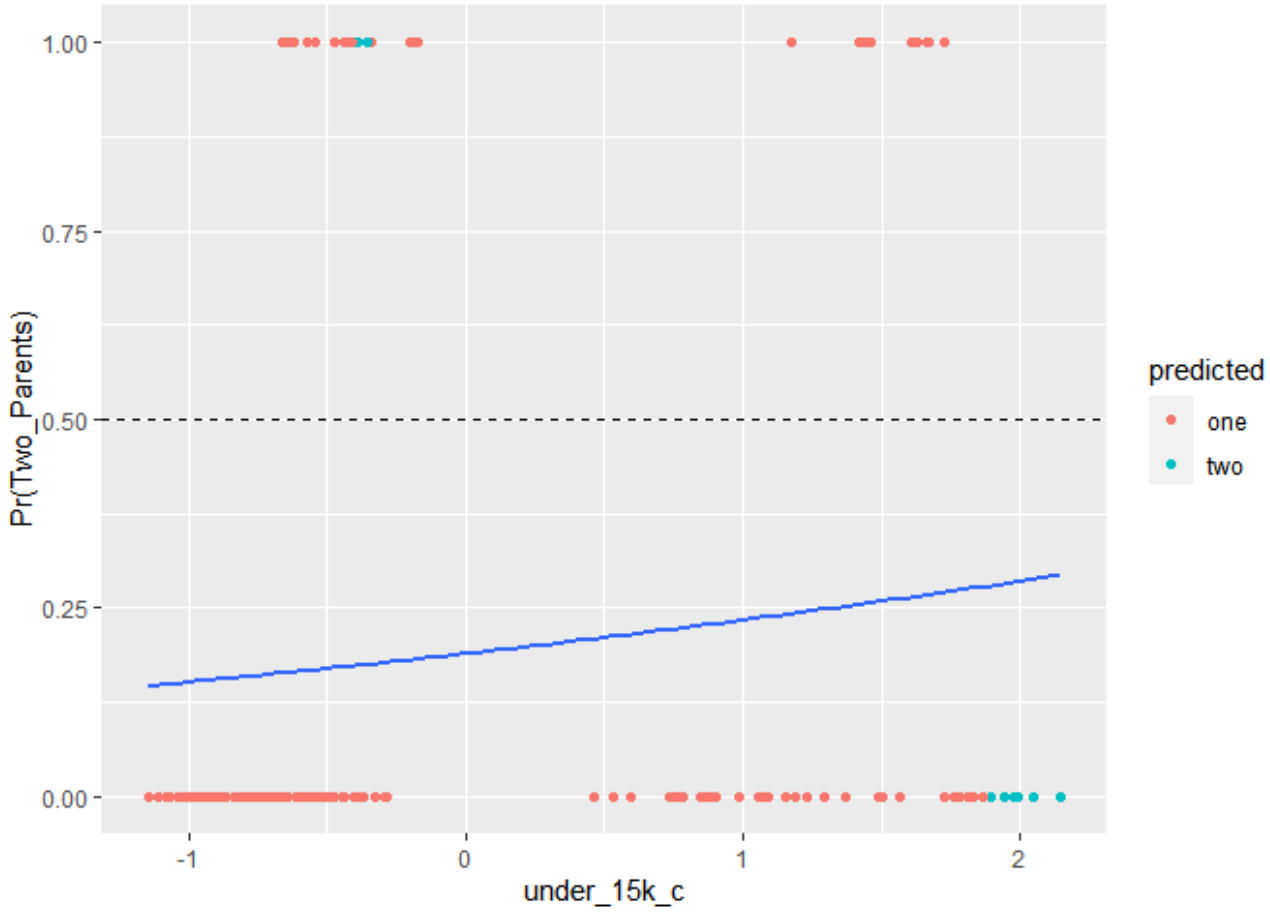
```
## -1.6420  -0.6415  -0.4396  -0.1662   2.4699
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.6168     0.5975   1.032 0.301936
## raceblack    -8.4886     2.2965  -3.696 0.000219 ***
## racewhite     1.0654     0.5858   1.819 0.068948 .
## under_15k_c   4.1526     1.0342   4.015 5.94e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 152.74  on 155  degrees of freedom
## Residual deviance: 128.67  on 152  degrees of freedom
## AIC: 136.67
##
## Number of Fisher Scoring iterations: 5
```

```
#odds
exp(coef(fit_log))
```

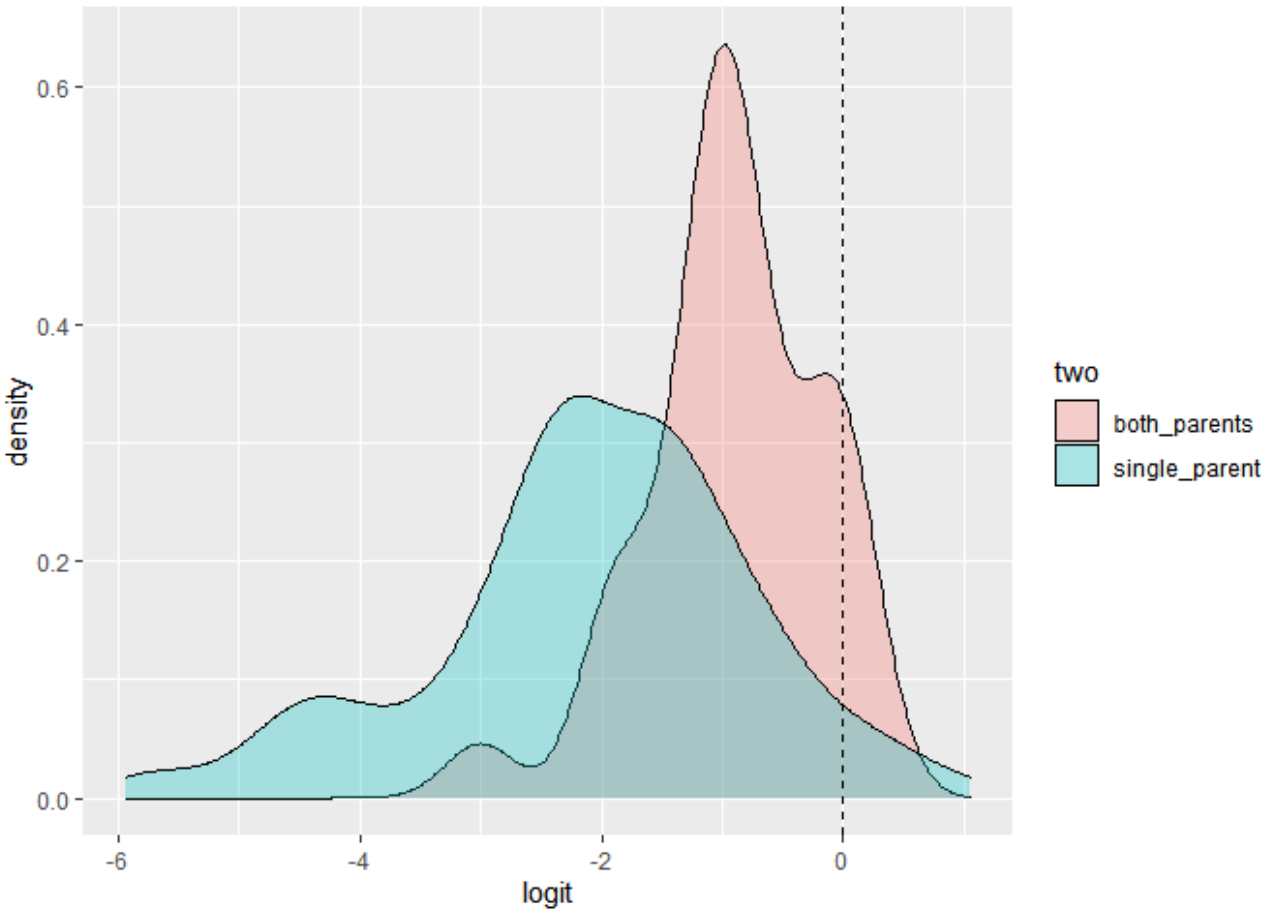
```
##  (Intercept)    raceblack    racewhite  under_15k_c
## 1.852991e+00 2.058109e-04 2.902010e+00 6.360101e+01
```

```
#Added this entire chunk for revised version

# Add predicted probabilities to the dataset

p2_for_log$prob <- predict(fit_log, type = "response")

# Predicted outcome is based on the probability of malignant
# if the probability is greater than 0.5, the clump is found to be malignant
p2_for_log$predicted <- ifelse(p2_for_log$prob > .5, "two", "one")

#plot the model
ggplot(p2_for_log, aes(under_15k_c,y)) +
  geom_jitter(aes(color = predicted), width = .005, height = 0) +
  stat_smooth(method="glm", method.args = list(family="binomial"), se = FALSE) +
  geom_hline(yintercept = 0.5, lty = 2) +
  ylab("Pr(Two_Parents)")
```

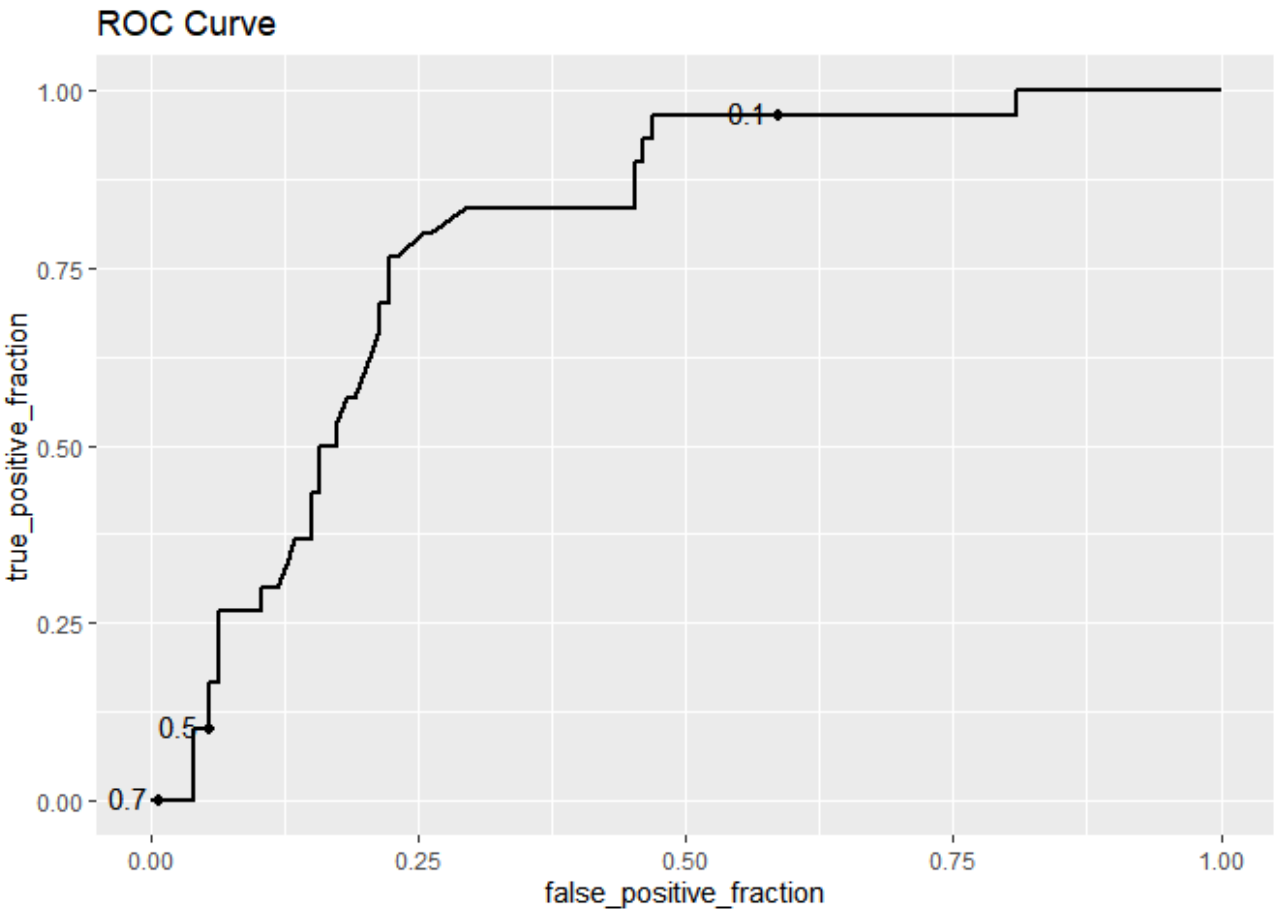
```
## `geom_smooth()` using formula 'y ~ x'
```




```
logit <- function(p) log(odds(p))
# Save the predicted log-odds in the dataset
p2_for_log$logit <- predict(fit_log)

# Compare to the outcome in the dataset with a density plot
ggplot(p2_for_log, aes(logit, fill = as.factor(dominant_arrangement))) +
  geom_density(alpha = .3) +
  geom_vline(xintercept = 0, lty = 2) +
  labs(fill = "two")
```




```
#Added this whole chunk for revied version of project

library(plotROC)


## Warning: package 'plotROC' was built under R version 4.0.4


#ROC plot
ROCplot1 <- ggplot(p2_for_log) +
  geom_roc(aes(d = y, m = prob), cutoffs.at = list(0.1, 0.5, 0.9))+ ggtitle("ROC Curv
ROCplot1
```




```
#AUC
calc_auc(ROCplot1)
```

```
##   PANEL group       AUC
## 1     1    -1 0.7940476
```

```
#storing probabilities
p2_for_log$prob <- predict(fit_log, type = "response")

#predictions
p2_for_log$predicted <- ifelse(p2_for_log$prob > .5, "1", "0")

#confusion matrix
table(actual = p2_for_log$y, prediction = p2_for_log$predicted)
```

```
##       prediction
## actual   0   1
##      0 119   7
##      1  27   3
```

```
#accuracy
(119+3)/(119+3+7+27 )
```

```
## [1] 0.7820513
```

```
#sensitivity
3/(27+3)
```

```
## [1] 0.1
```

```
#specificity
119/(119+7)
```

```
## [1] 0.9444444
```

```
#precision
3/(7+3)
```

```
## [1] 0.3
```