MACS PS2

Neta Grossfeld

10/17/2018

```
getwd()
```

```
## [1] "/Users/netagrossfeld/Desktop/persp-analysis_A18/Assignments/A2"
```

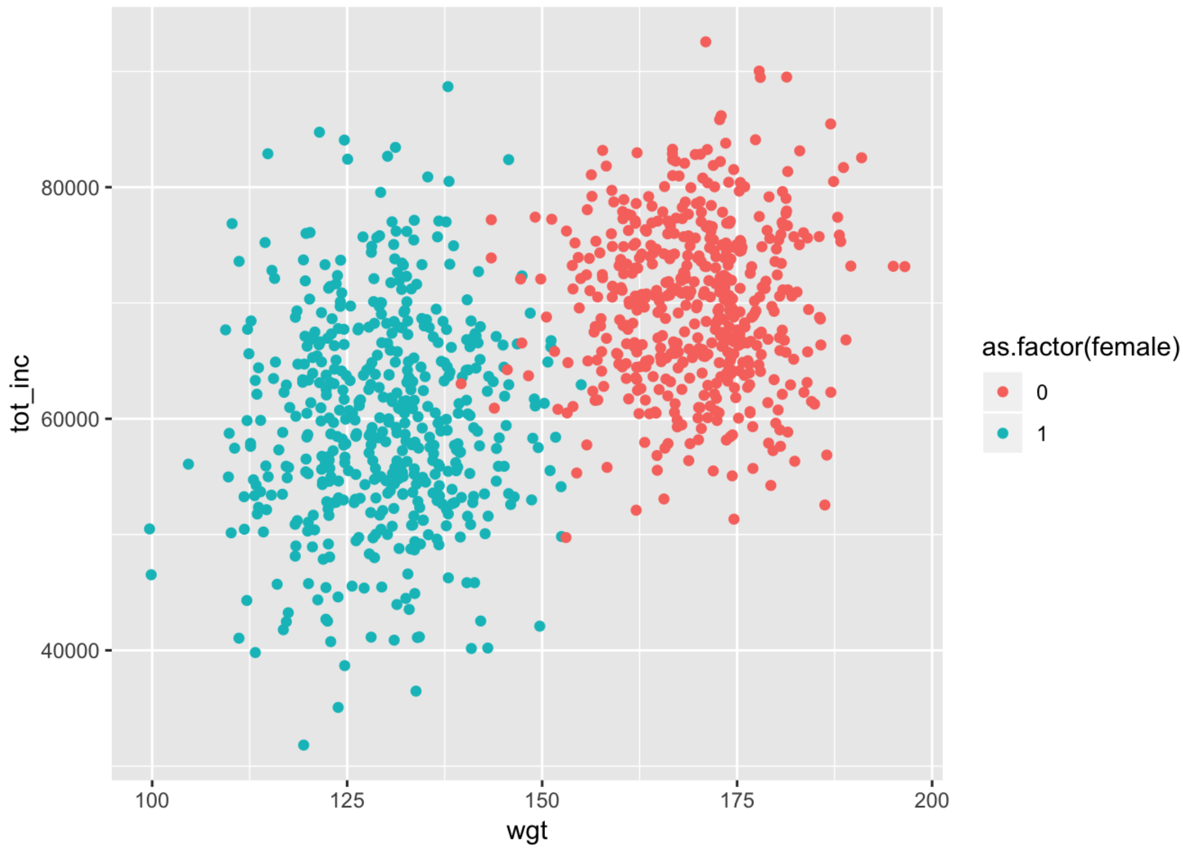
```
setwd("/Users/netagrossfeld/Desktop/persp-analysis_A18-master_2/Assignments/A2")
library(dplyr)
library(tidyverse)

#1. Imputing Age and Gender
best_income <- read_delim(file = 'BestIncome.txt', delim = ',',col_names = c("lab_inc", "cap_inc",
"hgt", "wgt"))
survey_income <- read_delim(file = 'SurvIncome.txt', delim = ',', col_names = c("tot_inc", "wgt",
"age", "female"))
summary(best_income)
```

```
##     lab_inc         cap_inc           hgt             wgt
##  Min.   :22918   Min.   : 1495   Min.   :58.18   Min.   :114.5
##  1st Qu.:51624   1st Qu.: 8612   1st Qu.:63.65   1st Qu.:143.3
##  Median :56969   Median : 9970   Median :65.00   Median :149.9
##  Mean   :57053   Mean   : 9986   Mean   :65.01   Mean   :150.0
##  3rd Qu.:62408   3rd Qu.:11340   3rd Qu.:66.36   3rd Qu.:156.7
##  Max.   :90060   Max.   :19882   Max.   :72.80   Max.   :185.4
```

```
summary(survey_income)
```

```
##     tot_inc           wgt              age            female
##  Min.   :31816   Min.   : 99.66   Min.   :25.74   Min.   :0.0
##  1st Qu.:58350   1st Qu.:130.18   1st Qu.:41.03   1st Qu.:0.0
##  Median :65281   Median :149.76   Median :44.96   Median :0.5
##  Mean   :64871   Mean   :149.54   Mean   :44.84   Mean   :0.5
##  3rd Qu.:71749   3rd Qu.:170.15   3rd Qu.:48.82   3rd Qu.:1.0
##  Max.   :92556   Max.   :196.50   Max.   :66.53   Max.   :1.0
```

```
#a) The scatterplot shows that the majority of females are 150 pounds or less, so we can impute gen
der based on whether or not the observation is 150 pounds or less. As for age, there is no clear tr
end, so we take the mean age and apply it to all observations.

ggplot(data=survey_income) +
  geom_point(mapping = aes(x=wgt, y=tot_inc, color = as.factor(female)))
```




```
ggplot(data=survey_income) +
  geom_point(mapping = aes(x=age, y=tot_inc, color = as.factor(female)))
```

```
#b)
best_income$gender <- ifelse(best_income$wgt < 150, 1, 0)
best_income$age <- mean(survey_income$age)

#c)
summary(best_income$gender)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  1.0000  0.5019  1.0000  1.0000
```

```
summary(best_income$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   44.84   44.84   44.84   44.84   44.84   44.84
```

```
sd(best_income$gender)
```

```
## [1] 0.5000214
```

```
sd(best_income$age)
```

```
## [1] 0
```

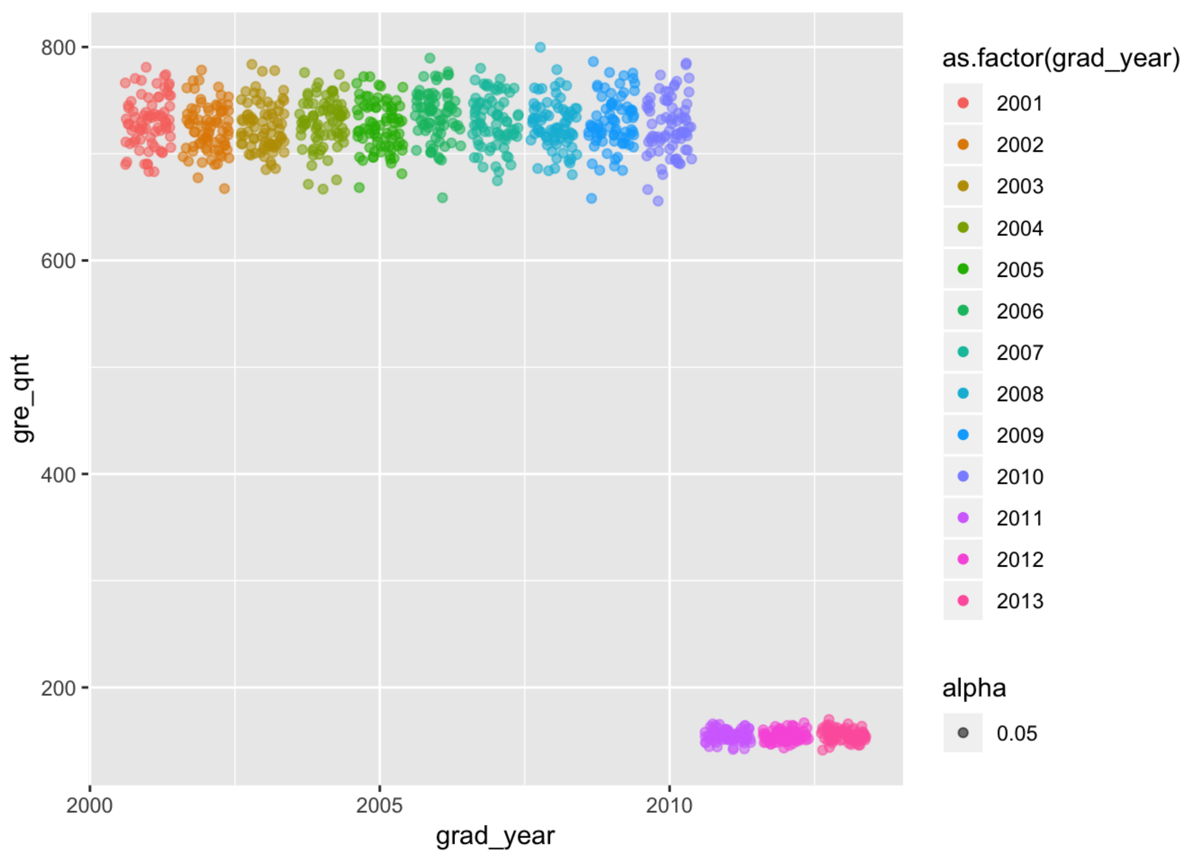
```
#d)
correlation <- cor(best_income)
round(correlation, 2)
```

```
##         lab_inc cap_inc   hgt   wgt gender age
## lab_inc    1.00    0.01  0.00  0.00  -0.01  NA
## cap_inc    0.01    1.00  0.02  0.01  -0.01  NA
## hgt        0.00    0.02  1.00  0.17  -0.14  NA
## wgt        0.00    0.01  0.17  1.00  -0.80  NA
## gender    -0.01   -0.01 -0.14 -0.80   1.00  NA
## age          NA      NA    NA    NA     NA   1
```

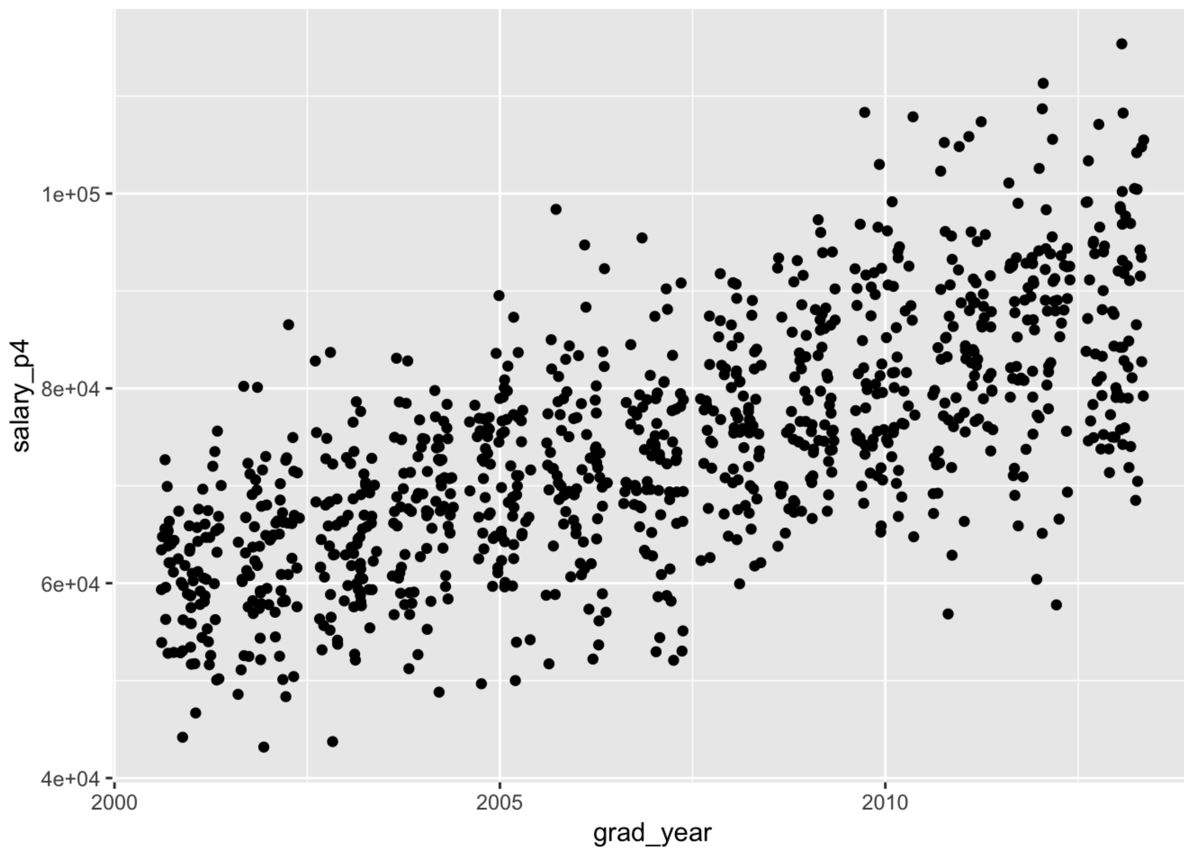
```
#2. Stationarity and Data Drift
income_intel <- read_delim(file = 'IncomeIntel.txt', delim = ',',col_names = c("grad_year", "gre_qn
t", "salary_p4"))

#a)
lm_s_g = lm(income_intel$salary_p4 ~ income_intel$gre_qnt)
summary(lm_s_g)
```

```
##
## Call:
## lm(formula = income_intel$salary_p4 ~ income_intel$gre_qnt)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -28761  -7049   -293   6549  37666
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           89541.293    878.764  101.89   <2e-16 ***
## income_intel$gre_qnt    -25.763      1.365  -18.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10460 on 998 degrees of freedom
## Multiple R-squared:  0.2631, Adjusted R-squared:  0.2623
## F-statistic: 356.3 on 1 and 998 DF,  p-value: < 2.2e-16
```

```
#b)
ggplot(data=income_intel) +
  geom_jitter(mapping = aes(x=grad_year, y=gre_qnt, color = as.factor(grad_year), alpha=.05))
```



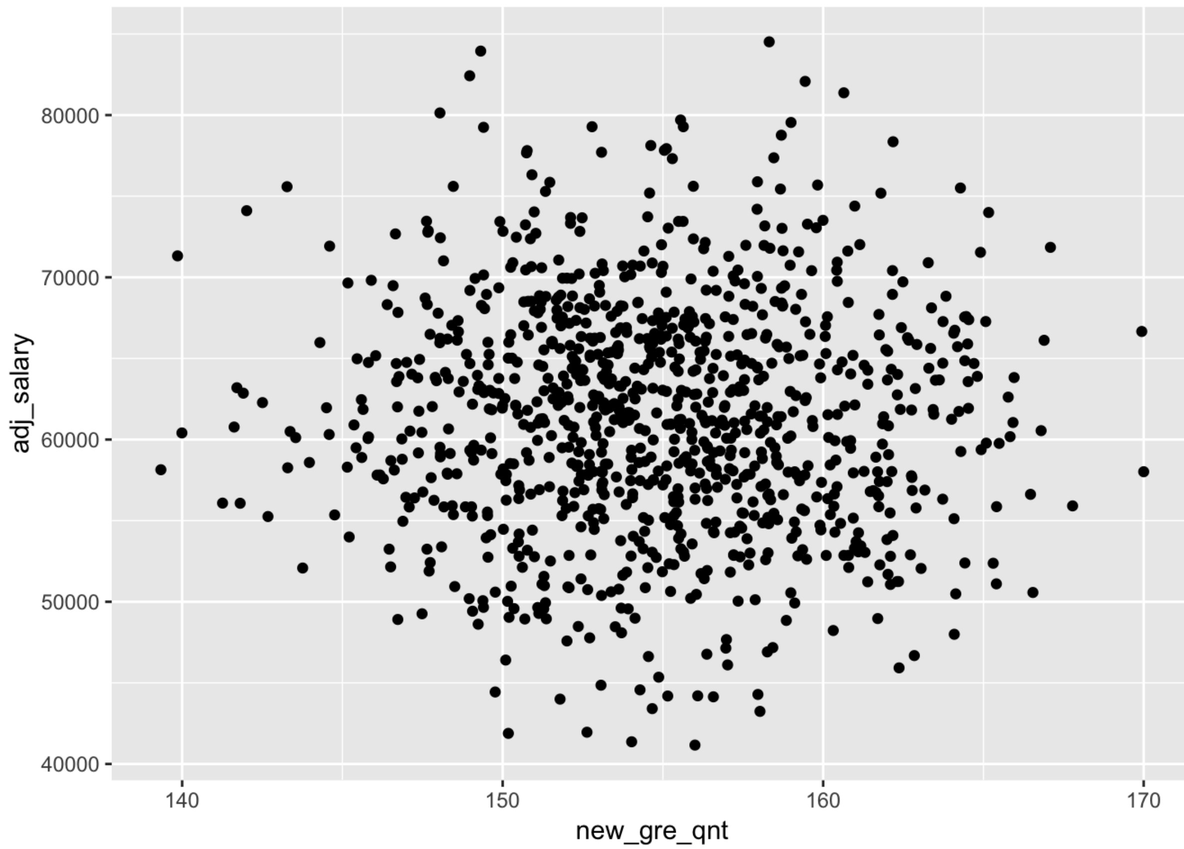
```
#The problem with using this variable to test my hypothesis is that the GRE quant scoring scale cha
nged in 2011. See below for the code that implements changing the scale for old scores.

income_intel$new_gre_qnt <- with(income_intel, ifelse(grad_year < 2011, gre_qnt * 170 / 800, gre_qn
t))

#c)
ggplot(data=income_intel) +
  geom_jitter(mapping = aes(x=grad_year, y=salary_p4))
```




```
# The problem is that inflation is not accounted for, since salaries have the same distribution but
 higher every year. I used Rick's solution to detrend the variable below.

by_grad_year <- group_by(income_intel, grad_year)
avg_inc_by_year <- summarise(by_grad_year, mean_salary=mean(salary_p4))

avg_growth_rate <- mean(diff(avg_inc_by_year$mean_salary, lag = 1, differences = 1)/((slice(avg_inc
_by_year, 1:12))$mean_salary))
avg_growth_rate
```

```
## [1] 0.03083535
```

```
income_intel$adj_salary <- income_intel$salary_p4/((1+avg_growth_rate)**(income_intel$grad_year - 2
001))

ggplot(data=income_intel) +
  geom_jitter(mapping = aes(x=new_gre_qnt, y=adj_salary))
```




```
#d)
new_lm_s_g = lm(income_intel$adj_salary ~ income_intel$new_gre_qnt)
summary(new_lm_s_g)
```

```
##
## Call:
## lm(formula = income_intel$adj_salary ~ income_intel$new_gre_qnt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20213.6  -4783.4    123.4   4793.5  23219.5
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              66834.37    6968.68   9.591   <2e-16 ***
## income_intel$new_gre_qnt   -34.97      44.99  -0.777    0.437
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7137 on 998 degrees of freedom
## Multiple R-squared:  0.0006052,  Adjusted R-squared:  -0.0003962
## F-statistic: 0.6043 on 1 and 998 DF,  p-value: 0.4371
```

3. Assessment of Kossinets and Watts (2009) (3 points). Read the paper, Kossinets and Watts (2009). Write a one-to-two page response to the paper that answers the following questions. Make sure that your response is a single flowing composition that follows the rules of spelling, grammar, and good writing.

In the *Origins of Homophily in an Evolving Social Network*, researchers Kossinets and Watts attempt to discover how individuals choose to form or break ties within their community, and how those individuals' choices reveal homophily. Specifically, to what extent do individual preferences and structural constraints affect homophily?

To answer this question, Kossinets and Watts created a database from three data sources. The first data source was logs of email interactions within the university, using university emails. The second data source was made up of individual attributes, such as status, gender, and age. The third data source consisted of class registration records. These three data sources include data from the course of two calendar years or six academic semesters. In the span of 270 days, researchers observed and created a cleaned dataset of 7,156,162 messages exchanged by 30,390 consistently active, or stable, email users. Appendix A in the article lists all variables with their description and definition.

There is, however, a potential problem that the data cleaning process used for the final dataset mentioned above may have introduced. A number of differing email addresses, and therefore email interactions or messages, were excluded from the final dataset. Such email addresses include those that were used by specific departments in the university, and thus had different email addresses that, while part of the university community, were unable to be matched with employee records. This subset of the data was excluded but could have provided the authors with a more detailed look at how structural opportunities may be induced when individuals are in the same department, and whether these individuals have similar attributes.

One weakness of matching the theoretical construct of "social relationships" to the data source of email logs is that emails aren't necessarily the communication method used by those with strong social relationship ties. Messages sent via cell phones or even phone calls may indicate a stronger social relationship than emails, simply because two individuals may need a stronger social relationship to exchange phone numbers in the first place. However, the email logs data source still captures the formation and maintenance of new social ties, albeit not as strong as other communication methods may be. Another weakness is that the email messages were not analyzed, due to privacy issues. Analysis of these messages may be a better indicator as to the type of social relationship the two email addresses are participating in. Fortunately, the authors are less interested in the structure of social networks and more interested in the evolution of the network itself, which these emails still capture.