STATS101A - HW2

```r
air <- read.csv("airfareW22.csv")
air
```

```
##      X City Fare Distance
## 1    1    9  309     1204
## 2    2    5   93      190
## 3    3    7  291     1102
## 4    4   15  231      818
## 5    5   17  429     1813
## 6    6   11   90      184
## 7    7    1  360     1463
## 8    8    6  141      393
## 9    9    6  141      393
## 10  10    1  360     1463
## 11  11    5   93      190
## 12  12   13  477     1828
## 13  13    4  111      270
## 14  14    5   93      190
## 15  15   11   90      184
## 16  16   11   90      184
## 17  17   13  477     1828
## 18  18    8  183      578
## 19  19    7  291     1102
## 20  20    4  111      270
## 21  21   13  477     1828
## 22  22    8  183      578
## 23  23   16   54       90
## 24  24   15  231      818
## 25  25   14   84      179
## 26  26   15  231      818
## 27  27    6  141      393
## 28  28   12  162      502
## 29  29    1  360     1463
## 30  30   17  429     1813
## 31  31    6  141      393
## 32  32   14   84      179
## 33  33    8  183      578
## 34  34   17  429     1813
## 35  35    8  183      578
## 36  36   10  300     1138
## 37  37   10  300     1138
## 38  38   14   84      179
## 39  39   13  477     1828
## 40  40   14   84      179
## 41  41    4  111      270
## 42  42   10  300     1138
```

```
## 43 43   17  429      1813
## 44 44    7  291      1102
## 45 45    1  360      1463
## 46 46   16   54        90
## 47 47    6  141       393
## 48 48   12  162       502
## 49 49   16   54        90
## 50 50    5   93       190
```

```r
#A) 50 rows x 4 columns
dim(air)
```

```
## [1] 50  4
```

```r
#C) The ordinary straight line regression model can be improved through a transformation by taking the
```

```r
#Question 2

diamond <- read.csv("DiamondsW22.csv")
diamond
```

```
##        X Size Price
## 1      1 0.17   353
## 2      2 0.20   498
## 3      3 0.25   750
## 4      4 0.25   655
## 5      5 0.15   316
## 6      6 0.15   315
## 7      7 0.17   350
## 8      8 0.12   223
## 9      9 0.12   223
## 10    10 0.17   350
## 11    11 0.16   336
## 12    12 0.16   339
## 13    13 0.18   462
## 14    14 0.20   498
## 15    15 0.17   346
## 16    16 0.17   346
## 17    17 0.16   339
## 18    18 0.18   468
## 19    19 0.25   750
## 20    20 0.18   462
## 21    21 0.23   553
## 22    22 0.18   468
## 23    23 0.25   678
## 24    24 0.25   678
## 25    25 0.25   655
## 26    26 0.18   443
## 27    27 0.12   223
## 28    28 0.15   298
## 29    29 0.16   328
## 30    30 0.15   316
## 31    31 0.29   860
```

```
## 32   32 0.33   945
## 33   33 0.18   468
## 34   34 0.15   316
## 35   35 0.17   352
## 36   36 0.18   419
## 37   37 0.17   346
## 38   38 0.17   345
## 39   39 0.23   595
## 40   40 0.23   553
## 41   41 0.18   462
## 42   42 0.17   318
## 43   43 0.15   316
## 44   44 0.27   720
## 45   45 0.17   350
## 46   46 0.25   675
## 47   47 0.23   595
## 48   48 0.32   918
## 49   49 0.15   287
## 50   50 0.20   498
## 51   51 0.33   945
## 52   52 0.17   318
## 53   53 0.27   720
## 54   54 0.15   316
## 55   55 0.15   316
## 56   56 0.32   918
## 57   57 0.32   919
## 58   58 0.23   553
## 59   59 0.16   339
## 60   60 0.18   438
## 61   61 0.17   350
## 62   62 0.25   655
## 63   63 0.16   336
## 64   64 0.17   345
## 65   65 0.23   553
## 66   66 0.27   720
## 67   67 0.29   860
## 68   68 0.17   350
## 69   69 0.18   438
## 70   70 0.18   438
## 71   71 0.15   316
## 72   72 0.15   322
## 73   73 0.23   595
## 74   74 0.15   315
## 75   75 0.17   346
## 76   76 0.15   316
## 77   77 0.18   468
## 78   78 0.27   720
## 79   79 0.28   823
## 80   80 0.20   498
## 81   81 0.15   316
## 82   82 0.16   338
## 83   83 0.16   338
## 84   84 0.20   498
## 85   85 0.16   339
```

```
## 86   86 0.33   945
## 87   87 0.18   325
## 88   88 0.17   346
## 89   89 0.15   316
## 90   90 0.25   675
## 91   91 0.26   663
## 92   92 0.16   339
## 93   93 0.29   860
## 94   94 0.25   675
## 95   95 0.25   750
## 96   96 0.25   655
## 97   97 0.23   595
## 98   98 0.15   316
## 99   99 0.28   823
## 100 100 0.16   336
```

```r
###Part 1
#A)
dim(diamond)
```

```
## [1] 100   3
```

```r
#B)
diamondlm <- lm(Price ~ Size, data = diamond)
summary(diamondlm)
```

```
##
## Call:
## lm(formula = Price ~ Size, data = diamond)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -90.183 -24.595  -2.981  10.621  78.608
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -243.64      12.19  -19.99   <2e-16 ***
## Size         3660.13      58.47   62.60   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.36 on 98 degrees of freedom
## Multiple R-squared:  0.9756, Adjusted R-squared:  0.9754
## F-statistic:  3919 on 1 and 98 DF,  p-value: < 2.2e-16
```

```r
#Our R^2 value is 0.9756 and adjusted R^2 value being 0.9754, meaning 97.54% of the data points accurat

#C)
#Weakness of this data would be the extreme negative values, represented by the negative intercept, min
```

```r
###Part 2
#A)
diamond$Size <- sqrt(diamond$Size)
diamond$price <- sqrt(diamond$Price)

lm2 <- lm(Price ~ Size, data = diamond)
summary(lm2)
```

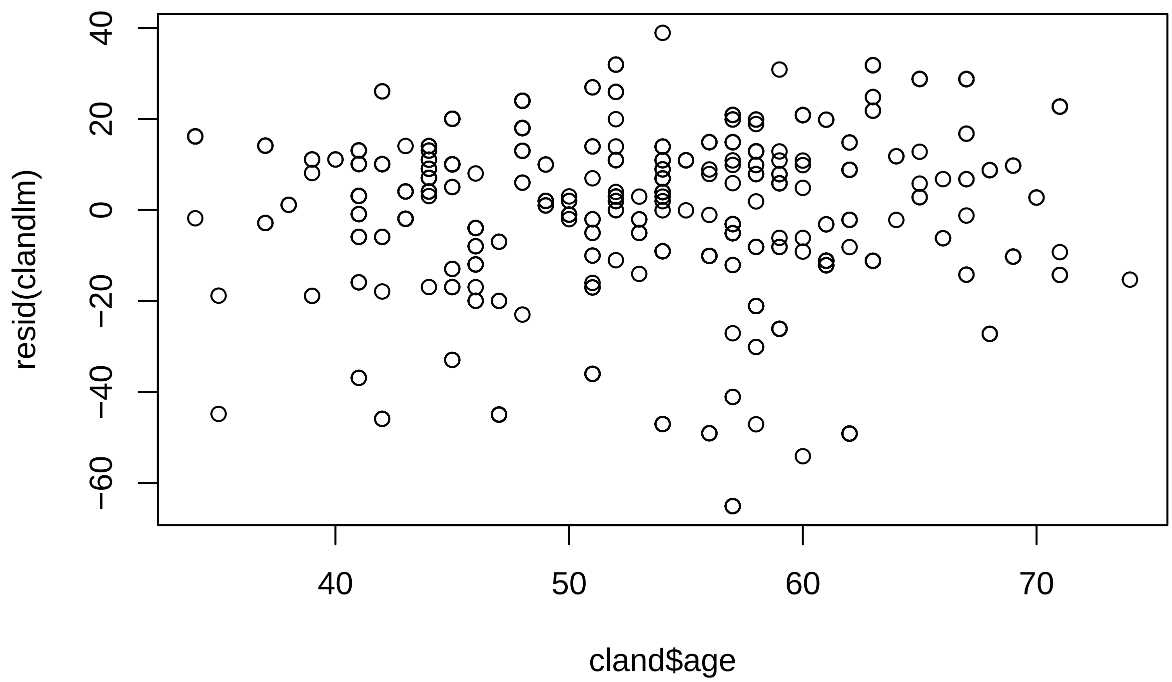
```
##
## Call:
## lm(formula = Price ~ Size, data = diamond)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -98.141 -26.134  -2.415  18.077  70.311
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1014.01      26.62  -38.10   <2e-16 ***
## Size         3387.40      59.31   57.11   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34.28 on 98 degrees of freedom
## Multiple R-squared:  0.9708, Adjusted R-squared:  0.9705
## F-statistic:  3262 on 1 and 98 DF,  p-value: < 2.2e-16
```

```r
#B)
#lm2 is extremely similar to lm1. This transformation that occured in lm2 was square rooting the size a

#C) Both models are fairly similar, but model 1 provides a more accurate and more reasonable simple lin
```

```r
#Question 3
library(ggplot2)
cland <- read.csv("ClevelandW22.csv")
clandlm <- lm(maxheartrate ~ age, data = cland)


plot(cland$age, resid(clandlm))
```




```r
#B)
anova(clandlm)
```

```
## Analysis of Variance Table
##
## Response: maxheartrate
##              Df Sum Sq Mean Sq F value    Pr(>F)
## age           1  30328 30328.1  80.503 < 2.2e-16 ***
## Residuals 398 149939   376.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#To calculate F value from R^2 we can use that we know R^2 is SSR/SST. We then divide our SSR/DF so 303

#Our null hypothesis for ANOVA is testing H_0: Beta_0 = Beta_1 = 0

#Based on our F table and ANOVA output, we will reject our null hypothesis.
#MSE and Se^2 = 376.7
var(cland$maxheartrate)* (1-.1682)
```

```
## [1] 375.8043
```

```r
# So it's quite a good approximation, with 375.8043 being the output
```

```r
#C)
# R^2 Adjusted is .1662 compared to R^2 .1682 so quite a small difference

#D)
# We can clearly see assumption of normality is violatied, the residuals are far off the line on the QQ

#E)
lp <- hatvalues(clandlm)
rs <- rstandard((clandlm))

#F)
#diagPlot(lp)
#diagPlot(rs)
```

```r
#A)
#SS 2671; 159236 - MS 1335.5; 53078
#B) Leverage points: Observation 2 - iii
#C) Outlier Observation 1,3 - ii
```