Investigating The Framingham Heart Study

By Matthew Curcio

Executive Summary

1. This report investigates data from the 1948 Framingham Heart Study.
    - This longitudinal study includes 4,133 participants with 13 factors total over 10 years.
    - Using this data, I investigate the risk factors for cardiovascular disease (CVD).
2. This report and my article Introduction to Logit display my understanding of logistic regression and R.
3. Seven (7) of the 13 factors have a strong correlation that leads to cardiovascular disease. The odds related to each factor are calculated from the study.

No.	Factors	Odds(Over Mean) of CVD When Risk Is Prevalent
1	Age (Odds Compared 80 yr Male to 20 yr Male)	28:1
2	Systolic Blood Pressure	7.8:1
3	Glucose Levels	2.5:1
4	Prevalence Of Stroke In Family History	2.4:1
5	Cigarettes Per Day	2.1:1
6	Male Vs Female	1.5:1
7	Prevalence Of Hypertension In Family History	1.3:1

2. Introduction

- Data can be found at Kaggle

The Framingham Heart Study began in 1948 by recruiting 5,209 men and women between the ages of 30 and 62 from the town of Framingham, Massachusetts. [These recruits] had not yet developed overt symptoms of cardiovascular disease or suffered a heart attack or stroke.

[The] study has since led to the identification of major CVD risk factors, as well as valuable information on the effects of these factors such as blood pressure, blood triglyceride and cholesterol levels, age, gender, and psychosocial issues.

https://www.framinghamheartstudy.org/fhs-about/

3. Exploratory Data Analysis

```
library(ggplot2)

# Create a dataframe for males and females
males <- df[df$male == 1, ]
females <- df[df$male == 0, ]

summary(males)
```

```
##  male          age       education   cigsPerDay   prevalentStroke prevalentHyp
##  0:   0   Min.   :33.0   0:1266    Min.   : 0.0   0:1756          0:1213
##  1:1766   1st Qu.:42.0   1: 500    1st Qu.: 0.0   1:  10          1: 553
##           Median :48.0             Median :15.0
##           Mean   :49.3             Mean   :13.5
##           3rd Qu.:56.0             3rd Qu.:20.0
##           Max.   :69.0             Max.   :70.0
##  diabetes    totChol        sysBP           diaBP            BMI
##  0:1715   Min.   :113   Min.   : 83.5   Min.   : 48.0   Min.   :15.5
##  1:  51   1st Qu.:206   1st Qu.:118.0   1st Qu.: 76.0   1st Qu.:23.9
##           Median :231   Median :128.0   Median : 82.0   Median :26.1
##           Mean   :233   Mean   :131.4   Mean   : 83.6   Mean   :26.2
##           3rd Qu.:258   3rd Qu.:141.0   3rd Qu.: 90.0   3rd Qu.:28.3
##           Max.   :453   Max.   :235.0   Max.   :136.0   Max.   :40.4
##    heartRate        glucose      TenYearCHD
##  Min.   : 44.0   Min.   : 40.0   0:1436
##  1st Qu.: 66.0   1st Qu.: 71.0   1: 330
##  Median : 75.0   Median : 79.0
##  Mean   : 74.3   Mean   : 82.1
##  3rd Qu.: 80.0   3rd Qu.: 86.0
##  Max.   :125.0   Max.   :394.0
```

```
# Find male min/max ages
min_age <- min(males$age)
max_age <- max(males$age)
# Find female min/max age
min_age <- min(females$age)
max_age <- max(females$age)
```

```
# Print the minimum and maximum ages
#print("Male Ages")
#print("---------")
#print(paste("Minimum =", min_age))
#print(paste("Maximum =", max_age))

# Print the minimum and maximum ages
print("Female Ages")
```

```
## [1] "Female Ages"
```

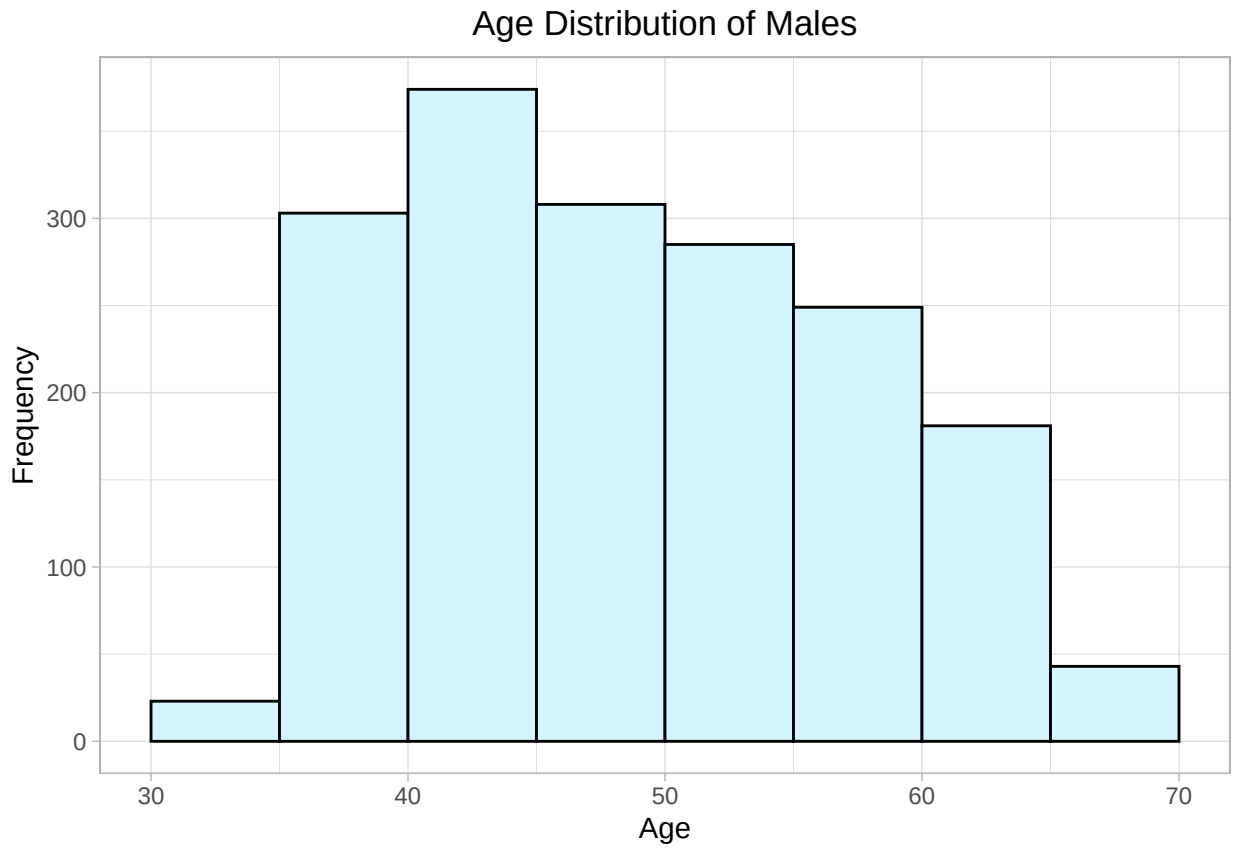
```
print(paste("Minimum =", min_age))
```

```
## [1] "Minimum = 32"
```

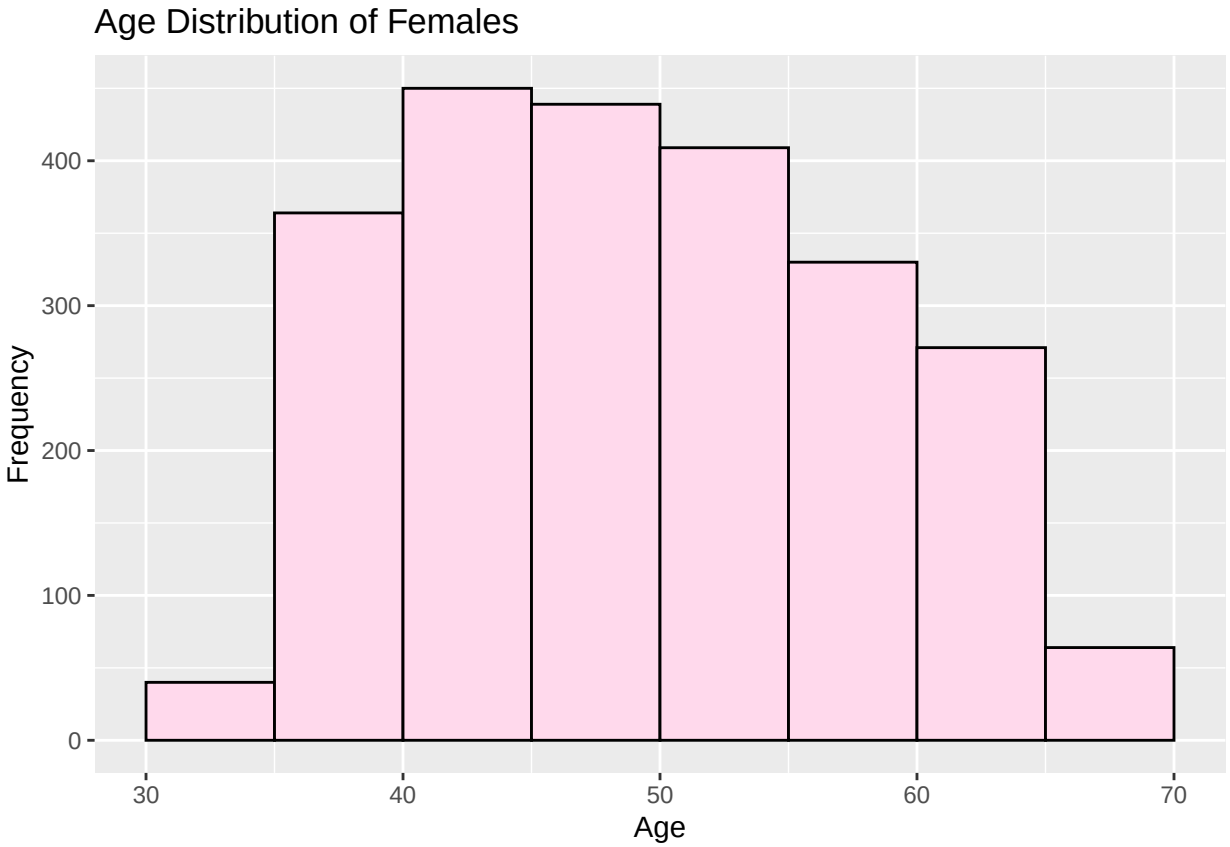
```
print(paste("Maximum =", max_age))
```

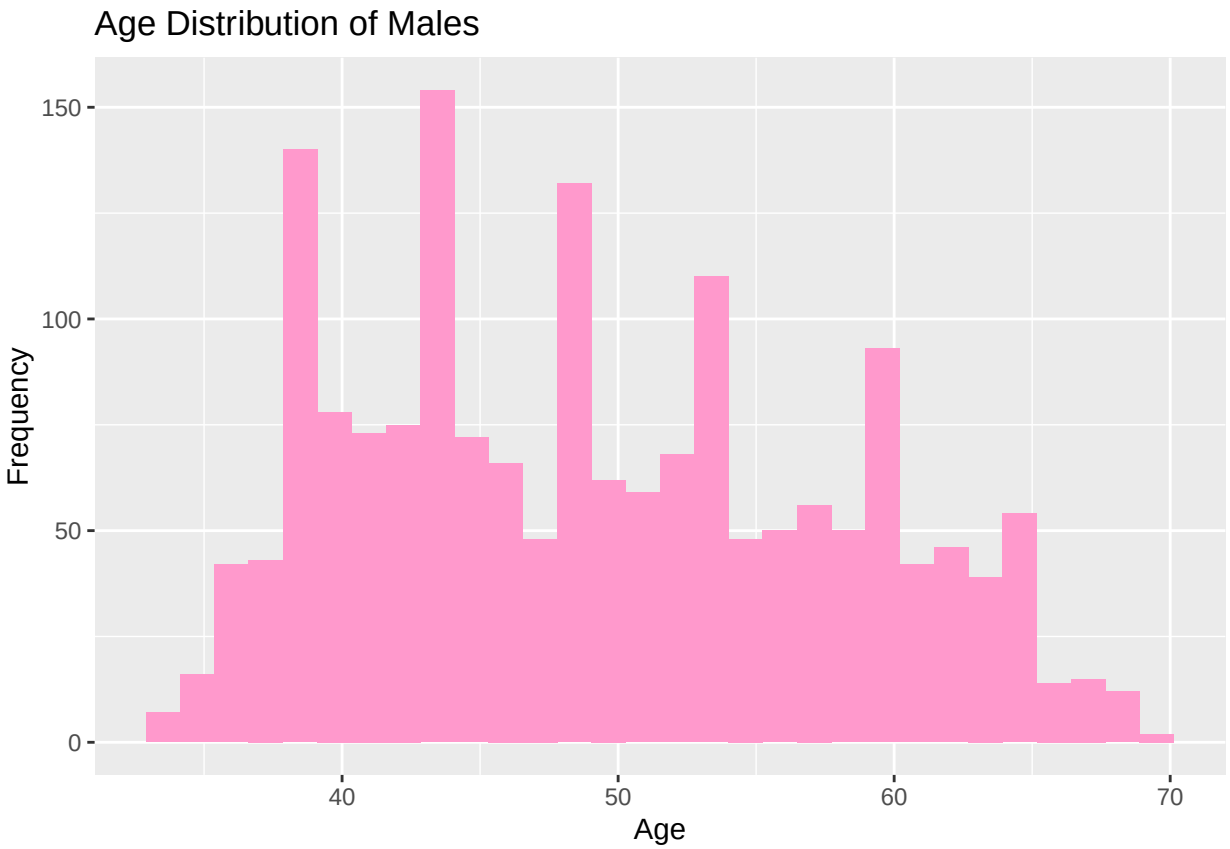
```
## [1] "Maximum = 70"
```

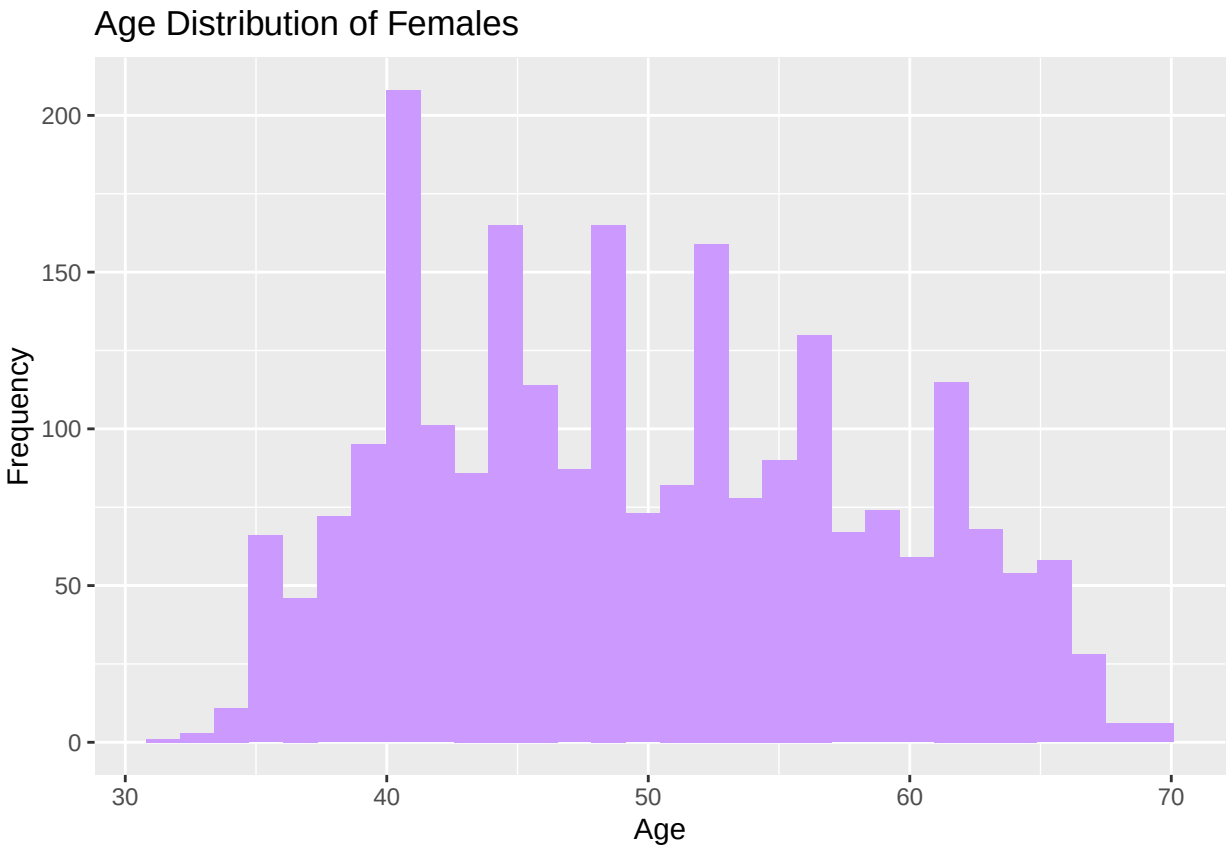
```r
# Create a histogram of ages for males
ggplot(males, aes(x = age)) +
  geom_histogram(breaks=seq(30, 70, by = 5),
                 color="black", fill = "#D4F5FF") +
  labs(title = "Age Distribution of Males",
       x = "Age",
       y = "Frequency")+
  theme_light() +
  theme(plot.title = element_text(hjust = 0.5))
```

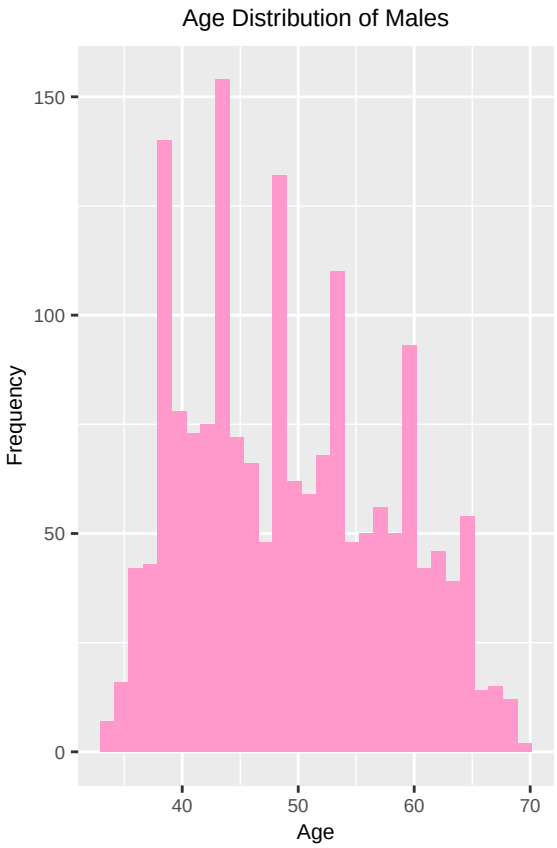


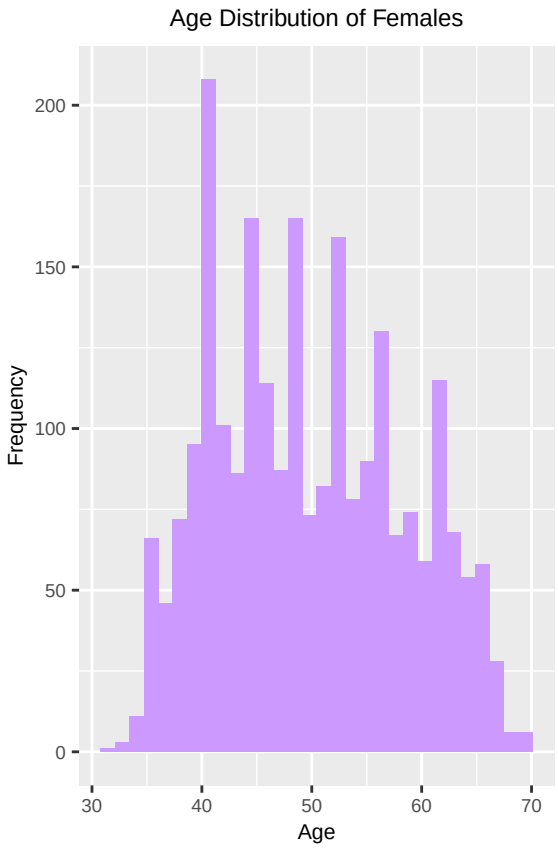

```r
# Create a histogram of ages for females
ggplot(females, aes(x = age)) +
  geom_histogram(breaks=seq(30, 70, by = 5),
                 color = "black", fill = "#FFD9EC") +
  labs(title = "Age Distribution of Females",
       x = "Age",
       y = "Frequency")
```


Age Distribution of Females
Frequency
400
300
200
100
0
30
40
50
60
70
Age

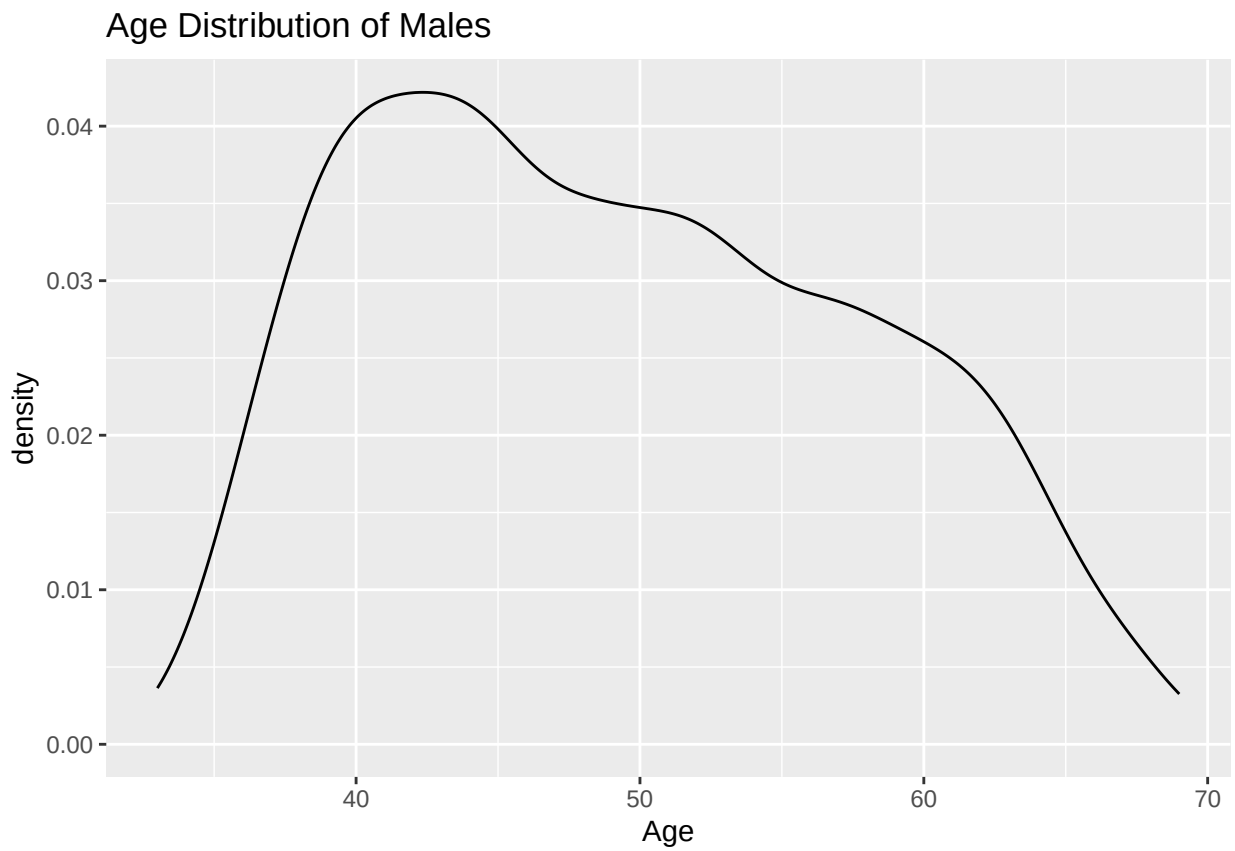

Age Distribution of Males
Frequency
150
100
50
0
40
50
60
70
Age


Age Distribution of Females
Frequency
200
150
100
50
0
30
40
50
60
70
Age

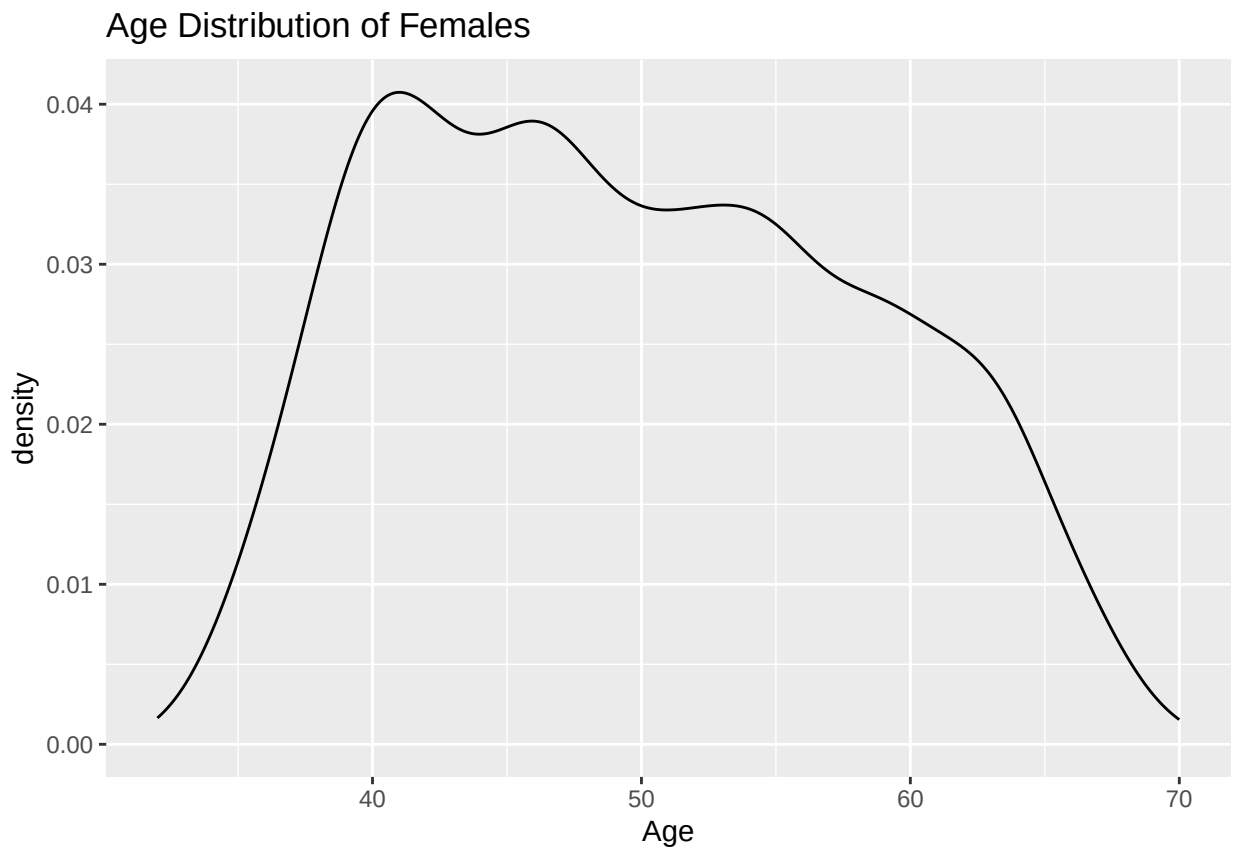




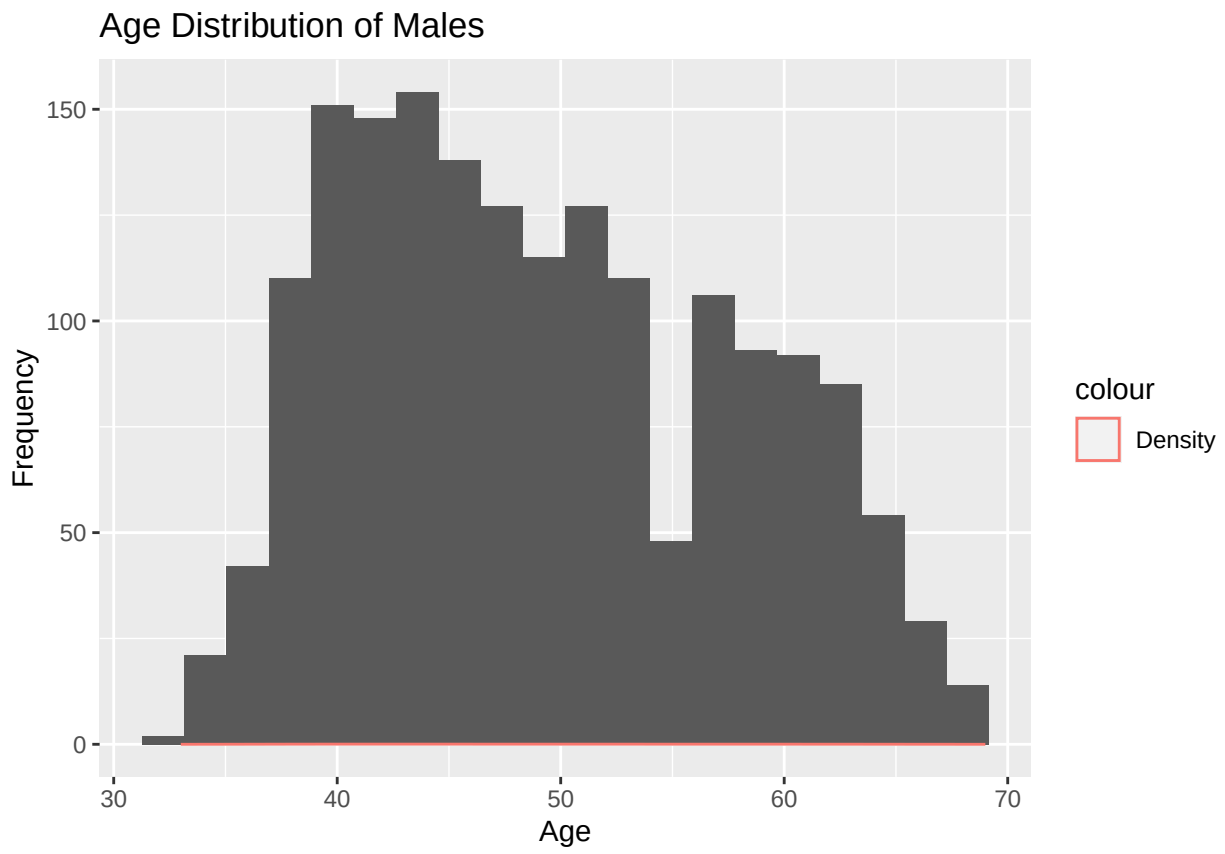

```r
# Create a density plot of ages for males
ggplot(males, aes(x = age)) +
  geom_density() +
  labs(title = "Age Distribution of Males",
       x = "Age")
```




```
# Create a density plot of ages for females
ggplot(females, aes(x = age)) +
  geom_density() +
  labs(title = "Age Distribution of Females",
       x = "Age")
```




```
# Create a histogram of ages for males
ggplot(males, aes(x = age)) +
  geom_histogram(bins = 20) +
  geom_density(aes(color = "Density")) +
  labs(title = "Age Distribution of Males",
       x = "Age",
       y = "Frequency")
```



Cigs per day for men and women

2. Results

2.1 Logistic Regression Model

```
mylogit <- glm(TenYearCHD ~ male + age + education +
                            cigsPerDay + prevalentStroke + prevalentHyp +
                            diabetes + totChol + sysBP + diaBP + BMI +
                            heartRate + glucose,
               data = df,
               family = "binomial")

summary(mylogit)
```

```
##
## Call:
## glm(formula = TenYearCHD ~ male + age + education + cigsPerDay +
##     prevalentStroke + prevalentHyp + diabetes + totChol + sysBP +
##     diaBP + BMI + heartRate + glucose, family = "binomial", data = df)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.964  -0.596  -0.432  -0.294   2.810
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)      -8.04990    0.64770  -12.43  < 2e-16 ***
## male1             0.48093    0.10163    4.73  2.2e-06 ***
## age               0.06263    0.00625   10.02  < 2e-16 ***
## education1        0.03031    0.10610    0.29    0.775
## cigsPerDay        0.02087    0.00397    5.25  1.5e-07 ***
## prevalentStroke1  1.00721    0.43923    2.29    0.022 *
## prevalentHyp1     0.25864    0.12955    2.00    0.046 *
## diabetes1         0.24052    0.29605    0.81    0.417
## totChol           0.00184    0.00106    1.73    0.083 .
## sysBP             0.01498    0.00355    4.22  2.5e-05 ***
## diaBP            -0.00386    0.00602   -0.64    0.521
## BMI               0.00212    0.01182    0.18    0.857
## heartRate        -0.00248    0.00393   -0.63    0.528
## glucose           0.00619    0.00215    2.88    0.004 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3521.9  on 4132  degrees of freedom
## Residual deviance: 3131.2  on 4119  degrees of freedom
## AIC: 3159
##
## Number of Fisher Scoring iterations: 5
```

- **7 most significant variables**
- Seven predictors have $\alpha < 0.05$. They are significant and associated with acquiring cardiovascular disease.

Rank	Risk Factor
1	Prevalence of Stroke1
2	Male1
3	Prevalence of Hypertension1
4	Age
5	Cigarettes Per Day
6	Systolic Blood Pressure
7	Glucose

2.2 Wald Test: Do The Seven Factors Fit Our Model

- The Wald Chi-Square Test can help determine if our proposed model is significant.
- The Wald test generates a *P-value « 0.001*.
- Therefore, we conclude the seven (7) parameters are significant and useful in describing cardiovascular disease.

2.3 Determination of Odds for Seven Variables

- We can calculate the odds of acquiring cardiovascular disease for each of the seven variables.
- By holding all other values constant we create a dataframe that investigates the odds given Prevalence of Stroke, for example.

```r
strok_test <- with(df, data.frame(male = "0",
                                  age = mean(age),
                                  education = "0",
                                  cigsPerDay = 0, # Non-smoker
                                  prevalentHyp = "0",
                                  diabetes = "0",
                                  totChol = mean(totChol),
                                  sysBP = mean(sysBP),
                                  diaBP = mean(diaBP),
                                  BMI = mean(BMI),
                                  heartRate = mean(heartRate),
                                  glucose = mean(glucose),
                                  prevalentStroke = c("0", "1"))
                   )

# Convert prevalentStroke from Numeric to FACTOR
strok_test$prevalentStroke <- as.factor(strok_test$prevalentStroke)
# str(strok_test)

strok_test$prevalentStroke <- predict(mylogit,
                                      newdata = strok_test,
                                      type = "response")
#strok_test$prevalentStroke
```

2.4 Odds Given Prevalence Of Stroke In family history.

1. WITH Prevalence of Stroke: 0.18761
2. NO Prevalence of Stroke: 0.07778

- **Odds = 2.4119**

2.5 Odds Given For Male Vs Female

```r
male_test <- with(df, data.frame(male = c("0","1"), # Factor of Interest
                                 age = mean(age),
                                 education = "0",
                                 cigsPerDay = 0,
                                 prevalentHyp = "0",
                                 diabetes = "0",
                                 totChol = mean(totChol),
                                 sysBP = mean(sysBP),
                                 diaBP = mean(diaBP),
                                 BMI = mean(BMI),
                                 heartRate = mean(heartRate),
                                 glucose = mean(glucose),
                                 prevalentStroke = "0"))

# REMEMBER convert male_test from numeric to FACTOR
male_test$male <- as.factor(male_test$male)

male_test$male <- predict(mylogit, newdata = male_test, type = "response")
```

1. Males: 0.12005
2. Female: 0.07778

- **Odds = 1.54343**

2.6 Odds Prevalence of Hypertension In Family History

```
hyperT_test <- with(df, data.frame(male = "0",
                                   age = mean(age),
                                   education = "0",
                                   cigsPerDay = 0,
                                   prevalentHyp = c("0","1"), # Factor of Interest
                                   diabetes = "0",
                                   totChol = mean(totChol),
                                   sysBP = mean(sysBP),
                                   diaBP = mean(diaBP),
                                   BMI = mean(BMI),
                                   heartRate = mean(heartRate),
                                   glucose = mean(glucose),
                                   prevalentStroke = "0"))

# REMEMBER convert male_test from numeric to FACTOR
hyperT_test$prevalentHyp <- as.factor(hyperT_test$prevalentHyp)

hyperT_test$prevalentHyp <- predict(mylogit, newdata = hyperT_test, type = "response")
```

1. WITH Prevalence of Hypertension: 0.09848
2. NO Prevalence of Hypertension: 0.07778

- **Odds = 1.2661**

2.7 Odds Given Age

```
age_test <- with(df, data.frame(male = "0",
                                   age = c(20,30,40,50,60,70,80),
                                   education = "0",
                                   cigsPerDay = 0,
                                   prevalentHyp = "0",
                                   diabetes = "0",
                                   totChol = mean(totChol),
                                   sysBP = mean(sysBP),
                                   diaBP = mean(diaBP),
                                   BMI = mean(BMI),
                                   heartRate = mean(heartRate),
                                   glucose = mean(glucose),
                                   prevalentStroke = "0"))

age_test$age <- predict(mylogit, newdata = age_test, type = "response")
```

Age (years)	Probability Given Age	Odds Compared to 20 yr old
20	0.01307	1
30	0.02418	1.84969
40	0.0443	3.38895
50	0.0798	6.1044
60	0.13958	10.67785
70	0.23282	17.81084
80	0.36214	27.70331

Age (years)	Probability Given Age	Odds Compared to 20 yr old

2.8 Odds Given Number Of Cigarettes Per Day

```
cigs_per_day <- with(df, data.frame(male = "0",
                                    age = mean(age),
                                    education = "0",
                                    cigsPerDay = c(0,10,20,30,40),
                                    prevalentHyp = "0",
                                    diabetes = "0",
                                    totChol = mean(totChol),
                                    sysBP = mean(sysBP),
                                    diaBP = mean(diaBP),
                                    BMI = mean(BMI),
                                    heartRate = mean(heartRate),
                                    glucose = mean(glucose),
                                    prevalentStroke = "0"))

cigs_per_day$cigsPerDay <- predict(mylogit, newdata = cigs_per_day, type = "response")

#cigs_per_day$cigsPerDay
```

1. A pack of cigarettes gave a person 45% increase of acquiring Cardiovascular disease, **using this data set. This seems oddly low.**

Age (years)	Probability Given Age	Odds Compared to Zero Cigarettes Per Day
0	0.07778	1
10	0.09414	1.21027
20	0.11351	1.45932
30	0.13627	1.7519
40	0.16275	2.09238

2.9 Odds Given Systolic Blood Pressure

```
summary(df$sysBP)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    83.5   117.0   128.0   132.4   144.0   295.0
```

```
# Min. 1st Qu.  Median    Mean  3rd Qu.  Max.
# 83.5   117.0   128.0   132.4   144.0   295.0

sysBP_calc <- with(df, data.frame(male = "0",
                                  age = mean(age),
                                  education = "0",
                                  cigsPerDay = 0,
                                  prevalentHyp = "0",
                                  diabetes = "0",
                                  totChol = mean(totChol),
                                  sysBP = c(117, 128, 144, 295),
                                  diaBP = mean(diaBP),
```

```
                                    BMI = mean(BMI),
                                    heartRate = mean(heartRate),
                                    glucose = mean(glucose),
                                    prevalentStroke = "0"))

sysBP_calc$sysBP <- predict(mylogit, newdata = sysBP_calc, type = "response")

#sysBP_calc$sysBP
```

Systolic BP	Probability Given Systolic BP	Odds Systolic BP
117	0.06279	1
128	0.07322	1.16607
144	0.09124	1.45318
Max 295	0.49104	7.8204

2.10 Odds Given Glucose Levels

```
summary(df$glucose)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      40      72      80      82      85     394
```

```
#    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#     40      72      80      82      85     394

glucose_calc <- with(df, data.frame(male = "0",
                                    age = mean(age),
                                    education = "0",
                                    cigsPerDay = 0,
                                    prevalentHyp = "0",
                                    diabetes = "0",
                                    totChol = mean(totChol),
                                    sysBP = mean(sysBP),
                                    diaBP = mean(diaBP),
                                    BMI = mean(BMI),
                                    heartRate = mean(heartRate),
                                    glucose = c(72, 80, 85, 394),
                                    prevalentStroke = "0"))

glucose_calc$glucose <- predict(mylogit, newdata = glucose_calc, type = "response")

# glucose_calc$glucose.
# 0.094843 0.100852 0.110194 0.239738
```

Glucose	Probabilities	Odds Given Glucose
72	0.094843	1
80	0.100852	1.06336
85	0.110194	1.16186
Max 394	0.239738	2.52774

3. Conclusion

1. We find seven (7) of the 13 factors lead to cardiovascular disease. The odds related to each factor were calculated from the study.

No.	Factors	Approximate Odds Over Mean
1	Prevalence Of Stroke In Family History	240%
2	Male Vs Female	150%
3	Prevalence Of Hypertension In Family History	130%
4	Age	$<$ 2,800%
5	Cigarettes Per Day	$<$ 210%
6	Systolic Blood Pressure	$<$ 780%
7	Glucose Levels	$<$ 250%

2. The Wald Chi-Square Test can help determine if our proposed model is valuable and significant. The Wald test generates a P-value « 0.001. Therefore, we conclude the seven (7) parameters are significant and useful in describing cardiovascular disease.

3. A pack of cigarettes gave a person 45% increase of acquiring Cardiovascular disease, **using this data set. This seems oddly low.**

Cigs Per Day	Probability Given Age	Odds Compared to Zero Cigarettes Per Day
0	0.07778	1
10	0.09414	1.21027
20	0.11351	1.45932
30	0.13627	1.7519
40	0.16275	2.09238

Notes

- For analysis help https://stats.idre.ucla.edu/r/dae/logit-regression/
- For interpretation help https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-interpret-odds-ratios-in-logistic-regression/.
- https://stats.oarc.ucla.edu/other/mult-pkg/faq/general/faqhow-are-the-likelihood-ratio-wald-and-lagrange-multiplier-score-tests-different-andor-similar/

Wald test info

- https://www.mbaskool.com/business-concepts/statistics/6916-wald-test.html
- https://www.statology.org/wald-test-in-r/
- https://handwiki.org/wiki/Wald_test
- https://questionerlab.com/what-is-the-use-of-wald-test-in-logistic-regression
- https://bookdown.org/mike/data_analysis/wald-test.html
- https://bookdown.org/mike/data_analysis/hypothesis-testing.html#wald-test