Solution to Assignment 1

MSBX-5310 (Customer Analyitcs)

Overview

The data sets used for this assignment are the same that as the ones we used in the in-class consulting project – Dating-men.csv and Dating-women.csv. Your in-class work was focused on the women's dataset. Here you will mainly work with the men's data. As with the women's data, here your goal is to understand what drives men's popularity on the dating website.

As a recap, the data were obtained from the user records of a large US dating website. We have information on how many "first-contact" e-mails a user received during a specific time period. A first-contact e-mail is an unsolicited e-mail from some other user on the site. We also have information describing the users, such as a rating of her or his looks based on the posted profile picture(s), height, body mass index, education level (measured by years of schooling), age, and the days the user was active on the dating site during a specific observation period.

- `emails`: Number of first-contact e-mails received
- `rating`: Rating of posted picture
- `height`: Measured in inches
- `bmi`: Body mass index
- `education`: Years of education
- `age`: Age group (categorical variable). 1 = 31-35 years, 2 = 36-40 years, 3 = 41-45 years
- `days_active`: Days user was active on the site during observation period

Using the data Dating-men.csv, answer the following questions.

Questions

Question 1 Read and summarize the data in R. (Hint: Use the describe command from the `psych` package.)

```
men_DF = read.csv("Dating-Men.csv")
women_DF = read.csv("Dating-Women.csv")
library("psych")
describe(men_DF)
```

```
##               vars    n  mean    sd median trimmed   mad   min    max
## emails           1 1507  7.69 10.26   4.00    5.62  4.45  0.00  88.00
## rating           2 1507 -0.09  0.54  -0.13   -0.11  0.51 -1.94   2.53
## height           3 1507 70.93  2.69  71.50   70.93  2.97 61.00  85.00
## bmi              4 1507 25.76  2.81  25.44   25.59  2.20 12.37  36.44
## yrs_education    5 1507 15.66  2.58  16.00   15.72  2.97  8.00  21.00
## age              6 1507  1.95  0.82   2.00    1.93  1.48  1.00   3.00
## days_active      7 1507 76.68 34.13  85.00   80.90 34.10  1.00 108.00
##                range  skew kurtosis   se
## emails         88.00  2.97    12.54 0.26
## rating          4.47  0.38     0.61 0.01
## height         24.00  0.06     0.64 0.07
## bmi            24.07  0.53     1.41 0.07
```

```
## yrs_education  13.00 -0.47     0.79 0.07
## age             2.00  0.10    -1.51 0.02
## days_active   107.00 -0.67    -0.91 0.88
```

```
describe(women_DF)
```

```
##               vars    n  mean    sd median trimmed   mad   min    max
## emails           1 1366 32.72 31.81  22.00   27.51 23.72  0.00 186.00
## rating           2 1366  0.11  0.58   0.08    0.09  0.57 -1.49   3.14
## height           3 1366 65.19  2.70  65.50   65.14  2.97 59.00  75.50
## bmi              4 1366 22.46  3.71  21.79   21.98  2.70 16.20  50.07
## yrs_education    5 1366 15.45  2.62  16.00   15.58  2.97  8.00  21.00
## age              6 1366  2.02  0.82   2.00    2.03  1.48  1.00   3.00
## days_active      7 1366 73.49 34.71  78.00   77.08 44.48  1.00 108.00
##                range  skew kurtosis   se
## emails        186.00  1.72     3.69 0.86
## rating          4.63  0.34     0.32 0.02
## height         16.50  0.10    -0.42 0.07
## bmi            33.87  2.20     8.54 0.10
## yrs_education  13.00 -0.60     0.51 0.07
## age             2.00 -0.04    -1.51 0.02
## days_active   107.00 -0.52    -1.09 0.94
```

Question 2 Compare the summary statistics (means of the different variables) for the men's data with the women's data we analyzed in class. Are they different? If so, how and on which variables? (Hint: Focus on the variables for which the means are very different. Ignore the variables with small differences in means.)

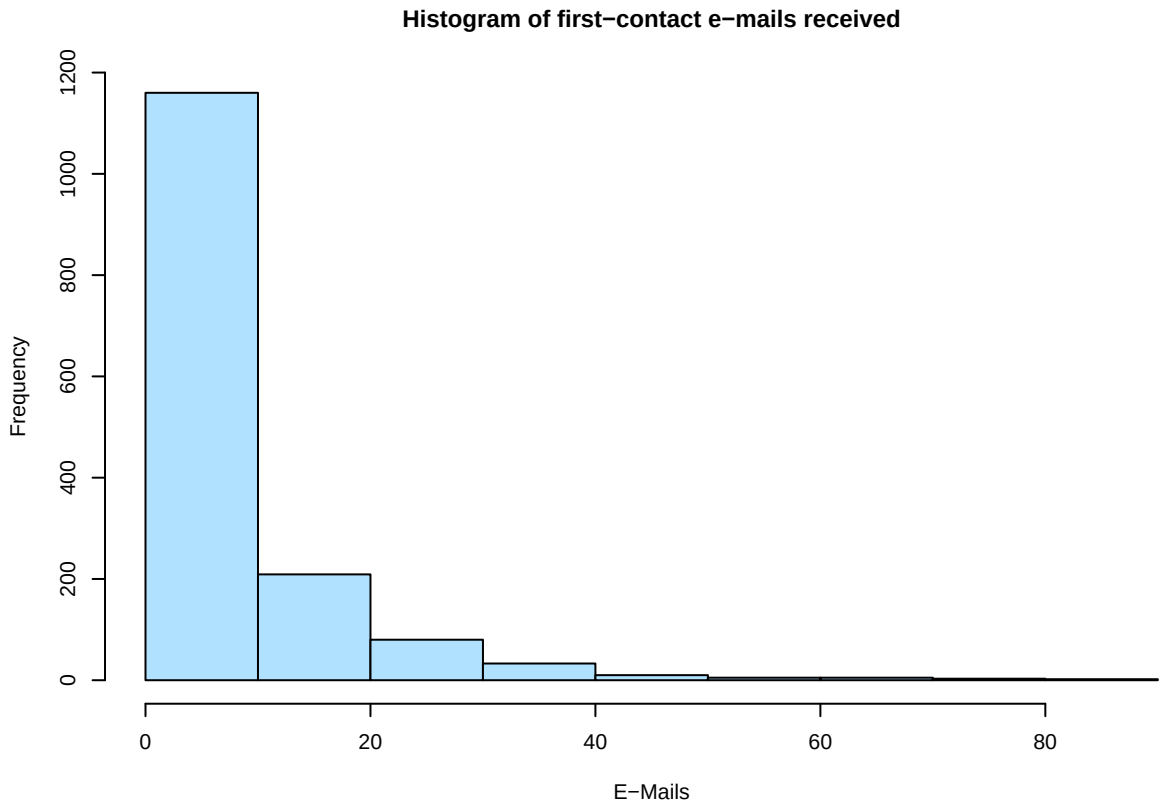
There are two main differences in the variables by gender:

1. On avergae, women get more first-contact emails (mean = 32.72) than men (mean = 7.69). This suggests that women are less likely to initiate first contacts on the dating site.
2. On average, women have lower BMIs then men (22.46 vs. 25.76)
3. As expected, men are taller then women on average (70.93 inches vs. 65.19 inches).

Question 3 Now perform some simple exploratory analysis of the data using graphs.

Question 3a Plot a histogram of the number of first-contact emails that men receive.

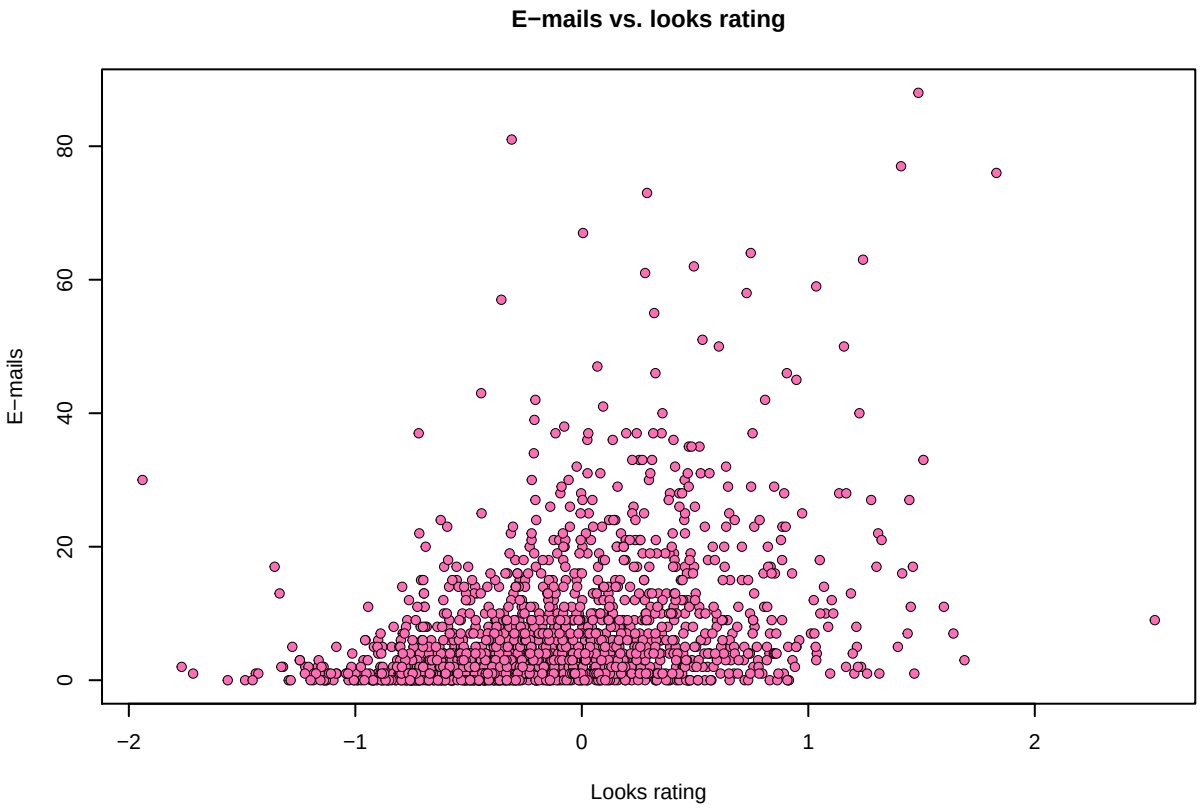
```
par(cex = 0.65)
hist(men_DF$emails,
     col  = "lightskyblue1",
     main = "Histogram of first-contact e-mails received",
     xlab = "E-Mails")
```



###

Question 3b Make a scatter plot to visualize the relationship between number of first contact emails that a man receives and his looks rating.

```
par(cex = 0.65)
plot(men_DF$rating, men_DF$emails,
     pch = 21, lwd = 0.4, bg = "hotpink1",
     main = "E-mails vs. looks rating",
     xlab = "Looks rating", ylab = "E-mails")
```



Question 3c What inferences can you draw about men's popularity from these pictorial representations of the data?

There seems to be a mildly positive relationship between looks ratings and emails received.

Question 4 Now run a linear regression with emails as the dependent variable and rating as the independent variable and show the results from this regression. Also plot the regression line on the scatter plot.

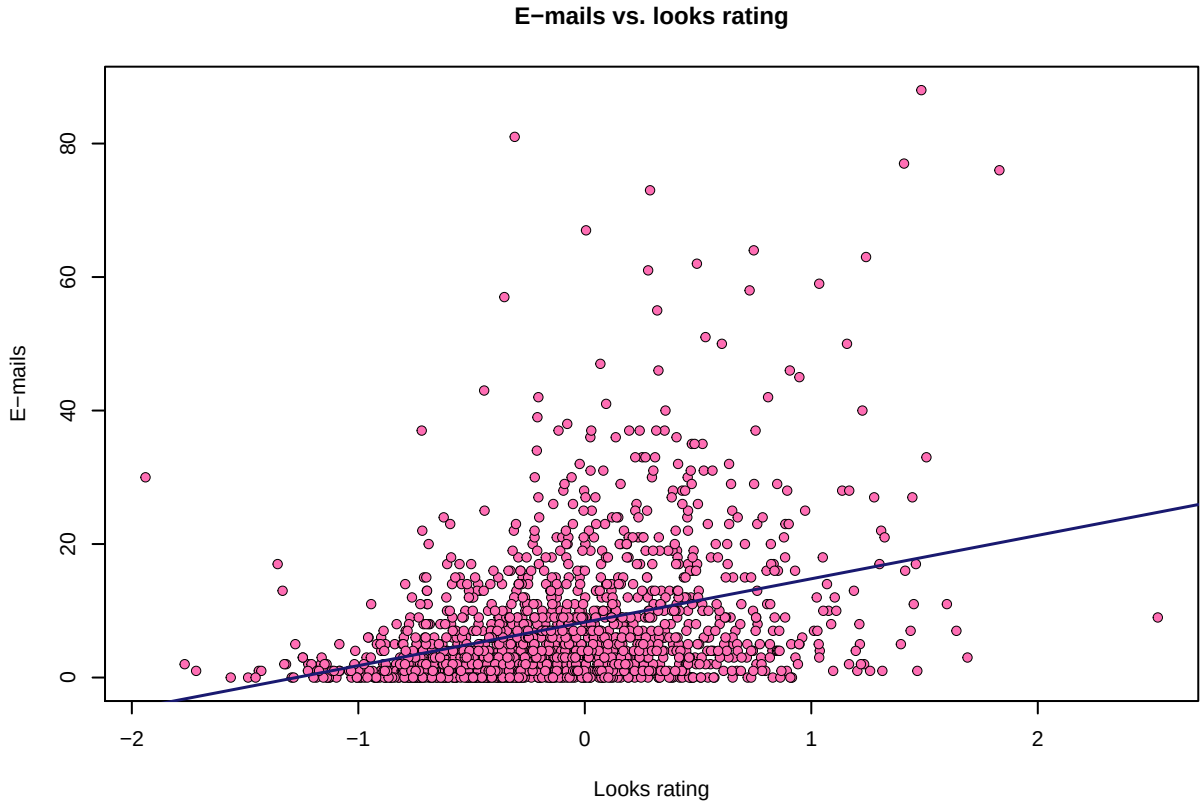
```
lm_fit_1 = lm(emails ~ rating, data = men_DF)
summary(lm_fit_1)
```

```
##
## Call:
## lm(formula = emails ~ rating, data = men_DF)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.842  -5.334  -2.391   2.217  74.724
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.2900     0.2521   32.88   <2e-16 ***
## rating        6.5061     0.4616   14.10   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 9.648 on 1505 degrees of freedom
## Multiple R-squared:  0.1166, Adjusted R-squared:  0.116
## F-statistic: 198.7 on 1 and 1505 DF,  p-value: < 2.2e-16
```

```r
par(cex = 0.65)
plot(men_DF$rating, men_DF$emails, pch = 21, lwd = 0.4, bg = "hotpink1",
     main = "E-mails vs. looks rating",
     xlab = "Looks rating", ylab = "E-mails")

abline(lm_fit_1,
       lwd = 1.5, col = "midnightblue")
```



Question 5 Based on the regression results from previous question, answer the following questions:

Question 5a What is the intercept and what is the coefficient of rating? Are these βs significant? How do you interpret these coefficients?

The intercept is 8.29 and the coefficient of ratings is 6.51. Both are significant. Interpretation: At a rating of zero, men in the data receive 8.29 emails on average. Further, for every unit increase in rating, the number of emails increases by 6.51.

Question 5b Compare the results of this regression with the results from the analogous regression on the women's data (which we analyzed in class). How does the effect of looks (rating) on women's popularity differ from the effect of looks on men's popularity. What does this comparison tell you about how women and respond to looks in the dating market?

```
lm_fit_1w = lm(emails ~ rating, data = women_DF)
summary(lm_fit_1w)
```

```
##
## Call:
## lm(formula = emails ~ rating, data = women_DF)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -72.703 -17.472  -6.665  11.727 158.936
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  30.1557     0.7845   38.44   <2e-16 ***
## rating       24.2450     1.3286   18.25   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.53 on 1364 degrees of freedom
## Multiple R-squared:  0.1962, Adjusted R-squared:  0.1956
## F-statistic:   333 on 1 and 1364 DF,  p-value: < 2.2e-16
```

Both the intercept and coefficient of the rating parameter are smaller in the men's data compared to the women's data. Thus, 1) men receive fewer emails on average, and 2) the number of emails that men get is less responsive to the looks rating. This suggests that looks play a smaller role in the men's popularity on the website compared to women. In other words, men are more likely to respond to women's looks than the other way around.

Question 6 Now run a multiple regression with emails as the dependent variable and both rating and bmi as independent variables.

```
lm_fit_2 = lm(emails ~ rating + bmi, data = men_DF)
summary(lm_fit_2)
```

```
##
## Call:
## lm(formula = emails ~ rating + bmi, data = men_DF)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.999  -5.348  -2.370   2.249  74.638
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.52310    2.34509   2.355   0.0186 *
## rating       6.63352    0.47383  14.000   <2e-16 ***
## bmi          0.10786    0.09089   1.187   0.2355
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.647 on 1504 degrees of freedom
## Multiple R-squared:  0.1174, Adjusted R-squared:  0.1163
## F-statistic: 100.1 on 2 and 1504 DF,  p-value: < 2.2e-16
```

Question 6a Interpret the coefficients.

The coefficient for rating is 6.63, indicating for every unit increase in rating, the number of emails expected will increase by 6.63. The coefficient for bmi is .108, indicating that for every unit increase in bmi, the expected number of emails will increase by .108. Interestingly, BMI has no significant effect on the number of emails men receive.

Question 6b Does the effect of rating change much compared to the regression with only rating as the independent variable?

We also find that the rating variable did not change significantly. This is because the BMI variable is not a significant predictor of emails. Hence there is no confound between bmi and rating.

Question 7 Now run a multiple regression with all the continuous variables (rating, height, bmi, education, days_active) as independent variables and emails as the dependent variable. Interpret the coefficients.

```
lm_fit_3 = lm(emails ~ rating + height + bmi + yrs_education + days_active, data = men_DF)
summary(lm_fit_3)
```

```
##
## Call:
## lm(formula = emails ~ rating + height + bmi + yrs_education +
##     days_active, data = men_DF)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.102  -5.196  -1.785   2.405  71.661
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -29.245527   7.041111  -4.154 3.46e-05 ***
## rating          6.449319   0.458182  14.076  < 2e-16 ***
## height          0.359369   0.089190   4.029 5.88e-05 ***
## bmi             0.133503   0.087594   1.524   0.1277
## yrs_education   0.198638   0.093653   2.121   0.0341 *
## days_active     0.071579   0.007012  10.208  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.278 on 1501 degrees of freedom
## Multiple R-squared:  0.1852, Adjusted R-squared:  0.1825
## F-statistic: 68.25 on 5 and 1501 DF,  p-value: < 2.2e-16
```

rating has a positive effect and this hasn't changed much from the previous regressions. Also, bmi remains insignificant. Height, years of education, and days_active all have a positive effect on the number of emails

received.

Question 8 Include age as a *categorical* variable. I.e., run a model that includes rating, height, bmi, education, days active and age as independent variables (with age as categorical) and emails as the dependent variable.

```
lm_fit_4 = lm(emails ~ rating + height + bmi + yrs_education + days_active + factor(age),
              data = men_DF)
summary(lm_fit_4)
```

```
##
## Call:
## lm(formula = emails ~ rating + height + bmi + yrs_education +
##     days_active + factor(age), data = men_DF)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.024  -5.024  -1.922   2.568  71.795
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -30.304611   6.997861  -4.331 1.59e-05 ***
## rating          7.017372   0.469123  14.958  < 2e-16 ***
## height          0.351147   0.088530   3.966 7.64e-05 ***
## bmi             0.159421   0.087073   1.831  0.06732 .
## yrs_education   0.187084   0.093043   2.011  0.04453 *
## days_active     0.068678   0.006985   9.832  < 2e-16 ***
## factor(age)2    1.515025   0.582677   2.600  0.00941 **
## factor(age)3    3.018363   0.598233   5.045 5.08e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.206 on 1499 degrees of freedom
## Multiple R-squared:  0.1988, Adjusted R-squared:  0.1951
## F-statistic: 53.15 on 7 and 1499 DF,  p-value: < 2.2e-16
```

Question 9 Based on the results from the previous regression, how do you interpret the effect of age on men's popularity?

Age has a positive effect on men's popularity! Men in age group 2 receive about 1.5 more emails then those in age group 1. Similarly, men in age group 3 receive about 3.0 emails more than those in age group 2.

Question 10 Compare the results for the regression model estimated in Question 8 (which uses all the variables in the data) for men and women. What are the differences? Based on your findings, what types of men and women do you want to attract to your site?

We first run the same model for women.

```
lm_fit_5 = lm(emails ~ rating + height + bmi + yrs_education + days_active + factor(age),
              data = women_DF)
summary(lm_fit_5)
```

```
##
## Call:
## lm(formula = emails ~ rating + height + bmi + yrs_education +
##     days_active + factor(age), data = women_DF)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.592 -17.506  -5.243  11.711 157.462
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   35.45650   20.03083   1.770  0.07694 .
## rating        20.51068    1.40165  14.633  < 2e-16 ***
## height         0.51520    0.27798   1.853  0.06405 .
## bmi           -1.54207    0.21530  -7.162 1.30e-12 ***
## yrs_education -0.70130    0.28841  -2.432  0.01516 *
## days_active    0.13117    0.02169   6.046 1.91e-09 ***
## factor(age)2  -2.04640    1.86802  -1.095  0.27349
## factor(age)3  -5.76479    1.90451  -3.027  0.00252 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.54 on 1358 degrees of freedom
## Multiple R-squared:  0.2542, Adjusted R-squared:  0.2504
## F-statistic: 66.14 on 7 and 1358 DF,  p-value: < 2.2e-16
```

The effects of the variables on men and women are quite different.

1. Women's emails are more responsive to their looks rating. A one unit increase in looks rating leads to an increase in 20 emails for women and only 7 emails for men.
2. For both men and women, there is a small positive effect of height. The effect is stronger for men.
3. BMI has no significant impact on the number of emails men get, but has negative effect for women.
4. Years of education has a positive effect for men, but has a negative effect on women.
5. Days active is very similar for both.
6. Age has a negative effect for women, but positive effect for men. Thus, younger women tend to get more emails, but older men get more emails.

To the extent the website wants to maximize communications, these differences suggest that the website should focus on attracting: 1) good looking, lower BMI, younger women, with lower education to appeal to their male consumers, and 2) good to moderate looking, older men with more education to appeal to their female consumers.

Question 11 Finally, explore the impact of interactions and non-linear regression terms for the men's dataset.

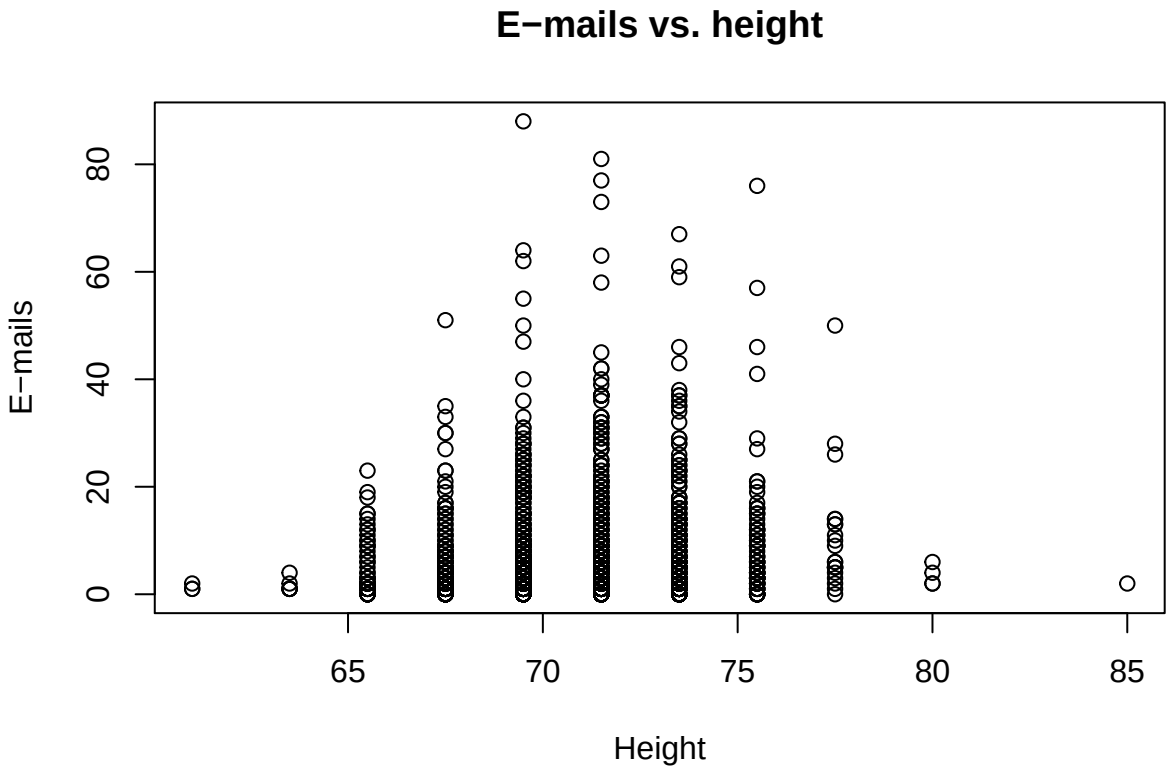
Question 11a First, run a model that adds a quadratic height term to the model in Question 8 (rating, height, bmi, education, days active and age (as categorical) as independent variables and emails as the dependent variable). Then, interpret the coefficients involving height. What pattern of height preferences corresponds to such coefficients? Can this pattern be seen in the data?

```
lm_fit_6= lm(emails ~ rating + height + bmi + yrs_education + days_active + factor(age)
             + I(height^2), data = men_DF)
summary(lm_fit_6)
```

```
##
## Call:
## lm(formula = emails ~ rating + height + bmi + yrs_education +
##     days_active + factor(age) + I(height^2), data = men_DF)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.051  -5.143  -1.913   2.491  71.536
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -235.92525  102.03907  -2.312  0.02091 *
## rating           6.99897    0.46873  14.932  < 2e-16 ***
## height           6.15365    2.87408   2.141  0.03243 *
## bmi              0.15791    0.08699   1.815  0.06967 .
## yrs_education    0.18342    0.09297   1.973  0.04867 *
## days_active      0.06832    0.00698   9.787  < 2e-16 ***
## factor(age)2     1.50649    0.58210   2.588  0.00975 **
## factor(age)3     3.02908    0.59764   5.068 4.51e-07 ***
## I(height^2)     -0.04085    0.02023  -2.020  0.04358 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.197 on 1498 degrees of freedom
## Multiple R-squared:  0.201,  Adjusted R-squared:  0.1968
## F-statistic: 47.11 on 8 and 1498 DF,  p-value: < 2.2e-16
```

Since height appears in the model with both linear and quadratic terms, we can no longer interpret the height coefficients separately. We can say that the contribution to the number of predicted emails from height h is $6.15h$ - 0.04h^2. Since the sign on the squared term is negative, the shape of the curve is a downward-facing parabola (inverted U), which implies there is a maximal point – we can think of this maximal point as the "ideal" height. This pattern may be seen in the data, by plotting emails vs height for the men_DF data:

```
plot(men_DF$height, men_DF$emails,
     main = "E-mails vs. height",
     xlab = "Height", ylab = "E-mails")
```



Question 11b Next, run a model that adds the interaction of age (as a categorical variable) with rating to the model in Question 8 (rating, height, bmi, education, days active and age (as categorical) as independent variables and emails as the dependent variable). Then, interpret the coefficients involving rating and age. What does the significance of the interaction coefficients indicate?

```
lm_fit_7= lm(emails ~ rating + height + bmi + yrs_education + days_active
             + factor(age)  + factor(age):rating, data = men_DF)
summary(lm_fit_7)
```

```
##
## Call:
## lm(formula = emails ~ rating + height + bmi + yrs_education +
##     days_active + factor(age) + factor(age):rating, data = men_DF)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.236  -5.093  -1.827   2.580  71.655
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -30.250127   7.000617  -4.321 1.66e-05 ***
## rating                7.530973   0.735750  10.236  < 2e-16 ***
## height                0.347217   0.088652   3.917 9.38e-05 ***
## bmi                   0.164605   0.087210   1.887   0.0593 .
## yrs_education         0.189714   0.093091   2.038   0.0417 *
## days_active           0.068848   0.006989   9.851  < 2e-16 ***
## factor(age)2          1.438759   0.587766   2.448   0.0145 *
```

```
## factor(age)3          3.106737   0.617969   5.027 5.57e-07 ***
## rating:factor(age)2  -1.283762   1.065426  -1.205   0.2284
## rating:factor(age)3  -0.283069   1.137914  -0.249   0.8036
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.208 on 1497 degrees of freedom
## Multiple R-squared:  0.1997, Adjusted R-squared:  0.1949
## F-statistic:  41.5 on 9 and 1497 DF,  p-value: < 2.2e-16
```

As discussed in class, the presence of interactions implies that coefficients contribute different amounts to predicted emails, depending on the value of the related data variables. We can make the following statements:

rating: for age group 1, increasing rating by 1 point increases emails by 7.53

factor(age)2: for rating=0, age group 2 members receive 1.44 more emails than group 1

factor(age)3: for rating=0, age group 3 members receive 3.11 more emails group 1

rating:factor(age)2: relative to age group 1, group 2 members receive -1.28 fewer emails for each additional rating point. I.e., group 2 members receive 7.53-1.28=6.25 more emails for each additional rating point.*

rating:factor(age)3: relative to age group 1, group 3 members receive -0.28 fewer emails for each additional rating point. I.e., group 2 members receive 7.53-0.28=7.25 more emails for each additional rating point.

The interpretations above are literal in the sense of ignoring the statistical significance. Since both interaction terms are not significant at the conventional 5% level, we conclude there is no evidence for the effect of rating being different across male age groups.