Week8

James Rundle, Brad Odac, Avi

10/8/2017

Upload Data

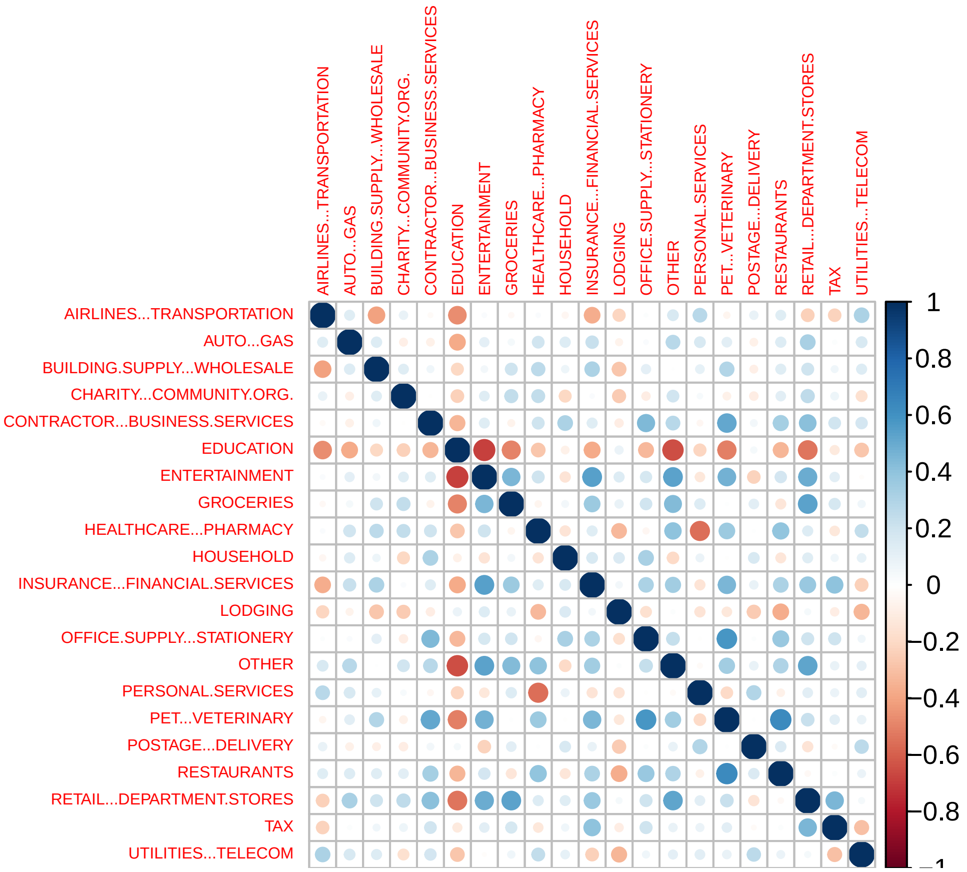
I used Des 2 Catagory and am looking for information on Groceries

```
library(reshape)
reshaped_df = cast(Fake_Data_and_Metadata_Final_no_pass, masked_id ~ Des2,
                   value = 'Payment', fun.aggregate=sum)
# when we do this it is a good idea to fix the names of the columns
tidy.name.vector <- make.names(colnames(reshaped_df), unique=TRUE)
colnames(reshaped_df) = tidy.name.vector
```

Looking for information on how to predict charges of "GROCERIES" in Des2 Catagory

We see that for GROCERIES, Retail, Insurance, other and entertainment had the highest positive correlation

```
library(corrplot)
M = cor(reshaped_df_norm)
corrplot(M,tl.cex = .5)
```



Now we can do some predictions

I found the biggest contributors to a high %varexplained was EDUCATION and INSURANCE. If we look again we can see that education was actually negatively correlated, but i guess that is also useful for the prediction

```
library(randomForest)
```

```
## randomForest 4.6-10
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
fit <- randomForest(GROCERIES ~ ENTERTAINMENT+INSURANCE...FINANCIAL.SERVICES
                    +GROCERIES+RETAIL...DEPARTMENT.STORES+OTHER+EDUCATION ,
                    data=reshaped_df_norm,
                    importance=TRUE,
                    ntree=2500)
print(fit)
```

```
##
## Call:
##  randomForest(formula = GROCERIES ~ ENTERTAINMENT + INSURANCE...FINANCIAL.SERVICES +      GROCERIES
##                Type of random forest: regression
##                      Number of trees: 2500
## No. of variables tried at each split: 2
##
##           Mean of squared residuals: 1.708094e-05
##                        % Var explained: 78.81
```

78 % is pretty darn good.

Competetion is wrapping up this week and while im not anywhere close to where I think we could be I'm hoping to package this as "We can see that if a person is making increasing purchases in these catagories they might be spending more on groceries soon"

Id really like to be able to tie this in to our idea of gamifying the rewards system, maybe including some of the descriptive statistics for individual masked ids we can provide some info that Wells Fargo will find some value in.