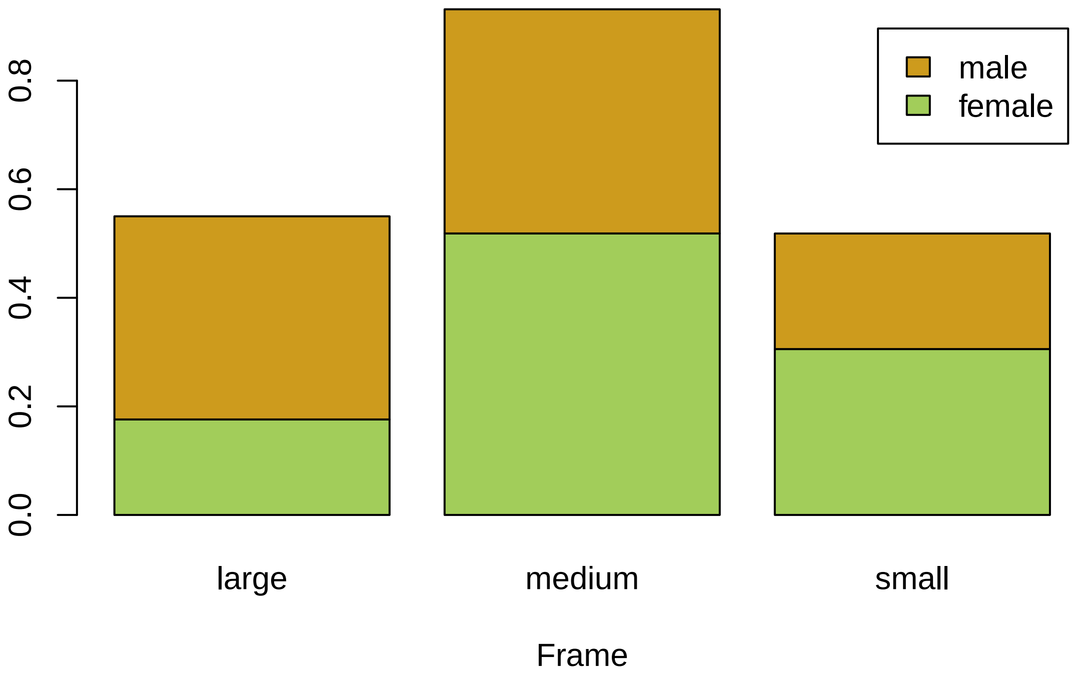
5. a)

```
diabetes <- read.csv("diabetes_fall2022.csv")
table1 <- table(diabetes$gender, diabetes$frame)
addmargins(table1)
```

```
##
##          large medium small Sum
##   female    38    112    66 216
##   male      58     64    33 155
##   Sum       96    176    99 371
```

5. b)

```
barplot(prop.table(table1,margin=1), xlab="Frame", legend.text = rownames(table1),
        col=c("darkolivegreen3", "goldenrod3"))
```



5. c)

H_0 : frame and gender are independent vs H_1 : frame and gender are associated $\alpha = 0.05$

```
chisq.test(table1, correct=F)
```

```
##
```

```
##  Pearson's Chi-squared test
##
## data:  table1
## X-squared = 18.734, df = 2, p-value = 8.548e-05
```

```
qchisq(1-0.05, df = (3-1)*(2-1))
```

```
## [1] 5.991465
```

We get a critical value of 5.9915 and a p-value of 0.00008548. Since $5.9915 < 18.734 = X^2$ and $0.00008548 < 0.05 = \alpha$, we reject the null hypothesis that frame and gender are independent and conclude that there is an association between frame and gender. ***6. a)***
$H_0 : \sigma_m^2 = \sigma_f^2$ vs $H_1 : \sigma_m^2 \neq \sigma_f^2$ $\alpha = 0.01$

```
s_m <- var(diabetes[diabetes$gender=="male",]$stab.glu)
n_m <- length(diabetes[diabetes$gender=="male",]$stab.glu)
s_f <- var(diabetes[diabetes$gender=="female",]$stab.glu)
n_f <- length(diabetes[diabetes$gender=="female",]$stab.glu)

f_obs <- s_m/s_f
f_obs
```

```
## [1] 2.114595
```

```
2*(1-pf(f_obs, n_m-1, n_f-1))
```

```
## [1] 4.123336e-07
```

Since p-value $= 4.1233 * 10^{-7} < 0.01 = \alpha$, we reject the null hypothesis that the variances of stabilized glucose for gender are the same.

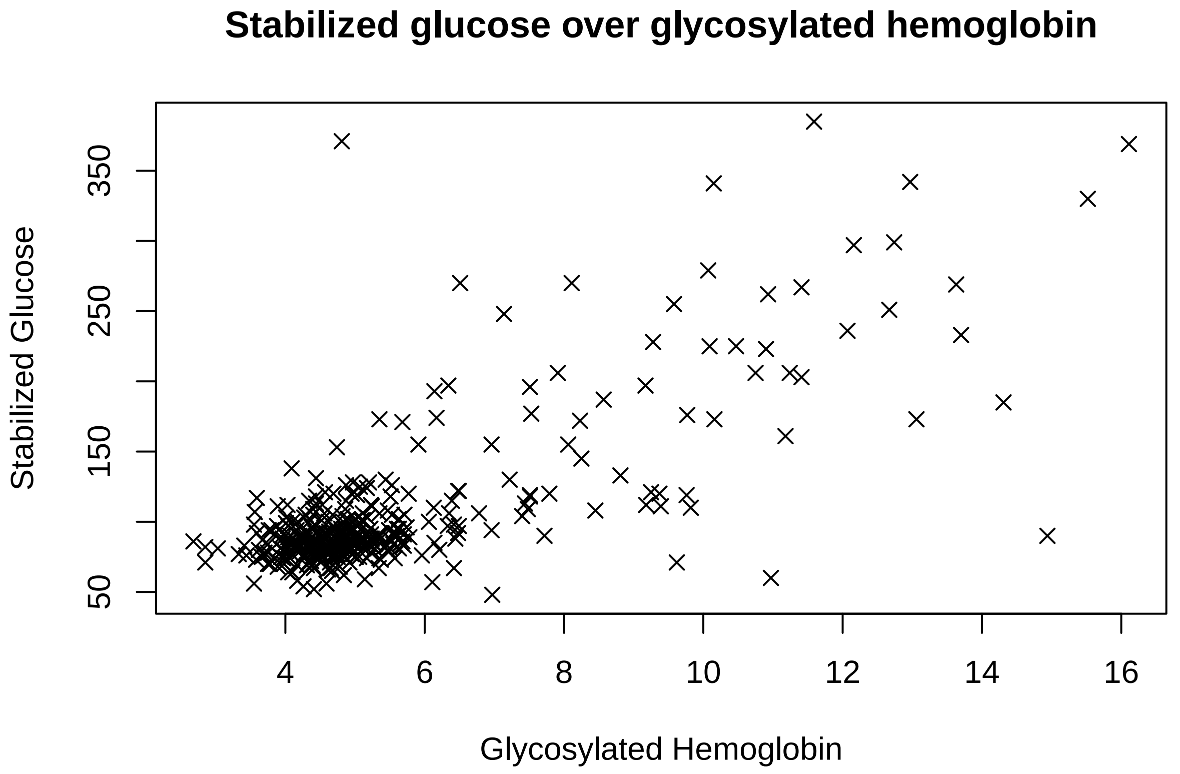
6. b)
$H_0 : \mu_m - \mu_f \leq 0$ vs $H_1 : \mu_m - \mu_f > 0$
$\alpha = 0.05$

```
t.test(diabetes[diabetes$gender=="male",]$stab.glu, diabetes[diabetes$gender=="female",]$stab.glu,
       alternative="greater", conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  diabetes[diabetes$gender == "male", ]$stab.glu and diabetes[diabetes$gender == "female", ]$sta
## t = 1.5167, df = 255.2, p-value = 0.06529
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.8043766        Inf
## sample estimates:
## mean of x mean of y
##  112.4194  103.3241
```

Since p-value $= 0.0653 > 0.05 = \alpha$, we fail to reject the null hypothesis that the mean stabilized glucose of males is less than or equal to the stabilized glucose of females. ***7. a)***

```
plot(diabetes$glyhb, diabetes$stab.glu, pch=4, main="Stabilized glucose over glycosylated hemoglobin",
     ylab = "Stabilized Glucose", xlab = "Glycosylated Hemoglobin")
```



```
cor(diabetes$glyhb, diabetes$stab.glu)
```

```
## [1] 0.7408235
```

There appears to be a moderately strong linear relationship between the two variables based on the correlation of 0.7408 and the scatterplor, indicating linear regression to be appropriate.

7. b) $H_0 : \rho = 0$ vs $H_1 : \rho \neq 0$ $\alpha = 0.05$

```
cor.test(diabetes$glyhb, diabetes$stab.glu)
```

```
##
##  Pearson's product-moment correlation
##
## data:  diabetes$glyhb and diabetes$stab.glu
## t = 21.186, df = 369, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
```

```
##  0.6911384 0.7835390
## sample estimates:
##       cor
## 0.7408235
```

```r
qt(1-(0.05/2), df = length(diabetes$glyhb)-2)
```

```
## [1] 1.966414
```

We find the p-value to be less than $2.2*10^{-16}$ and the test statistic to be 21.186. Since $2.2*10^{-16} < 0.05 = \alpha$, we reject the null hypothesis that stabilized glucose and glycosylated hemoglobin are independent. ***7. c)***

```r
lm1 <- lm(diabetes$stab.glu~diabetes$glyhb)
lm1
```

```
##
## Call:
## lm(formula = diabetes$stab.glu ~ diabetes$glyhb)
##
## Coefficients:
##    (Intercept)  diabetes$glyhb
##          6.945          17.930
```

The regression equation is $y = 17.930x + 6.945$

7. d) $\alpha = 0.01$

```r
confint(lm1, level=0.99)
```

```
##                     0.5 %   99.5 %
## (Intercept)    -6.229659 20.11944
## diabetes$glyhb 15.739136 20.12179
```

We are 99% confident that the slope for the regression equation falls between 15.7391 and 20.1218.

7. e)

```r
summary(lm1)
```

```
##
## Call:
## lm(formula = diabetes$stab.glu ~ diabetes$glyhb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -184.826  -15.391   -3.784   10.282  277.810
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       6.9449     5.0882   1.365    0.173
## diabetes$glyhb   17.9305     0.8463  21.186   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 36.2 on 369 degrees of freedom
## Multiple R-squared:  0.5488, Adjusted R-squared:  0.5476
## F-statistic: 448.9 on 1 and 369 DF,  p-value: < 2.2e-16
```

54.88% of the variability in the glycosylated hemoglobin is explained by its linear relationship with stabilized glucose.

7. f)

```
lm2 <- lm(diabetes$stab.glu~diabetes$glyhb+diabetes$bmi+diabetes$frame)
lm2
```

```
##
## Call:
## lm(formula = diabetes$stab.glu ~ diabetes$glyhb + diabetes$bmi +
##     diabetes$frame)
##
## Coefficients:
##          (Intercept)        diabetes$glyhb          diabetes$bmi
##               9.3176               17.7060                0.1715
## diabetes$framemedium   diabetes$framesmall
##              -9.2878               -6.2408
```

$y_{large} = 17.7060x_{glyhb} * 0.1715x_{bmi} + 9.3176$ $y_{medium} = 17.7060x_{glyhb} * 0.1715x_{bmi} + 0.0298$ $y_{small} = 17.7060x_{glyhb} * 0.1715x_{bmi} + 3.0768$ The slope for all the equations is the same but each different frame has a different intercept.

7. g)

```
summary(lm2)
```

```
##
## Call:
## lm(formula = diabetes$stab.glu ~ diabetes$glyhb + diabetes$bmi +
##     diabetes$frame)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -178.827  -15.397   -4.319   11.285  278.504
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)            9.3176    11.7298   0.794   0.4275
## diabetes$glyhb        17.7060     0.8583  20.630   <2e-16 ***
## diabetes$bmi           0.1715     0.3137   0.547   0.5849
## diabetes$framemedium  -9.2878     4.7065  -1.973   0.0492 *
## diabetes$framesmall   -6.2408     5.7661  -1.082   0.2798
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
```

```
## Residual standard error: 36.1 on 366 degrees of freedom
## Multiple R-squared:  0.5548, Adjusted R-squared:   0.55
## F-statistic:   114 on 4 and 366 DF,  p-value: < 2.2e-16
```

Multiple R-squared is 0.5548 and the adjusted R-squared is 0.55, since the adjusted R-squred is lower, that means that we added a variable that is not impactful on the output.

7. h) H_0 : The bmi variable does not predict stabilized glucose vs H_1 : The bmi variable significantly predicts stabilized glucose H_0 : The frame variable does not predict stabilized glucose vs H_1 : The frame variable significantly predicts stabilized glucose $\alpha = 0.05$

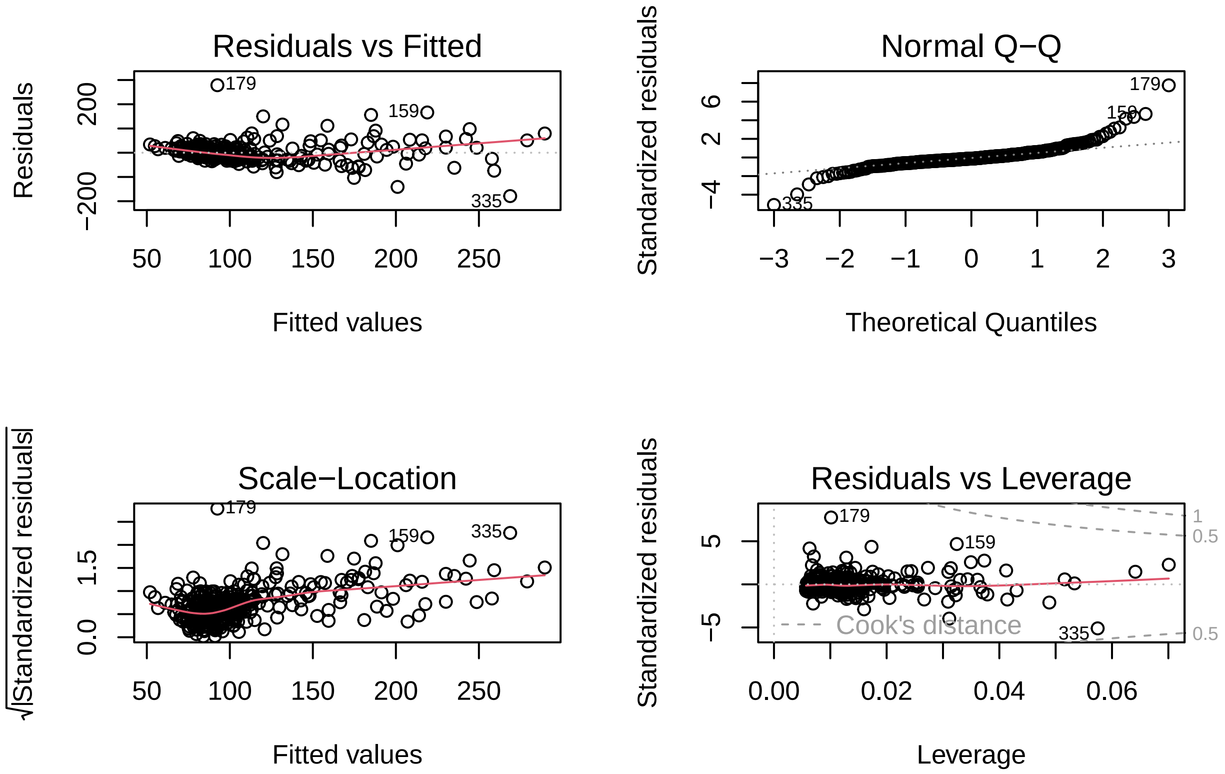
```
lm3 <- lm(diabetes$stab.glu~diabetes$bmi+diabetes$frame)
anova(lm3)
```

```
## Analysis of Variance Table
##
## Response: diabetes$stab.glu
##                  Df  Sum Sq Mean Sq F value   Pr(>F)
## diabetes$bmi      1   19271   19271  6.8555 0.009203 **
## diabetes$frame    2   20566   10283  3.6580 0.026727 *
## Residuals       367 1031652    2811
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We find the test statistic for bmi to be 6.8555 and the test statistic for frame to be 3.6580. Since the bmi p-value $= 0.009203 < 0.05 = \alpha$, we reject the null hypothesis that the bmi variable does not predict stabilized glucose. Since the frame p-value $= 0.026726 < 0.05 = \alpha$, we reject the null hypothesis that the frame variable does not predict stabilized glucose.

7. i)

```
par(mfrow=c(2,2))
plot(lm2)
```



In the residuals vs fitted graph, we see increasing variability which indicates that the relationship is potentially non-linear and our regression line is not appropriate for modeling the data. For our homoscedasticity, we do not see a constant variance across all values of x. The QQ plot shows that the data skews right which means that our values do not have a normal distrbution.