Obj1.2

Reilly Maw

6/7/2018

I would first load the data into R in order to explore/test the data.

```
rm(list=ls())
data<-read.csv("jpmc_rse_assignment_data.csv",sep=",",header=T)
dim(data)
```

```
## [1] 27261    75
```

Three feasible issues include

a) Unreliable, incorrect, or impossible entries.

b) Seeing how many NA answers were given since some variables such as ER34209 and ER34210 have options to not answer. This will require compiling a list of variables that have this "opt-out" option and testing the percentage of data points that responded in this way. If the percentage is high, that could raise questions about the integrity/projectability of the data.

ii) Check the file for entries tagged as "possible listing error" if the number is high enough, then it could suggest that the data was disturbed in some way. This would again require going through every variable and seeing where this terminology arises.

iii) Checking for age outliers is important to verify that the data wasn't entered incorrectly. It is very unlikely that 0-5 or 105+ year old's are taking the survey. This option is very quick and easy. There are almost no costs associated and the benefit is that the survey can be trusted in the most rudimentary capacity.

```
### Test option iii)
# ER34104 is the age of the individual as of the 2011 interview, ER34204 is the age of
# individual as of the 2013 interview

age_11 <- data[7]
age_13 <- data[67]
summary(age_11)
```

```
##     ER34104
##  Min.   :   0.00
##  1st Qu.:  10.00
##  Median :  27.00
##  Mean   :  29.57
##  3rd Qu.:  46.00
##  Max.   :1004.00
```

```
summary(age_13)
```

```
##     ER34204
##  Min.   :  0.00
##  1st Qu.: 10.00
##  Median : 27.00
##  Mean   : 29.66
##  3rd Qu.: 46.00
##  Max.   :999.00
```

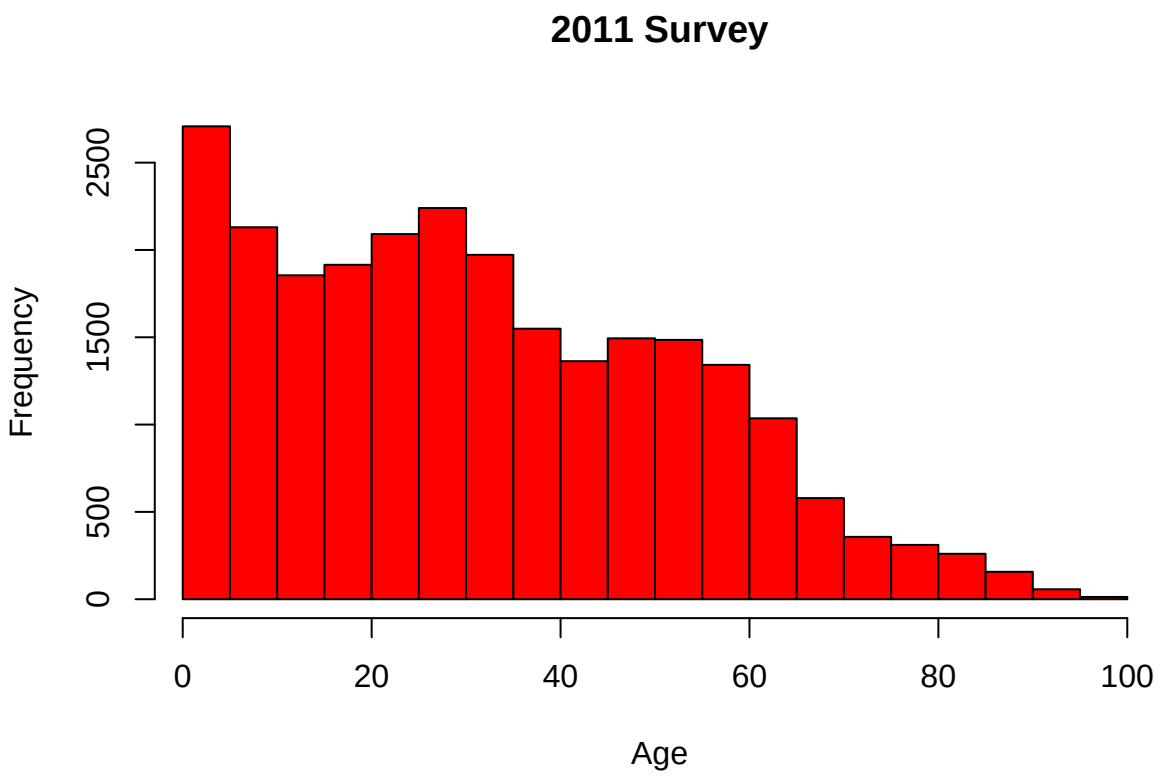
```
# One interesting find here is that in 2011 an individual was entered as 1,004 years old,
# which is obviously a mistake. This data point should probably be removed. It does prove
# that the document is not absent of data  entry issues.

# In the survey, 999 (refusal, NA, DK) and 0 (Inap.) represents an unknown age. So we will
# remove these entries.
clean_11 <- as.data.frame(age_11[! age_11$ER34104 %in% c(0,999,1004), ])
clean_13 <- as.data.frame(age_13[! age_13$ER34204 %in% c(0,999), ])

# 1stQ = 14, Median = 29, Mean = 32.15, 3rdQ = 48
summary(clean_11)
```

```
##  age_11[!age_11$ER34104 %in% c(0, 999, 1004), ]
##  Min.   :  1.00
##  1st Qu.: 14.00
##  Median : 29.00
##  Mean   : 32.15
##  3rd Qu.: 48.00
##  Max.   :100.00
```

```
hist(clean_11[,1],col="red",freq=T,xlim=c(1,100),main ='2011 Survey', xlab = "Age")
```

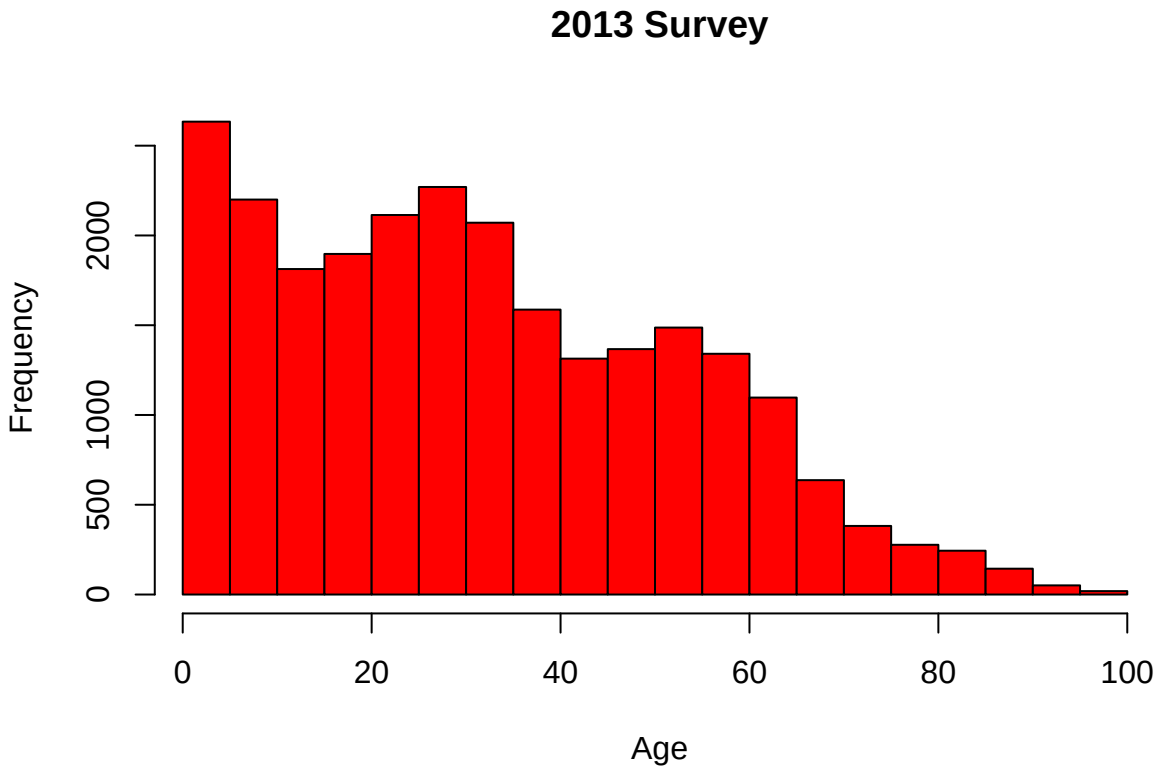



```
# 1stQ = 14, Median = 29, Mean = 32.17, 3rdQ = 49
summary(clean_13)
```

```
##  age_13[!age_13$ER34204 %in% c(0, 999), ]
##  Min.   :  1.00
##  1st Qu.: 14.00
##  Median : 29.00
##  Mean   : 32.17
```

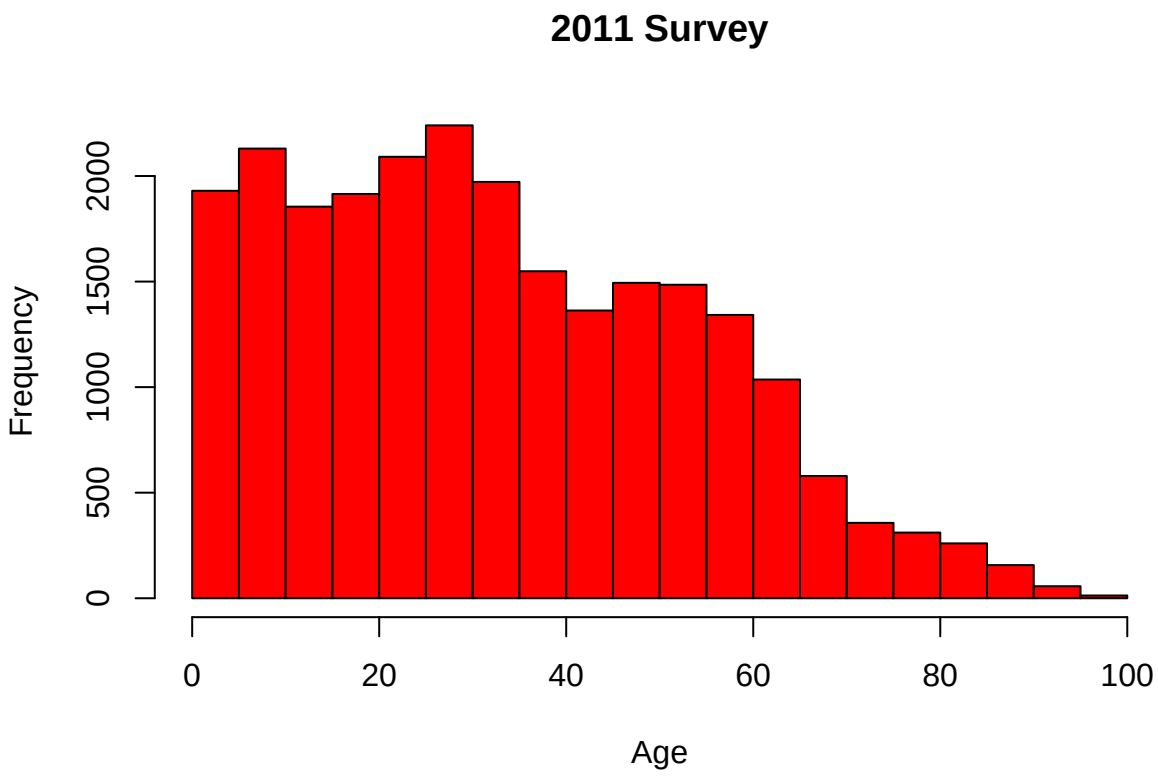
```
##  3rd Qu.: 49.00
##  Max.   :100.00
```

```
hist(clean_13[,1],col="red",freq=T,xlim=c(1,100),main ='2013 Survey', xlab = "Age")
```

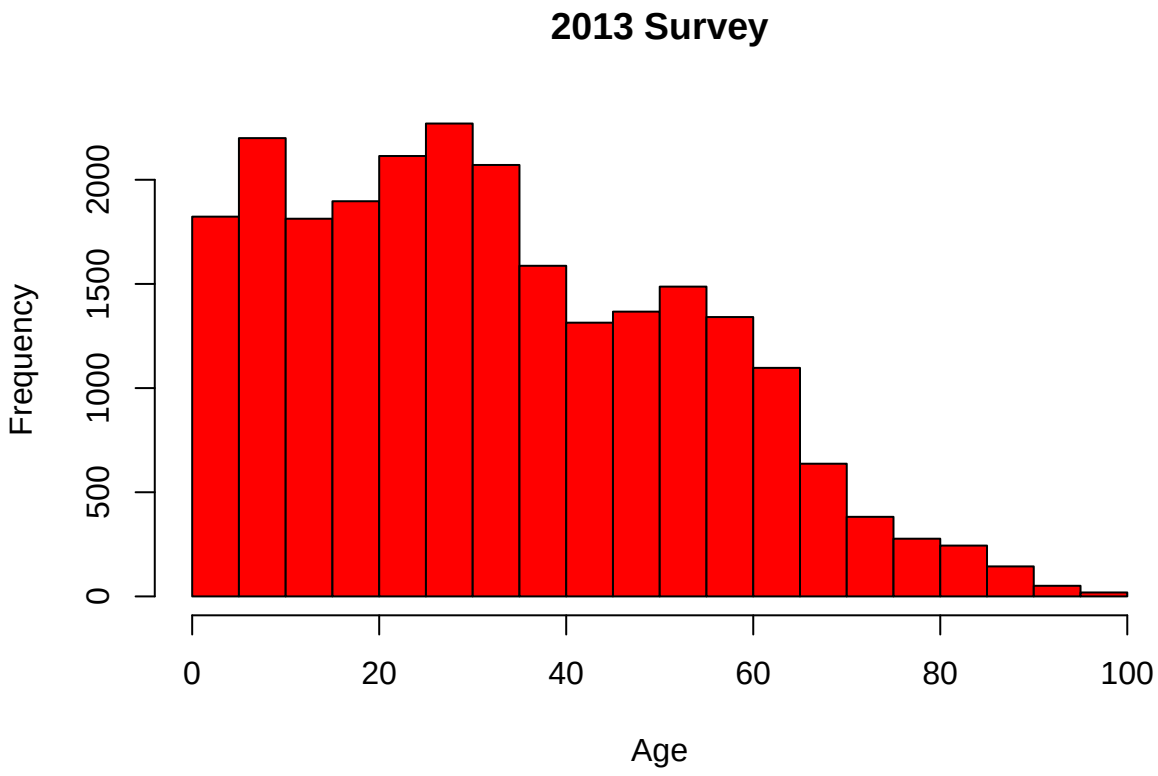



```
# As we can see from the histograms above, the bucket "1" represents a good portion of
# surveys in both years.  This "1" maps to an individual from the age of 0-2 years old.
# After  doing some research, it seems that this is  an overstatement of the US population
# in this age band.  This is concerning and may require more investigation.  If we remove
# the 0-2 year old's from the data, we get a better representation of the population, as
# seen below.

clean_11 <- as.data.frame(age_11[! age_11$ER34104 %in% c(0,999,1004,1), ])
clean_13 <- as.data.frame(age_13[! age_13$ER34204 %in% c(0,999,1), ])
hist(clean_11[,1],col="red",freq=T,main ='2011 Survey', xlab = "Age")
```




```
hist(clean_13[,1],col="red",freq=T,main ='2013 Survey', xlab = "Age")
```




```
# It seems that after some cleansing the two years of interviews are be mostly
# representative of our population.
```

b) Confirmation Bias is something that can be an issue even with anonymized data.

i) Checking the income distribution of bands to make sure that the data is representative of society, and not just a certain group of people that may have been chosen to prove a point. This would be a similar exercise as the one above. It is relatively quick, but very important in order to generalize findings of a study to the population.

ii) Checking the correlation of data, by running the data set through correlation tests to confirm that there is a good distribution of people involved in the survey. If non-numeric information is removed, a PCA model could reveal which variables might explain trends. While both models can benefit the user by gaining insight into which variables may be worth investigating further PCA may not be ideal for clarity when presenting the study to a non-technical audience.

iii) Checking the percentage of inputs that are imputed manually to make sure they are not overwhelming. If so one would have cause for further investigation. This would require compiling a list of which variables have the option of being manually imputed. It would be important to find out how the manually imputed values were derived and if it was the correct method.

```
### Test option iii)
# ER34144 represents the accuracy of the social security income, i.e. if it
# was imputed manually, and how. ER34144B represents the same information for total labor
# income. There seems to be four other variables (ER34144A,ER34144C,ER34144D,ER34144E)
# that may, or may not be manually altered. Unlike the first two, there is no explanation
# variable to say exactly how. I will try to work these four variables into my analysis.

total <- nrow(data)
accSocialInc <- data[47]
totLaborInc<- data[48]
accLaborInc<- data[49]
totAssetInc<- data[50]
totTaxIn<- data[51]
totTranferIn<- data[52]

# Remove the values that weren't imputed manually.

manualPerSocial<- nrow(as.data.frame(accSocialInc[! accSocialInc$ER34144 %in% 0, ]))/total
manualPerLabor<- nrow(as.data.frame(accLaborInc[! accLaborInc$ER34144B %in% 0, ]))/total

# As we can see the percentage of manually altered data in these two variables are pretty
#low around .07% for Social Income and 1.5% for Labor Income.

manualPerSocial
```

```
## [1] 0.007079711
```

```
manualPerLabor
```

```
## [1] 0.01581013
```

```
# It seems that these variables have a high amount of "0"s or unknown information. Let's
# find the exact percentage. This could be an issue if we try to draw conclusions from
# this information.

unkTotLaborInc<- nrow(as.data.frame(totLaborInc[totLaborInc$ER34144A %in% 0,]))/total
unkTotAssetInc<- nrow(as.data.frame(totAssetInc[ totAssetInc$ER34144C %in% 0, ]))/total
unkTaxInc<- nrow(as.data.frame(totTaxIn[ totTaxIn$ER34144D %in% 0,]))/total
unkTotTranInc <-nrow(as.data.frame(totTranferIn[totTranferIn$ER34144E %in% 0,]))/total

# These numbers suggest that conclusions including total Labor Income, total Asset Income,
```

```
# total Taxable Income,or, total Transfer Income should be scrutinized, since about 95% of
# values were represented as "0".

unkTotLaborInc
```

```
## [1] 0.9326877
```

```
unkTotAssetInc
```

```
## [1] 0.9980558
```

```
unkTaxInc
```

```
## [1] 0.9319908
```

```
unkTotTranInc
```

```
## [1] 0.9815121
```

```
# For example, it could be misleading to project conclusions found from the numbers below
# because 93.27% of the observations were labeled "0".

LabIncome <- as.data.frame(totLaborInc[! totLaborInc$ER34144A %in% 0, ])
summary(LabIncome)
```

```
##  totLaborInc[!totLaborInc$ER34144A %in% 0, ]
##  Min.   : -8500
##  1st Qu.:  3628
##  Median :  8516
##  Mean   : 15045
##  3rd Qu.: 18014
##  Max.   :600000
```

```
# A recommendation I would make would be to have 0 represent an income of 0, and find
# another number to represent missing information. This way, I would feel better about the
# strength of conclusions made.
```

c) Transforming the data could provide issues of matching surveys from various years to individuals.

i) A way to check this could be to match the birthdate information given about the individual from each interview. This would be easy to do and easy to understand why a row would be deleted if the information doesn't match. I will elaborate in the code below.

ii) When someone turns over 50, they may have a different individual to attend the second interview. This could provide another instance for data to be corrupted, especially when transformed. A way to check this could be to check the age and add two years. If it is inconsistent, that data file may be flagged as a candidate for deletion. This process wouldn't be too complicated, and it would allow a user to filter out questionable data points.

iii) Check to see if the sequence number difference between two interviews are outside a "buffer" range to see if these data points need to be flagged for further review. For example if 7 people have left the household in 2 years, it may be worth putting together the whole story of this data point to see if it is reasonable or not. It would be pretty easy to do the initial flag, but the extra scrutiny may prove to be impossible with anonymized data.

```
### Test option i)
# ER34104, ER34105, and ER34106 map to age, month, and year of birth from 2011.
# Similarly, ER34204, ER34205, and ER34206 map to the same information from the 2013
# interview. I will also pull in ER34103 and ER34203, which is the relationship to head at
# time of 2011 and 2013 interview.
```

```r
ageCheck <- data[c(6:9,66:69)]

# The first step is to remove any row of data where either relationship to head is 0
# because we won't be able to compare birthdates with empty values.

ageCheck <- ageCheck[ageCheck$ER34103!=0 & ageCheck$ER34203!=0, ]

# Now I will add a conditional flag column to the end indicating if the relationship to
# head is the same in each year.

ageCheck$RelHeadFlag <- 1
for (i in 1:nrow(ageCheck)){
  if (ageCheck$ER34103[i] == ageCheck$ER34203[i] ){
    ageCheck$RelHeadFlag[i] <- 0
  }
}
sum(ageCheck$RelHeadFlag)/nrow(ageCheck)
```

```
## [1] 0.08813529
```

```r
# 8.8% of the data remaining wither theoretically had a different person present for
# the second interview, or the relationship to head changed. Some data points tagged 1
# (indicating relationship change or new individual) have similar birthdates. I'm assuming
# some individual's relationship to head were either incorrectly labeled, or the
# relationship changed. Since the data is anonymized, there is no way for us to tell if
# the therelationship was entered incorrectly, or the relationship to head actually
# changed. There isn't a hard and fast rule as far as age goes, because depending on what
# date the interviews were conducted we could see differences of 1, 2, or 3 years of age.

# It seems that year of birth is shifted 5 years, which might be explained by the
# difference between the 2009 and 2013 survey. I will flag any entries that are labeled as
# the same person, but have a difference other than 5  years in birthdate, indicating that
# some piece of the information may have been entered incorrectly.

ageCheck$BirthYearFlag <- 0
for (i in 1:nrow(ageCheck)){
  # if the people at each interview are labeled with same relationship to head
  if (ageCheck$RelHeadFlag[i] == 0){
    # but they have different birth years, label them 1
    if(abs(ageCheck$ER34106[i] - ageCheck$ER34206[i]) !=5){
    ageCheck$BirthYearFlag[i] <- 1
    }
  }
}

sum(ageCheck$BirthYearFlag)
```

```
## [1] 39
```

```r
# It seems that 39 rows of data have the birthdate entered incorrectly. From an initial
# look, most are due to  the birth year being labeled "9999" or NA;DK. These data points
# may be candidates for deletion to strengthen findings.

# Overall, I would say that the data has been transformed/matched up pretty well. There
```

```
# are some rows that may require further investigation, or simply just deletion.
```

Five non-feasible solutions include

As far as non-computer algorithm feasibility is concerned :

a) Staffing from data source could be concerning. If employees are consistently careless and enter data incorrectly it could change the findings completely. For example entering a 3 vs 4 in ER34116 (employment status) would correspond to 'Looking for work' and 'Retired', which imply very different situations.

b) Response Bias could be a huge issue. It is hard for the survey implementors themselves to combat this issue. It would be exponentially harder for a secondary source to find, prove, and resolve the issue, especially if the data is anonymized.

Computer non-feasibility

c) Merging large amounts of data could prove to be an issue. It would take a really long time to simply create the master dataset if one was to look at all available data since 1968.

d) If someone takes on a project to show how the country's economic dynamic has changed, transforming the data into a consistent database will be next to impossible without having incomplete information. The form of the survey completed in 1968 asks very different questions than the most recent survey. Another big issue arises when you think about dealing with inflation. A question posed in 1968 proves both points; "Did your family spend more than $50 on your clothing bill?". Grabbing each data point's survey year and then linking that with a calculation for inflation would be no easy task.

e) With giant datasets, Machine Learning models will be hard to test and improve. For example if you are building boosted tree models and want to test tree depth, iterations, or bag fractions it could take weeks to find the best error rates.