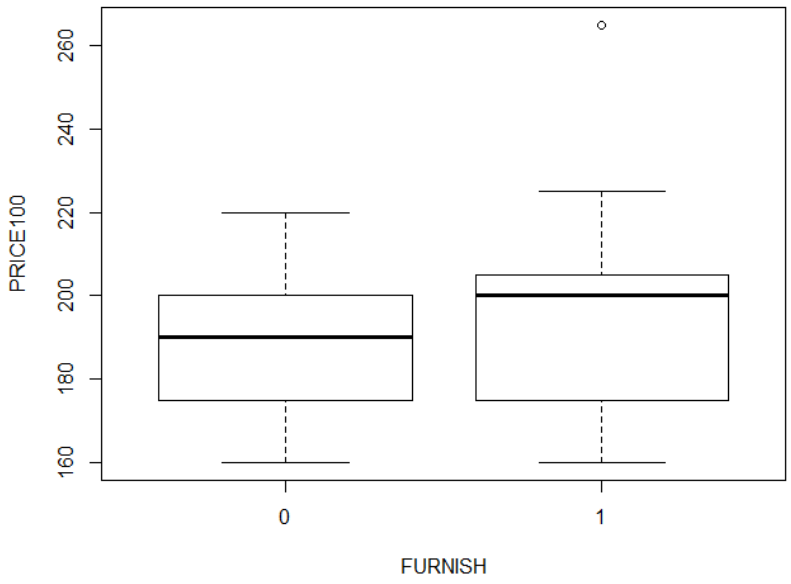
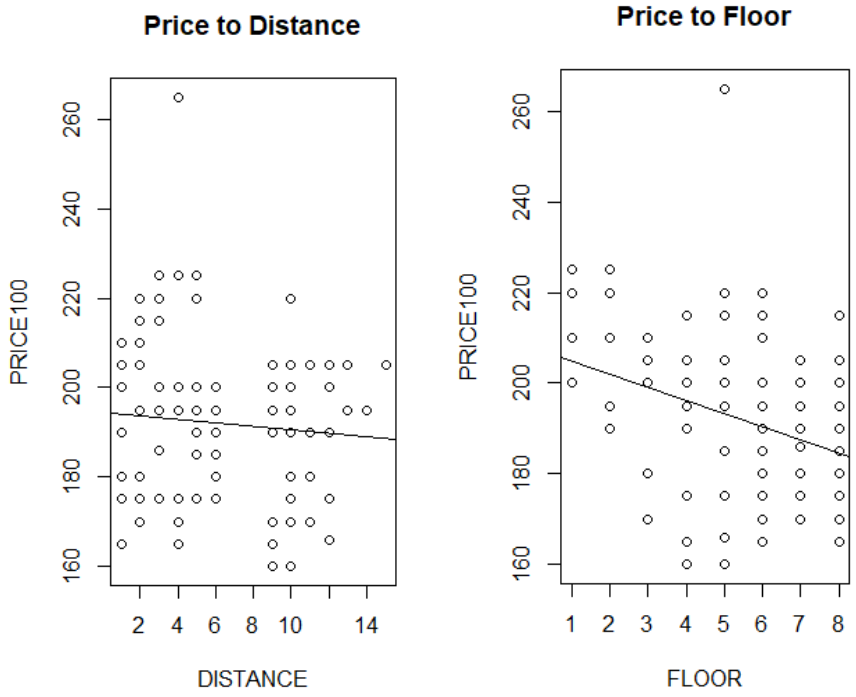
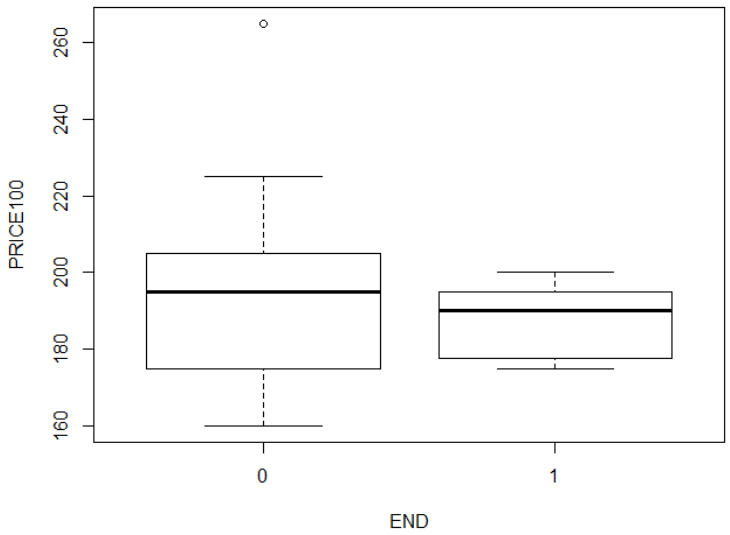
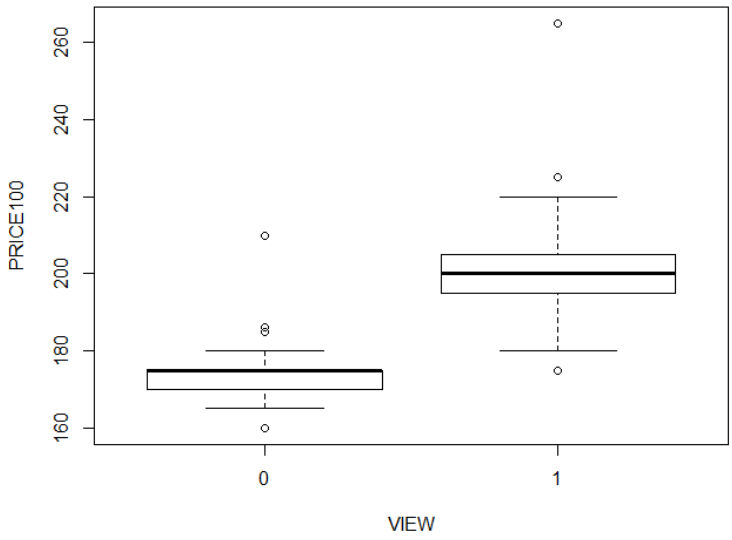
Stat 384 Group Project:

Problem 1






PRICE100
160
180
200
220
240
260
0
1
END


PRICE100
160
180
200
220
240
260
0
1
VIEW

> summary(fit)

Call:
lm(formula = PRICE100 ~ FLOOR + DISTANCE + VIEW + END + FURNISH)

Residuals:
Min 1Q Median 3Q Max
-20.953 -5.877 -0.929 4.793 53.109

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 177.7035 4.1684 42.631 < 2e-16 *
FLOOR -0.7151 0.5308 -1.347 0.18090
DISTANCE -0.8733 0.2449 -3.565 0.00056 ***
VIEW 31.2728 2.2312 14.016 < 2e-16 ***
END -17.8078 3.9820 -4.472 2.05e-05 ***
FURNISH 9.9838 2.0515 4.867 4.24e-06 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**

**Residual standard error: 9.905 on 100 degrees of freedom
Multiple R-squared: 0.7058, Adjusted R-squared: 0.6911
F-statistic: 47.98 on 5 and 100 DF, p-value: < 2.2e-16**

When alpha = 0.05

- **The Intercept is Significant**
- **Floor is NOT Significant**
- **Distance is significant**
- **View is significant**
- **End is significant**
- **Furnish is significant**

Distance	**R2 = 0.7%**	
Distance + View	**R2 = 56%**	

Distance + End	**R2 = 1.3%**	
Distance + Furnish	**R2 = 2.9%**	
Distance + View + End	**R2 = 61%**	
Distance + View + Furnish	**R2 = 64%**	
Distance + End + Furnish	**R2 = 3.4%**	
Distance + End + View + Furnish	**R2 = 70.%**	

Multicollinearity Test:

```
> X<-cbind(Floor,Distance,View,End,Furnish)
> library(mctest)
> imcdiag(X,y)

Call:
imcdiag(x = X, y = y)


All Individual Multicollinearity Diagnostics Result

            VIF    TOL     Wi     Fi Leamer   CVIF Klein
Floor    1.1787 0.8484 4.5124 6.0761 0.9211 1.0192     0
Distance 1.0535 0.9493 1.3498 1.8175 0.9743 0.9109     0
View     1.2064 0.8289 5.2124 7.0187 0.9104 1.0432     0
End      1.0567 0.9464 1.4314 1.9274 0.9728 0.9137     0
Furnish  1.1110 0.9001 2.8023 3.7733 0.9487 0.9607     0

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test

Floor , coefficient(s) are non-significant may be due to multicollinearity

R-square of y on all x: 0.7058

* use method argument to check which regressors may be the reason of collinearity
==================================
```

2nd-order terms:

```
> fit2<-lm(Price~Distance+Floor+DistanceSQ+FloorSQ)
> summary(fit2)

Call:
lm(formula = Price ~ Distance + Floor + DistanceSQ + FloorSQ)

Residuals:
    Min      1Q  Median      3Q     Max
-31.978 -12.769   0.545  11.667  75.997

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 232.0639    10.7067  21.675   <2e-16 ***
Distance     -3.5166     1.7635  -1.994   0.0488 *
Floor       -10.1620     4.0034  -2.538   0.0127 *
DistanceSQ    0.2388     0.1230   1.942   0.0549 .
FloorSQ       0.7197     0.3924   1.834   0.0696 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.65 on 101 degrees of freedom
Multiple R-squared:  0.1605,    Adjusted R-squared:  0.1272
F-statistic: 4.826 on 4 and 101 DF,  p-value: 0.001329
```

> summary(fit2)

Call:
lm(formula = PRICE100 ~ DISTANCE + VIEW + END + FURNISH)

Residuals:

Min	1Q	Median	3Q	Max
-21.100	-5.566	-1.461	4.856	52.714

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	173.1461	2.4460	70.788	< 2e-16	***
DISTANCE	-0.9189	0.2436	-3.773	0.000272	***
VIEW	32.1431	2.1443	14.990	< 2e-16	***
END	-17.2065	3.9728	-4.331	3.51e-05	***
FURNISH	10.6724	1.9948	5.350	5.50e-07	***

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.944 on 101 degrees of freedom
Multiple R-squared:  0.7005,          Adjusted R-squared:  0.6886
F-statistic: 59.04 on 4 and 101 DF,  p-value: < 2.2e-16
```

- **All variables now significant**
- **70% of the Price can be explained through the regression**
- **Can we improve the model by adding interaction variables?**

```
summary(fit2)

Call:
lm(formula = PRICE100 ~ DISTANCE + FURNISH + END + VIEW + DistFurn +
   Distview + EndFurn + FurnView)

Residuals:
   Min     1Q  Median     3Q    Max
-20.492  -5.252  -0.777   4.602  47.671

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 170.18682    4.13516  41.156  < 2e-16 ***
DISTANCE      0.02011    0.53654   0.037 0.970181
FURNISH      12.49584    4.47556   2.792 0.006310 **
END         -17.37692    5.08012  -3.421 0.000916 ***
VIEW         33.14292    4.54743   7.288 8.45e-11 ***
DistFurn     -1.17509    0.52838  -2.224 0.028472 *
Distview     -0.80391    0.55222  -1.456 0.148684
EndFurn      -1.03465    7.84140  -0.132 0.895299
FurnView      9.33936    4.22289   2.212 0.029342 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.678 on 97 degrees of freedom
Multiple R-squared:  0.7275,          Adjusted R-squared:  0.705
F-statistic: 32.37 on 8 and 97 DF,  p-value: < 2.2e-16
```

Fit5:

```
Call:
lm(formula = Price ~ Distance + DistSQ + Floor + FloorSQ + Furnish +
    End + View + DistFurn + DistView + FurnView + EndFurn + FurnView +
    FloorFurn + FloorEnd + FloorView + DistFloor)

Residuals:
    Min      1Q  Median      3Q     Max
-16.950  -4.169  -0.854   2.284  50.111

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 163.38138   14.24704  11.468  < 2e-16 ***
Distance     -2.33445    1.28407  -1.818   0.0724 .
DistSQ        0.17019    0.08041   2.117   0.0371 *
Floor        -0.25688    3.36686  -0.076   0.9394
FloorSQ       0.25996    0.25233   1.030   0.3056
Furnish      14.90567    7.95278   1.874   0.0641 .
End         -12.91407    8.60093  -1.501   0.1367
View         57.87623   11.28475   5.129 1.66e-06 ***
DistFurn     -1.07042    0.51418  -2.082   0.0402 *
DistView     -1.02095    0.56536  -1.806   0.0743 .
FurnView      6.09379    4.72321   1.290   0.2003
EndFurn      -1.52608    7.50119  -0.203   0.8392
FloorFurn     0.14292    1.08085   0.132   0.8951
FloorEnd     -0.65866    1.58205  -0.416   0.6782
FloorView    -3.81792    1.54838  -2.466   0.0156 *
DistFloor     0.03864    0.13567   0.285   0.7765
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.137 on 90 degrees of freedom
Multiple R-squared:  0.7747,    Adjusted R-squared:  0.7371
F-statistic: 20.63 on 15 and 90 DF,  p-value: < 2.2e-16
```

Conclusion: The highest R-square obtained is 0.7747 with the following predictors: Distance + Distance Square + Floor + Floor Square + Furnish + End + View + Distance/Furniture + Furniture/View + End/Furniture + Furniture/View + Floor/Furniture + Floor/End + Floor/View + Distance/Floor)

Problem 2

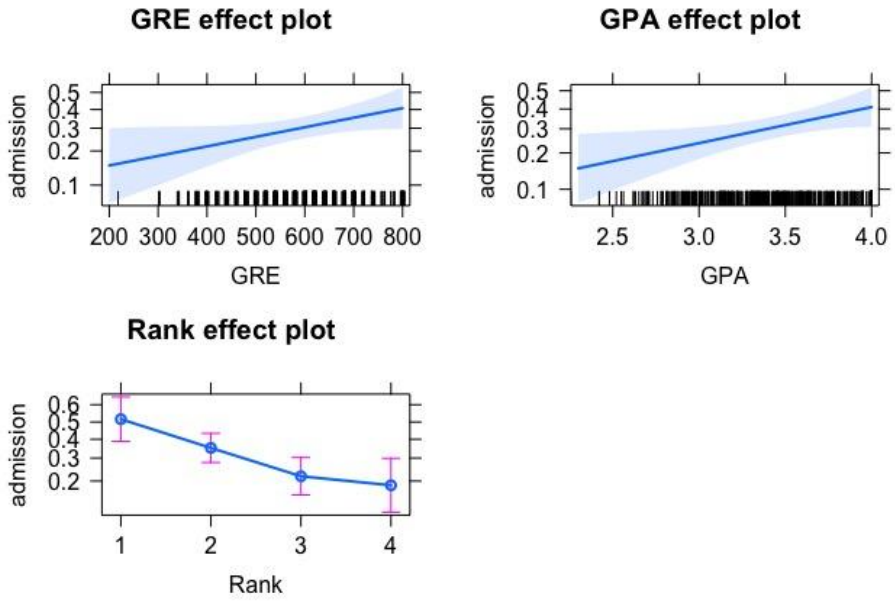
> library(Amelia)

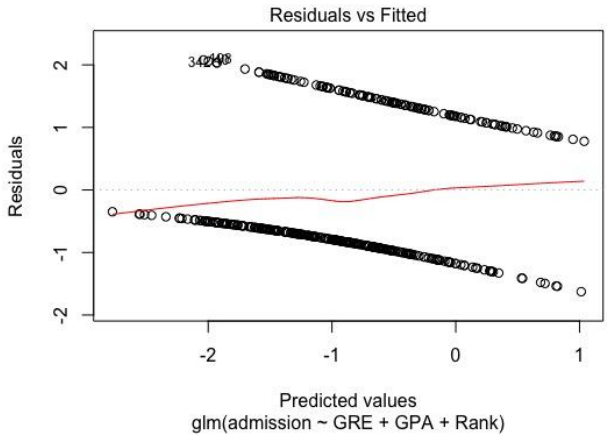
```
> library(effects)
> library(carData)
> library(lmtest)
> library(RVAideMemoire)
> library(DescTools)
> library(readxl)
> Admission <- read_excel("~/Desktop/Admission.xlsx")
> View(Admission)
> missmap(Admission, main = "Missing values vs observed")
> library(aod)
> library(ggplot2)
> summary(Admission)
   admission          GRE             GPA             Rank
 Min.   :0.0000   Min.   :220.0   Min.   :2.260   Min.   :1.000
 1st Qu.:0.0000   1st Qu.:520.0   1st Qu.:3.130   1st Qu.:2.000
 Median :0.0000   Median :580.0   Median :3.395   Median :2.000
 Mean   :0.3175   Mean   :587.7   Mean   :3.390   Mean   :2.485
 3rd Qu.:1.0000   3rd Qu.:660.0   3rd Qu.:3.670   3rd Qu.:3.000
 Max.   :1.0000   Max.   :800.0   Max.   :4.000   Max.   :4.000

> sapply(Admission, sd)
  admission         GRE         GPA        Rank
  0.4660867 115.5165364   0.3805668   0.9444602

> xtabs(~admission + Rank, data = Admission)
                    Rank
admission    1     2     3     4
        0   28    97    93    55
        1   33    54    28    12

> table <- xtabs(~admission + Rank, data = Admission)
> summary(table)
Call: xtabs(formula = ~admission + Rank, data = Admission)
Number of cases in table: 400
Number of factors: 2
Test for independence of all factors:
        Chisq = 25.242, df = 3, p-value = 1.374e-05

> Admission$Rank <- factor(Admission$Rank)
> mylogit <- glm(admission ~ GRE + GPA + Rank, data = Admission, family = "binomial")
```

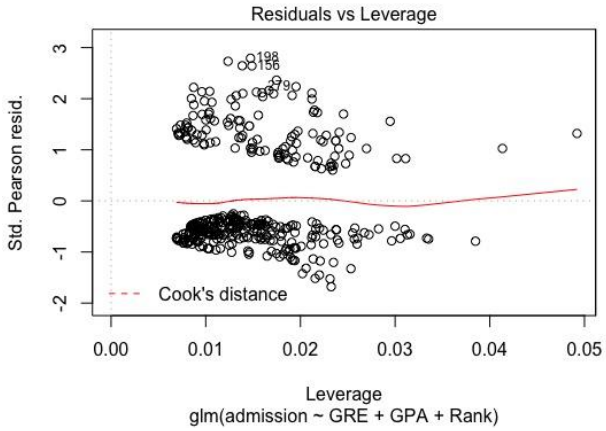
```
> summary(mylogit)

Call:
glm(formula = admission ~ GRE + GPA + Rank, family = "binomial",
    data = Admission)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6268  -0.8662  -0.6388   1.1490   2.0790

Coefficients:
             Estimate   Std. Error  z value  Pr(>|z|)
(Intercept) -3.989979   1.139951    -3.500   0.000465 ***
GRE          0.002264   0.001094     2.070   0.038465 *
GPA          0.804038   0.331819     2.423   0.015388 *
Rank2       -0.675443   0.316490    -2.134   0.032829 *
Rank3       -1.340204   0.345306    -3.881   0.000104 ***
Rank4       -1.551464   0.417832    -3.713   0.000205 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 499.98  on 399  degrees of freedom
Residual deviance: 458.52  on 394  degrees of freedom
AIC: 470.52

Number of Fisher Scoring iterations: 4

> anova(mylogit)
Analysis of Deviance Table

Model: binomial, link: logit

Response: admission

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev
```

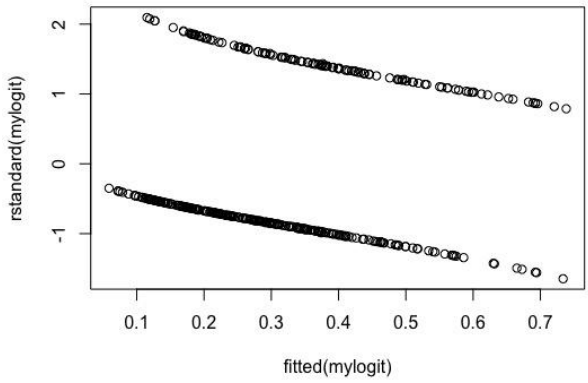
```
NULL                399     499.98
GRE   1  13.9204      398     486.06
GPA   1   5.7122      397     480.34
Rank  3  21.8265      394     458.52


> exp(coef(mylogit))
(Intercept)         GRE         GPA           Rank2           Rank3           Rank4
  0.0185001   1.0022670   2.2345448     0.5089310       0.2617923       0.2119375


> wald.test(b = coef(mylogit), Sigma = vcov(mylogit), Terms = 4:6)
Wald test:
----------

Chi-squared test:
X2 = 20.9, df = 3, P(> X2) = 0.00011

> anova(mylogit,update(mylogit, ~1),test="Chisq")
Analysis of Deviance Table

Model 1: admission ~ GRE + GPA + Rank
Model 2: admission ~ 1
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       394     458.52
2       399     499.98 -5  -41.459 7.578e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


> lrtest(mylogit)
Likelihood ratio test

Model 1: admission ~ GRE + GPA + Rank
Model 2: admission ~ 1
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   6 -229.26
2   1 -249.99 -5 41.459  7.578e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

> plot(allEffects(mylogit))









Report for problem 2:

The regression objective is to investigate and determine which factors have an effect on admission into graduate school. Based on the given data, fitting a logistic model was done using admission to graduate school as the dichotomous response variable (0 = not admitted, 1 = admitted) and GRE, GPA, and Rank as predictor variables. More specifically, GRE and GPA are the continuous variables and Rank is treated as a factor taking on the values 1 through 4.

A two-way contingency table of admission by prestige level, or Rank, was created. To test whether the two classifications are independent, a chi-square test that was conducted. The p-value of 1.374e-05 for this test shows that admission and prestige level are independent classifications.

To test the overall effect of the Rank variable, the Wald Chi-Squared test was conducted. The p-value of 0.00011 shows that the overall effect of Rank is statistically significant.

Higher GPA increases the likelihood of being admitted since the odds ratio is >=1. Ranks 2, 3, and 4 all have odds ratios <1, resulting in lower odds of being admitted if the undergraduate institution has rank 2, 3, or 4. GRE has an odds ratio of about 1, which doesn't affect the odds of being admitted.