Two Sigma Data Science Challenge - Section 2

Code

Marcelo Rodrigues dos Santos

20 maio 2019

Hide

```
library(dplyr)
```

Hide

```
data_folder <- paste(getwd(), '/../datasets/', sep='')
```

Oklahoma State Spending

Preparing data

Opening dataset

Hide

```
dfpurorig <- read.csv(paste(data_folder,"res_purchase_2014.csv",sep=''), sep=',',dec='.')
```

Hide

```
dfpurorig$Transaction.Date<-as.POSIXct(strptime(as.character(dfpurorig$Transaction.Date), "%m/%d/%Y"))
dfpurorig$Posted.Date<-as.POSIXct(strptime(as.character(dfpurorig$Posted.Date), "%m/%d/%Y"))
dfpurorig$Description<-as.factor(toupper(dfpurorig$Description))
dfpurorig$Cardholder.Last.Name <-as.factor(toupper(dfpurorig$Cardholder.Last.Name))
dfpurorig$Agency.Name <-as.factor(toupper(dfpurorig$Agency.Name))
dfpurorig$Vendor <-as.factor(toupper(dfpurorig$Vendor))
dfpurorig$Cardholder.First.Initial <- as.factor(toupper(dfpurorig$Cardholder.First.Initial))
dfpurorig$Merchant.Category.Code..MCC. <- as.factor(toupper(dfpurorig$Merchant.Category.Code..MCC.))
dfpur<-dfpurorig
```

Cleaning dataset

Hide

```
a<-as.data.frame(as.numeric(as.character(dfpur$Amount)))
names(a)<-'value'
dfpur[is.na(a$value),'Amount']
```

```
[1] ($29.99)    $572.27     $12.90      452.91 zero
90449 Levels: -0.01 0.01 -0.02 0.02 0.03 -0.04 0.04 -0.05 0.05 -0.06 0.06 ... 999.99
```

The field Amount includes some wrong numeric values… let us correct it.

Hide

```
dfpur$Amount<-as.numeric(as.character(dfpur$Amount))
dfpur[is.na(a$value),'Amount']<-c(29.99,572.27,12.90,452.91)
```

Verifying all fields.

Hide

```
summary(dfpur)
```

```
  Year.Month     Agency.Number                                    Agency.Name
Min.   :  -999   Min.   : 1000   OKLAHOMA STATE UNIVERSITY            :115995
1st Qu.:201309   1st Qu.: 1000   UNIVERSITY OF OKLAHOMA               : 76143
Median :201401   Median :47700   UNIV. OF OKLA. HEALTH SCIENCES CENTER: 58247
Mean   :201090   Mean   :42786   DEPARTMENT OF CORRECTIONS            : 22322
3rd Qu.:201404   3rd Qu.:76000   DEPARTMENT OF TOURISM AND RECREATION : 17232
Max.   :201900   Max.   :98000   DEPARTMENT OF TRANSPORTATION         : 15689
                                 (Other)                              :136829
              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      : 35401
CARDHOLDER              :  3789    S      : 35081
HINES                   :  3423    C      : 33213
(Other)                 :408984    (Other):203360
                      Description         Amount
GENERAL PURCHASE             :247186   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
0                            :  1828   Max.   :1903858.4
(Other)                      :141077
                     Vendor       Transaction.Date
STAPLES                 : 14842   Min.   :2013-04-17 00:00:00
AMAZON MKTPLACE PMTS    : 12197   1st Qu.:2013-09-25 00:00:00
WW GRAINGER             : 12076   Median :2014-01-06 00:00:00
AMAZON.COM              : 10766   Mean   :2013-12-28 12:36:37
BILL WARREN OFFICE PRODUC:  4479   3rd Qu.:2014-04-02 00:00:00
LOWES #00241            :  4231   Max.   :2014-06-30 00:00:00
(Other)                 :383866
 Posted.Date
Min.   :2013-07-01 00:00:00
1st Qu.:2013-09-26 00:00:00
Median :2014-01-07 00:00:00
Mean   :2013-12-30 09:39:08
3rd Qu.:2014-04-03 00:00:00
Max.   :2014-06-30 00:00:00
```

```
                                        Merchant.Category.Code..MCC.
 STATIONERY, OFFICE SUPPLIES, PRINTING AND WRITING PAPER: 24860
 BOOK STORES                                            : 21981
 INDUSTRIAL SUPPLIES NOT ELSEWHERE CLASSIFIED           : 21668
 DENTAL/LABORATORY/MEDICAL/OPHTHALMIC HOSP EQIP AND SUP.: 20183
 GROCERY STORES,AND SUPERMARKETS                        : 17152
 MISCELLANEOUS AND SPECIALTY RETAIL STORES              : 13335
 (Other)                                                :323278
```

Year.month field seems to include wrong “-999” values.

Hide

```
table(dfpur$Year.Month)
```

```
  -999 201307 201308 201309 201310 201311 201312 201401 201402 201403 201404
   586  37635  39314  38762  40266  34275  26969  37230  35830  37720  39249
201405 201406 201900
 36022  37955    644
```

Also “201900” seems a mistake.

Checking if year.months values can be generated through Transaction.date or Posted.Date fields.

Checking consistence of Posted.Date and Transaction.Date. It means verifying if all Posted.Date is equal or after Transaction.Date…

Hide

```
count(dfpur[dfpur$Posted.Date<dfpur$Transaction.Date,])
```

n
<int>
0

1 row

Hide

```
a<-dfpur[dfpur$Year.Month!=format(dfpur$Posted.Date,'%Y%m'),]
nrow(a)
```

```
[1] 1230
```

Hide

```
a<-dfpur[dfpur$Year.Month!=format(dfpur$Transaction.Date,'%Y%m'),]
nrow(a)
```

```
[1] 24771
```

We can assume that Posted.Date is better than Transaction.Date to regenerate Year.month field. Also, it is possible to say that we have two main wrong values on Year.Month field: “-999” and “201900”.

Updating Year.Month based on Posted.Date field.

Hide

```
dfpur$Year.Month<-as.factor(format(dfpur$Posted.Date,'%Y%m'))
table(dfpur$Year.Month)
```

```
201307 201308 201309 201310 201311 201312 201401 201402 201403 201404 201405
 37635  39314  38762  40266  34275  26969  37230  35830  38188  39249  36784
201406
 37955
```

Checking how Amount values are distributed…

Hide

```
summary(dfpur$Amount)
```

```
    Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-42863.0     30.9     104.9     425.0     345.0 1903858.4
```

It seems that there are some outliers after the 75o percentile. Checking top 20 amounts.

Hide

```
(dfpur%>%select(Vendor, Transaction.Date, Amount)%>%filter(Amount>359)%>%top_n(20))%>%arrange(desc(Amount))
```

Vendor <fctr>	**Transaction.Date** <S3: POSIXct>	**Amount** <dbl>
PAYMENT ADJUSTMENT	2013-08-21	1903858.4
PAYMENT ADJUSTMENT	2013-07-19	1750380.0
PELCO STRUCTURAL LLC	2013-10-25	1089180.0
PELCO STRUCTURAL LLC	2014-02-28	855343.0
EMC CORPORATION	2013-10-03	814934.8
EMC CORPORATION	2013-10-15	403490.8
TK CONSTRUCTIO US LLC	2014-04-04	373150.3
NORTH AMERICAN SALT CO	2014-05-12	348053.8
MOTOROLA, INC. - ONLINE	2013-09-24	345176.0
PAYMENT ADJUSTMENT	2013-06-13	343148.5

1-10 of 20 rows Previous **1** 2 Next

There are some suspect values for a vendor named “PAYMENT ADJUSTMENT”. Exploring “PAYMENT ADJUSTMENT” vendor registers.

Hide

```
(a<-dfpur%>%select(Vendor, Transaction.Date, Amount)%>%filter(Vendor=='PAYMENT ADJUSTMENT'))
```

Vendor <fctr>	Transaction.Date <S3: POSIXct>	Amount <dbl>
PAYMENT ADJUSTMENT	2013-08-15	29728.88
PAYMENT ADJUSTMENT	2013-08-15	30018.34
PAYMENT ADJUSTMENT	2013-07-19	1750379.98
PAYMENT ADJUSTMENT	2013-08-21	1903858.37
PAYMENT ADJUSTMENT	2013-06-13	343148.50
PAYMENT ADJUSTMENT	2014-03-10	4626.46

6 rows

It is clear that these registers are not real purchases.

Hide

```
sum(a$Amount)
```

```
[1] 4061761
```

As you can see, $4,061,761 are related to this “PAYMENT ADJUSTMENT” vendor. Also, we realized some negative amounts. Checking for negative numbers…

Hide

```
(b<-dfpur%>%select(Vendor, Transaction.Date, Amount)%>%filter(Amount<=0))%>%arrange(Amount)
```

Vendor <fctr>	Transaction.Date <S3: POSIXct>	Amount <dbl>
SUNSHINE INDUSTRIES INC	2014-02-26	-42863.04
BIO RAD NORMN-40001124	2014-03-20	-41740.00
ORACL OPN	2013-07-06	-38506.87

Vendor <fctr>	Transaction.Date <S3: POSIXct>	Amount <dbl>
C P INTEGRATED SERVICES	2014-02-14	-34108.00
MINICK MATERIALS COMPA	2013-09-09	-33075.32
ROBERTS TRUCK CENTER	2013-09-27	-30076.45
HERTZ EQUIPMENT	2014-01-07	-27864.00
CONSTRUCTION DIVISION	2014-06-19	-21000.00
REDWOOD TOXICOLOGY	2013-06-27	-20000.00
IP NETWORKS	2013-08-01	-18899.00

1-10 of 14,531 rows Previous **1** 2 3 4 5 6 … 100 Next

Hide

```
sum(b$Amount)
```

```
[1] -3562604
```

The total of Negative numbers is -$3,562,604.

As we do not have any instruction related to “PAYMENT ADJUSTMENTS” and negative amounts, we will not use these amounts for answering questions.

Removing these registers…

Hide

```
dfpur<-subset(dfpur,Vendor!='PAYMENT ADJUSTMENT')
dfpur<-subset(dfpur,Amount>0)
```

Reducing the name of the field “Merchant.Category.Code..MCC.” to “Merchant.Category”

Hide

```
names(dfpur)[11]<-'Merchant.Category'
```

Checking for top 10 descriptions.

Hide

```
group_by(dfpur,Description)%>%summarize(c=n())%>%top_n(10)%>%arrange(desc(c))
```

Description <fctr>	c <int>
GENERAL PURCHASE	236155
AIR TRAVEL	28097
ROOM CHARGES	17472
AT&T SERVICE PAYMENT ITM	2657
001 PRIORITY 1LB PCE	2005
0	1828
PRODUCTS AND SERVICES EA	1264
SHIPPING CHARGES	1202
001 STANDARD 1LB PCE	738
JANITORIAL SUPPLIES NMB	605

1-10 of 10 rows

Checking for other strange descriptions.

Hide

```
head(dfpur%>%group_by(Description)%>%summarize(c=n()),20)
```

Description <fctr>	c <int>

Description <fctr>	c <int>
	333
0	1828
00000000000000000000000 |00000000000000000000000	1
00000000000000000000002 |PARTS SET	1
00000000000000000000000 164|MINI SCISSOR 8.5CM ST	1
00000000000000000000000 200897004000|SUMMARY EA	1
00000000000000000000000 300120001000|SUMMARY EA	1
00000000000000000000000 3107966|CC/1590064513862	1
00000000000000000000000 313784001000|SUMMARY EA	1
00000000000000000000000 315294001000|SUMMARY EA	1

1-10 of 20 rows Previous **1** 2 Next

There are several registers with strange descriptions (e.g.,"",0,0000000000, etc.). However, the amounts and other data are correct. We will keep these registers.

Question #1:

What is the total amount of spending captured in this dataset?

Hide

```
sum(dfpur$Amount)
```

```
[1] 187541509
```

Question #2:

How much was spent at WW GRAINGER?

Hide

```
a<-dfpur%>%select(Vendor,Posted.Date,Description,Amount)%>%filter(Vendor=='WW GRAINGER')%>%arrange(Posted.Date)
```

Hide

```
sum(a$Amount)
```

```
[1] 5225095
```

Question #3:

How much was spent at WM SUPERCENTER?

Hide

```
a<-dfpur%>%select(Vendor,Posted.Date,Description,Amount)%>%filter(Vendor=='WM SUPERCENTER')%>%arrange(Posted.Dat
e)
```

Hide

```
sum(a$Amount)
```

```
[1] 31777.83
```

Question #4:

What is the standard deviation of the total monthly spending in the dataset?

Hide

```
(a<-group_by(dfpur,Year.Month)%>%summarize(mean=mean(Amount),sd=sd(Amount),count=n()))
```

Year.Month <fctr>	mean <dbl>	sd <dbl>	count <int>
201307	432.2208	1758.829	36437
201308	432.6268	2073.386	38027
201309	422.6527	2600.290	37541
201310	459.7899	7415.509	38842
201311	399.2878	2539.912	33182
201312	459.0846	2792.663	26095
201401	416.9907	2864.179	36156
201402	411.0198	2529.528	34639
201403	487.1979	5298.356	36933
201404	448.0597	3306.698	37929

1-10 of 12 rows Previous **1** 2 Next

Hide

```
sd(a$mean)
```

```
[1] 25.22905
```

Question #5:

Describe the process you would follow to build a model on this dataset to make predictions about the stock market.

1. meet with user (Client/Product Owner) to understand business questions and expectations;
2. understand the business concepts behind this dataset; invest time for cleaning and preparing data; checking for outliers; review progress and clarify points on dataset and business concepts with user; research on analysis perspectives that could potentially be interesting for investidors from the specific stock market;

3. explore the dataset to capture business behavior and verify possible correlations among variables; maybe apply some clustering methods or decision tree for better understanding relationship among variables and understanding preliminary patterns; review progress and clarify points with user;
4. define statistics/Machine learning approaches, develop algorithms, apply proper cross-validation methods and metrics for evaluating generated models; if required, repeat activities from previous steps and this step until achieving best results; review progress and clarify points with user;
5. review final jupiter/R notebook to make sure it includes relevant steps; present final results to user and deliver notebook.

Question #6:

What biases might this dataset have if you tried to use it to model equities?

It is important to consider that this dataset includes information on the purchase/billing perspective. Of course, to be assertive for modeling equities, other perspectives (kind of information) are very important and must be avaiable (e.g., costs, cash flow, balance Sheet, income statement, etc.).