Statistical inference with the GSS data

Setup

Load packages

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.2
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.2
```

```
library(statsr)
library(Hmisc)
```

```
## Warning: package 'Hmisc' was built under R version 3.4.2
```

```
## Warning: package 'Formula' was built under R version 3.4.1
```

Load data

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

Part 1: Data

The GSS is an observational study, since the data is collected in such a way that it did not directly interfere with how the data arise.

There was a random selection of households in this survey, so as to represent a cross-section of the USA. The scientific sample was designed to ensure that all households from across the country had an equal probability of being selected i.e. Simple Random Sampling.

This survey is subject to non-response bias; the data that was collected comes from randomly selected people willing to answer a survey.

Since the households selected were random, and there was no random assignment, we have that the study was not causal. The study was rather generalisable.

In this report, we will be focusing on participants from the South Atlantic region.

```
data <- gss %>%
  filter(region == "South Atlantic")
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.1
```

```
dim(data)
```

```
## [1] 10977   114
```

Part 2: Research question

How do High School students and Graduate students compare with respect to their political party affiliation? Are there similarities between political party affiliation and their political views?

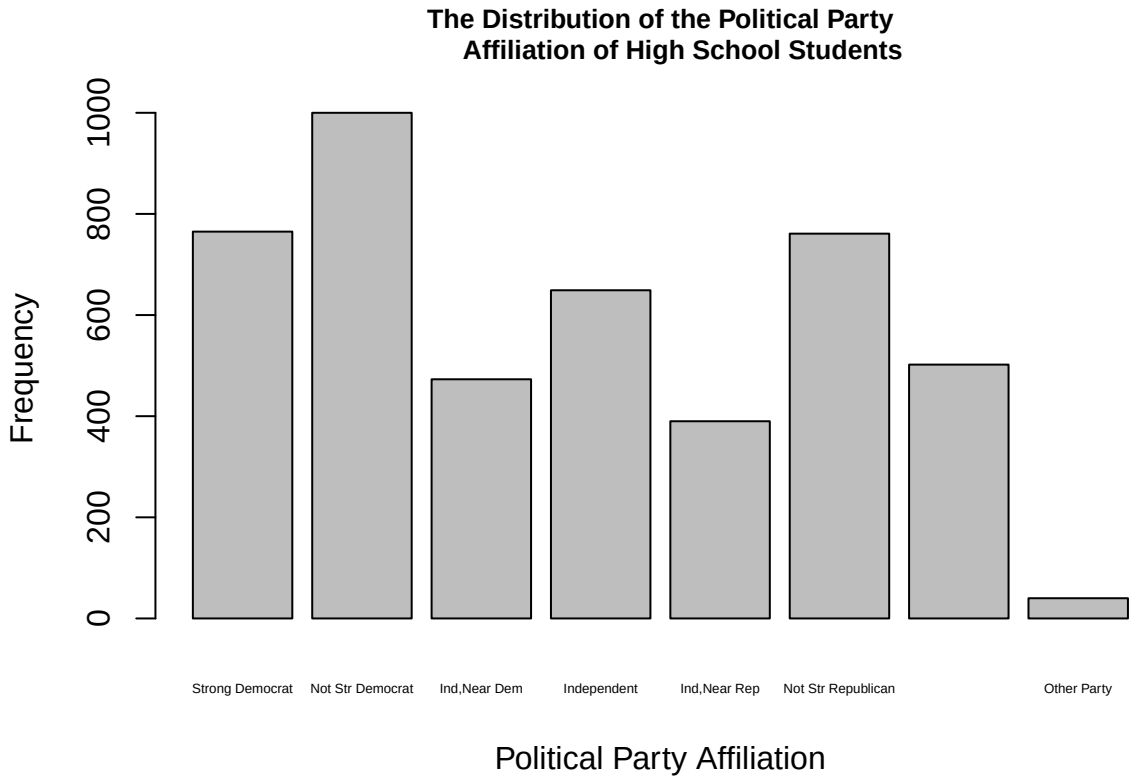
Researching the political views of high school students compared to graduate students will provide valuable information that can be useful in a variety of areas. For example: political campaigns, government advertising, general advertising etc.

Part 3: Exploratory data analysis

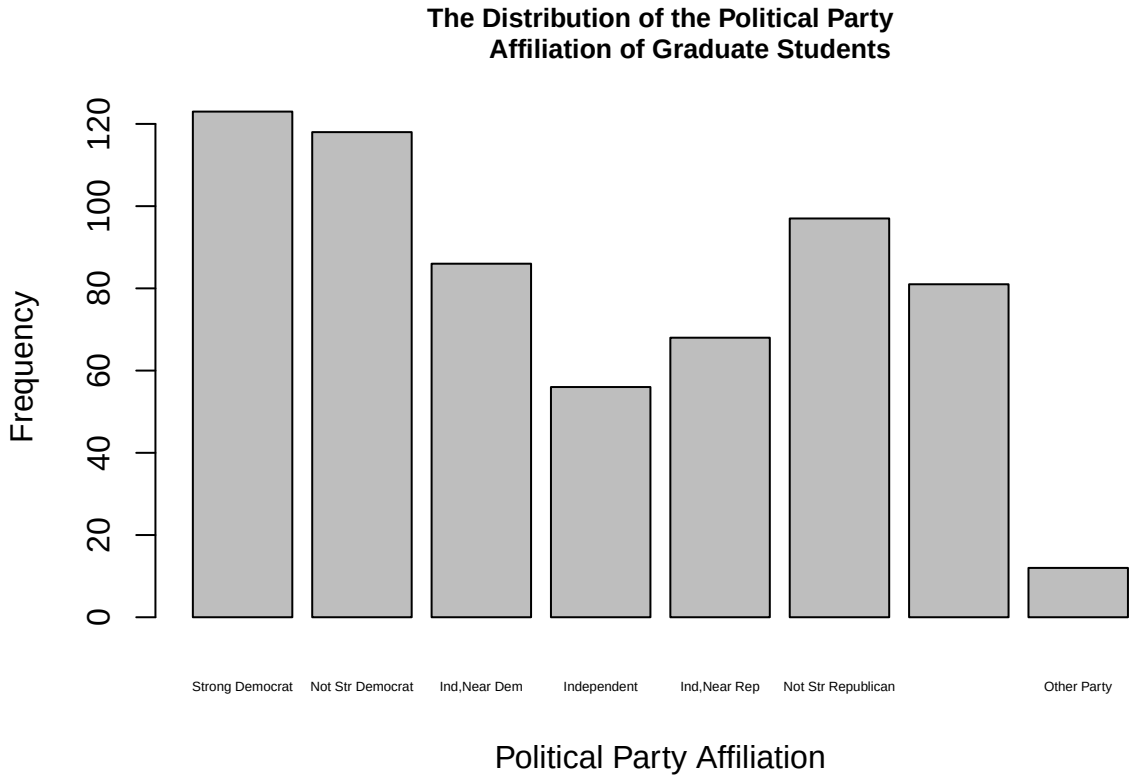
First compare the Political Party Affiliation between High School students and Graduate students.

```
df <- data[, c("degree", "partyid", "polviews")]
mydata <- na.omit(df)

hs <- subset(mydata, degree == "High School")
      barplot(table(hs$partyid),main = "The Distribution of the Political Party
      Affiliation of High School Students", xlab = "Political Party Affiliation",
      ylab = "Frequency", cex.main = 0.8, cex.names = 0.45)
```



```
grad <- subset(mydata, degree == "Graduate")
        barplot(table(grad$partyid),main = "The Distribution of the Political Party
        Affiliation of Graduate Students", xlab = "Political Party Affiliation",
        ylab = "Frequency", cex.main = 0.8, cex.names = 0.45)
```

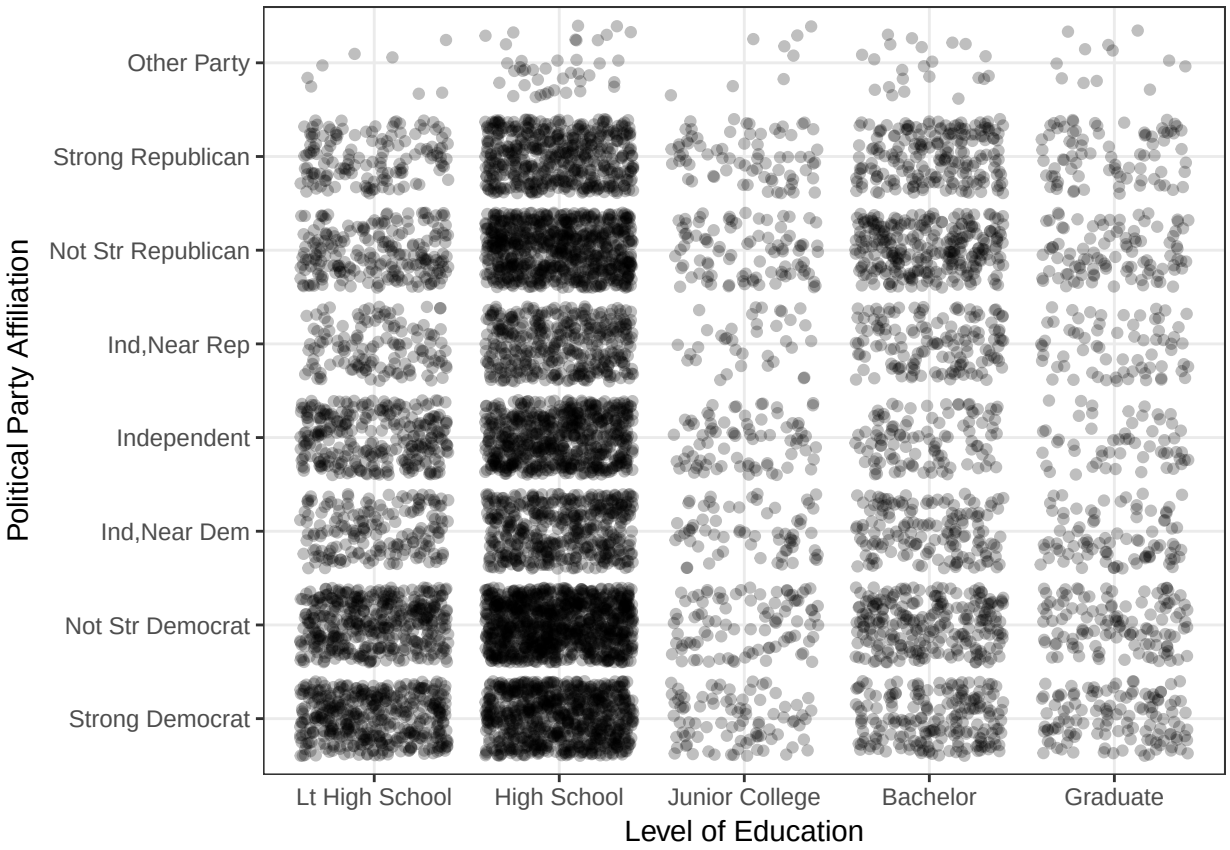


As one can see, the barplot for High School students shows a large variation, whereas for Graduate students there is more consistency. Majority of High School and Graduate students associate with the Democratic Party.

Now, let us compare the students political party affiliation to their political views.

```
ggplot(mydata, aes(x=degree, y=partyid)) +
 theme_bw() +
  geom_jitter(alpha=0.25) +
   geom_smooth() +
    labs(x="Level of Education", y="Political Party Affiliation")
```

```
## `geom_smooth()` using method = 'gam'
```

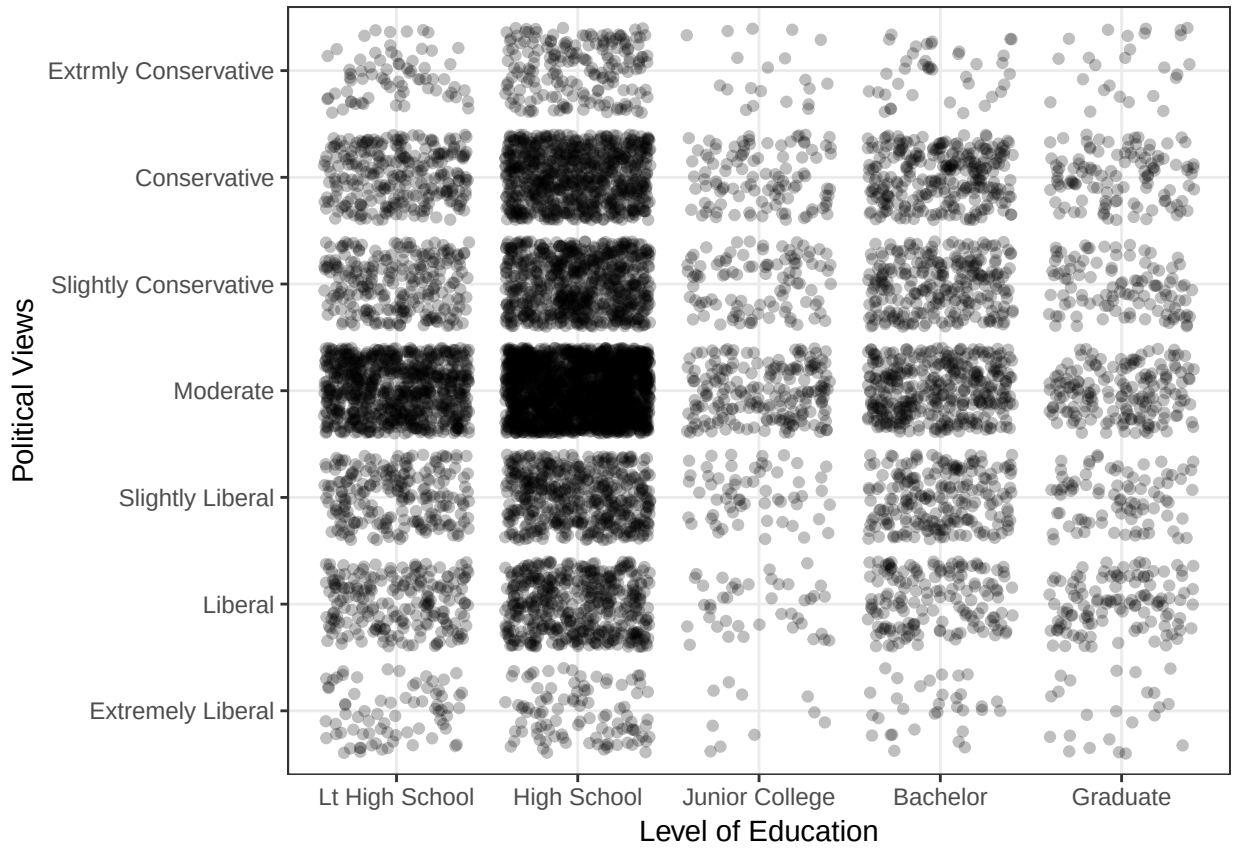



```
cor(as.numeric(mydata$degree), as.numeric(mydata$partyid), use = "everything")
```

```
## [1] 0.09991226
```

```
ggplot(mydata, aes(x=degree, y=polviews)) +
 theme_bw() +
  geom_jitter(alpha=0.25) +
   geom_smooth() +
    labs(x="Level of Education", y="Political Views")
```

```
## `geom_smooth()` using method = 'gam'
```




```
cor(as.numeric(mydata$degree), as.numeric(mydata$polviews), use = "everything")
```

```
## [1] 0.006408485
```

As can be seen by the graphs above, and their relative coefficient of correlation, we have that the variables 'degree' and 'partyid' are more correlated (r = 0.0999) than the variables 'degree' and 'polviews' (r = 0.0064).

Part 4: Inference

A hypothesis test was setup, at 5% significance level, to compare High School students and Graduate students with respect to their political party affiliation.

H_0: There is no significant difference in political party affiliation between High School students and Graduate students

H_A: There is a significant difference in political party affiliation between High School students and Graduate students

Conditions for the Chi-Square Test:

1. Independence: The sampled observations are independent

- As discussed in the description of the data above, this is a random sample of data.

- When looking at the table constructed below, one can see that each case only contributes to one cell i.e. there are no overlapping cells, and each cell corresponds to their own variables.

2. Sample Size: Each cell has at least 5 expected cases, as can be seen in the table constructed below.

i.e. All conditions are met.

The method we will be using is the Chi-Square Test of Independence, because we will be evaluating the relationship between two categorical variables (where one has more than two levels).

```
final_data <- mydata %>%
  filter(degree == c("High School", "Graduate"))

final_df <- final_data[c("degree", "partyid")]

x<-table(final_df$partyid, final_df$degree, exclude = c("Lt High School",
        "Junior College", "Bachelor"))
x
```

```
##
##                      High School Graduate
##   Strong Democrat            381       63
##   Not Str Democrat           529       67
##   Ind,Near Dem               239       46
##   Independent                323       26
##   Ind,Near Rep               202       35
##   Not Str Republican         389       43
##   Strong Republican          236       45
##   Other Party                 18        6
```

```
chisq.test(x)
```

```
## Warning in chisq.test(x): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  x
## X-squared = 23.889, df = 7, p-value = 0.001192
```

As the p-value = 0.001192 $<$ 0.05, we reject the null hpothesis that there is no significant difference in political party affiliation between High School students and Graduate students.

The method of Chi-Square Test of Independence was used, therefore no other methods were used. Since only one method of inference was used, there is no comparison available.

Conclusion: Based on the results obtained from both the Exploratory Data Analysis and the Inference, High School students and Graduate Students differ significantly in terms of their political party affiliation. However, this might be due to students at High School level being uncertain of their choices.
There are clear similarities between political party affiliation and political views, which implies that they are not independent of one another.