ECOM20001 ASSIGNMENT 1 COVER PAGE

Name	Student ID Number
Gia Han Ly	1074109
Rui Zheng	1084537

1. Summary statistics for amount ($), share_under25 and young

	Variables		
	Amount ($)	Share_under25	Young
Mean	319.220	0.473	0.301
Standard Deviation	374.967	0.045	0.459
Minimum	0.0	0.250	0.0
Maximum	2000.0	0.674	1.0

The average amount donated to the Democratic Party during the two elections is $319.22, and mean voters under the age of 25 is 0.473. Voted counties with a more youthful demographic (over 50% of voters under 25) have a mean of 0.301.

Furthermore the mean for Young implies that there are more results where Young=0, hence there is not an even distribution of 0 and 1 in the sample. As a result the sample median for Young would not be 0.5.

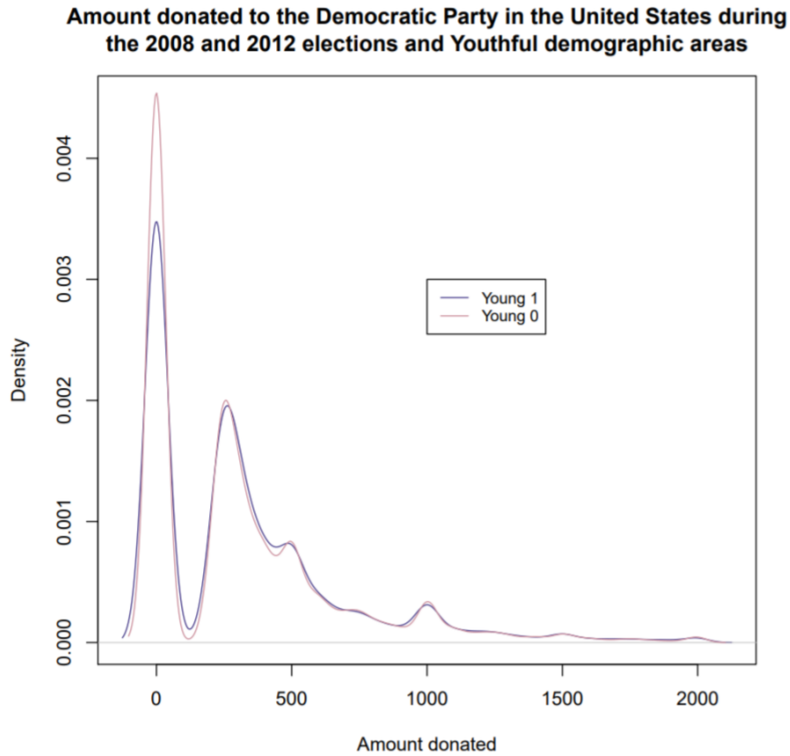
2. 95% Confidence interval (C.I)= [$\overline{x}$ - 1.960SE($\overline{x}$), $\overline{x}$ + 1.960SE($\overline{x}$)]

Mean(amount) = 319.2199
SE(amount) = 1.2715
95% C.I for amount
=[319.2199-1.960(1.2715), 319.2199+1.960(1.2715)
=[316.728, 321.712]

Mean(share_under25) = 0.47335
SE(share_under25) = 0.00015
95% C.I for share_under25
= [0.47335-1.960(0.00015), 0.47335+1.960(0.00015)]
= [0.473, 0.474]

Mean(young) = 0.3012
SE(young) = 0.0016
95% C.I for young
= [0.3012-1.960(0.0016), 0.3012+1.960(0.0016)]
= [0.298, 0.304]

3. Both densities graphs appear to be positively skewed, and bimodal at amount $0 and approximately $250. The densities exhibit the largest difference at amount = 0. Densities for young = 1 is approximately 0.0035, and young = 0 is approximately 0.045.



4. Hypothesis testing for difference in means at 5% significance level

H0: mean(amount if young=1) = mean(amount if young=0)

H1: mean(amount if young=1) mean(amount if young=0)

Hypothesis testing for difference in mean	***Actual results***
Mean(amount if Young=1)	327.403
Mean(amount if Young=0)	315.693
Difference in means	11.709
Degrees of freedom	49440
t-statistic	4.217
p-value	2.4780e-05
95% Confidence Interval	[6.267, 17.151]

T-statistic |4.217| > critical value (1.960), suggesting statistical significance; p-value0.0 < 0.05 also implies a strong evidence against the null hypothesis. Additionally, we are 95% confident that the difference in mean is between [6.267, 17.151], which excludes zero. Therefore, we reject null hypothesis at 5% significance level.

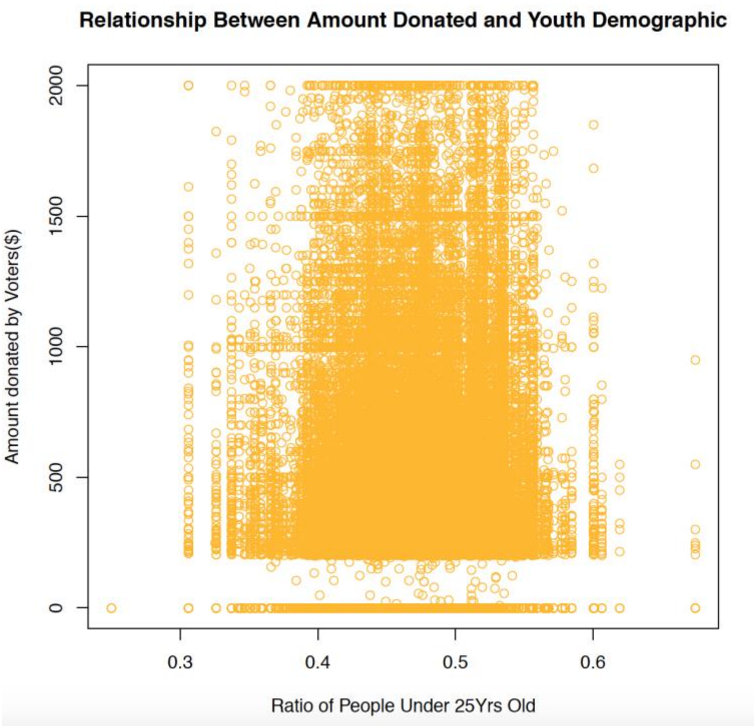
5. Hypothesis testing for difference in means at 5% significance level

H0: mean(amount_zero if young=1) = mean(amount_zero if young=0)
H1: mean(amount_zero if young=1) mean(amount_zero if young=0)

Hypothesis testing for difference in means	*Actual results*
Mean(amount_zero if young=1)	0.365
Mean(amount_zero if young=0)	0.392
Difference in means	0.0259
Degrees of freedom	50262
t-statistic	-7.255
p-value	4.087e-13
95% Confidence Interval	[-0.033, -0.019]

Absolute t-statistic |-7.255| > critical value (1.96); p-value 4.087e-13 < 0.05 which is within the rejection region and thus, there is sufficient evidence at 5% significance level to reject the null hypothesis. There is 95% confidence that the difference in mean is between [-0.033, -0.019].

6. The scatter plot appears to have cloud-like correlation, showing no observable relationship or correlation between the two variables; amount and share_under25. This is supported by the correlation coefficient (0.0153) of close to zero.



Correlation coefficient = 0.0153

7.

Single Linear Regression	$Amount_i$=259.0545+127.1045$share_under25_i$
sd(share_under25)	0.0452

When the demographic of under 25 equals to zero, the estimated amount donated is $259.05. For every 0.01 unit (1%) increase in share_under25, the amount donated is estimated to increase by $1.27. Lastly, the amount increases by $5.75 for every one standard deviation (0.0452) increase in the ratio of under 25 year old.

In the regression model, there exists only one independent variable (Share_under25i). In this case, one-unit increase implies a 1% increase in the ratio of under 25 year old, with no other independent variables or unit of measurements.

One-unit increase may also be changed to a 0.1 unit or 0.001 unit increase, increasing the amount donated by $12.71 or $0.13 respectively. Through this per unit interpretation, we can better identify how each incremental increase, in the ratio of under 25 year old, influences the donation amount towards the Democratic Party.

Furthermore, the standard deviation (sd) interpretation would be a better choice if there were multiple independent variables with different units of measurements. As the current regression coefficients are not standardized, we have to constantly multiply the sd with the regression coefficient to know the estimated increased amount.

Therefore, in this regression model, the sd approach does not provide more information and would require more calculations. Hence, in analysing the relationship between age levels and its impact on donation amounts, we would be better off using the ‘one-unit increase’ interpretation of the regression.

Appendix: R-code

```
#create binary variable young
as1_obama=read.csv("as1_obama.csv")
as1_obama$young=1*(as1_obama$share_under25>0.5)

#Question 1
summary(as1_obama)
sd(as1_obama$amount)
sd(as1_obama$young)
sd(as1_obama$share_under25)

#Question 2
#95% confidence interval of amount
#sample mean of amount
amount_mu=mean(as1_obama$amount)
#number of observation in amount
amount_nobs=length(as1_obama$amount)
#sample standard deviation of amount
amount_sd=sd(as1_obama$amount)
#standard error of sample mean of amount
amount_se=amount_sd/sqrt(amount_nobs)
#lower bound of the 95% CI
amount_CI95_low=amount_mu-1.96*amount_se
#upper bound of the 95% CI
amount_CI95_high=amount_mu+1.96*amount_se

paste("95% CI Lower Bound:",amount_CI95_low)
paste("95% CI Upper Bound:",amount_CI95_high)

#95% confidence interval of share_under25
#sample mean of share_under25
share_mu=mean(as1_obama$share_under25)
#number of observation in share_under25
share_nobs=length(as1_obama$share_under25)
#sample standard deviation of share_under25
share_sd=sd(as1_obama$share_under25)
#standard error of sample mean of share_under25
share_se=share_sd/sqrt(share_nobs)
#lower bound of the 95% CI
share_CI95_low=share_mu-1.96*share_se
```

```
#upper bound of the 95% CI
share_CI95_high=share_mu+1.96*share_se

paste("95% CI Lower Bound:",share_CI95_low)
paste("95% CI Upper Bound:",share_CI95_high)

#95% confidence interval of young
#sample mean of young
young_mu=mean(as1_obama$young)
#number of observation in young
young_nobs=length(as1_obama$young)
#sample standard deviation of young
young_sd=sd(as1_obama$young)
#standard error of sample mean of young
young_se=young_sd/sqrt(young_nobs)
#lower bound of the 95% CI
young_CI95_low=young_mu-1.96*young_se
#upper bound of the 95% CI
young_CI95_high=young_mu+1.96*young_se

paste("95% CI Lower Bound:",young_CI95_low)
paste("95% CI Upper Bound:",young_CI95_high)


#Question 3
#create pdf file
pdf("amount_v_shareunder25.pdf")
#plot density graphs for amount given young =1 and amount given young = 0
plot(density(as1_obama$amount[as1_obama$young==1]),
main = "Amount donated to the Democratic Party in the United States during the 2008 and 2012 elections and Youthful demographic areas",
ylim=c(0, 0.0045),
xlab="Amount donated",
ylab="Density",
col='darkslateblue')
lines(density(as1_obama$amount[as1_obama$young==0]),col='pink3')
legend(1000, 0.003, legend=c("Young 1", "Young 0"),
       col=c("darkslateblue", "pink3"), lty=1:1, cex=0.8,
        text.font=1, bg='white')
dev.off()
```

```
#Question 4
#mean of amount if young==1
mean(as1_obama$amount[as1_obama$young==1])
#mean of amount if young==0
mean(as1_obama$amount[as1_obama$young==0])
#difference of sample mean
mean(as1_obama$amount[as1_obama$young==1])-
mean(as1_obama$amount[as1_obama$young==0])
#t-test for difference of sample mean
t.test(as1_obama$amount[as1_obama$young==1],as1_obama$amount[as1_obama
$young==0])


#Question 5
#create binary variable amount_zero (amount >0 or amount =0)
as1_obama$amount_zero=1*(as1_obama$amount==0)
summary(as1_obama$amount_zero)
#mean of amount_zero given young =1
mean(as1_obama$amount_zero[as1_obama$young==1])
#mean of amount_zero given young =0
mean(as1_obama$amount_zero[as1_obama$young==0])
#difference of sample mean
mean(as1_obama$amount_zero[as1_obama$young==0])-
mean(as1_obama$amount_zero[as1_obama$young==1])
#t-test for difference of sample mean
t.test(as1_obama$amount_zero[as1_obama$young==1],as1_obama$amount_zero[
as1_obama$young==0])


#Question 6
#Create PDF file
pdf("as1_obama_Q6")
#Scatter plot with Amount on Y-axis & Share_25 on X-axis
plot(as1_obama$share_under25,as1_obama$amount,
    main="Relationship Between Amount Donated and Youth Demographic",
    xlab="Ratio of People Under 25Yrs Old",
    ylab="Amount donated by Voters($)",
    col="darkgoldenrod1",
    pch=1)
dev.off()
#Correlation Coefficient between (Amount, Share_25)
cor(as1_obama$share_under25,as1_obama$amount)
```

```
#Question 7
#Single linear regression of amount on demographic of youth
amount_reg=lm(amount~share_under25,data=as1_obama)
#Print regression output
summary(amount_reg)
#Standard deviation of share_under25
sd(as1_obama$share_under25)
```