1. Read in the data from the csv file

```r
reactionTimeDf <- read.csv(file = 'D21125621.csv')
```

—

2. Show summaries the data and define the different data types in the file.

```r
summary(reactionTimeDf)
```

```
##        X               Age             Sex                Height
##  Min.   :  1.00   Min.   :20.01   Length:300         Min.   :151.5
##  1st Qu.: 75.75   1st Qu.:33.73   Class :character   1st Qu.:162.4
##  Median :150.50   Median :44.02   Mode  :character   Median :167.8
##  Mean   :150.50   Mean   :44.90                      Mean   :169.7
##  3rd Qu.:225.25   3rd Qu.:56.76                      3rd Qu.:175.7
##  Max.   :300.00   Max.   :69.78                      Max.   :196.8
##   ReactionTime    AGE_GROUP
##  Min.   :160.2   Length:300
##  1st Qu.:393.6   Class :character
##  Median :472.4   Mode  :character
##  Mean   :472.9
##  3rd Qu.:555.8
##  Max.   :831.1
```

- X: Ordinal Data. Those are integers numbers that seems to represent an ID of each measurement so although numeric, the numeric difference between them doesn't seem to be usable to infer any extra information apart from an order.
- Age: Interval data. Represent a number in years.
- Sex: Categorical data that has the character values "F" and "M" representing Female and Male respectively

```r
reactionTimeDf$Sex[!duplicated(reactionTimeDf$Sex)]
```

```
## [1] "F" "M"
```

- Height: Interval data. Person's height in centimeters.
- ReactionTime: Interval data. Time to take a reaction in ms (time unit was inferred from question 6).
- Age Group: Categorical data. One person can only belong to one of those groups, either "20-40" or "40-70"

—

3. Calculate the mean, and standard deviation of the age, height and reaction time data.

```r
meansAndSdsForIntervals <- list(
ageMean = mean(reactionTimeDf$Age),
ageStandardDeviation = sd(reactionTimeDf$Age),

heightMean = mean(reactionTimeDf$Height),
heightStandardDeviation = sd(reactionTimeDf$Height),

reactionTimeMean = mean(reactionTimeDf$ReactionTime),
reactionTimeStandardDeviation = sd(reactionTimeDf$ReactionTime)
)
print(meansAndSdsForIntervals)
```

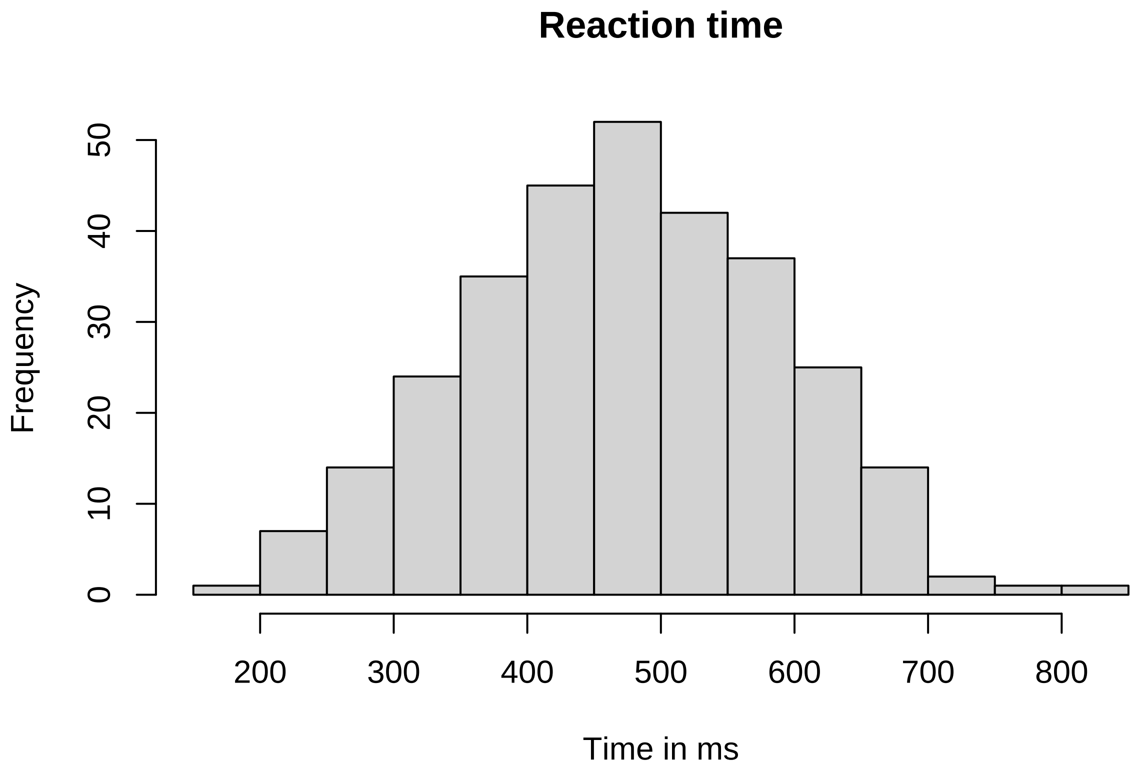
```
## $ageMean
## [1] 44.8983
##
## $ageStandardDeviation
## [1] 14.1418
##
## $heightMean
## [1] 169.6758
##
## $heightStandardDeviation
## [1] 9.29366
##
## $reactionTimeMean
## [1] 472.9268
##
## $reactionTimeStandardDeviation
## [1] 114.5484
```

—

4. a) Plot a histogram of the

i) reaction time data

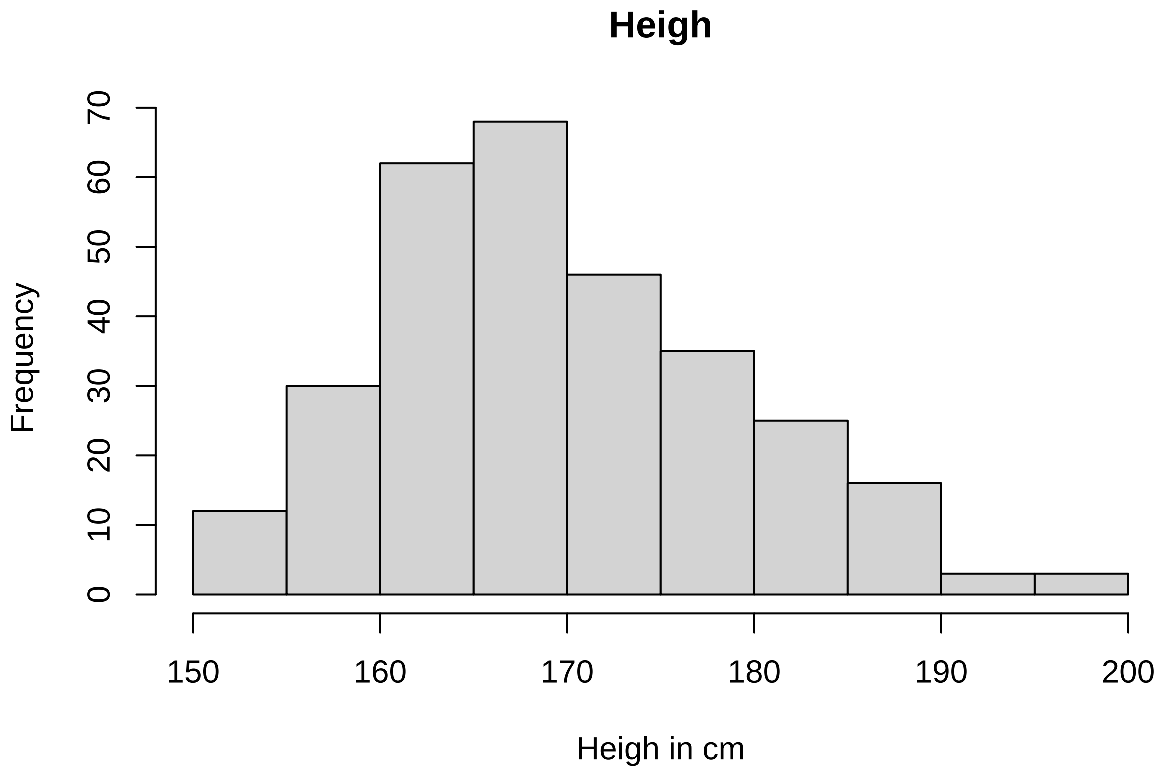
```r
hist(reactionTimeDf$ReactionTime, main = "Reaction time", xlab = "Time in ms")
```



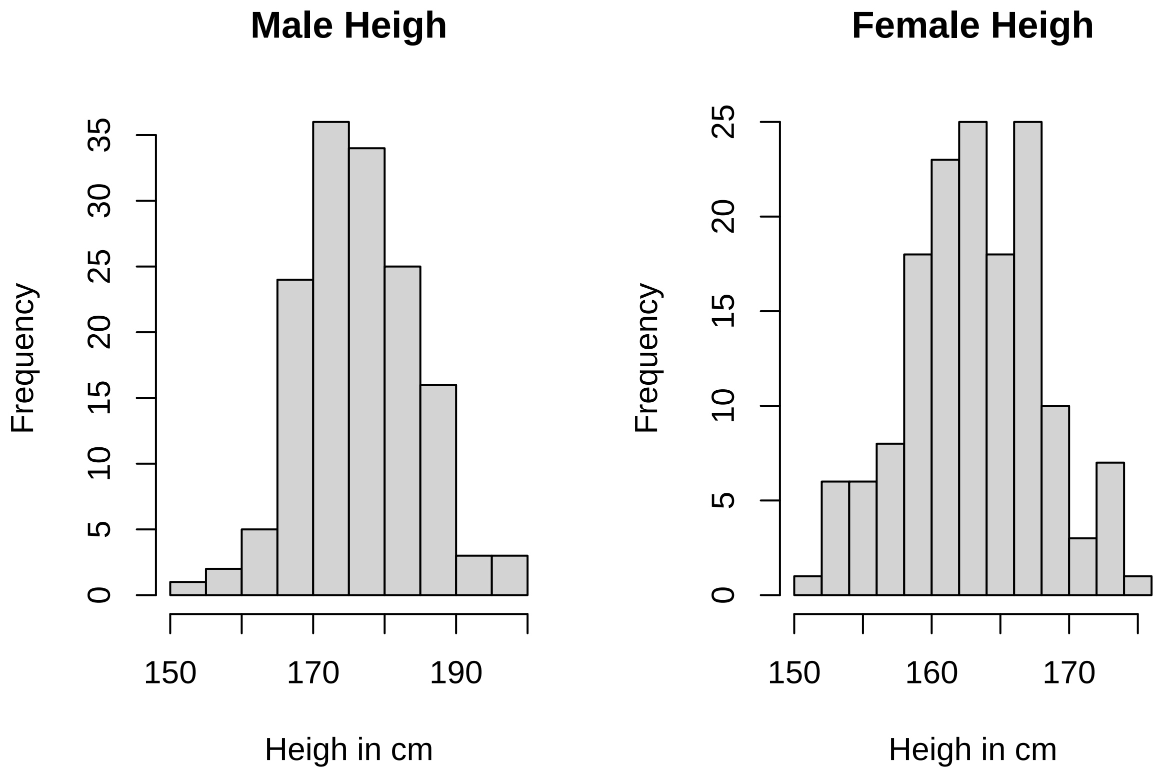
- 4.a.i.i Reaction Time measurements seem to obey a normal distribution with mean and median pretty close to each other.

ii) height data

```
hist(reactionTimeDf$Height, main = "Heigh", xlab = "Heigh in cm")
```



```
par(mfrow = c(1, 2))
hist(reactionTimeDf$Height[reactionTimeDf$Sex == 'M'], main = "Male Heigh", xlab = "Heigh in cm")
hist(reactionTimeDf$Height[reactionTimeDf$Sex == 'F'], main = "Female Heigh", xlab = "Heigh in cm")
```

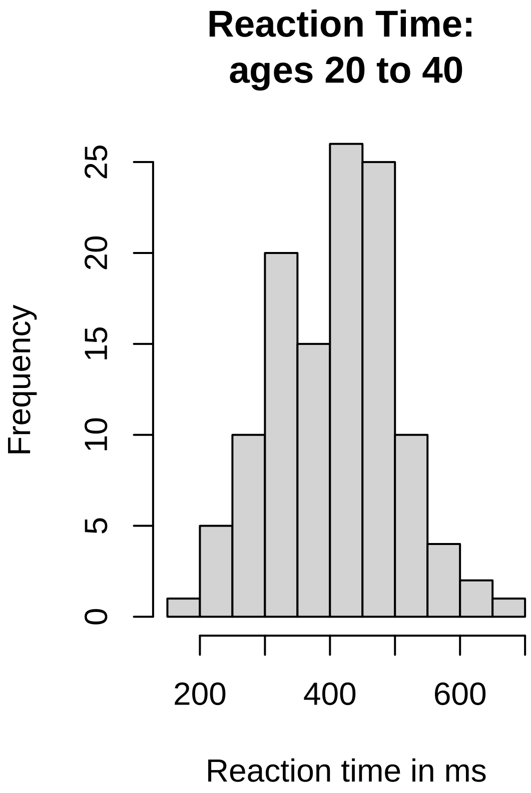


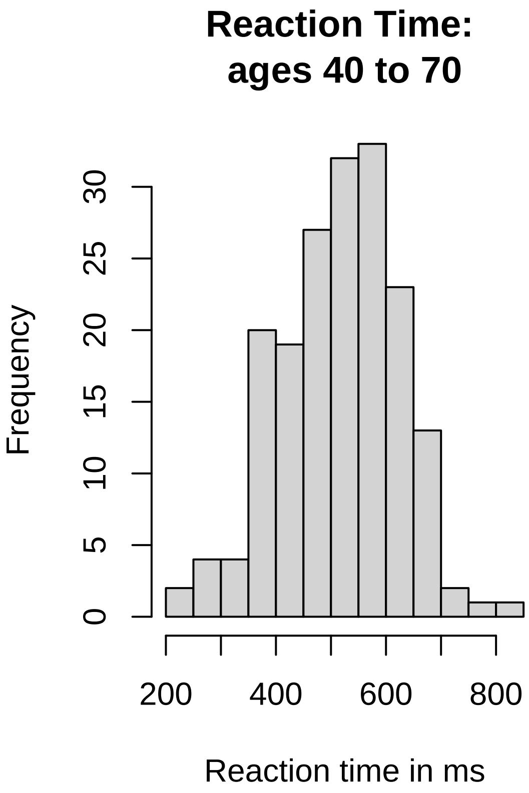
- 4.a.ii.i Plotting Height Data for the entire data set show a right skewed distribution but, if we plot separately male and female height we can notice a normal distribution for each revealing that males are, on average, taller than females.

---

4. b) Plot a histogram of the reaction time data for each age group (20-40yrs, 40-70yrs)

```
par(mfrow = c(1, 2))
hist(reactionTimeDf$ReactionTime[reactionTimeDf$AGE_GROUP == '20-40'], main = "Reaction Time: \nages 20
hist(reactionTimeDf$ReactionTime[reactionTimeDf$AGE_GROUP == '40-70'], main = "Reaction Time: \nages 40
```





- 4.b.i. Both distributions look like a normal distribution, we can notice that the average reaction time on the older group is considerably longer that on the younger group.

6. Hypothesis testing: Given that the historical populations mean reaction time is 450ms. Conduct a hypothesis test stating the Null H0 and alternative $H\alpha$, calculating the test statistics, stating your criteria for rejection, and your conclusion for the reaction time data of the:

c) 20-40 year old group.

1. H_0 The average reaction time is equal to 450ms
2. H_α The average reaction time for the 20 to 40 years old population is less than the historical average of 450
3. Test Statistic:

```
t.test(reactionTimeDf$ReactionTime[reactionTimeDf$AGE_GROUP == '20-40'], mu=450, alternative = "less", 
```

```
##
##  One Sample t-test
##
## data:  reactionTimeDf$ReactionTime[reactionTimeDf$AGE_GROUP == "20-40"]
## t = -4.7763, df = 118, p-value = 2.59e-06
## alternative hypothesis: true mean is less than 450
## 95 percent confidence interval:
##      -Inf 422.5934
## sample estimates:
## mean of x
##  408.0232
```

- Also Done step by step:

```
mu <- 450
reaction_20_40 <- reactionTimeDf$ReactionTime[reactionTimeDf$AGE_GROUP == '20-40']
x_bar_20_40 <- mean(reaction_20_40)
s_20_40 <- sd(reaction_20_40)
n_20_40 <- length(reaction_20_40)
z = (x_bar_20_40 - mu) / (s_20_40 / sqrt(n_20_40))
```

4. Calculate rejection region:

- For 95% one tailed test

```
qnorm(.05)
```

```
## [1] -1.644854
```

5. State conclusions:

- The test statistic -4.776 is smaller than the -1.645 found for the rejection region therefore, the null hypothesis (H_0) got rejected.

d) 40-70 year old group.

1. H_0 The average reaction time is equal to 450ms

2. H_α The average reaction time for the 40 to 70 years old population is more than the historical average of 450

3. Test Statistic:

```
t.test(reactionTimeDf$ReactionTime[reactionTimeDf$AGE_GROUP == '40-70'], mu=450, alternative = "greater
```

```
##
##  One Sample t-test
##
## data:  reactionTimeDf$ReactionTime[reactionTimeDf$AGE_GROUP == "40-70"]
## t = 8.3561, df = 180, p-value = 8.531e-15
## alternative hypothesis: true mean is greater than 450
## 95 percent confidence interval:
##  502.6187      Inf
## sample estimates:
## mean of x
##  515.5982
```

- Also Done step by step:

```r
mu <- 450
reaction_40_70 <- reactionTimeDf$ReactionTime[reactionTimeDf$AGE_GROUP == '40-70']
x_bar_40_70 <- mean(reaction_40_70)
s_40_70 <- sd(reaction_40_70)
n_40_70 <- length(reaction_40_70)
z = (x_bar_40_70 - mu) / (s_40_70 / sqrt(n_40_70))
z
```

```
## [1] 8.356071
```

4. Calculate rejection region:

- For 95% one tailed test

```r
qnorm(.95)
```

```
## [1] 1.644854
```

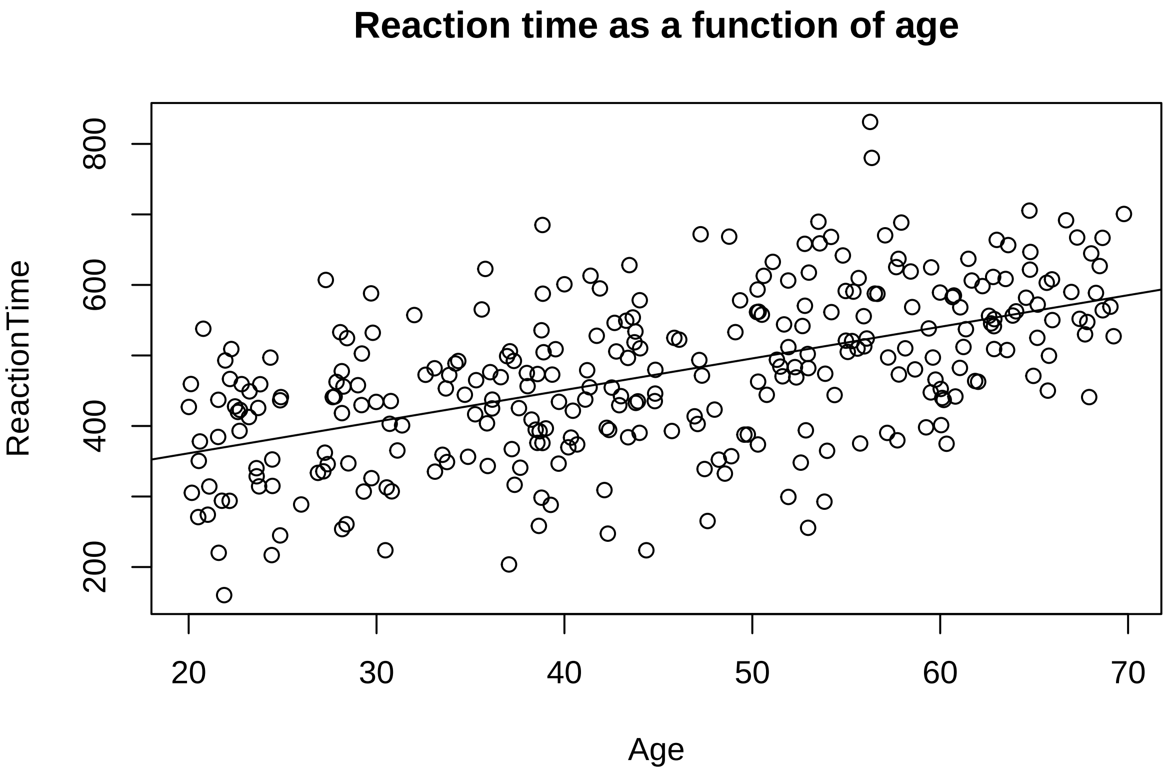
5. State conclusions:

- The test statistic 8.356071 is greater than the 1.645 found for the rejection region therefore, the null hypothesis (H_0) got rejected.

7.e) Plot a scatterplot of the reaction time data (y-axis) as a function of age (x-axis).

```r
reactionTimeModel<-lm(ReactionTime~Age,data=reactionTimeDf)
plot(ReactionTime~Age,data=reactionTimeDf, main = "Reaction time as a function of age")

abline(reactionTimeModel)
```



f) Conduct a linear regression analysis of reaction time as a function age and interpret the results.

```
summary(reactionTimeModel)
```

```
##
## Call:
## lm(formula = ReactionTime ~ Age, data = reactionTimeDf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -253.489  -66.477    0.642   62.272  307.249
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 271.6667    18.3942   14.77   <2e-16 ***
## Age           4.4826     0.3908   11.47   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 95.57 on 298 degrees of freedom
## Multiple R-squared:  0.3063, Adjusted R-squared:  0.3039
## F-statistic: 131.6 on 1 and 298 DF,  p-value: < 2.2e-16
```

- The model explains 30.63% of the variance of reaction time and, as p-value is < 2.2e-16, it’s statistically relevant. It’s also shown that the Intercept and Age are statistically significant with $* * *$.

- The intercept is at 271.6667 ms and each extra year a person has adds more 4.4826 to the predicted value so, the model indicates a positive correlation between age and reaction time. We can also observe this correlation present in the plot.
- The model also shows the standard error for the intercept and age estimates. With that, we can assert, with 95% confidence, that those estimates are between the following values:

$$interceptConfInt = 271.6667 \pm 1.96(18.3942) = (235.6141, 307.7193)$$

$$ageConfInt = 4.4826 \pm 1.96(0.3908) = (3.716632, 5.248568)$$

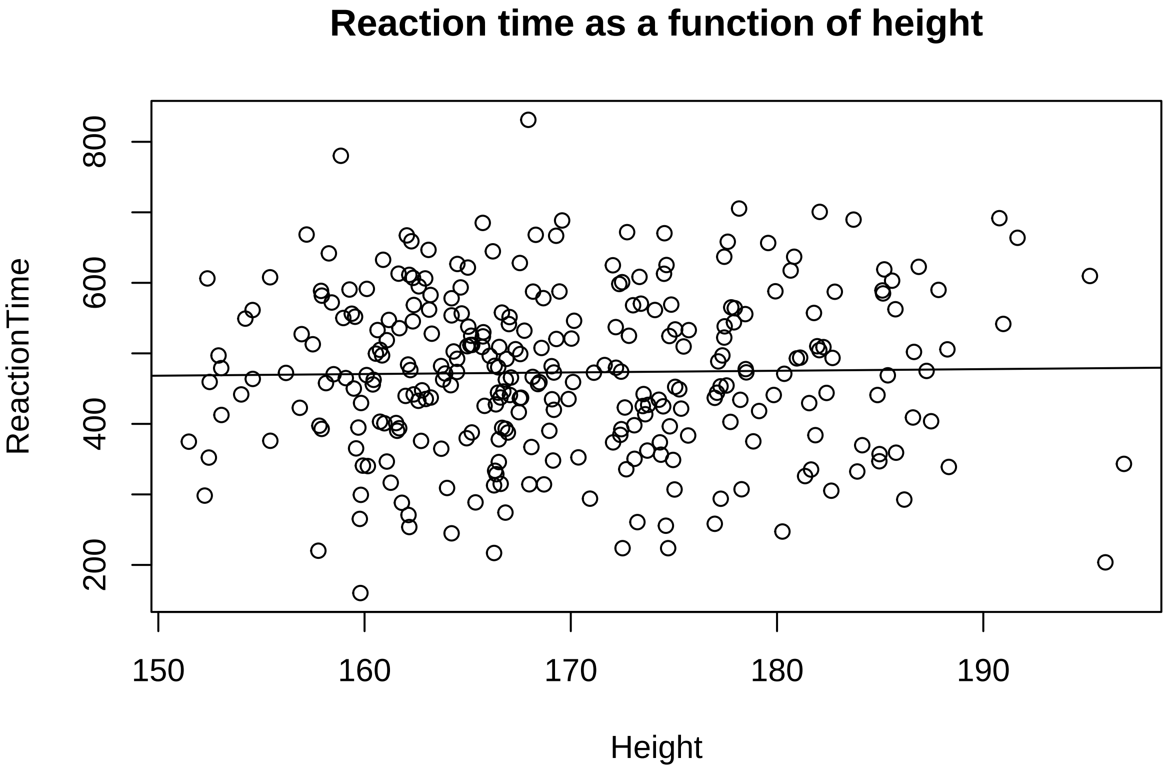
- With the data given by this model, we can predict the Reaction Time of a person using the following formula:

$$personReactionTime = 271.6667 + 4.4826 * PersonAge$$

7.a) Plot a scatterplot of the reaction time data (y-axis) as a function of height (x-axis).

```r
reactionTimeModelHeight<-lm(ReactionTime~Height,data=reactionTimeDf)
plot(ReactionTime~Height,data=reactionTimeDf, main = "Reaction time as a function of height")

abline(reactionTimeModelHeight)
```



7. b) Conduct a linear regression analysis of reaction time as a function of height and interpret the results.

```
summary(reactionTimeModelHeight)
```

```
##
## Call:
## lm(formula = ReactionTime ~ Height, data = reactionTimeDf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -310.46  -77.64   -0.37   81.70  358.63
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 433.4502   121.3067   3.573 0.000411 ***
## Height        0.2327     0.7139   0.326 0.744718
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 114.7 on 298 degrees of freedom
## Multiple R-squared:  0.0003563,  Adjusted R-squared:  -0.002998
## F-statistic: 0.1062 on 1 and 298 DF,  p-value: 0.7447
```

- The model p-value of 0.7447 shows that it’s not statistically relevant so there is not much point in investigate other attributes present here.
- The scatter plot, with a near straight line, also suggests that is not possible to predict Reaction Time based on Height.

9. g) Conduct a Logistic regression analysis of age group as a function of reaction time and interpret the results.

```
ageGroupLog <- glm(as.factor(AGE_GROUP) ~ ReactionTime,family=binomial("logit"),data=reactionTimeDf)
summary(ageGroupLog)
```

```
##
## Call:
## glm(formula = as.factor(AGE_GROUP) ~ ReactionTime, family = binomial("logit"),
##     data = reactionTimeDf)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3587  -0.9873   0.5096   0.8608   2.0763
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -4.336847   0.665361  -6.518 7.12e-11 ***
## ReactionTime  0.010299   0.001435   7.176 7.15e-13 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 402.98  on 299  degrees of freedom
## Residual deviance: 332.24  on 298  degrees of freedom
## AIC: 336.24
##
## Number of Fisher Scoring iterations: 4
```

- This model was created by putting age group variable as a factor creating two levels 20-40 and 40-70. The function as.factor, by default, orders levels by ascending order so, belong to 20-40 group is considered our 0 (fail), and belong to 40-70 is considered our 1 (success).
- Z value on Reaction Time is fairly high, indicating that we can predict age group using Reaction Time.
- We can predict the the age group using the following calculation.

1. Calculate the log odds:
$$\eta_i = -4.336847 + (0.010299) * ReactionTime)$$
2. Calculating the odds:
$$e^{\eta_i}$$
3. Calculating the probability:
$$p_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}}$$

- For a person with reaction time of 200. This would be the probability that it’s in age group of 40-70

```
log_odd <- -4.336847 + (0.010299) * 200
odd <- exp(log_odd)
prob40_70 <- odd / (1 + odd)
paste("Probability of being on age group 40-70:", prob40_70)
```

```
## [1] "Probability of being on age group 40-70: 0.0930418429110131"
```

- When η_i is equal to zero, we have a 50/50 chance of a person belonging to the age group 20-40 or 40-70. In this model this happens when the reaction time is:

$$0 = -4.336847 + (0.010299) * ReactionTime- > ReactionTime = \frac{4.336847}{0.010299}$$

```
paste(4.336847/0.010299, "ms")
```

```
## [1] "421.093989707739 ms"
```

- The model also shows the standard error for the intercept and ReactionTime estimates. With that, we can assert, with 95% confidence, that those estimates are between the following values:

$$interceptConfInt = -4.336847 \pm 1.96(0.665361) = (-5.640955, -3.032739)$$

$$reactionTimeConfInt = 0.010299 \pm 1.96(0.001435) = (0.0074864, 0.0131116)$$

9. h) Conduct a Logistic regression analysis of age group as a function of height and interpret the results.

```
log_oddH <- glm(as.factor(AGE_GROUP) ~ Height,family=binomial("logit"),data=reactionTimeDf)
summary(log_oddH)
```

```
##
## Call:
## glm(formula = as.factor(AGE_GROUP) ~ Height, family = binomial("logit"),
##     data = reactionTimeDf)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.420  -1.353   0.985   1.012   1.041
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.420967   2.166334  -0.194    0.846
## Height       0.004954   0.012755   0.388    0.698
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 402.98  on 299  degrees of freedom
## Residual deviance: 402.83  on 298  degrees of freedom
## AIC: 406.83
##
## Number of Fisher Scoring iterations: 4
```

- As this model suggests the estimates not being significant, we can double check by calculating the 95% confidence interval:

$$heightConfInt = 0.004954 \pm 1.96(0.012755) = (-0.0200458, 0.0299538)$$

- As we saw that zero is contained within the 95% confidence interval of Height, we can assert that we can't predict age group by height and doesn't make sense to further analyse this model.