# DATA 602 HW1

Kane Smith

2022-09-18

## Contents



## Question 1

### a.

```
0.6^2
```

```
## [1] 0.36
```

There is a 36% chance that both would want to change their undergraduate major.

### b.

```
(1-0.6)^2
```

```
## [1] 0.16
```

There is a 16% chance that neither would want to change their undergraduate major.

### c.

```
(1-0.6)*0.6 + 0.6*(1-0.6)
```

```
## [1] 0.48
```

There is a 48% chance that at least one would want to change their undergraduate major.

### d.

The probability that at least one Canadian wants to change their degree is equal to 1 - the probability that no Canadians want to change their degree. Below is the how n is derived:

$1 - P(x = 0)$

$1 - nC0 * (0.4)^0 * (0.6)^n$

$1 - (0.6)^n = 0.95$

$0.05 = 0.6^n$

$ln(0.05)/ln(0.6) = n$

```
log(0.05)/log(0.6)
```

```
## [1] 5.864491
```

You would need to select approximately 5.8645 Canadians. However, since you cannot select a fraction of a person, you would need to select at least 6 Canadians to have a probability of 0.95 that at least one of them would want to change their degree.

## Question 2

Step 1:

```
nsims = 1000
outcome = numeric(nsims)
```

Step 2:

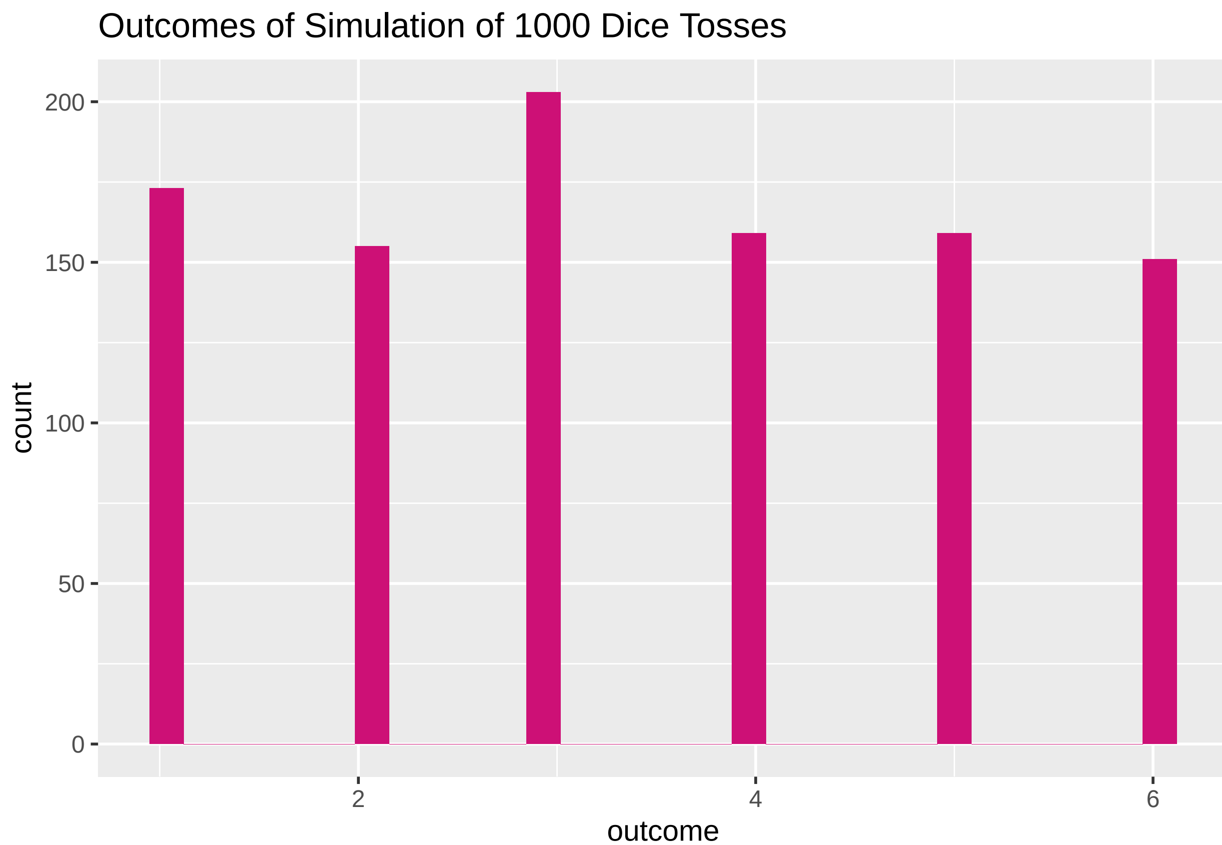
```
for(i in 1:nsims){
  outcome[i] = sample(c(1,2,3,4,5,6), 1, replace=FALSE)
}

simresult = data.frame(outcome)
head(simresult,3)
```

```
##   outcome
## 1       3
## 2       5
## 3       5
```

Step 3:

```
library(ggplot2)
ggplot(simresult, aes(x = outcome)) + geom_histogram(fill = 'deeppink3') + ggtitle("Outcomes of Simulati
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




This is consistent with what I expected of the simulation of X. If the dice is fair and each side is equally likely, we should expect a discrete uniform distribution where all outcomes have the same likelihood. If we were to do more simulations, I would expect the counts of each outcome to approach even closer to equal.

Step 4:

```
nsims = 3000
outcome = numeric(nsims)
fivesix = numeric(nsims)
```

simulation:

```
for(i in 1:nsims){
  outcome[i] = sample(c(1,2,3,4,5,6), 1, replace=FALSE)
  fivesix[i] = if (outcome[i] == 5 || outcome[i] == 6) 1 else 0
}

simresult = data.frame(outcome)
```

```r
nsims = 3000
outcome = numeric(nsims)
sum_fourteen = numeric(nsims)
```

```r
for(i in 1:nsims){
  tosses = sample(c(1,2,3,4,5,6,1,2,3,4,5,6,1,2,3,4,5,6), 3, replace=FALSE)
  outcome[i] = sum(tosses)
  sum_fourteen[i] = if (outcome[i] >= 14) 1 else 0
}

simresult = data.frame(outcome, sum_fourteen)
head(simresult)
```

```
##   outcome sum_fourteen
## 1      11            0
## 2      13            0
## 3      14            1
## 4       8            0
## 5      14            1
## 6       6            0
```

## Question 3

Our sample space is \$ 20C5 = 15504 \$ Since there is one 10, J, Q, K and Ace of each suit, the event of choosing a hand of the same suit is 5C5 = 1. There are 4 different suits so you multiply it by 4.

#### a.

```r
4/choose(20,5)
```

```
## [1] 0.0002579979
```

The event of getting a 3 of a kind for each card is $4C1 * 3C1 * 2C1 = 24$ There are 5 different types of card, so we multiply it by 5 to get 120.

```r
(5*(choose(4,1) * choose(3,1) * choose(2,1)))/choose(20,5)
```

```
## [1] 0.007739938
```

#### c.

**Assumption**: You must observe exactly 2 Aces and 2 Diamonds.

**Case 1:** Neither aces are diamonds: $3C2 * 4C2 * 12C1$

**Explanation:**

Out of the 3 other aces, you choose 2. Out of the 4 other diamonds you choose 2. Out of the 12 other cards that are neither diamonds nor aces, you choose 1.

**Case 2:** 1 ace is a diamond: $1C1 * 3C1 * 4C1 * 12C2$

**Explanation:**

There is only one way to choose the ace which is a diamond. Out of the 3 other aces, you choose 1. Out of the 4 other diamonds, you choose 1. Out of the 12 other cards that are neither diamonds nor aces, you choose 2.

```
(choose(3,2)*choose(4,2)*choose(12,1) + choose(1,1)*choose(3,1)*choose(4,1)*choose(12,2))/choose(20,5)
```

```
## [1] 0.06501548
```

## Question 4

$P(AA) = 0.15$

$P(UA) = x$

$P(D) = 3x$

$P(AA) + P(UA) + P(D) = 0.15 + x + 3x = 1$

$4x + 0.15 = 1$

$4x = 0.85$

$x = 0.2125$

Therefore $P(AA) = 0.15, P(UA) = 0.2125, P(D) = 0.6375$.

Probability that the executive called from Chicago or is flying UA is $(ChicagoUA)/[(ChicagoUA) + (DallasAA) + (MinneapolisD)]$

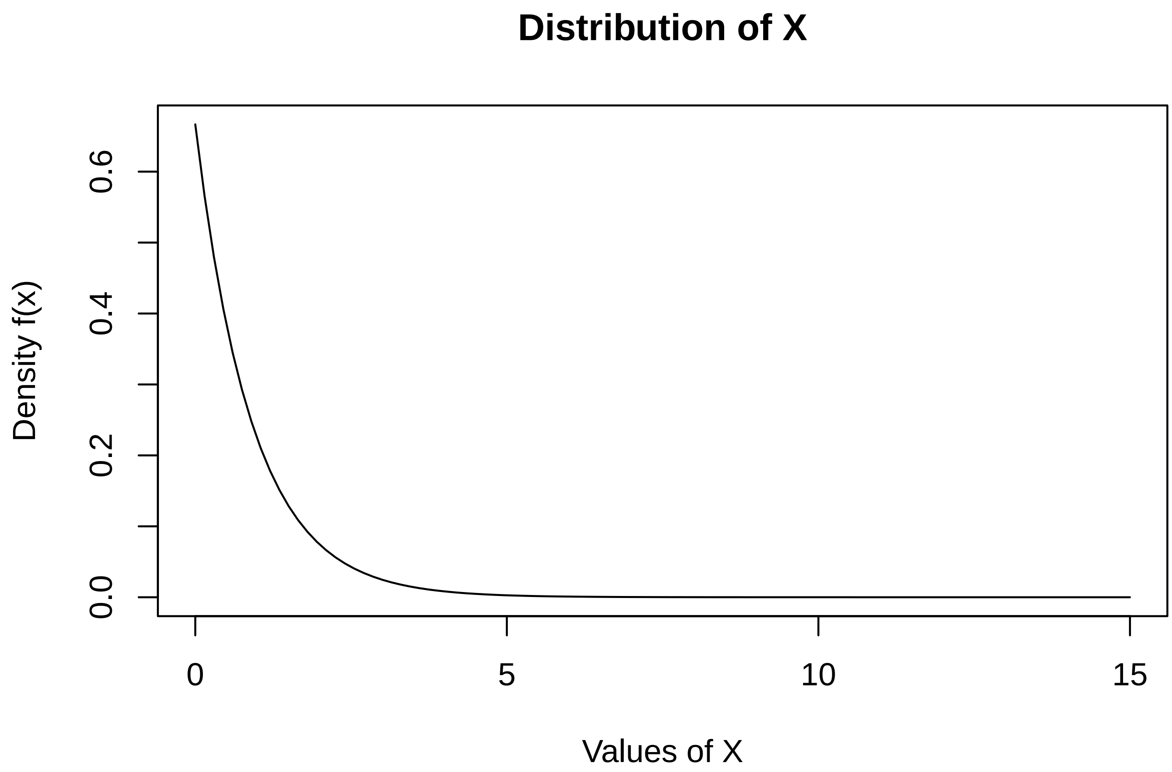
```
(0.2125*0.3)/((0.2125*0.3)+(0.15*0.15)+(0.6375*0.1))
```

```
## [1] 0.425
```

## Question 5

### a.

```
q5_func = function (x) {
  2/(3^(x+1))
}
plot(q5_func, 0, 15, xlab="Values of X", ylab="Density f(x)", main="Distribution of X")
```




**b.**

Taking the integral of the function $2/(3^{x+1})$ gives us $-(2 * 3^{-x-1})/ln(3))$. Plugging x=4 and x= 0 into $(-(2 * 3^{-x-1})/ln(3))$ will give the area under the curve for x values less than 4. We can then subtract this from 1 to get the probability that x > 4.

```
1- ((-((2*3^(-4-1))/log(3)))- (-((2*3^(-0-1))/log(3))))
```

```
## [1] 0.4006655
```

**c.**

```
q5_vec = seq(0, 100, by =1)
mean(q5_func(q5_vec))
```

```
## [1] 0.00990099
```

**d.**

```r
sd(q5_func(q5_vec))
```

```
## [1] 0.07000707
```

Using the method from b) to find the area under the curve:

#### e.

```r
mean_minus_sd <-mean(q5_func(q5_vec)) - sd(q5_func(q5_vec))
mean_plus_sd <- mean(q5_func(q5_vec)) + sd(q5_func(q5_vec))
(-((2*3^(-mean_plus_sd-1))/log(3))) - (-((2*3^(-mean_minus_sd-1))/log(3)))
```

```
## [1] 0.092424
```

## Question 6

#### a.

```r
dbinom(35, 50, 0.4)
```

```
## [1] 1.249428e-05
```

#### b.

The 40% statistic seems inaccurate. The probability for us to observe 35 out of the 50 believe truckers deserve no sympathies is 0.0012% which is extremely unlikely.

#### c.

We must assume that in the first 29 people, 9 of them have no sympathies for truckers. To do this we use a binomial distribution where there are 9 successes, n=29 and probability of success is 0.40. Multiplying this by 0.4 will be the probability that the 10th person is not sympathetic.

```r
pbinom(9,29,0.40)*0.40
```

```
## [1] 0.08587264
```

## Question 7

```r
tosses <- seq(0,10000,1)
toss_probs <-((0.6875)^(tosses-1))*(0.3125)
sum(tosses*toss_probs)
```

```
## [1] 3.2
```

The expected value of X is 3.2 tosses. This is obtained by finding the probability of each value of x and multiply x that probability. This means that x = 10 tosses is much greater than the expected amount of tosses. (x= 10 has a probability of approx. 0.0107).

## Question 8

### a.

To find the area under the curve between 1.91L and 1.83L, we subtract the pnorms for the respective values.

```
pnorm(1.91, mean = 1.89, sd = 0.05) - pnorm(1.83, mean = 1.89, sd = 0.05)
```

```
## [1] 0.5403521
```

### b.

To find the 90th percentile, we use qnorm and pass 0.90 into it. The 90th percentile means that 90% of soft-drinks fall under 1.9541L.

```
qnorm(0.90, mean = 1.89, sd = 0.05)
```

```
## [1] 1.954078
```

### c.

To find the proportion of bottle that will overflow, we find the area under curve where x values are greater than 2L.

```
1 - pnorm(2, mean = 1.89, sd = 0.05)
```

```
## [1] 0.01390345
```

### d.

```
pnorm(1.85, mean = 1.89, sd = 0.05)
```

```
## [1] 0.2118554
```

The probability that a single drink falls below 1.85L is ~0.2119. If we think of a drink falling below 1.85L as a "success", we can use a binomial distribution to find the probability of it between 5 and 10 drinks.

```
pbinom(10, size = 50, prob = pnorm(1.85, mean = 1.89, sd = 0.05)) - pbinom(5, size = 50, prob = pnorm(1
```

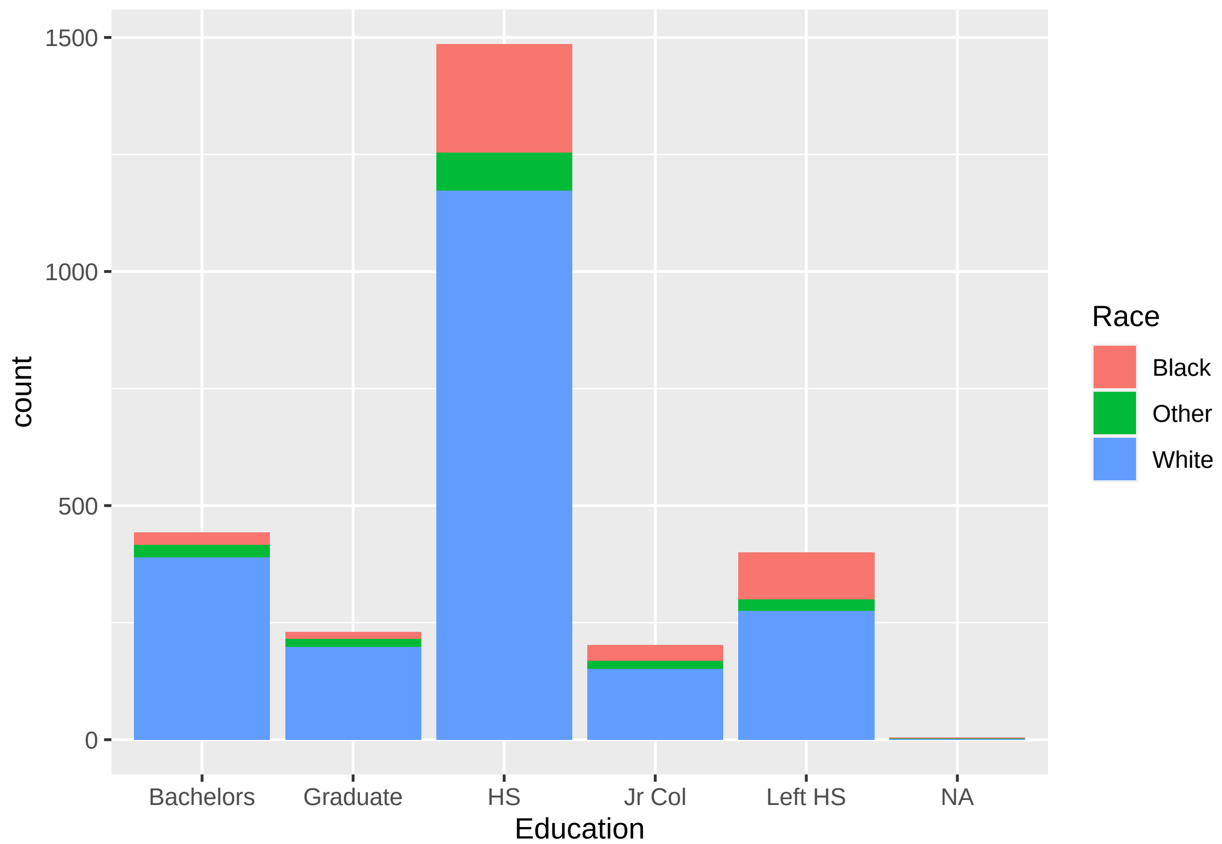
```
## [1] 0.4691327
```

## Question 9

```
gss = read.csv("http://people.ucalgary.ca/~jbstall/DataFiles/GSS2002.csv")
```

### a.

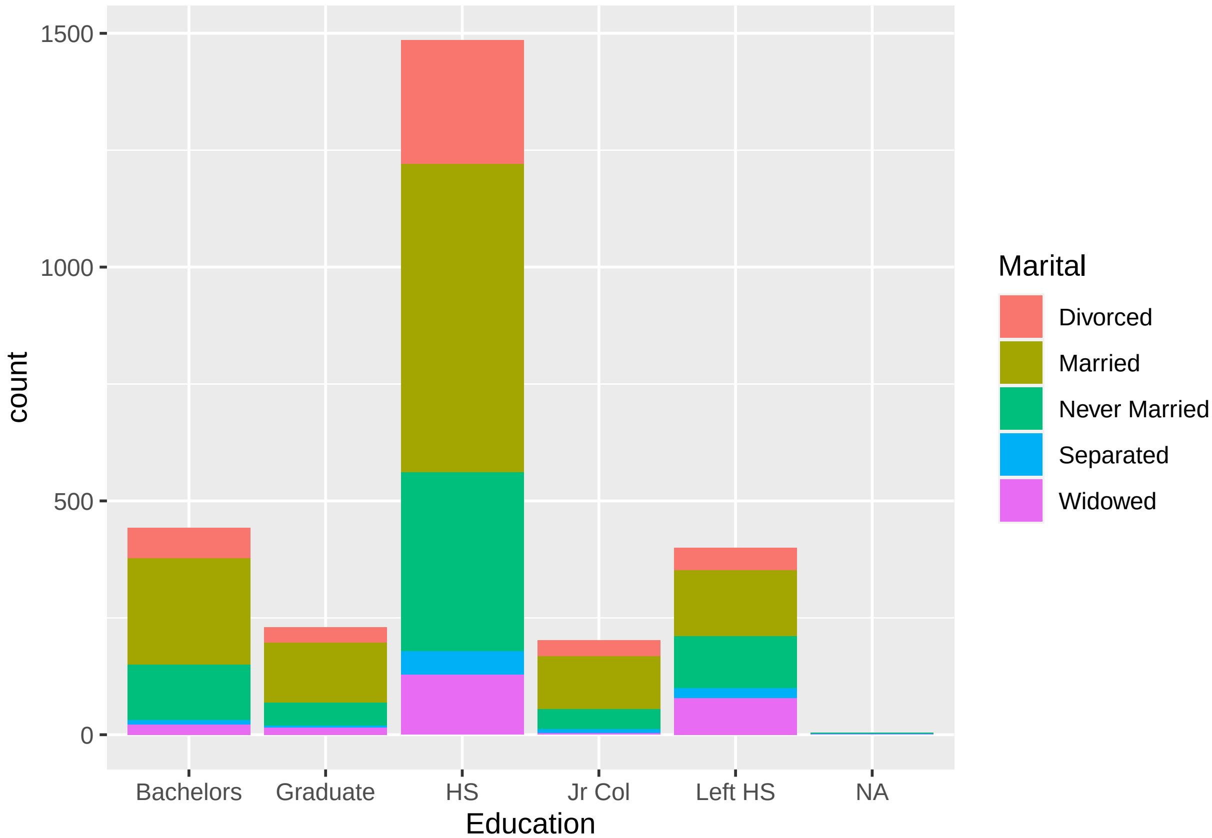
```r
library(ggplot2)
ggplot(gss, aes(x = Education, fill = Race)) + geom_bar(, na.rm = TRUE)
```




White people seem to be the most prevalent in the survey for all levels of education. Most people have at most a high school diploma. The majority of bachelor and graduate degree holders are white. There was a significant amount of black people who left HS compared to other education levels (besides HS).

**b.**

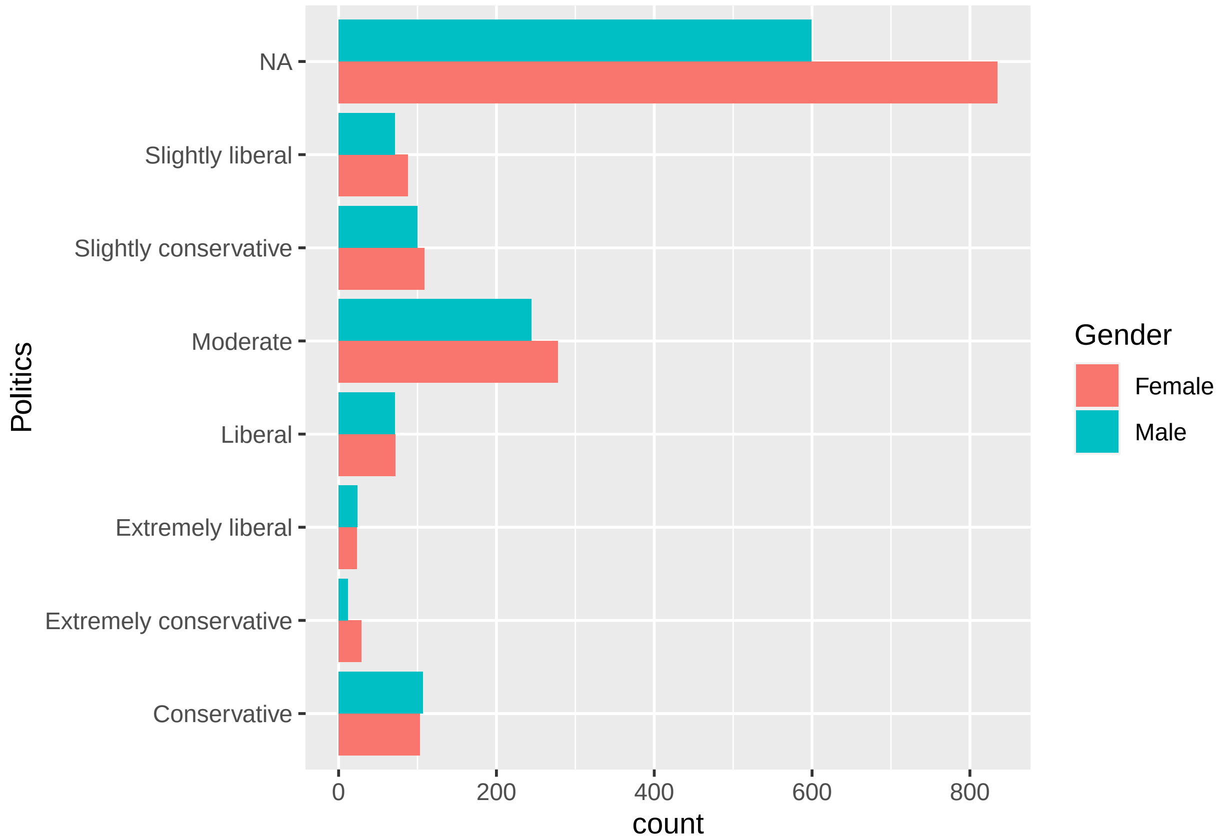
```r
ggplot(gss, aes(x = Education, fill = Marital)) + geom_bar(na.rm = TRUE)
```




Based on this visualization, divorced individuals mostly have at most a high school diploma. Widowed individuals are most prevelant in 'Left HS' and 'HS' categories. It looks like almost no seperated indiviuals went past high school.

**c.**

```
ggplot(gss, aes(y = Politics, fill = Gender)) + geom_bar(position = 'dodge',na.rm = TRUE)
```




Most people, regardless of gender do not have a political affiliation. Females are more represented in this visualization, as they have higher counts for almost every political category except for Conservative, which is more popular among men.

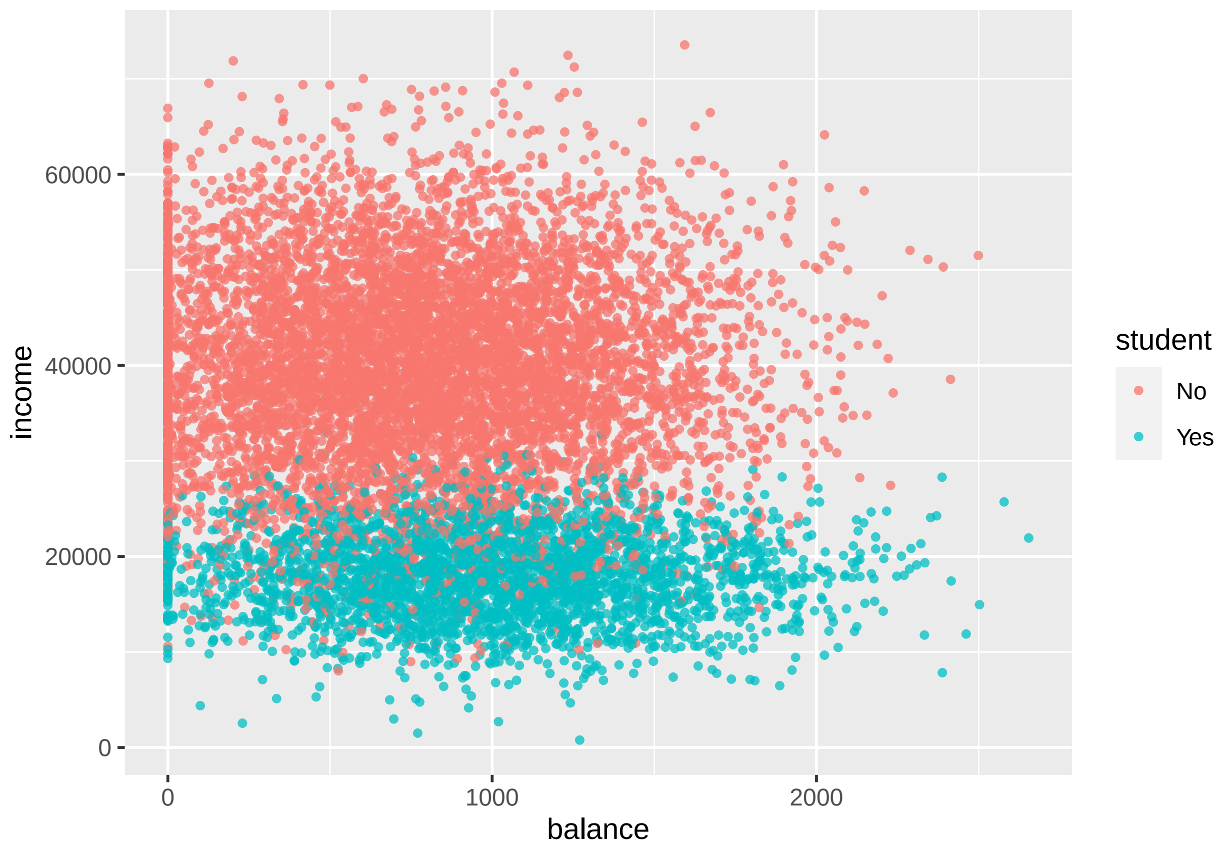
### Question 10

```
library(ISLR)
head(Default)
```

```
##   default student   balance    income
## 1      No      No  729.5265 44361.625
## 2      No     Yes  817.1804 12106.135
## 3      No      No 1073.5492 31767.139
## 4      No      No  529.2506 35704.494
## 5      No      No  785.6559 38463.496
## 6      No     Yes  919.5885  7491.559
```

**a.**

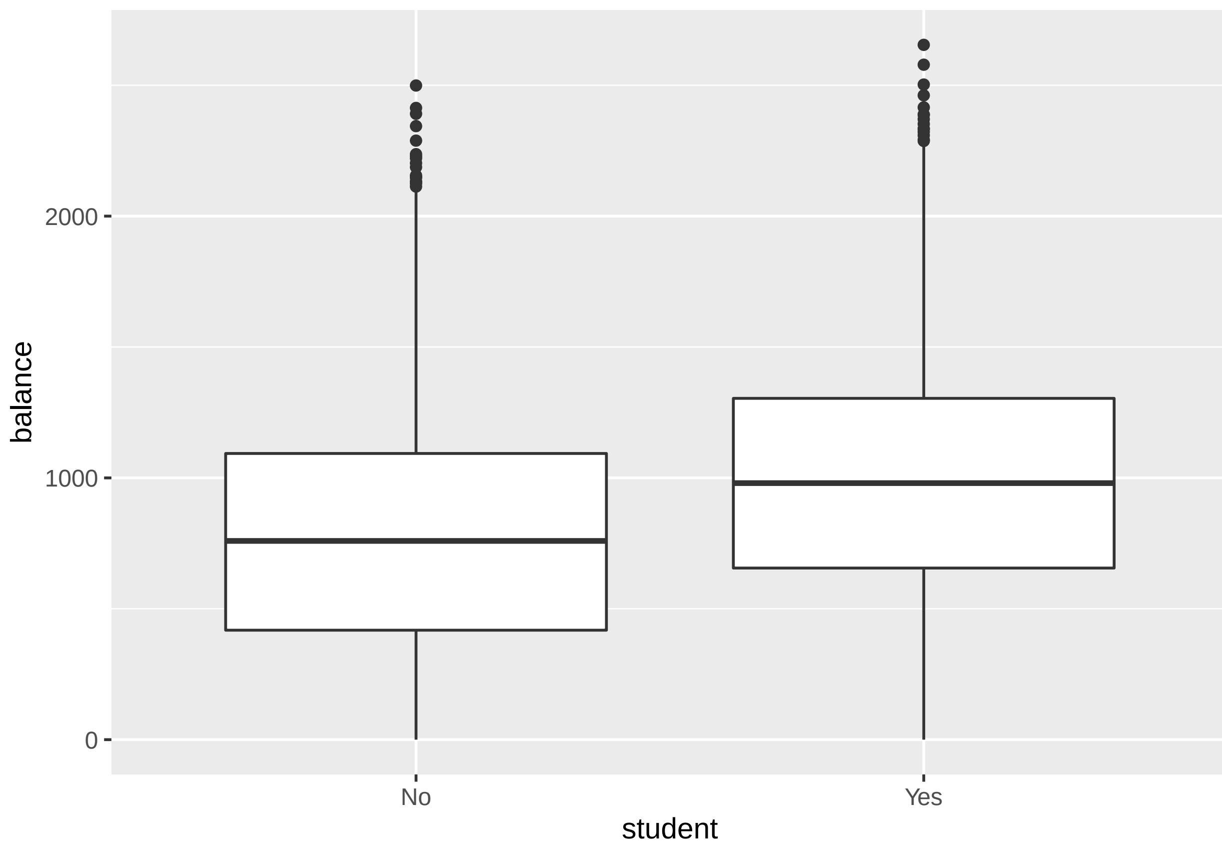
```
ggplot(Default, aes(x=balance, y=income, color = student)) + geom_point(position="jitter", alpha = 0.75
```




There seems to be 2 main clusters around a balance of 1,000 and income of 20,000, which is students, as well as a balance of 1,000 and income of 40,000, which is non-students.

**b.**

```
ggplot(Default, aes(x = student, y = balance ))+ geom_boxplot()
```




Students seem to have a slightly higher balance than non-students. Both students and non-students seem to have a similar IQR.

**c.**

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
default_student <- Default %>% group_by(student)
default_student %>% summarise(mean_balance = mean(balance), median_balance = median(balance), st.dev_ba
```

```
## # A tibble: 2 x 6
##   student mean_balance median_balance st.dev_balance percentile_05 percentile_95
```

```
##   <fct>            <dbl>          <dbl>          <dbl>          <dbl>          <dbl>
## 1 No                772.           759.           470.             0           1582.
## 2 Yes               988.           980.           483.           173.          1812.
```

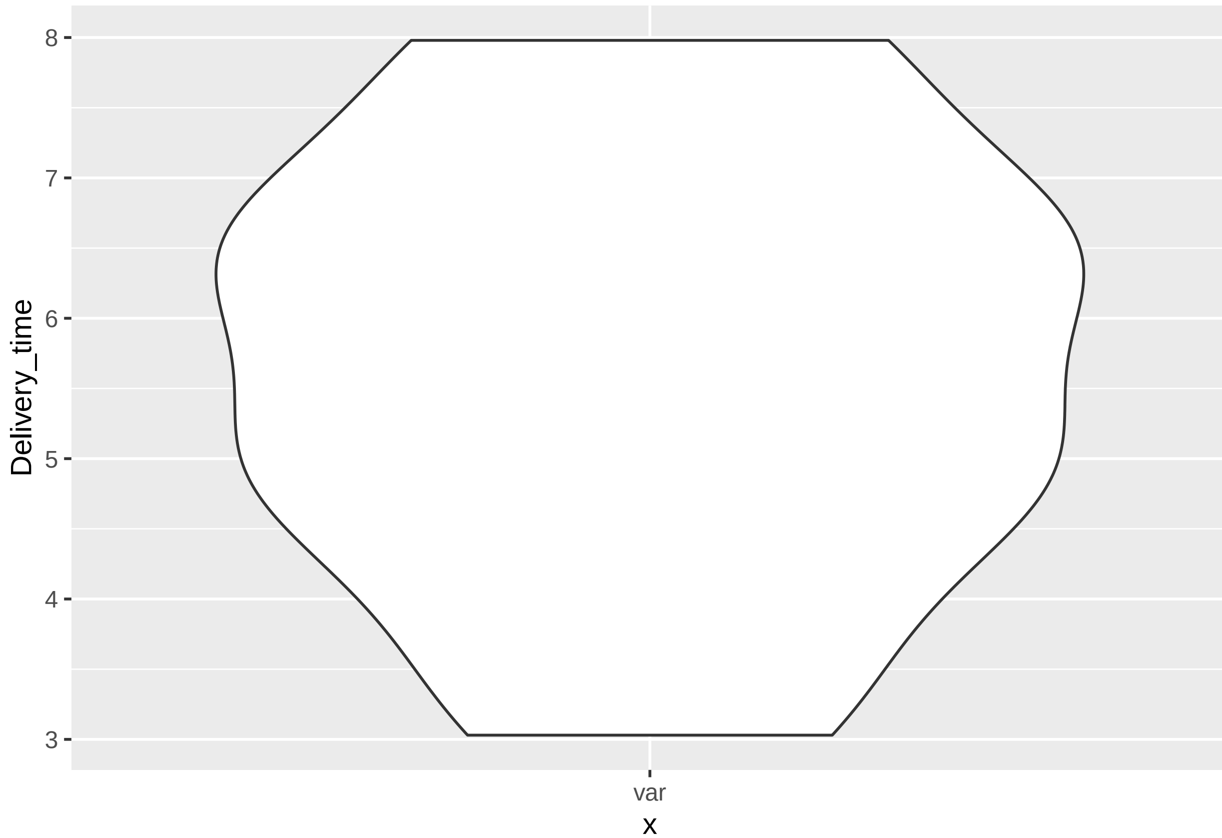
### Question 11

```r
q11_df <- read.csv("http://people.ucalgary.ca/~jbstall/DataFiles/Data602Assignment1Question11.csv")
q11_df
```

```
##    Delivery_time
## 1           3.03
## 2           6.33
## 3           6.50
## 4           5.22
## 5           3.56
## 6           6.76
## 7           7.98
## 8           4.82
## 9           7.96
## 10          4.54
## 11          5.09
## 12          6.46
```

#### a.

```r
ggplot(q11_df, aes(x = 'var', y = Delivery_time)) + geom_violin()
```




**b.**

```
mean(q11_df$Delivery_time)
```

```
## [1] 5.6875
```

```
median(q11_df$Delivery_time)
```

```
## [1] 5.775
```

```
sd(q11_df$Delivery_time)
```

```
## [1] 1.580369
```

```
quantile(q11_df$Delivery_time, probs = c(0.25, 0.75, 0.99))
```

```
##    25%    75%    99%
## 4.7500 6.5650 7.9778
```

**c.**

```
quantile(q11_df$Delivery_time,0.99)
```

```
##    99%
## 7.9778
```

The point of refund would be the 99th percentile, as this would leave the other 1% of deliveries to be refunded. The 99th percentile is 7.9778.