# APAN5420 — HW 3

*Megan Wilder*

*6/11/18*

## Contents



---

## 1 Load Data

```
#load packages
library(dplyr)
library(DataExplorer)
library(ggplot2)
library(plotly)
library(xts)
library(zoo)

#load data
ccard <- read.csv("res_purchase_card.csv")
```

## 2 Explore the DataFrame

```
#explore data
dim(ccard)
```

```
## [1] 442458     11
```

```
summary(ccard)
```

```
##    Year.Month     Agency.Number
##  Min.   :201307   Min.   : 1000
##  1st Qu.:201309   1st Qu.: 1000
##  Median :201401   Median :47700
##  Mean   :201357   Mean   :42786
##  3rd Qu.:201404   3rd Qu.:76000
##  Max.   :201406   Max.   :98000
##
##                                  Agency.Name
##  OKLAHOMA STATE UNIVERSITY             :115995
##  UNIVERSITY OF OKLAHOMA                : 76143
##  UNIV. OF OKLA. HEALTH SCIENCES CENTER: 58247
```

```
##  DEPARTMENT OF CORRECTIONS            : 22322
##  DEPARTMENT OF TOURISM AND RECREATION : 17232
##  DEPARTMENT OF TRANSPORTATION         : 15689
##  (Other)                              :136830
##                Cardholder.Last.Name Cardholder.First.Initial
##  JOURNEY HOUSE TRAVEL INC: 10137    J      : 55031
##  UNIVERSITY AMERICAN     :  7219    G      : 42251
##  JOURNEY HOUSE TRAVEL    :  4693    D      : 38120
##  Heusel                  :  4212    M      : 35352
##  Hines                   :  3423    S      : 34698
##  Bowers                  :  2448    C      : 33213
##  (Other)                 :410326    (Other):203793
##                         Description         Amount
##  GENERAL PURCHASE             :247187   Min.   : -42863.0
##  AIR TRAVEL                   : 29584   1st Qu.:     30.9
##  ROOM CHARGES                 : 18120   Median :    104.9
##  AT&T SERVICE PAYMENT ITM     :  2657   Mean   :    425.0
##  001 Priority          1LB PCE:  2005   3rd Qu.:    345.0
##  00000000000000000000000      :  1828   Max.   :1903858.4
##  (Other)                      :141077
##                        Vendor                     Transaction.Date
##  STAPLES                  : 14842   09/11/2013 12:00:00 AM:  2122
##  AMAZON MKTPLACE PMTS     : 12197   08/07/2013 12:00:00 AM:  2108
##  WW GRAINGER              : 12076   01/14/2014 12:00:00 AM:  2059
##  Amazon.com               : 10766   01/16/2014 12:00:00 AM:  2009
##  BILL WARREN OFFICE PRODUC:  4479   09/05/2013 12:00:00 AM:  1999
##  LOWES #00241             :  4231   10/01/2013 12:00:00 AM:  1996
##  (Other)                  :383867   (Other)               :430165
##                  Posted.Date
##  01/13/2014 12:00:00 AM:  3256
##  04/14/2014 12:00:00 AM:  3163
##  03/10/2014 12:00:00 AM:  3139
##  03/03/2014 12:00:00 AM:  3101
##  09/16/2013 12:00:00 AM:  3062
##  01/20/2014 12:00:00 AM:  3032
##  (Other)                :423705
##                                          Merchant.Category.Code..MCC.
##  STATIONERY, OFFICE SUPPLIES, PRINTING AND WRITING PAPER: 24860
##  BOOK STORES                                            : 21981
##  INDUSTRIAL SUPPLIES NOT ELSEWHERE CLASSIFIED           : 21669
##  DENTAL/LABORATORY/MEDICAL/OPHTHALMIC HOSP EQIP AND SUP.: 20183
##  GROCERY STORES,AND SUPERMARKETS                        : 17152
##  MISCELLANEOUS AND SPECIALTY RETAIL STORES              : 13335
##  (Other)                                                :323278
```

```
colnames(ccard)
```

```
##  [1] "Year.Month"                  "Agency.Number"
##  [3] "Agency.Name"                 "Cardholder.Last.Name"
##  [5] "Cardholder.First.Initial"    "Description"
##  [7] "Amount"                      "Vendor"
##  [9] "Transaction.Date"            "Posted.Date"
## [11] "Merchant.Category.Code..MCC."
```

```r
#change col names
colnames(ccard) <-
c(
'Year_Month',
'Agency_Number',
'Agency_Name',
'Cardholder_Last_Name',
'Cardholder_First_Initial',
'Description',
'Amount',
'Vendor',
'Transaction_Date',
'Posted_Date',
'Merchant_Category'
)

colnames(ccard)
```

```
##  [1] "Year_Month"               "Agency_Number"
##  [3] "Agency_Name"              "Cardholder_Last_Name"
##  [5] "Cardholder_First_Initial" "Description"
##  [7] "Amount"                   "Vendor"
##  [9] "Transaction_Date"         "Posted_Date"
## [11] "Merchant_Category"
```

```r
#number of rows
nrow(ccard)
```

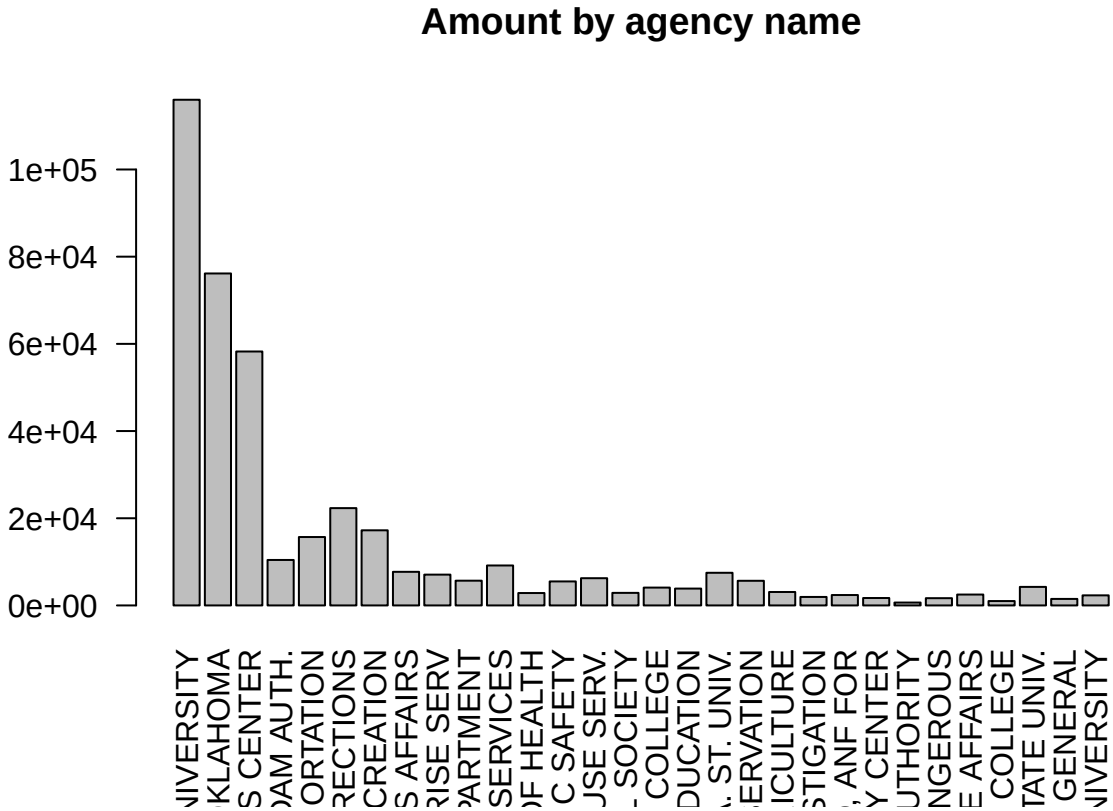
```
## [1] 442458
```

```r
# Count of agencies
# Spent by agency
# Count by merchant.Category.Code
# Simple Bar Plot

#Create new DF grouped by Agency Name with summary statistics, arrange in descending order by amount
stat_by_agency <- ccard %>% group_by(Agency_Name) %>%
summarise(
count = n(),
amount = sum(Amount),
mean = mean(Amount),
min = min(Amount),
max = max(Amount)
) %>%
arrange(desc(amount)) %>% ungroup()

#add number to beginning of Agency name enabling ranking based on amount, add percent column
stat_by_agency <- stat_by_agency %>%
mutate(
row = rep(1:nrow(stat_by_agency)),
Agency_Name_ind = paste(row, Agency_Name, sep = "_"),
percent = amount / sum(amount)
) %>%
select(Agency_Name_ind, count, amount, percent, mean, min, max)
```

```r
head(stat_by_agency)
```

```
## # A tibble: 6 x 7
##                            Agency_Name_ind  count   amount    percent
##                                      <chr>  <int>    <dbl>      <dbl>
## 1             1_OKLAHOMA STATE UNIVERSITY 115995 33778840 0.17963575
## 2                2_UNIVERSITY OF OKLAHOMA  76143 24886383 0.13234570
## 3 3_UNIV. OF OKLA. HEALTH SCIENCES CENTER  58247 24527325 0.13043623
## 4                 4_GRAND RIVER DAM AUTH.  10427 22213829 0.11813306
## 5          5_DEPARTMENT OF TRANSPORTATION  15689 14399262 0.07657522
## 6             6_DEPARTMENT OF CORRECTIONS  22322 13988872 0.07439277
## # ... with 3 more variables: mean <dbl>, min <dbl>, max <dbl>
```

```r
#create df with top 30 agencies ranked by amount
df_30 <- stat_by_agency[1:30, ]
#plot
barplot(
df_30$count,
names.arg = df_30$Agency_Name_ind,
main = "Amount by agency name",
las = 2
)
```



# 3 Feature Creation

## 3.1 Recency

Lead: I'm going to calculate the time since last transaction across all transactions for each Agency and time since last transaction for each Agency at a particular merchant category.

```
#create new DF group by agency, with Recency column (time since last transaction)
time_by_agency <- ccard %>% group_by(Agency_Name) %>%
mutate(Transaction_Date = as.Date(Transaction_Date, format = "%m/%d/%Y %H:%M")) %>%
arrange(Agency_Name, Transaction_Date) %>%
mutate(Recency = Transaction_Date - lag(Transaction_Date))

time_by_agency[, c("Agency_Number", "Agency_Name", "Transaction_Date", "Recency")]
```

```
## # A tibble: 442,458 x 4
## # Groups:   Agency_Name [124]
##    Agency_Number              Agency_Name Transaction_Date Recency
##            <int>                    <fctr>           <date>  <time>
##  1         26500 `DEPARTMENT OF EDUCATION       2013-06-29 NA days
##  2         26500 `DEPARTMENT OF EDUCATION       2013-07-01  2 days
##  3         26500 `DEPARTMENT OF EDUCATION       2013-07-01  0 days
##  4         26500 `DEPARTMENT OF EDUCATION       2013-07-03  2 days
##  5         26500 `DEPARTMENT OF EDUCATION       2013-07-03  0 days
##  6         26500 `DEPARTMENT OF EDUCATION       2013-07-03  0 days
##  7         26500 `DEPARTMENT OF EDUCATION       2013-07-03  0 days
##  8         26500 `DEPARTMENT OF EDUCATION       2013-07-03  0 days
##  9         26500 `DEPARTMENT OF EDUCATION       2013-07-03  0 days
## 10         26500 `DEPARTMENT OF EDUCATION       2013-07-03  0 days
## # ... with 442,448 more rows
```

```
#filter to make sure first recency for each agency is NA
time_by_agency %>% filter(Agency_Number == 4000) %>% group_by(Vendor, Merchant_Category)
```

```
## # A tibble: 5,470 x 12
## # Groups:   Vendor, Merchant_Category [1,330]
##    Year_Month Agency_Number                Agency_Name Cardholder_Last_Name
##         <int>         <int>                      <fctr>               <fctr>
##  1     201307          4000 DEPARTMENT OF AGRICULTURE                 Irby
##  2     201307          4000 DEPARTMENT OF AGRICULTURE            Marquardt
##  3     201307          4000 DEPARTMENT OF AGRICULTURE                James
##  4     201307          4000 DEPARTMENT OF AGRICULTURE                James
##  5     201307          4000 DEPARTMENT OF AGRICULTURE              Belcher
##  6     201307          4000 DEPARTMENT OF AGRICULTURE               Bourns
##  7     201307          4000 DEPARTMENT OF AGRICULTURE                 Carr
##  8     201307          4000 DEPARTMENT OF AGRICULTURE                Davis
##  9     201307          4000 DEPARTMENT OF AGRICULTURE                Lerch
## 10     201307          4000 DEPARTMENT OF AGRICULTURE               Lester
## # ... with 5,460 more rows, and 8 more variables:
## #   Cardholder_First_Initial <fctr>, Description <fctr>, Amount <dbl>,
## #   Vendor <fctr>, Transaction_Date <date>, Posted_Date <fctr>,
## #   Merchant_Category <fctr>, Recency <time>
```

```
#create new DF grouped by agencies and by Merchant_Category,
#with Recency column (time since last transaction)
```

```
time_by_Merchant_Category <-
ccard %>% group_by(Agency_Name, Merchant_Category) %>%
mutate(Transaction_Date = as.Date(Transaction_Date, format = "%m/%d/%Y %H:%M")) %>%
arrange(Agency_Name, Merchant_Category, Transaction_Date) %>%
mutate(Recency = Transaction_Date - lag(Transaction_Date))

head(time_by_Merchant_Category[, c("Agency_Number",
"Agency_Name",
"Merchant_Category",
"Transaction_Date",
"Recency")])
```

```
## # A tibble: 6 x 5
## # Groups:   Agency_Name, Merchant_Category [2]
##   Agency_Number              Agency_Name    Merchant_Category
##           <int>                    <fctr>               <fctr>
## 1         26500 `DEPARTMENT OF EDUCATION ADVERTISING SERVICES
## 2         26500 `DEPARTMENT OF EDUCATION ADVERTISING SERVICES
## 3         26500 `DEPARTMENT OF EDUCATION ADVERTISING SERVICES
## 4         26500 `DEPARTMENT OF EDUCATION ADVERTISING SERVICES
## 5         26500 `DEPARTMENT OF EDUCATION ADVERTISING SERVICES
## 6         26500 `DEPARTMENT OF EDUCATION    AMERICAN AIRLINES
## # ... with 2 more variables: Transaction_Date <date>, Recency <time>
```

```
#sort by recency
Recency_cat_sorted <-
time_by_Merchant_Category %>% arrange(Merchant_Category, Recency) %>% na.omit

#filter OKLA. PANHANDLE STATE UNIV.
Recency_cat_OKLA <-
Recency_cat_sorted  %>% filter(Agency_Name == "OKLA. PANHANDLE STATE UNIV.")
Recency_cat_OKLA <-
Recency_cat_OKLA %>% arrange(Recency) %>% na.omit
OKLA_head <- head(Recency_cat_OKLA)
OKLA_tail <- tail(Recency_cat_OKLA)
#ACCOUNTING,AUDITING AND BOOKKEEPING SERVICES had the greatest recency
#AMUSEMENT PRKS,CIRCUSES,CARNIVLS,AND FORTUNE TELLERS
#had the greatest amount of time between transactions
```

Analysis: Recency represents the time since the previous transaction. For example at OKLA Panhandle State University, the time between accounting and bookkeeping services charges was small, indicating that these are typical transactions for the university, which seems logical. In contrast, the time between charges at amusement parks was significant, 351 days, as this is not a typical charge.

Conclusion: Going forward, this variable can be used to see if future credit transactions fit the normal customer profile.

## 3.2 Monetary

Lead: I'm going to aggregate data into the past 3, 7 and 30 transactions grouped by Agency. I'm then going to calculate the average, sum and max amount for these aggregated transactions.

```
#Aggregrate data into past 3, 7 and 30 transactions by Agency
#create sum function
rollag <- function(x, i) {
```

```
lagsum = 0
for (u in 1:i) {
lagsum = lagsum + lag(x, u)
}
lagsum
}

#create avg function
rollave <- function(x, i) {
lagsum = 0
for (u in 1:i) {
lagsum = lagsum + lag(x, u)
lagave = lagsum / i
}
lagave
}


#create new DF group by agency, with lagged sum amount,
#average amount and max amount for past 3 transactions,
#7 transactions and 30 transacitons
time_by_agency_lag <- time_by_agency %>% group_by(Agency_Name) %>%
arrange(Agency_Name, Transaction_Date) %>%
mutate(
Last3sum = rollag(Amount, 3),
Last7sum = rollag(Amount, 7),
Last30sum = rollag(Amount, 30)
) %>%
mutate(
Last3ave = rollave(Amount, 3),
Last7ave = rollave(Amount, 7),
Last30ave = rollave(Amount, 30)
)  %>%
mutate(
Last3max = rollapplyr(Amount, 3, max, partial = TRUE),
Last7max = rollapplyr(Amount, 7, max, partial = TRUE),
Last30max = rollapplyr(Amount, 30, max, partial = TRUE)
)

#filter to make sure first lag for each agency is NA or first transaction for max
time_by_agency_lag %>% filter(Agency_Number == 4000) %>% group_by(Vendor, Merchant_Category)
```

```
## # A tibble: 5,470 x 21
## # Groups:   Vendor, Merchant_Category [1,330]
##    Year_Month Agency_Number               Agency_Name Cardholder_Last_Name
##         <int>         <int>                     <fctr>               <fctr>
##  1     201307          4000 DEPARTMENT OF AGRICULTURE                 Irby
##  2     201307          4000 DEPARTMENT OF AGRICULTURE            Marquardt
##  3     201307          4000 DEPARTMENT OF AGRICULTURE                James
##  4     201307          4000 DEPARTMENT OF AGRICULTURE                James
##  5     201307          4000 DEPARTMENT OF AGRICULTURE              Belcher
##  6     201307          4000 DEPARTMENT OF AGRICULTURE               Bourns
##  7     201307          4000 DEPARTMENT OF AGRICULTURE                 Carr
##  8     201307          4000 DEPARTMENT OF AGRICULTURE                Davis
```

```
##  9     201307          4000 DEPARTMENT OF AGRICULTURE                Lerch
## 10     201307          4000 DEPARTMENT OF AGRICULTURE               Lester
## # ... with 5,460 more rows, and 17 more variables:
## #   Cardholder_First_Initial <fctr>, Description <fctr>, Amount <dbl>,
## #   Vendor <fctr>, Transaction_Date <date>, Posted_Date <fctr>,
## #   Merchant_Category <fctr>, Recency <time>, Last3sum <dbl>,
## #   Last7sum <dbl>, Last30sum <dbl>, Last3ave <dbl>, Last7ave <dbl>,
## #   Last30ave <dbl>, Last3max <dbl>, Last7max <dbl>, Last30max <dbl>
```

```
#filter by 3 transaction average to find findings
time_sorted_3ave <-
time_by_agency_lag %>% arrange(desc(Last3ave)) %>% na.omit
#filter UNIV. OF OKLA. HEALTH SCIENCES CENTER
time_sorted_OKHS_avg <-
time_sorted_3ave  %>% filter(Agency_Name == "UNIV. OF OKLA. HEALTH SCIENCES CENTER")
OKHS_avg_head <-
head(time_sorted_OKHS_avg) #largest 3 transaction average was $634,751.0
# compare to avg and max of all transactions at UNIV. OF OKLA. HEALTH SCIENCES CENTER
OKHS <-
stat_by_agency  %>% filter(Agency_Name_ind == "3_UNIV. OF OKLA. HEALTH SCIENCES CENTER")
#average transaction size is $421.0916

#filter by 3 transaction max to find findings
time_sorted_3max <-
time_by_agency_lag %>% arrange(desc(Last3max)) %>% na.omit
#filter UNIV. OF OKLA. HEALTH SCIENCES CENTER
time_sorted_OKHS_max <-
time_sorted_3max  %>% filter(Agency_Name == "UNIV. OF OKLA. HEALTH SCIENCES CENTER")
OKHS_max_head <-
head(time_sorted_OKHS_max) #max out of rolling 3 transactions was $1,903,858
```

Analysis: Monetary value is the amount spent on a credit transaction. For example at OKLA Health Sciences Center, the largest 3 transaction average was $634,751. This is compared to the average transaction size of $421.0916 for the organization.

Conclusion: As V. Van Vlasselaer et al. found in their study "Decision Support Systems", the contrast between current and past purchasing patterns enable a model to correctly estimate fraud. Going forward, this variable can be used to see if future credit transactions fit the normal customer profile.

### 3.3 Frequency

Lead: I'm going to aggregate data into 1 day time periods and count the number of transactions. I subset the data for Oklahoma State University but this analysis can be applied to all Agencies in the data set.

```
#subset OKLAHOMA STATE UNIVERSITY
OSU_freq <-
time_by_agency  %>% filter(Agency_Name == "OKLAHOMA STATE UNIVERSITY")  %>% arrange(Transaction_Date)

#convert DF to XTS
OSU_xts <-
xts(OSU_freq,
as.POSIXct(OSU_freq$Transaction_Date, format = "%m/%d/%Y"))

# count the number of observations each day
tdd <- apply.daily(OSU_xts$Transaction_Date, length)
```

```r
#convert to DF
OSU_df <- as.data.frame(tdd)

#change col names
colnames(OSU_df) <-
c('Daily_Count')

summary(OSU_df)
```

```
##   Daily_Count
##  Min.   :  1.0
##  1st Qu.:112.0
##  Median :380.0
##  Mean   :307.7
##  3rd Qu.:455.0
##  Max.   :568.0
```

Analysis: Frequency is the number of transactions over a certain time period. For Oklahoma State University, the max number of transactions in one day is 568 and the average is 380.

Conclusion: Again this can be used to evaluate fraud by contrasting current and past purchasing behavior. Going forward, this variable can be used to see if future credit transactions fit the normal customer profile.