Assignment 8

Matthew McAvoy

September 27, 2016

We begin my downloading the data and reading it in as a dataframe.

```
universities <- read.csv("C:/Users/homur/OneDrive/New College/EDA/Week 5/fl_university_system.csv")
```

1.

We want to know how many people work in the Florida state university system. Looking at the head of the data, we can see there are duplicates. Michael Abazinge with position # 18703000 works in two budget entities, but is most likely the same person. I will begin by saying the position number is a unique identifier for each person, then checking to see how well it works in this role.

```
library(sqldf)
```

```
## Loading required package: gsubfn
```

```
## Loading required package: proto
```

```
## Loading required package: RSQLite
```

```
## Loading required package: DBI
```

Since later commands will use SQL, we need to rename columns with a period in it, since SQL doesn't like them.

```
names(universities) <- sub("^Position.Number$", "Position_Number", names(universities))
names(universities) <- sub("^First.Name$", "First_Name", names(universities))
names(universities) <- sub("^Last.Name$", "Last_Name", names(universities))
names(universities) <- sub("^Employee.Type$", "Employee_Type", names(universities))
names(universities) <- sub("^Class.Code$", "Class_Code", names(universities))
names(universities) <- sub("^Class.Title$", "Class_Title", names(universities))
names(universities) <- sub("^Annual.Salary$", "Annual_Salary", names(universities))
names(universities) <- sub("^OPS.Term.Amount$", "OPS_Term_Amount", names(universities))
names(universities) <- sub("^Budget.Entity$", "Budget_Entity", names(universities))
```

```
posnum <- sqldf("SELECT count(Position_Number) FROM universities")
```

```
## Loading required package: tcltk
```

```
distinct_posnum <- sqldf("SELECT count(Position_Number) FROM (SELECT DISTINCT Position_Number FROM univ
posnum; distinct_posnum
```

```
##   count(Position_Number)
## 1                  86021
```

```
##   count(Position_Number)
## 1                  52730
```

This is promising, but looking at people that have a position number of zero's is numerous.

```
zero_posnum <- sqldf("SELECT count(Position_Number) FROM universities WHERE Position_Number = '0' ")
zero_posnum
```

```
##   count(Position_Number)
## 1                    437
```

This tells us there are quite a number of people that don't have position numbers correctly labeled for each person.

Perhaps the best way with the data we have to identify unique individuals is to match on position number, first name, last name, middle inital; or just on the latter three. Lets look at if we get any duplicates in each case.

```
sqldf("SELECT First_Name, Last_Name, MI, Position_Number, Annual_Salary FROM universities ORDER BY First
```

```
##     First_Name   Last_Name MI Position_Number Annual_Salary
## 1            A   ALBERTSON  M        00016290        119159
## 2            A        EJAZ  A        00002441         26100
## 3            A        EJAZ  A        00000000        240399
## 4            A    HECHICHE  .        00004325         93157
## 5     A COSKUN       SAMLI  .          316780        130356
## 6      A DAVID       KLINE  .          926100            NA
## 7  A H M ANWAR     SADMANI  .           37056         80000
## 8      A LEROY        ODOM  .           51918         95007
## 9      A YOUNG         KIM  .           61683         45175
## 10      A'NAJA     NEWSOME  M        00005983         46000
## 11      A'RION     RAYMOND  M           00000            NA
## 12      A'RION     RAYMOND  M           00000            NA
## 13       AAMIR        SOFI  A        00000000            NA
## 14       AARON     APONICK  .        00008033         82307
## 15       AARON     ARNETTE  D        00014332         25152
## 16       AARON    BANFIELD  G           42363         31363
## 17       AARON    BLACKHAM  U        00000000            NA
## 18       AARON    BLUMBERG  P        30951000         33400
## 19       AARON     BRAFMAN  F        00000000            NA
## 20       AARON      BROOME  D        00018796         46294
## 21       AARON      CAINES  R        00018179         54601
## 22       AARON     COUTURE  .        00027324         29105
## 23       AARON     CRELLER  B          441400         35000
## 24       AARON      DENSON  C        00000000            NA
## 25       AARON       DUMAS  D        42466000         44000
## 26       AARON       EVANS  H        01003757            NA
## 27       AARON         FAY  T           40191         93340
## 28       AARON      FISHER  C           00000            NA
## 29       AARON      FOWLER  W        00018176         21838
## 30       AARON      FOWLER  W        00018176          7279
## 31       AARON      FRANKE  J        10000000            NA
## 32       AARON      FRANKE  J        00000000            NA
```

```
## 33       AARON       GETER  .          51478         24199
## 34       AARON     GREASER  W       00006628         37080
## 35       AARON     HACKMAN  .         992485         55000
## 36       AARON      HAMLIN  E          54598         24941
## 37       AARON      HAYDEN  J       19617000         33280
## 38       AARON      HAYDEN  M       00004387         21953
## 39       AARON    HILLIARD  L       19325000        104502
## 40       AARON          HO  .       00000000            NA
## 41       AARON     HOLLAND  R       00000000            NA
## 42       AARON     HOLLAND  R       00023403         52000
## 43       AARON      HOOVER  H       00006867         95800
## 44       AARON        HOSE  D          40234         62970
## 45       AARON JOHNSON III  A       70004668         31616
## 46       AARON      KEYSER  G          36365         74049
## 47       AARON       KLINE  D       10000000            NA
## 48       AARON       KLINE  D       00000000            NA
## 49       AARON        KULA  D       01003849            NA
## 50       AARON        KULA  D         991172          5008
```

It seems the best way to identify a person is purely on their full name and not their position number as it might be different. For example Aaron R Holland has two position numbers 00000000 and 00023403, yet this is most likely the same person.

We can count how many people there are now, which appears to be 53512. When there are 86021 entries in our starting data, this might be accurate since this would mean about 40% of people work in more than one department, which in todays multi-facted world where professors need to wear multiple hats, seems pretty true to me.

```
sqldf("SELECT count(*) FROM (SELECT DISTINCT First_Name, Last_Name, MI FROM universities)")
```

```
##   count(*)
## 1    53512
```

Storing the professors based on name into a new dataframe.

```
profs <- sqldf("SELECT * FROM universities GROUP BY First_Name, Last_Name, MI")
```

2.

We can use the following SQL command to see what professor titles there are.

```
sqldf("SELECT DISTINCT Class_Title FROM profs WHERE Class_Title like '%PROFESSOR%' ORDER BY Class_Title
```

```
##                          Class_Title
## 1                 ASSISTANT PROFESSOR
## 2                    ASSOC PROFESSOR
## 3                 ASSOCIATE PROFESSOR
## 4             DEAN/ASSOCIATE PROFESSOR
## 5              DISTINGUISHED PROFESSOR
## 6         GRADUATE RESEARCH PROFESSOR
## 7                           PROFESSOR
```

```
## 8               PROVOST/ASSOC. PROFESSOR
## 9  UNIVERSITY SCHOOL ASSISTANT PROFESSOR
## 10 UNIVERSITY SCHOOL ASSOCIATE PROFESSOR
## 11           UNIVERSITY SCHOOL PROFESSOR
```

It looks like what we want are labeled only 'Professor'. We can use SQL again to count the number of professors with this title. We see there are 3926 full professors.

```
sqldf("SELECT count(*) FROM profs WHERE Class_Title like 'PROFESSOR'")
```

```
##   count(*)
## 1     3926
```

Storing information about full professors

```
full_profs <- sqldf("SELECT * FROM profs WHERE Class_Title like 'PROFESSOR'")
```

Confirming we have people only labeled as professors. Returns only our desired class_title.

```
sqldf('SELECT DISTINCT Class_Title FROM full_profs')
```

```
##   Class_Title
## 1   PROFESSOR
```

3

To look at median salary, we can simply use summary on our full professors salary column. We see the median salary is $103,700.

```
summary(full_profs$Annual_Salary)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       3   68380  103700  109900  141300  953900     186
```

4.

Since there is no distinguishing feature other than names in this data, perhaps the best way to identify females is to look for full professors with female names.

I will use names gathered from the site: http://deron.meranda.us/data/census-dist-female-first.txt. I then used the following command in the command line. This formats the file into a way we can use it. 'cat female_names.txt | awk'{print $1}'> female_names.csv' Now I read in the data and match where full professors have a matching first name.

```
female_names <- read.csv("C:/Users/homur/OneDrive/New College/EDA/Week 5/female_names.csv")
```

```
sqldf("SELECT count(*) FROM full_profs WHERE First_Name IN female_names")
```

```
##   count(*)
## 1     2544
```

Depending on how well made the female names file is, will determine the measure of confidence of my results. As to whether they are full professors, that I am pretty confident to say, as they are all marked as Professors.