Statistical inference with the GSS data

Setup

Load packages

```
library(ggplot2)
library(dplyr)
library(statsr)
```

Load data

```
load("gss.Rdata")
```

Part 1: Data

For more than four decades, the General Social Survey (GSS) has studied the growing complexity of American society. It is the only full-probability, personal-interview survey designed to monitor changes in both social characteristics and attitudes currently being conducted in the United States.
The data is collected via random observation.This is not a random assignment experiment.So whatever inferential statistic we perform we can only generalize it the population of Ameica.
The sampling method implemented is Stratified Sampling. In order to reduce this bias, the interviewers are given instructions to canvass and interview only after 3:00 p.m. on weekdays or during the weekend or holidays.

Part 2: Research question

We will use the GSS data set to analyze three situations in the American Society.

1. Whether highest degree of a person and how likely a person is to loose his or her job are dependent or independent.

2. The average age of people when the first child is born same accross all the classes or is it different.

3. Lastly we would like to analyze whether the proportion of people who thinks “On the average (negroes/blacks/African-Americans) have worse jobs, income, and housing than white people due to racial discrimination” is same or different among the Black and White Races.

Part 3: Exploratory data analysis

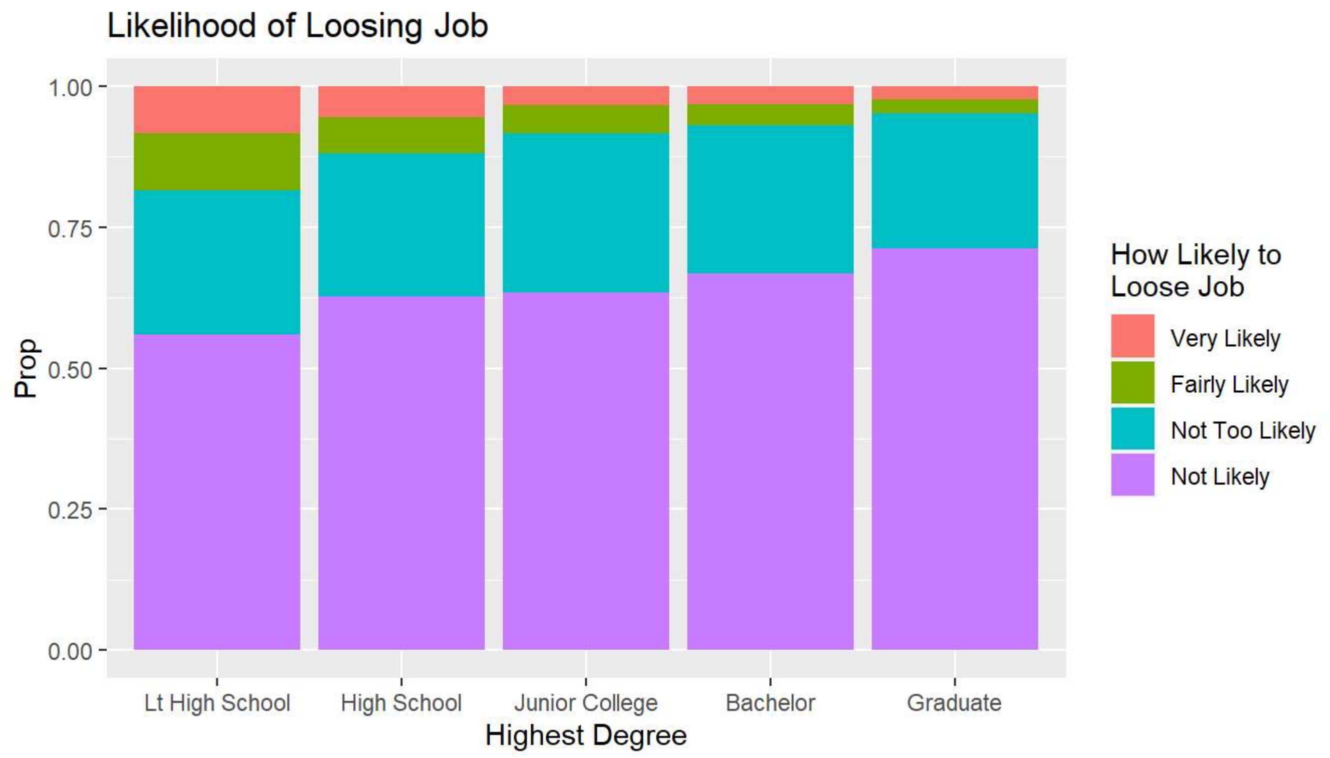
Research quesion 1:

We will create a new data frame named “job_degree” which is a subset of original data set containg only ‘degree’ and ‘jobloose’ data. We will eliminate all the NULL values. Nest we will drop the Leaving Labor Force of ‘joblose’ as it represent only few people.
We have also created a table named “job_degree_table” to create a Chi Square Test Table.

```
job_degree<- data.frame(gss$joblose,gss$degree)

job_degree<- job_degree%>%
filter(!is.na(gss.degree),!is.na(gss.joblose))

job_degree<- job_degree%>%
filter(gss.joblose!="Leaving Labor Force")

job_degree_table<-table(job_degree)

job_degree_table<- job_degree_table[1:4,]
```

We will now do preliminary visual analysis using Bar Plot.

```
ggplot(data = job_degree,aes(x=gss.degree,fill=factor(gss.joblose)))+
geom_bar(position = "fill")+
labs(fill="How Likely to\nLoose Job",x="Highest Degree",y="Prop",title = "Likelihood of Loosing Job")
```



From the above plot we can see that proportion of “Not likely to loose job” increases as the level of highest degree increse. Thus it is clear that proportion of people “likelihood of loosing job” decreases with incrse in Highest Degree Level.
Now will do Chi Square Test of Independence to analyze whether Highest Degree and Likelihood of loosing job are independent or not.

Before we can proceed with the analysis we nee to check whther the criteria for Chi Square test of Independence are satisfied or not.

Condition For Chi Square Test
1. Indepndence:-It is random sampling and size is <10% of the entire population.
2. Sample Size:-Each cell is having more than 5 expected cases as seen in the below table.

```
job_degree_table
```

```
##                 gss.degree
## gss.joblose      Lt High School High School Junior College Bachelor Graduate
##   Very Likely               184         552             45      106       40
##   Fairly Likely             219         639             66      127       44
##   Not Too Likely            563        2557            377      877      408
##   Not Likely               1224        6296            844     2225     1223
```

Now We will perform the Chi Square Test of Indepndence.

$$Ho = Highest\ Degree\ of\ Education\ and\ Likeliness\ to\ loose\ Job\ are\ indepndent.$$

$$H1 = Highest\ Degree\ of\ Education\ and\ Likeliness\ to\ loose\ Job\ are\ not\ indepndent.$$

```
X_sq<- chisq.test(job_degree_table)
```

Observed Value

```
X_sq$observed
```

```
##                 gss.degree
## gss.joblose      Lt High School High School Junior College Bachelor Graduate
##   Very Likely               184         552             45      106       40
##   Fairly Likely             219         639             66      127       44
##   Not Too Likely            563        2557            377      877      408
##   Not Likely               1224        6296            844     2225     1223
```

Expected Value

```
X_sq$expected
```

```
##                 gss.degree
## gss.joblose      Lt High School High School Junior College  Bachelor   Graduate
##   Very Likely          109.0530    500.1498       66.32810  166.0692   85.39992
##   Fairly Likely        128.8166    590.7918       78.34873  196.1659  100.87693
##   Not Too Likely       562.5580   2580.0606      342.15857  856.6808  440.54201
##   Not Likely          1389.5724   6372.9979      845.16459 2116.0840 1088.18113
```

```
print(X_sq)
```

```
##
##  Pearson's Chi-squared test
##
## data:  job_degree_table
## X-squared = 284.69, df = 12, p-value < 2.2e-16
```

From the above result we see that

$$P < 0.05$$

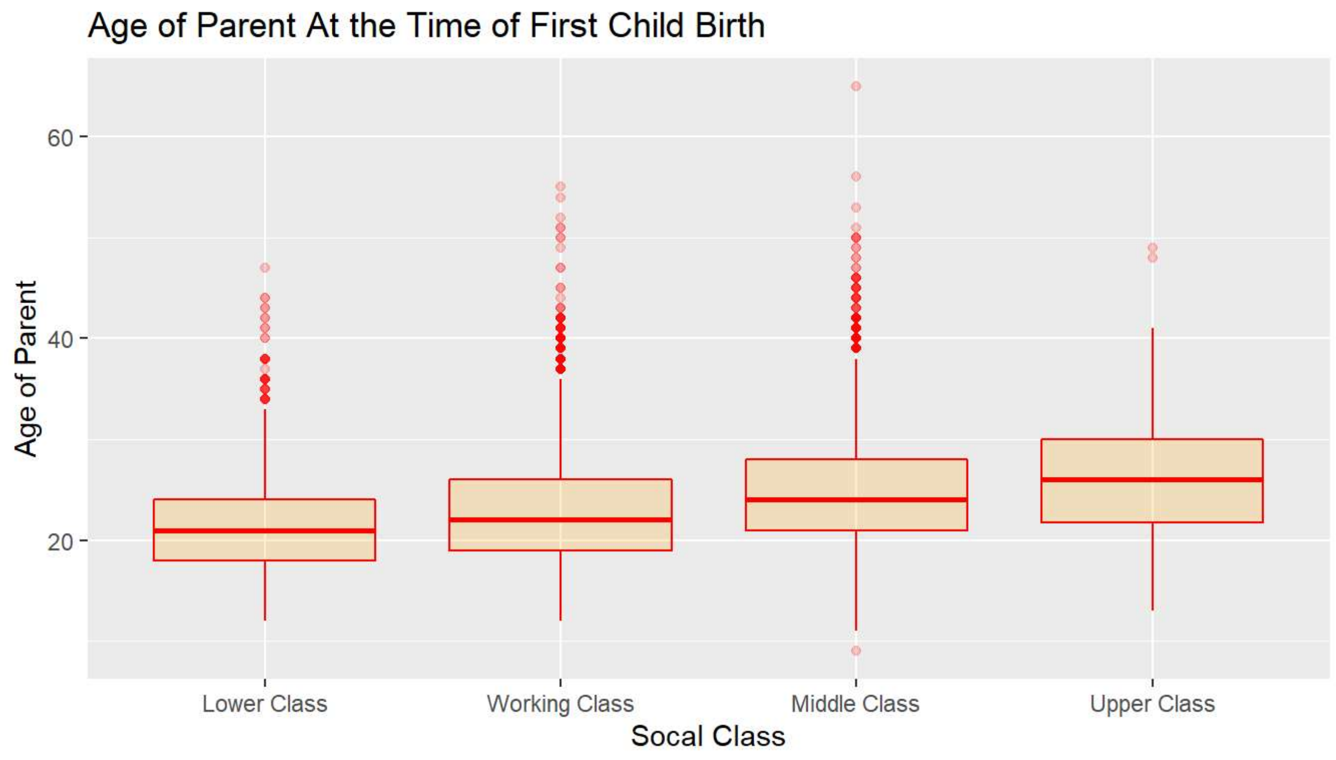
so we reject the NULL Hypothesis in Favour of Alternate Hypothesis.

Research quesion 2:

Let us check the average age of parent when first child is born accross all the social classes.

```
class_child<- gss%>%
filter(!is.na(class),!is.na(agekdbrn))%>%
select(class,agekdbrn)
```

```
ggplot(data = class_child,aes(x=class,y=agekdbrn))+
geom_boxplot(color="red", fill="orange", alpha=0.2)+
labs(x="Socal Class",y="Age of Parent",title = "Age of Parent At the Time of First Child Birth")
```



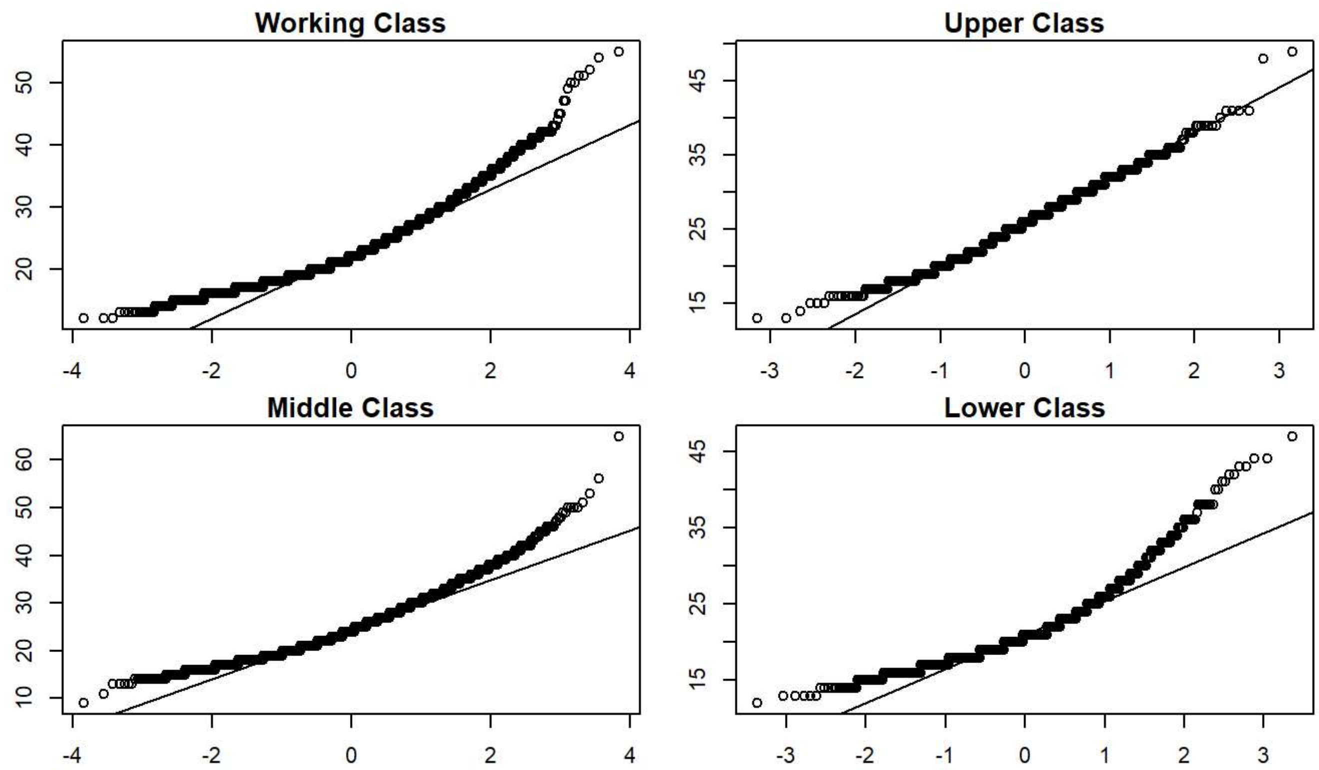
There seems to be a difference of average age of parent when first child born among the social classes.
We will perform ANOVA test to find out whether average age of parent when first child born is same or different for the social classes.

Condition For Chi Square Test
1. Indepndence:-It is random sampling and size is <10% of the entire population.The samples are also independent between the groups.
2. Approx Normality:-The sample distribution must be approximately normal for all the groups.
3. Homoscedasticity:- The Standard Deviation should be constant for all the groups.
Let’s check if condition are satisfied or not.

```
WC<- class_child%>%
filter(class=='Working Class')
UC<- class_child%>%
filter(class=='Upper Class')
MC<- class_child%>%
filter(class=='Middle Class')
LC<- class_child%>%
filter(class=='Lower Class')
```

Now we will check whether the age distribution is approx Normal in all the classes.

```
par(mfrow=c(2,2))
par(cex=0.7, mai=c(0.3,0.3,0.2,0.2))
qqnorm(WC$agekdbrn,main="Working Class")
qqline(WC$agekdbrn)
qqnorm(UC$agekdbrn,main="Upper Class")
qqline(UC$agekdbrn)
qqnorm(MC$agekdbrn,main="Middle Class")
qqline(MC$agekdbrn)
qqnorm(LC$agekdbrn,main="Lower Class")
qqline(LC$agekdbrn)
```

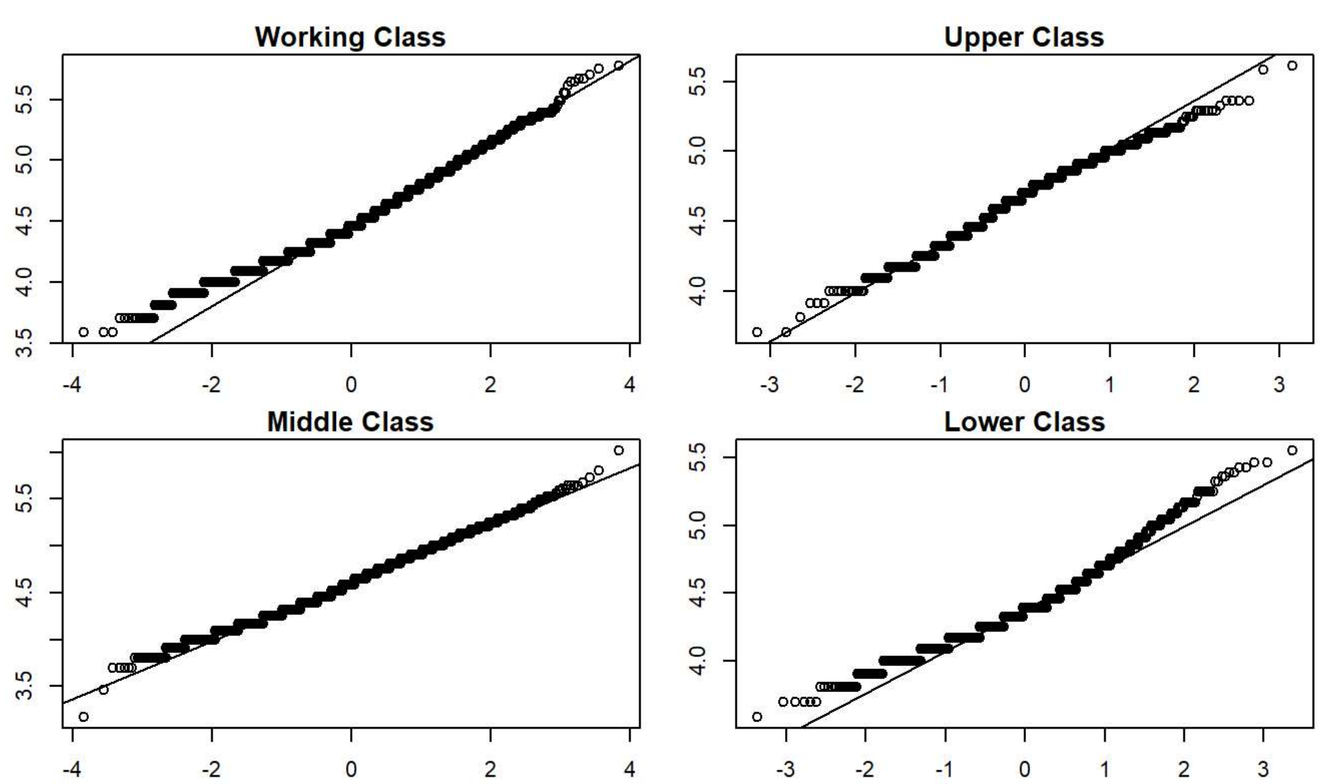


From the Normal Probability plot for each class we see that the Approx Normal distribution is not satisfied by the Working Class and Lower Class.
So we will apply Logarithmic Transformation and see if Normal Distribution is satis fied ot not.

```
class_child<- class_child%>%
mutate(log_tra = log2(agekdbrn))
WC<- class_child%>%
filter(class=='Working Class')
UC<- class_child%>%
filter(class=='Upper Class')
MC<- class_child%>%
filter(class=='Middle Class')
LC<- class_child%>%
filter(class=='Lower Class')
```

Let us now check the Normal Probabilty plot for each class.

```
par(mfrow=c(2,2))
par(cex=0.7, mai=c(0.3,0.3,0.2,0.2))
qqnorm(WC$log_tra,main="Working Class")
qqline(WC$log_tra)
qqnorm(UC$log_tra,main="Upper Class")
qqline(UC$log_tra)
qqnorm(MC$log_tra,main="Middle Class")
qqline(MC$log_tra)
qqnorm(LC$log_tra,main="Lower Class")
qqline(LC$log_tra)
```



Now the Approx Normality in Each grop condition is satisfied for ANOVA.
Lest check if the Homoscedasticity condition is satisfied.

```
class_child%>%
group_by(class)%>%
summarise(log_sd = sd(log_tra))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 4 x 2
##   class         log_sd
##   <fct>          <dbl>
## 1 Lower Class    0.307
## 2 Working Class  0.299
## 3 Middle Class   0.313
## 4 Upper Class    0.323
```

We see almost constant Standard Deviation so Homoscedasticity is satisfied.
Now we will perform ANOVA.

$$Ho = Average\ age\ of\ parent\ when\ first\ child\ is\ born\ is\ same\ accross\ all\ the\ classes.$$

$$H1 = Average\ age\ of\ parent\ when\ first\ child\ is\ born\ is\ different\ for\ atleat\ one\ class.$$

```
anova_one_way<- aov(log_tra ~ class,class_child)
summary(anova_one_way)
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## class           3     96   32.00   339.7 <2e-16 ***
## Residuals   18011   1696    0.09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the ANOVA test

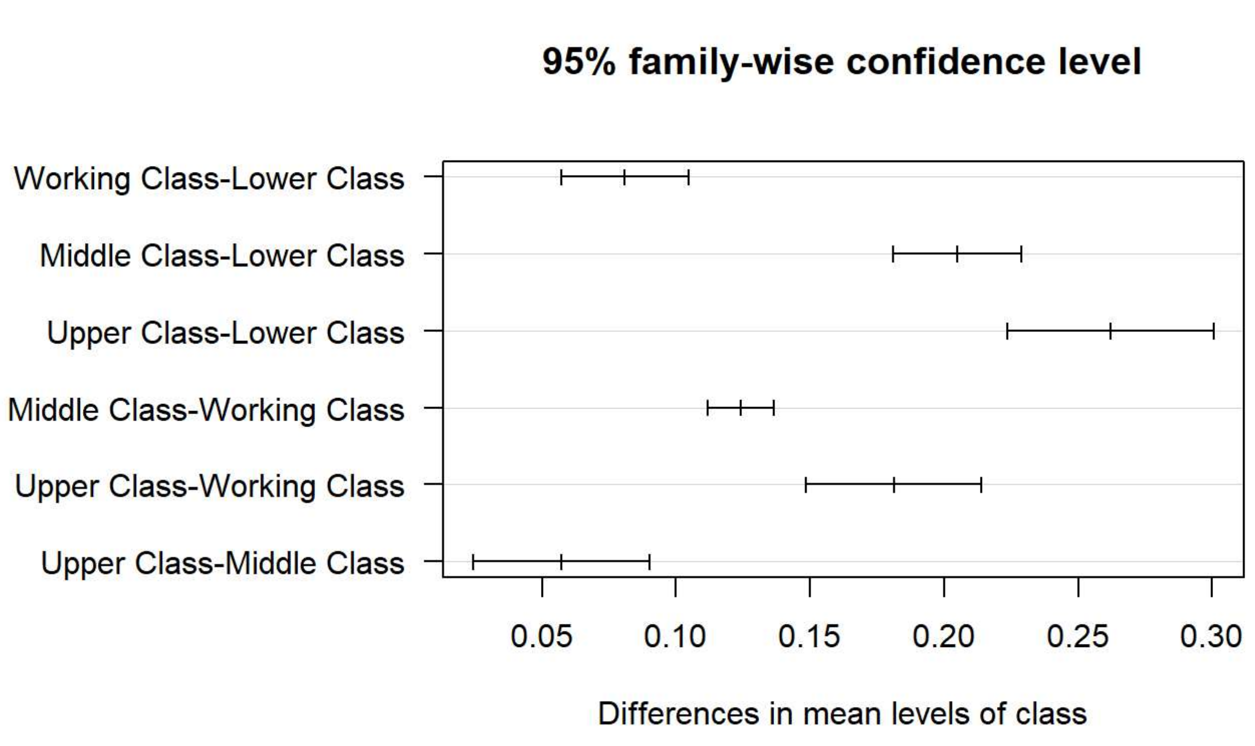
$$P < 0.05$$

so we reject the NULL Hypothesis in favour of Alternate Hypothesis. Let us now perform pair wise comparison to see for which classes the average age is differnent.

```
TukeyHSD(anova_one_way,conf.level = 0.95,p.adjust = "bonf")
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = log_tra ~ class, data = class_child)
##
## $class
##                                   diff        lwr        upr    p adj
## Working Class-Lower Class  0.08087197 0.05708259 0.10466135 0.00e+00
## Middle Class-Lower Class   0.20497657 0.18113899 0.22881415 0.00e+00
## Upper Class-Lower Class    0.26207483 0.22343835 0.30071132 0.00e+00
## Middle Class-Working Class 0.12410460 0.11168456 0.13652465 0.00e+00
## Upper Class-Working Class  0.18120287 0.14835769 0.21404804 0.00e+00
## Upper Class-Middle Class   0.05709826 0.02421816 0.08997837 4.84e-05
```

```
par(oma=c(0,8,0,0))
plot(TukeyHSD(anova_one_way),las = 1)
```



We see that

$$P < 0.05$$

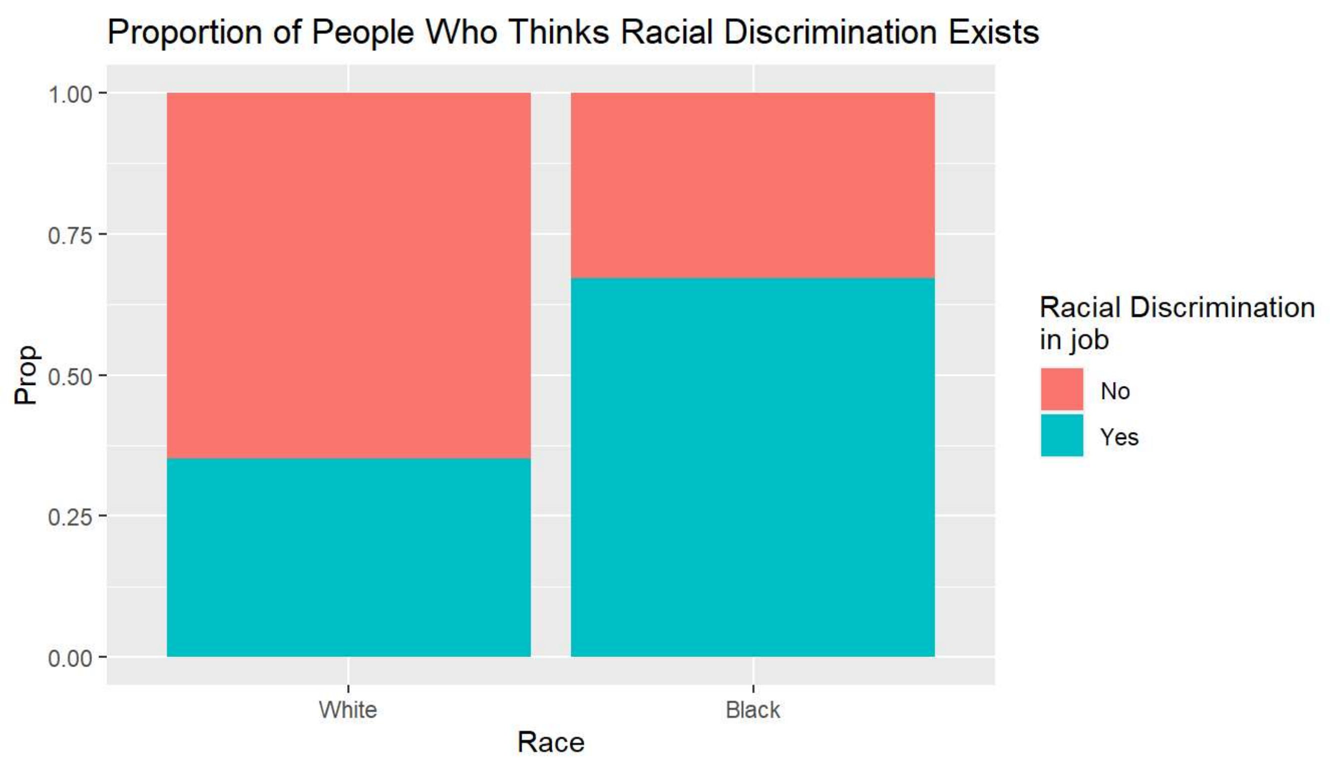
for all pairwise comparison. Hense the average age of parent wen first child is born is different accross all the classes.From the Figure also we see that none of the CI of difference in mean between classes include 0.

Research quesion 3:

We will nalyze whether the proportion of people who thinks “On the average (negroes/blacks/African-Americans) have worse jobs, income, and housing than white people due to racial discrimination” is same or different among the Black and White Races.

```
race_discr<- gss%>%
filter(!is.na(race),!is.na(racdif1),race!='Other')%>%
select(race,racdif1)
```

```
ggplot(data = race_discr,aes(x=race,fill=factor(racdif1,levels = c("No","Yes"))))+
geom_bar(position = "fill")+
labs(fill="Racial Discrimination\nin job",x="Race",y="Prop",
     title = "Proportion of People Who Thinks Racial Discrimination Exists")
```



We see that the proportion is differnet between the two race. We will analyze if the difference in proportion of people who thinks “(negroes/blacks/African-Americans) have worse jobs, income, and housing than white people due to racial discrimination” is zero or not. First let us check the condition for ineferential statistic for proportion on Two Categorical variable and Twol Level:Success and Failure.

Condition For above Infernce
1. Independence:-It is random sampling and size is <10% of the entire population.The samples are also independent between the groups.
2. Sample size:- There should be at least 10 success and 10 failures in the samples for each categorical variable.

```
race_discr_table<- table(race_discr)
race_discr_table<- race_discr_table[1:2,]
race_discr_table
```

```
##        racdif1
## race      Yes    No
##   White  6918 12712
##   Black  2047  1005
```

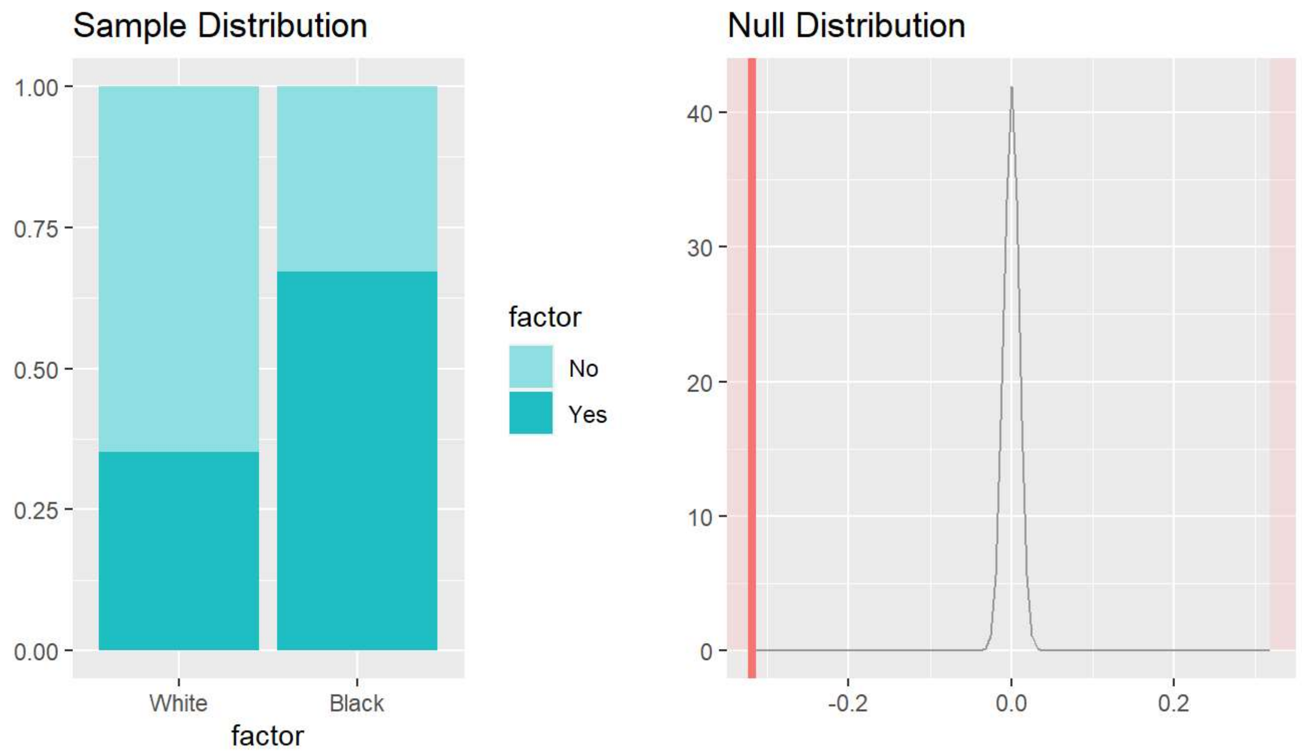
From the above table we see that the required conditions are satisfied. Let’s perform the infernce. Now we will perform Inference on Two Independent Proportions using Theoretical Method since we have hige number of observations.

$$Ho = Difference\ in\ proportion\ of\ people\ who\ thinks\ racial\ discrimination\ exists\ in\ job\ opportunity\ is\ zero.$$

$$H1 = Difference\ in\ proportion\ of\ people\ who\ thinks\ racial\ discrimination\ exists\ in\ job\ opportunity\ is\ not\ zero.$$

```
inference(data = race_discr,type = "ht",method = "theoretical",
          statistic = "proportion",null = 0,alternative = "twosided",
          x=factor(race,levels=c("White","Black")),y=factor(racdif1,levels=c("No","Yes")),
          success = "Yes",conf_level = 0.95)
```

```
## Response variable: categorical (2 levels, success: Yes)
## Explanatory variable: categorical (2 levels)
## n_White = 19630, p_hat_White = 0.3524
## n_Black = 3052, p_hat_Black = 0.6707
## H0: p_White =  p_Black
## HA: p_White != p_Black
## z = -33.4587
## p_value = < 0.0001
```



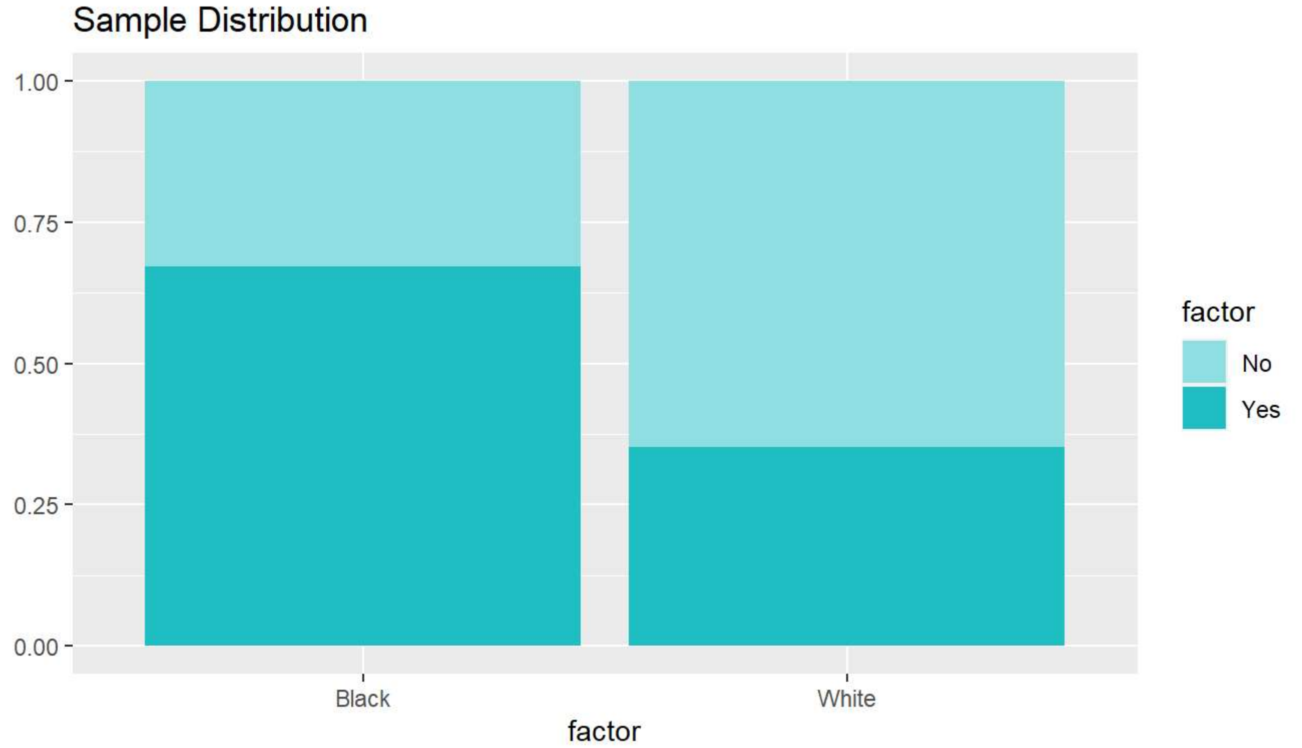
From the hypothesis test we see that

$$P < 0.05$$

so we reject the NULL hypothesis in favour of alternate hypothesis.
In support of our answer we will also calculate the 95% CI for differnce in Proportion and check if 0 is present in that interval.

```
inference(data = race_discr,type = "ci",method = "theoretical",
          statistic = "proportion",x=factor(race,levels=c("White","Black")),
          y=factor(racdif1,levels=c("No","Yes")),success = "Yes",conf_level = 0.95,
          order = TRUE)
```

```
## Response variable: categorical (2 levels, success: Yes)
## Explanatory variable: categorical (2 levels)
## n_Black = 3052, p_hat_Black = 0.6707
## n_White = 19630, p_hat_White = 0.3524
## 95% CI (Black - White): (0.3003 , 0.3363)
```



We see that 0 is Not included in 95% CI hence our which support our above hypothesis.

Part 4: Inference

From the above three Analysis we can conclude the following:-

1.How likely a person is to loose his or her job and highest degree of that person are not independent i.e. Higher the degree less likely a person is to loose his/her job.

2.The average age of people when the first child is born is different in all Social Classes.

3.The proportion of people who thinks “On the average (negroes/blacks/African-Americans) have worse jobs, income, and housing than white people due to racial discrimination” is different among the Black and White Races where the Black Race proportion is 0.3003 to 0.3363 higher than White Race

All the above Inference are generalization to the American Society and NOT CAUSATION.