Study of How People Donates

Our sample has four variables: income, sex, age and amount. The summary of these variables is as following:

```
##      income           sex                 age            amount      
##  Min.   : 51200   Length:94          Min.   :25.00   Min.   : 58.79  
##  1st Qu.: 69200   Class :character   1st Qu.:41.25   1st Qu.:109.03  
##  Median : 82400   Mode  :character   Median :49.00   Median :133.94  
##  Mean   : 90417                      Mean   :47.61   Mean   :147.89  
##  3rd Qu.:105200                      3rd Qu.:53.00   3rd Qu.:167.46  
##  Max.   :204800                      Max.   :71.00   Max.   :323.41  
```

1. Under the Null hypothesis that Donation amount increases with the age, we want to test the relationship between donation amount of young(<50) and old(>50) with 95% confidence.

```
##  [1] 133.23 146.87 101.57 141.22  68.67 264.51 108.84  58.79 114.96 128.74
## [11]  84.58 153.16 106.85 187.17 133.85  83.69 175.91 278.16 247.97 107.86
## [21] 130.66 127.42 117.78  94.06 106.11 147.14 135.41 119.63 105.89 120.16
## [31] 159.60 124.95  93.02 171.17  88.84 126.49  87.65 164.59 232.99 123.30
```

```
##  [1] 182.15  97.46 154.53 139.93 162.42 121.43 323.41 138.25 164.98 161.95
## [11] 159.80 205.14  95.83 198.13 159.64 159.16 214.78 134.02 230.02 144.73
## [21] 117.98 128.96 136.82 270.58 103.96 168.28 170.18  82.87 136.55 275.89
## [31] 109.60 100.37 305.82 123.49 117.74 127.08 231.63 225.64  97.21 146.20
## [41] 130.88  69.35 227.79 117.06 123.15 100.47 219.44  89.88 136.20
```

Since this is not a paired data and we do not have any information about population variances, we use two-sample t-test for the given data set.

```
## 
##  Welch Two Sample t-test
## 
## data:  Donation_amount_older and Donation_amount_younger
## t = -1.9912, df = 86.776, p-value = 0.0248
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -3.770952
## sample estimates:
## mean of x mean of y 
##  135.0865  157.9353 
```

The $t-statistics$ for this test is -1.9912 and $p-value$ is 0.0248. Since the $p-value < alpha$, we reject the null hypothesis. Thus, we have sufficient evidence to conclude that the donation amount doesn't increase with age. In other word, our finding is donation amount decreases with age.

2. We want to compute the Correlation coefficient between age and donation amount and perform the regression analysis to interpolate the donation amount of other ages within the range of our sample.

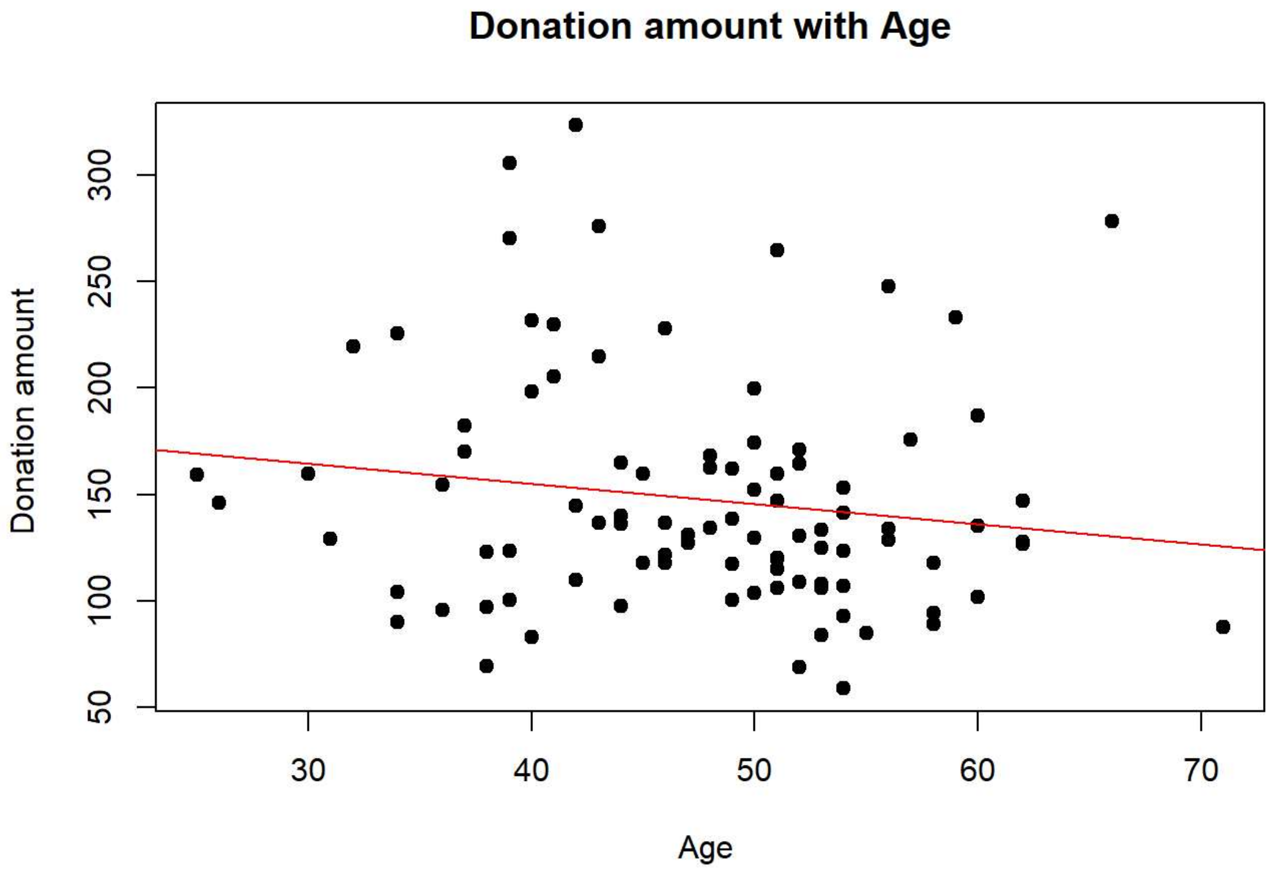
Since we have sufficient evidence to conclude that the donation amount doesn't increase with age, we want to observe the linear relationship between donation amount and age. For that, we calculate Pearson's correlation coefficient.

```
## [1] -0.1541701
```

We find the correlation coefficient is -0.154 which represents a weak negative correlation between age and donation amount. Although we find that the donation amount decreases with age, the correlation between donation amount and age is very weak (almost zero). To validate furthur if the result of our sample is applicable to the population, we want to test it against the null hypothesis that correlation coefficient of population is 0 with 95% confidence.

```
## 
##  Pearson's product-moment correlation
## 
## data:  donors$age and donors$amount
## t = -1.4966, df = 92, p-value = 0.1379
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.34597952  0.05000904
## sample estimates:
##        cor 
## -0.1541701
```

The $t - statistics$ for the test is -1.4966 and $p - value$ is 0.1379. Since the $p - value > alpha$, we fail to reject the null hypothesis. Thus, we don't have sufficient evidence to conclude that the there is no correlation between age and donation amount. To visualize the weak negative relationship between age and donation amount, we plot a scatterplot between age and donation amount.



3. Under the null hypothesis that the sex and donation amount are associated, we want to test against the hypothesis with 95% confidence.

We divide our sample into two donation amount groups: lower(<150) and higher(>150). Among the sample of donors, we want to find whether the donation amount and sex are independent. We make a contingency table of income level with sex as follows:

```
## [1] 31
```

```
## [1] 28
```

```
## [1] 24
```

```
## [1] 11
```

Donation_Sex	Female	Male	Total
Lower	31	28	59
Higher	24	11	35
Total	55	39	94

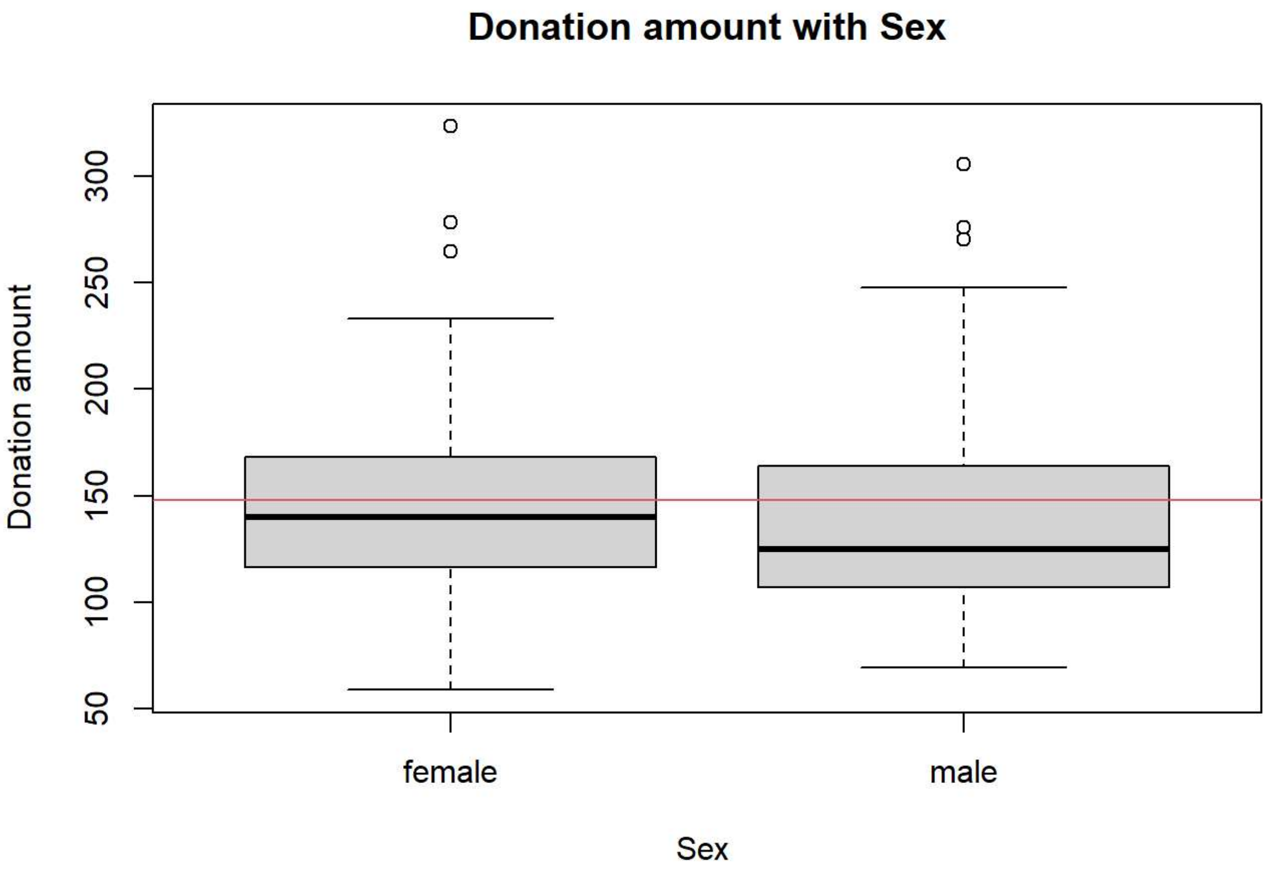
```
##      [,1] [,2]
## [1,]   31   28
## [2,]   24   11
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  observed
## X-squared = 2.3251, df = 1, p-value = 0.1273
```

The Chi-squared value is 2.3251 and $p - value$ is 0.1273. Since $p - value > \alpha$, we fail to reject null hypothesis. Thus, we do not have sufficient evidence to conclude that sex and donation amount are associated.

4. Under the Null hypothesis that female donates less, we want to test the relationship between donation amount of male and female with 95% confidence.

Before beginning the hypothesis testing, we summarize the male and female donation amount using boxplot as follows:



From this boxplot, we find that the median donation amount of female is more than donation amount of male. Also, median donation amount of female is close to the mean donation amount as shwon by red line.

Now, we test our claim against null hypothesis that the median difference between donation amount by female and male is same using Wilcoxon signed rank test.

```
##
##  Wilcoxon rank sum test
##
## data:  donors$amount[donors$sex == "female"] and donors$amount[donors$sex == "male"]
## W = 1203, p-value = 0.3166
## alternative hypothesis: true location shift is not equal to 0
```

The smallest sum for the test is 1203 and the $p - value$ is 0.3166. Since, $p - value > \alpha$, we fail to reject null hypothesis. Thus, we don't have sufficient evidence to conclude that the median difference between male and female donation amount is differet.

In order to furthur robust our hypothesis test, we test against the null hypothesis that female and male donates equally using t-test.

```
##
##  Welch Two Sample t-test
##
## data:  donors$amount[donors$sex == "female"] and donors$amount[donors$sex == "male"]
## t = 0.18226, df = 75.403, p-value = 0.8559
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -21.25121  25.53189
## sample estimates:
## mean of x mean of y
##  148.7785  146.6382
```

The $t - statistics$ for the test is 0.18226 and the $p - value$ is 0.8559. Since the $p - value > alpha$, we fail to reject the null hypothesis. Thus, we don't have sufficient evidence to conclude that the mean difference in donation amount between male and female is different.

From both Wilcoxon rank test and Welch two-sample test we find that there is no significant difference between donation amount of male and female.

5. Under the null hypothesis that the mean donation amount of different income levels are same, we want to use ANOVA test to claim against this hypothesis.

For this test, we divide our sample into three income levels: low(<75000), medium(75000-100000) and high(>100000). After that, we use ANOVA to compare mean donation amount of different income levels and test if at least on of them is different.

```
##  [1]  97.46 101.57  68.67  58.79 103.43  95.83  84.58 117.98  83.69 128.96
## [11] 107.86 130.66 103.96  82.87 136.55 109.60 100.37  94.06 106.11 127.08
## [21] 119.63  97.21  69.35 124.95 123.15 100.47  89.88  88.84  87.65
```

```
##  [1] 182.15 133.23 154.53 146.87 139.93 162.42 129.77 141.22 121.43 138.25
## [11] 108.84 114.96 128.74 159.64 159.16 106.85 134.02 144.73 133.85 136.82
## [21] 168.28 170.18 127.42 117.78 123.49 117.74 147.14 135.41 105.89 146.20
## [31] 130.88 120.16  93.02 117.06 126.49 136.20 164.59 152.20 123.30
```

```
##  [1] 174.25 323.41 164.98 264.51 161.95 159.80 205.14 198.13 214.78 199.77
## [11] 153.16 230.02 187.17 175.91 278.16 247.97 270.58 275.89 305.82 231.63
## [21] 225.64 159.60 227.79 219.44 171.17 232.99
```

The F-value using ANOVA test is 1.3817 and $p - value$ is 0.02428. Since $p - value > \alpha$, we reject null hypothesis. Thus, we have sufficient evidence to conclude that at least one of the mean donation amount is different.

6. At last, we want to compute a prediction equation for donation amount as a function of income, age and sex and interpolate the result for different age and sex within the range of our sample.

We model a system where donation amount is an independent variable depending on other dependent variables income, age and sex.

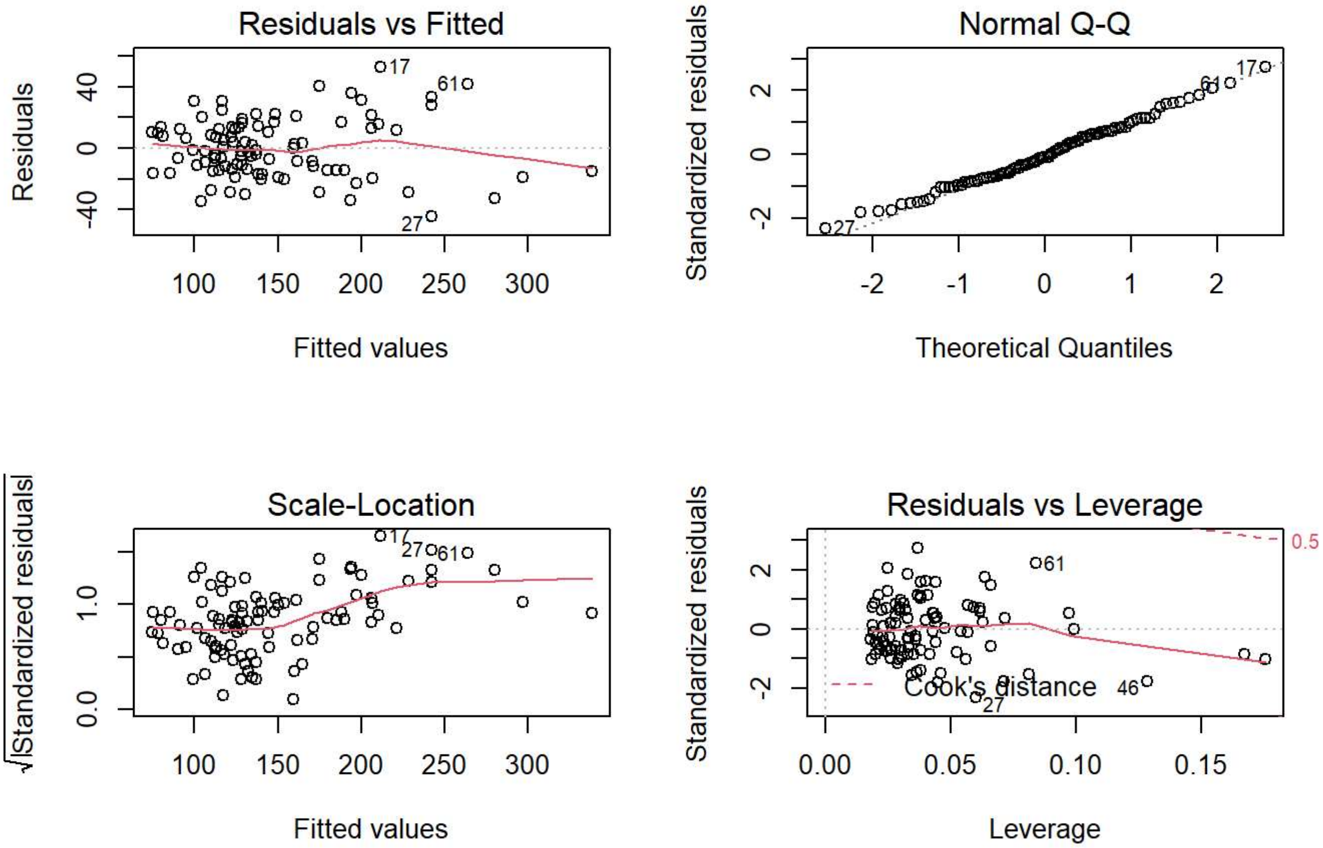
```
##
## Call:
## lm(formula = y ~ x + a + s)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.316 -14.255  -1.613  13.330  52.723
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.566e+01  1.324e+01   3.449 0.000859 ***
## x            1.636e-03  6.632e-05  24.669  < 2e-16 ***
## a           -1.003e+00  2.336e-01  -4.292 4.45e-05 ***
## smale        4.913e+00  4.199e+00   1.170 0.245100
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.67 on 90 degrees of freedom
## Multiple R-squared:  0.8745, Adjusted R-squared:  0.8703
## F-statistic: 209.1 on 3 and 90 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## x          1 234124  234124 605.0371 < 2.2e-16 ***
## a          1   8036    8036  20.7667 1.626e-05 ***
## s          1    530     530   1.3688    0.2451
## Residuals 90  34826     387
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We find the regression equation for this model is $y = 45.66 + 0.001636x - 1.003a + 4.913s$ where variables y, x, a and s are donation amount, income, age and sex respectively. The 95% confidence interval for each of the slope values are:- Income: (0.00150421, 0.001767707), Age: (-1.466510751, -0.538511895) and Sex: (-3.429511180 13.255492157). Based on our model, we test the hypothesis that $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ at the $\alpha = 0.05$ significance level for each slope value and find the $p - value$ for slope of the income, sex and age to be 2.2e-16, 1.626e-05 and 0.2451 respectively. The $p - value$ for income and age slope is less than $alpha$ whereas that of sex is greater than α. Thus, we have sufficient evidence to conclude $\beta_l \neq 0$ for income and age but we do not have sufficient evidence to conclude $\beta_l \neq 0$ for sex.

```
##                   2.5 %       97.5 %
## (Intercept) 19.357386316 71.961466037
## x            0.001504214  0.001767707
## a           -1.466510751 -0.538511895
## smale       -3.429511180 13.255492157
```

In order to test the regression fit of our model, we plot following plots:



From the normal qq plot, it can be seen that the assumption regarding normality holds true.The residual plot shows a small deviation from homoscedasticity. For smaller values of x,a and s; residuals are randomly distributed about the zero residual line while they are distant from the line for higher values. So applying linear regression model in this case needs caution.

We have performed series of statistical investigation to understand how people donates based on age, sex and income. The whole investigation process is divided into three subsections: 1. Inestigation of relationship between age and donation amount, 2. Investigation of relationship between sex and donation amount and 3. Affect of income along with age and sex on donation amoount. In the first subsection, we use t-test to test against the null hypothesis that donation amount increases with age and we find sufficient evidence to void this claim. We also use correlation analysis to validate our result. Both of these test result a consistent output. There could be several reasoning for the it. One of the major reason could be young people being a working age group obviously earns more. The more a person earns, the more is the tendency to donate. In the second subsection, we use Chi-squared test and Wilcoxon signed-test along with Welch two-sample test to test against the null hypothesis that female donates less and we find sufficient evidence to void this claim. It’s surprising how both parametric and non-parametric test resulted the same result. In this modern age, female are as empowered as male both economically and in social status. This result is one of the clear illustration that the taboo of female donating less has become a story. In the last subsection, we divide the whole income sample into three category: low(<75000), medium(75000-100000) and high(>100000). We use one way ANOVA test to test against the null hypothesis that donation amount of each income category is same. We collect sufficient evidence against this hypothesis. We also perform regression analysis to look for the affect of income, age and sex on donation amount. After all of the analysis we find that both income and age affects the donation amount but sex doesn’t. This result is consistent with the above two analyses and an obvious result for the modern society where people with more income and working age intends to donate more and females have equal capacity to earn and donate as male.

Overall, this study resulted a mirror of the modern day society that people with more income and working age intends to donate more whereas sex of a person is not a parameter anymore to distinguish the donation tendency.