There is a relation between race and income?

30/10/2015

```
load(url("http://bit.ly/dasi_gss_data"))
```

Introduction:

The project aims to study the relationship between the race of the respondents and their income. Does race influences in how much money a person makes? And also if in case there is a relationship between race and income, how is this relationship.

The source of research data is the General Social Survey (GSS), which is a sociological survey applied on US residents in order to collect data on demographic characteristics and behavior.By studying the survey, one could learn some interesting insights of American society.

Data:

The General Social Survey (GSS) data is composed of 57,061 cases (rows) and 114 variables (columns) provide by this course. The GSS data was collected by computer-assisted personal interview (CAPI), face-to-face interview and telephone interview of adults (18+) in randomly selected households.. Each row corresponds to a person surveyed. Of this data frame only two of those 114 variables will be used, these variables are: declared race (race, categorical) and the total family income in constant US dollars of the respondent (coninc, continuous numerical).

gss$race: categorical Race of respondent

What race do you consider yourself? VALUE LABEL NA IAP 1 WHITE 2 BLACK 3 OTHER

Data type: numeric Missing-data code: 0 Record/column: 1/17

---

gss$coninc: continuous numerical

Family income in constant US dollars

Inflation-adjusted family income. VALUE LABEL NA IAP NA DONT KNOW NA NA

Data type: numeric Missing-data codes: 0,999999,999998 Record/columns: 1/72-77

This report is the result of an observational study, because it can establish only correlation between the variables examined and not causation. It‘s opposed to an experiment, only correlations can be found. However, any strong association between income and vote would be useful to infer. The generalization could be applied for all the US population, because the GSS data way gather using random sample from the US population. The bias could be present on the respondent. Its not possible establish causality between the variables, and also for this been an observational study its not possible exists any casual relation between variables.

The data cut used in this research is first made a cut in the original frame data by selecting to dt2, the columns income and race. And in a second step the lines with NA values are removed from the data frame forming a subset. As is presented below:

```
dt2<-gss[,c("coninc","race")]
dt <- subset(dt2, race != "NA")
dt <- subset(dt, coninc != "NA")
```

After the preparation of the data frame, can be noted that of the 57.061 cases more than 5,000 were removed by having a NA value, during the previous step. This step which results in a data frame containing 51.232 rows and 2 columns, as can be observed with the following command:

```
str(dt)
```

Results on:

```
## 'data.frame':    51232 obs. of  2 variables:
##  $ coninc: int  25926 33333 33333 41667 69444 60185 50926 18519 3704 25926 ...
##  $ race  : Factor w/ 3 levels "White","Black",..: 1 1 1 1 1 1 1 1 2 2 ...
```

Exploratory data analysis:

We can see that the distribution of race column, which is the variable self declaration of their race, and it has a largest concentration in white.

This "summary" shows the distribution in numeric format from the levels of race column:

```
summary(dt$race)
```

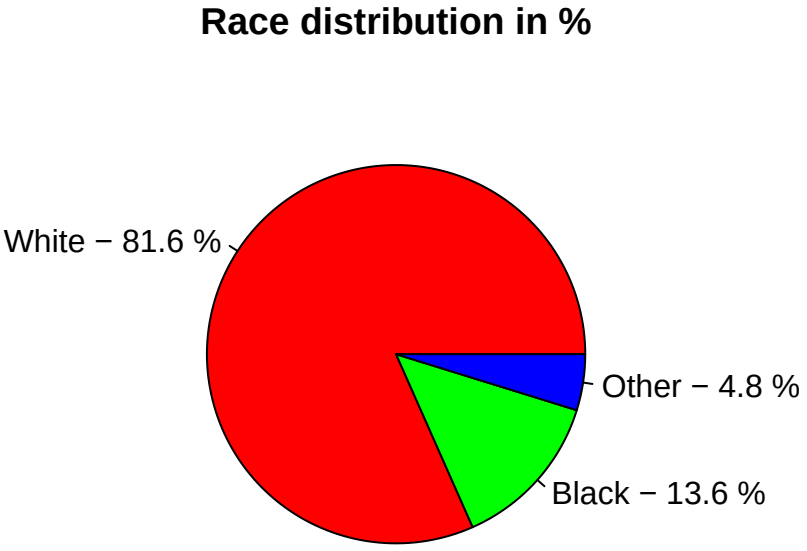
Results on:

```
## White Black Other
## 41824  6956  2452
```

This pie plot presents in visual form the command summary presented above in order to facilitate understanding of representation as a percentage of the variable "race".

```
eda1<-summary(dt$race)
pie(eda1, labels = paste(levels(dt$race), " - ", round(eda1/sum(eda1)*100, digit = 1), " %", sep=""),col
```

Results on:



On the other hand, the column "coninc" which is the family income of the respondent in US dollars. With the lower income the amount of US $ 320 and higher value as US $ 180,400, in addition, the mean is in US $ 45,210. More info can be seen below:

```
summary(dt$coninc)
```

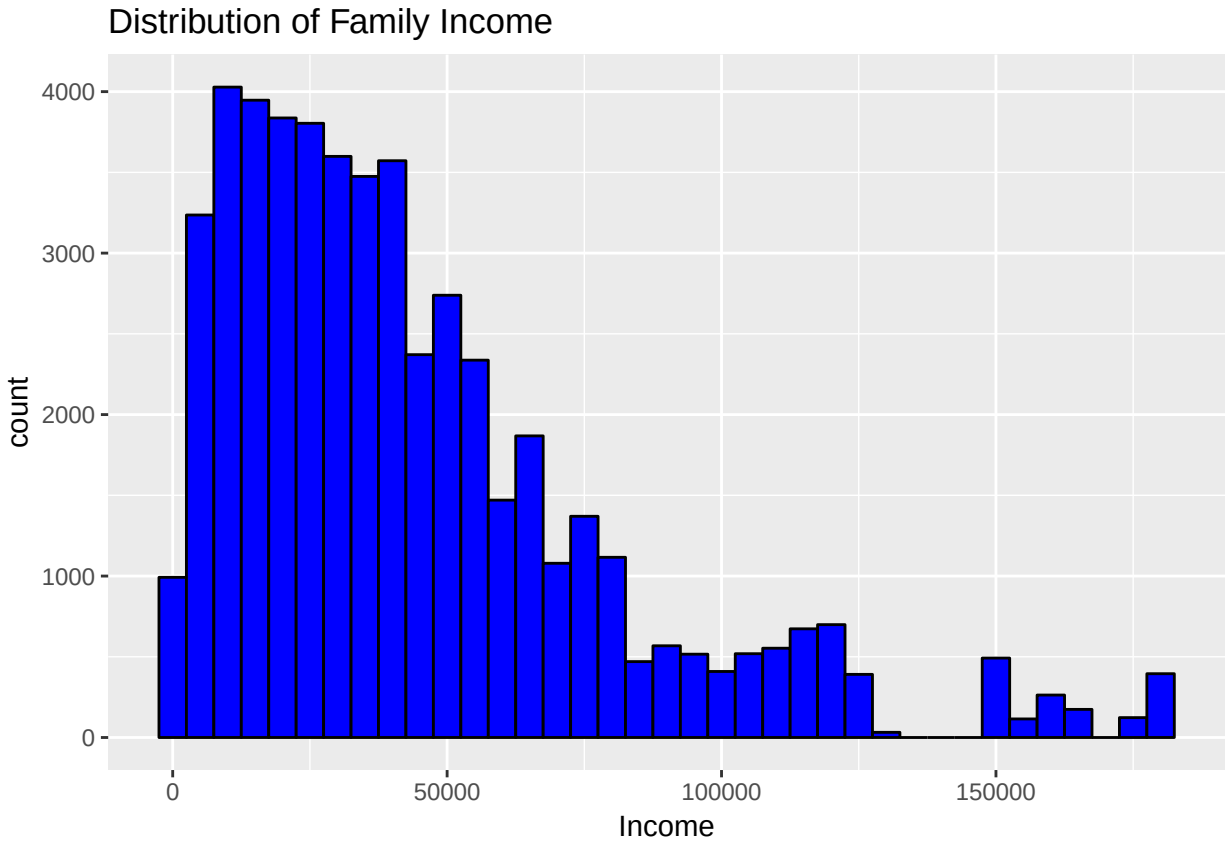
Results on:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     383   18445   35602   44503   59542  180386
```

We observe that the distribution for the family income is right-skewed and limited to zero on the left. Here again, both of these observations are expected as one cannot have a negative income and we expect the count of respondents to decrease as the income increases:

```
library(ggplot2)
ggplot(dt, aes(x=coninc)) +
    geom_histogram(binwidth=5000, colour="black", fill="blue") +
    xlab("Income") +
    ggtitle("Distribution of Family Income")
```

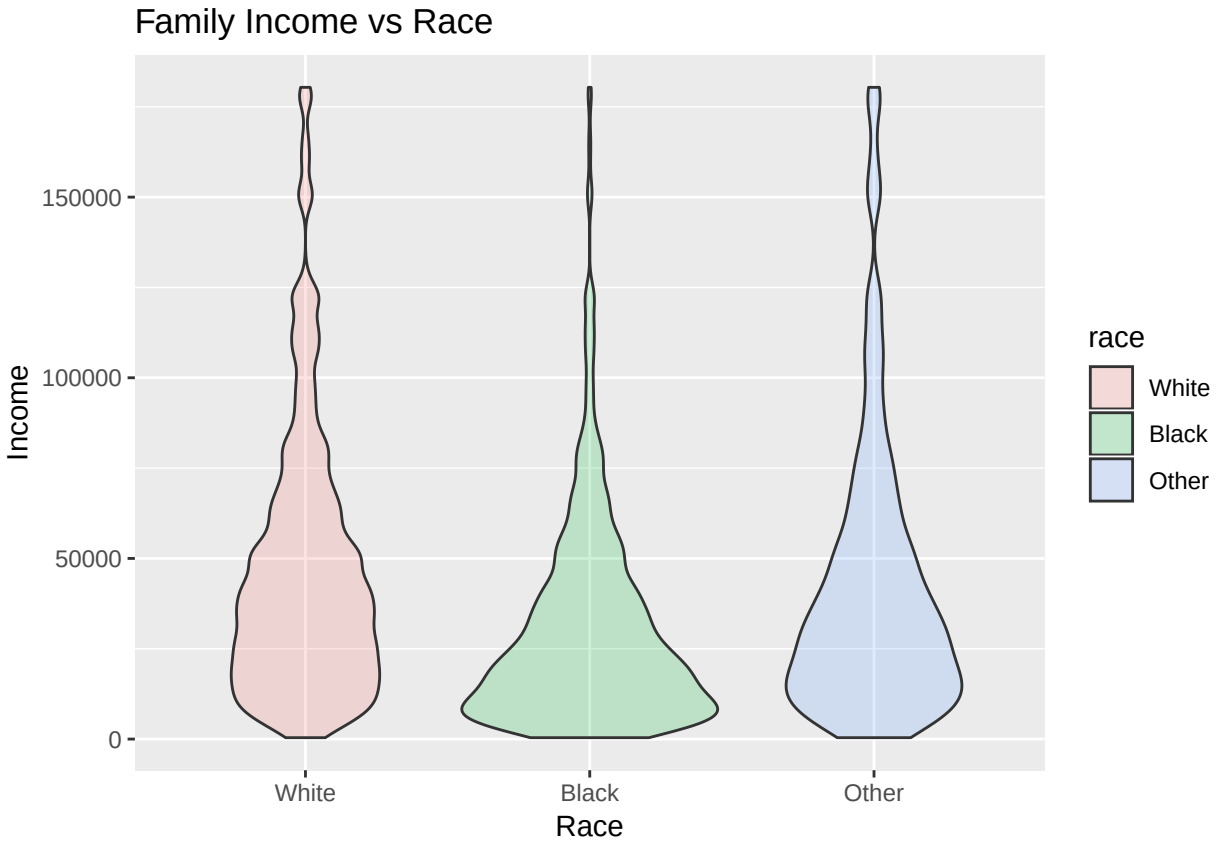
Results on:



Through the use of violin and overlapping graphical representations, it can be observed a great similarity in the relationship between income and races.

```
ggplot(dt, aes(x=race, y=coninc, fill=race)) +
    geom_violin(alpha=0.2) +
    xlab("Race") +
    ylab("Income") +
    ggtitle("Family Income vs Race")
```

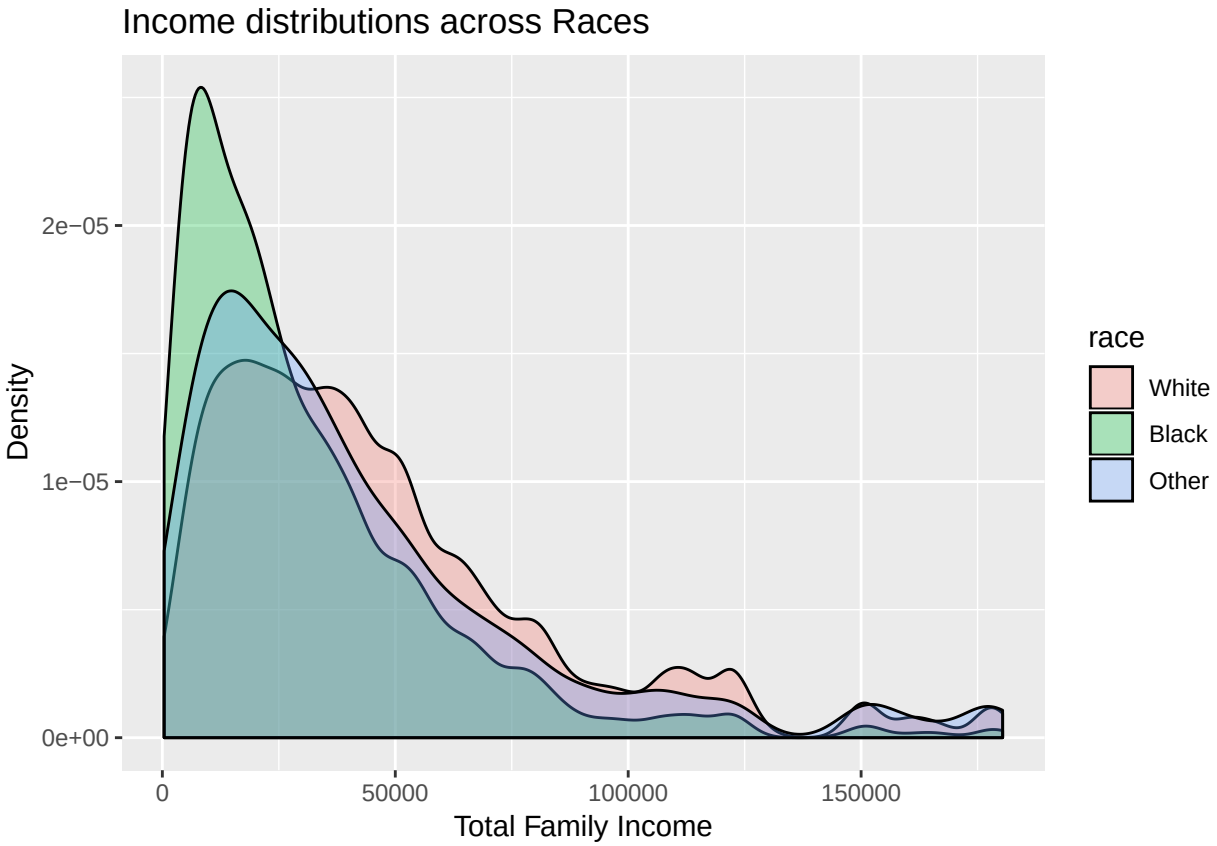
Results on:



The overlapping distribution plots:

```
g <- ggplot(dt, aes(coninc, fill = race))
g + geom_density (alpha = 0.3) + labs(title = "Income distributions across Races") + labs(x = "Total Fa
```

Results on:



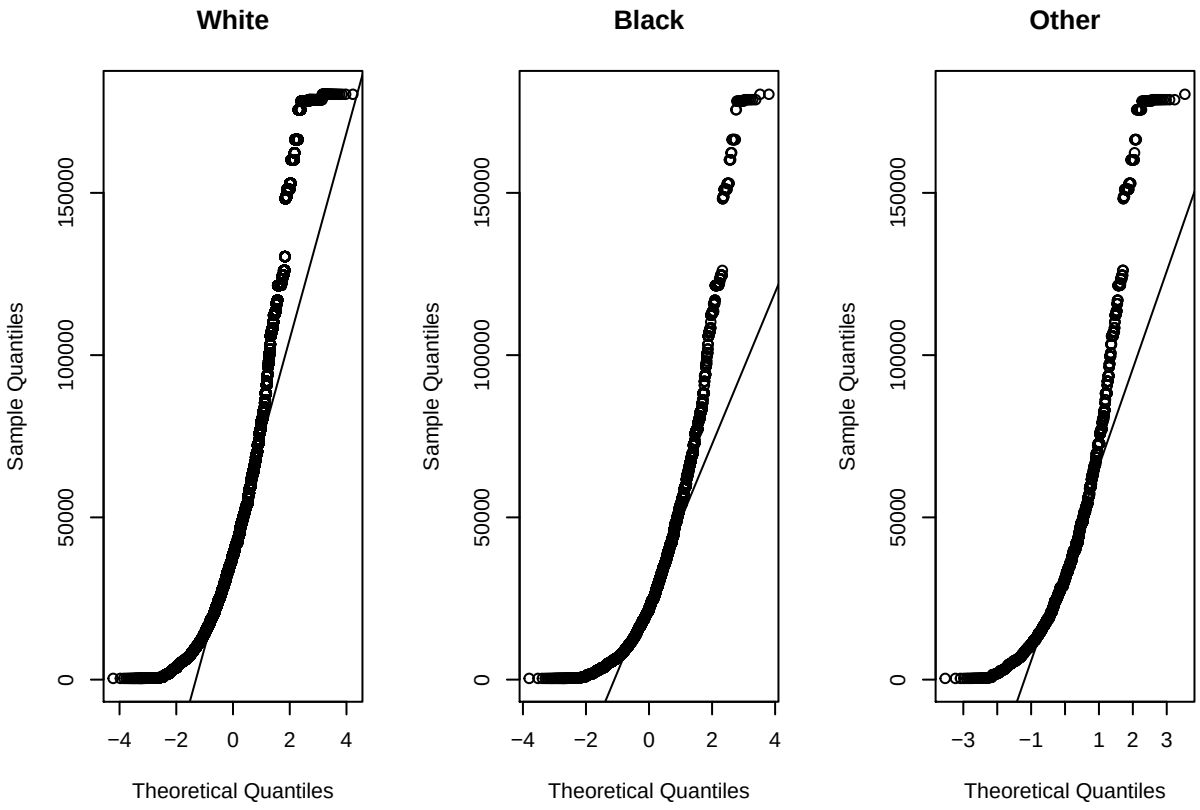
Inference:

The study want to explore if there is a statistical significant difference between the mean family income in constant dollars of United States resident as respect to their race.

State hypothesis H0 (null hypothesis): all means (?) of each race are equal, aka. ?1=?2=?3 Alternate hypothesis HA: the average income in constant dollar varies across some (or all) groups

```
par(mfrow = c(1,3))
qqnorm(dt$coninc[dt$race == "White"], main = "White")
qqline(dt$coninc[dt$race == "White"])
qqnorm(dt$coninc[dt$race == "Black"], main = "Black")
qqline(dt$coninc[dt$race == "Black"])
qqnorm(dt$coninc[dt$race == "Other"], main = "Other")
qqline(dt$coninc[dt$race == "Other"])
```

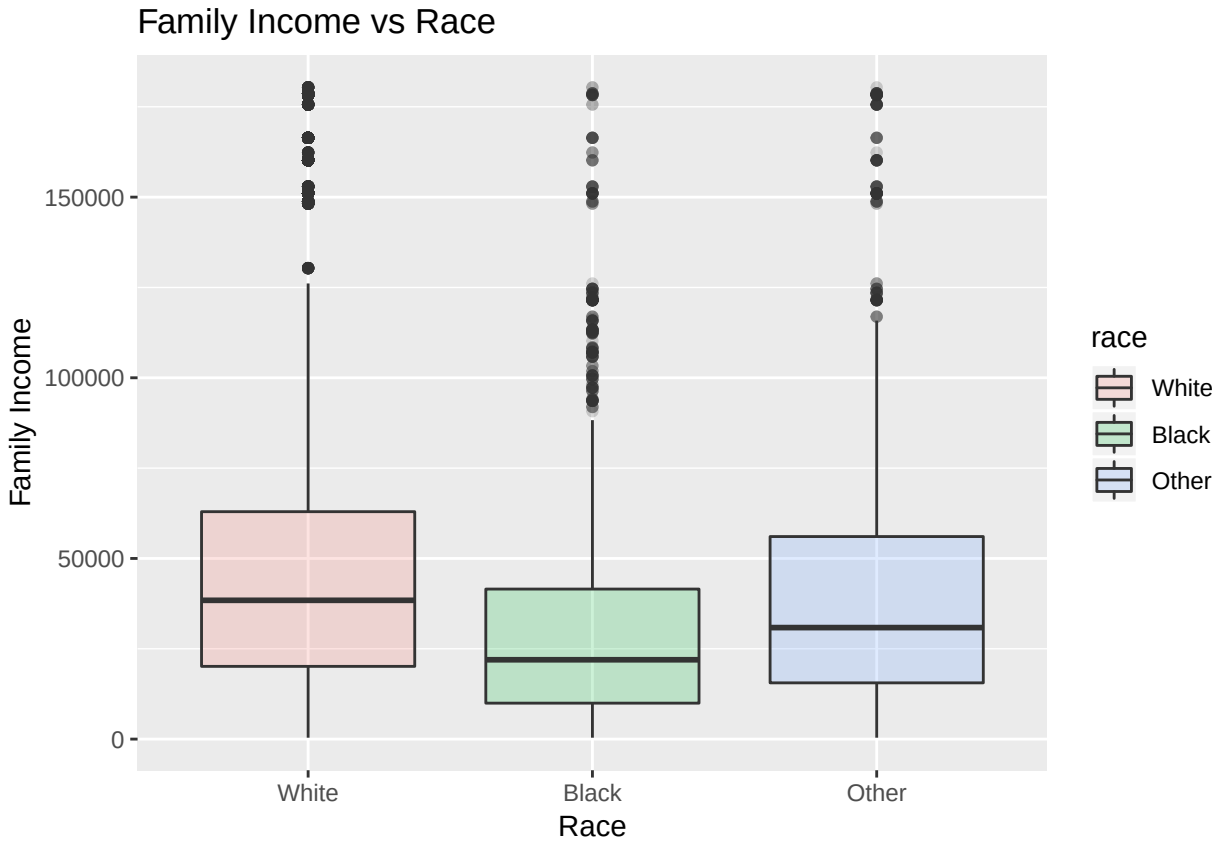
Results on:



Constant variance: we can check variability using boxplot below; total range and interquantile range of 3 race groups are roughly similar.

```
ggplot(dt, aes(x=race, y=coninc, fill=race)) +
    geom_boxplot(alpha=0.2) +
    xlab("Race") +
    ylab("Family Income") +
    ggtitle("Family Income vs Race")
```

Results on:



Although the conditions on normality and constant variance are not fully respected, we will use ANOVA in our hypothesis test and report the uncertainty in final results.

Since we are workng with categorical variables with more than 2 levels, we will use ANOVA test to check whether means across 3 groups are equal. If we can reject the null hypothesis, then results of pairwise comparison can be conducted with Bonferroni method to control Type I error.

ANOVA uses F statistics, which represents a standardized ratio of variability in sample means, relative to variability within the group. The larger the observed variability, the larger F will be, and the stronger the evidence against the null hypothesis. As presented below:

```
anova(lm(coninc ~ race, data=dt))
```

Results on:

```
## Analysis of Variance Table
##
## Response: coninc
##               Df     Sum Sq    Mean Sq F value    Pr(>F)
## race           2 1.6989e+12 8.4944e+11  675.08 < 2.2e-16 ***
## Residuals 51229 6.4461e+13 1.2583e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA reports a F statistics of 675.08 and a p-value of approximately zero. This mean that the probability of observing a F value of 675.08 or higher, if the null hypothesis is true, is very low. So we can reject the null hypothesis and say that family income in constant dollar varies statistically significant among groups.

Since the null hypothesis has been rejected, we can do a pairwise comparison to find out which groups have different means. For every possible pair of groups, we use a t test statistic to confirm the null hypothesis that the means of the two groups are equal or the alternative hypothesis that they are different.

To avoid the increase of Type I error (rejecting a true null hypothesis), we apply a Bonferroni correction to the p-values which are multiplied by the number of comparison. With this correction, the difference of the means has to be bigger to reject the null hypothesis.

```
pairwise.t.test(dt$coninc, dt$race, p.adj="bonferroni")
```

Results on:

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  dt$coninc and dt$race
##
##       White   Black
## Black < 2e-16 -
## Other 1.4e-09 < 2e-16
##
## P value adjustment method: bonferroni
```

Conclusion:

After analysis of the data it may be concluded that while family income between the races being relatively similar, there is a tendency for black respondents have a family income lower than the respondents of other races.

This fact can be observed in several graphs from explotatoria analyze data until the inference, through various forms of representation of information.

References:

General Social Survey Cumulative File, 1972-2012 Coursera Extract. Modified for Data Analysis and Statistical Inference course (Duke University).

R dataset could be downloaded at http://bit.ly/dasi_gss_data.

Original data: Smith, Tom W., Michael Hout, and Peter V. Marsden. General Social Survey, 1972-2012 [Cumulative File]. ICPSR34802-v1. Storrs, CT: Roper Center for Public Opinion Research, University of Connecticut /Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributors], 2013-09-11. doi:10.3886/ICPSR34802.v1

Persistent URL: http://doi.org/10.3886/ICPSR34802.v1

General Social Survey (GSS) FAQ. URL: http://publicdata.norc.org:41000/gssbeta/faqs.html. Accessed 10/27/2015

Comparing many means with ANOVA. In Diez M David, Barr D Christopher, ?etinkaya-Rundel Mine (2015), OpenIntro Statistics, Third Edition, URL: http://www.openintro.org/stat/textbook.php. Accessed 10/27/2015

Appendix:

```
head(dt, n=50)
```

Results on:

```
##    coninc  race
## 1   25926 White
```

```
## 2   33333 White
## 3   33333 White
## 4   41667 White
## 5   69444 White
## 6   60185 White
## 7   50926 White
## 8   18519 White
## 9    3704 Black
## 10  25926 Black
## 11  18519 Black
## 12  18519 Black
## 13  18519 Black
## 14  18519 Black
## 15  25926 Black
## 16  18519 Black
## 17  33333 White
## 18  25926 White
## 19  60185 White
## 20  69444 White
## 21  50926 White
## 22  83333 White
## 23  18519 White
## 24  25926 White
## 25  41667 White
## 26  41667 White
## 27  41667 White
## 28  41667 White
## 30  41667 White
## 31  33333 White
## 32  33333 White
## 33  41667 White
## 34   3704 White
## 35  18519 White
## 36  41667 White
## 37  69444 White
## 38  41667 White
## 39  25926 White
## 40  18519 White
## 41   3704 White
## 42  41667 White
## 43  18519 White
## 44  25926 White
## 45 101852 White
## 47  83333 White
## 48  25926 White
## 49  41667 White
## 50  83333 White
## 51  18519 White
## 52  41667 White
```

```
tail(dt, n=50)
```

Results on:

```
##       coninc  race
```

```
## 57006  10533 White
## 57008  24895 Black
## 57009  76600 White
## 57011   8618 Black
## 57012  16278 White
## 57013  10533 Black
## 57014  51705 White
## 57015  63195 White
## 57016  18193 White
## 57017  16278 White
## 57018   3447 White
## 57020  51705 White
## 57021  51705 White
## 57022  91920 White
## 57023  42130 White
## 57024  24895 White
## 57025  76600 White
## 57026  24895 White
## 57027  63195 White
## 57028   8618 White
## 57029   8618 White
## 57030   1532 Black
## 57031   2681 White
## 57032  12448 Black
## 57033  63195 White
## 57034  24895 Black
## 57035  16278 White
## 57037  63195 White
## 57038  14363 White
## 57039  51705 White
## 57040  34470 White
## 57041  28725 White
## 57043  51705 White
## 57044  34470 White
## 57045 178712 White
## 57046  63195 White
## 57047  51705 White
## 57048 178712 White
## 57049  28725 White
## 57050   6894 Other
## 57051   4213 White
## 57052  63195 White
## 57053  10533 White
## 57054  14363 White
## 57056    383 Other
## 57057  14363 Other
## 57058    383 Other
## 57059  76600 White
## 57060  14363 Other
## 57061    383 Other
```