APAN5420 — HW 6

Megan Wilder

7/3/18

Contents

---

1 Load Data

```
library(dplyr)
library(DataExplorer)
library(ggplot2)
library(plotly)
library(kableExtra)
library(reshape2)

#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."
```

```
nrow(ccard)
```

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

```
#change column 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'
)

#view head of ccard
kable(head(ccard)) %>% kable_styling(latex_options = "scale_down")
```

Year_Month	Agency_Number	Agency_Name	Cardholder_Last_Name	Cardholder_First_Initial	Description	Amount	Vendor	Transaction_Date	Posted_Date	Merchant_Category
201307	1000	OKLAHOMA STATE UNIVERSITY	Mason	C	GENERAL PURCHASE	890.00	NACAS	07/30/2013 12:00:00 AM	07/31/2013 12:00:00 AM	CHARITABLE AND SOCIAL SERVICE ORGANIZATIONS
201307	1000	OKLAHOMA STATE UNIVERSITY	Mason	C	ROOM CHARGES	368.96	SHERATON HOTEL	07/30/2013 12:00:00 AM	07/31/2013 12:00:00 AM	SHERATON
201307	1000	OKLAHOMA STATE UNIVERSITY	Massey	J	GENERAL PURCHASE	165.82	SEARS.COM 9300	07/29/2013 12:00:00 AM	07/31/2013 12:00:00 AM	DIRCT MARKETING/DIRCT MARKETERS NOT ELSEWHERE CLASSIFIED
201307	1000	OKLAHOMA STATE UNIVERSITY	Massey	T	GENERAL PURCHASE	96.39	WAL-MART #0137	07/30/2013 12:00:00 AM	07/31/2013 12:00:00 AM	GROCERY STORES,AND SUPERMARKETS
201307	1000	OKLAHOMA STATE UNIVERSITY	Mauro-Herrera	M	HAMMERMILL COPY PLUS COPY EA	125.96	STAPLES DIRECT	07/30/2013 12:00:00 AM	07/31/2013 12:00:00 AM	STATIONERY, OFFICE SUPPLIES, PRINTING AND WRITING PAPER
201307	1000	OKLAHOMA STATE UNIVERSITY	Mauro-Herrera	M	GENERAL PURCHASE	394.28	KYOCERA DOCUMENT SOLUTION	07/29/2013 12:00:00 AM	07/31/2013 12:00:00 AM	OFFICE, PHOTOGRAPHIC, PHOTOCOPY, AND MICROFILM EQUIPMENT

```
#view count for each month
kable(table(ccard$Year_Month))
```

Var1	Freq
201307	37635
201308	39314
201309	38762
201310	40266
201311	34275
201312	26969
201401	37230
201402	35831
201403	38188
201404	39249
201405	36784
201406	37955

3 Feature Creation

3.1 Monetary Feature

```
#Add Monetary feature
#Add Max, Average and Median Amount Ratio Features by agency_name and merchant category
avg_agency <- ccard %>% group_by(Agency_Name, Merchant_Category) %>%
summarise(
mean_category_amount = mean(Amount),
```

```
median_category_amount = median(Amount),
mean_count_trans = n()
)
#view head of avg_agency
kable(head(avg_agency)) %>% kable_styling(latex_options = "scale_down")
```

Agency_Name	Merchant_Category	mean_category_amount	median_category_amount	mean_count_trans
'DEPARTMENT OF EDUCATION	ADVERTISING SERVICES	508.48600	415.85	5
'DEPARTMENT OF EDUCATION	AMERICAN AIRLINES	492.98809	620.60	89
'DEPARTMENT OF EDUCATION	BEST WESTERN	94.81143	83.00	7
'DEPARTMENT OF EDUCATION	BOOK STORES	131.82414	80.66	29
'DEPARTMENT OF EDUCATION	BOOKS, PERIODICALS AND NEWSPAPERS	275.00000	275.00	1
'DEPARTMENT OF EDUCATION	BUSINESS SERVICES NOT ELSEWHERE CLASSIFIED	235.88067	300.00	15

```
# Append the max, average and median statistics back to the data to derive the ratios.
# Select the most recent 4 transactions
per_agency_category <-
ccard %>% group_by(Agency_Name, Merchant_Category, Year_Month) %>%
summarise(
max_amount = max(Amount),
mean_amount = mean(Amount),
median_amount = median(Amount),
count_trans = n()
) %>%
left_join(avg_agency, by = c('Agency_Name', 'Merchant_Category')) %>%
mutate(
max_amount_ratio = max_amount / mean_category_amount,
mean_amount_ratio = mean_amount / mean_category_amount,
median_amount_ratio = median_amount / median_category_amount,
mean_count_ratio  = count_trans / mean_count_trans
) %>%
select(
-mean_category_amount,
-median_category_amount,
-mean_count_trans,-max_amount,
-mean_amount,
-median_amount,
-count_trans
) %>%
top_n(-4)  # Use top_n(xx) to select the top xx rows, and top_n(-xx) for the bottom xx rows

#summary
summary(per_agency_category)
```

```
##                                Agency_Name
##  OKLAHOMA STATE UNIVERSITY            : 1359
##  UNIVERSITY OF OKLAHOMA               : 1080
##  UNIV. OF OKLA. HEALTH SCIENCES CENTER:  749
##  DEPARTMENT OF TOURISM AND RECREATION :  739
##  DEPARTMENT OF CORRECTIONS            :  713
##  GRAND RIVER DAM AUTH.                :  682
##  (Other)                              :20764
##                                          Merchant_Category
##  STATIONERY, OFFICE SUPPLIES, PRINTING AND WRITING PAPER:  438
##  MISCELLANEOUS AND SPECIALTY RETAIL STORES              :  435
##  BUSINESS SERVICES NOT ELSEWHERE CLASSIFIED             :  401
```

```
##  GOVERNMENT SERVICES--NOT ELSEWHERE CLASSIFIED          :  381
##  CHARITABLE AND SOCIAL SERVICE ORGANIZATIONS            :  363
##  AMERICAN AIRLINES                                      :  350
##  (Other)                                                :23718
##    Year_Month     max_amount_ratio mean_amount_ratio median_amount_ratio
##  Min.   :201307   Min.   : -Inf    Min.   :  -Inf    Min.   :  -Inf
##  1st Qu.:201310   1st Qu.:0.590    1st Qu.:0.4802    1st Qu.:0.7564
##  Median :201401   Median :1.001    Median :0.9675    Median :1.0000
##  Mean   :201357   Mean   :  NaN    Mean   :   NaN    Mean   :   NaN
##  3rd Qu.:201404   3rd Qu.:1.899    3rd Qu.:1.2534    3rd Qu.:1.4870
##  Max.   :201406   Max.   :  Inf    Max.   :   Inf    Max.   :   Inf
##                                    NA's   :22        NA's   :22
##  mean_count_ratio
##  Min.   :0.002833
##  1st Qu.:0.062500
##  Median :0.111111
##  Mean   :0.248170
##  3rd Qu.:0.333333
##  Max.   :1.000000
##
```

```r
#some category summations equaled zero, resulted in INF ratio outputs, remove
per_agency_category <-
per_agency_category[is.finite(per_agency_category$max_amount_ratio),]

#view head of per_agency_category
kable(head(per_agency_category)) %>% kable_styling(latex_options = "scale_down")
```

Agency_Name	Merchant_Category	Year_Month	max_amount_ratio	mean_amount_ratio	median_amount_ratio	mean_count_ratio
'DEPARTMENT OF EDUCATION	ADVERTISING SERVICES	201307	1.4822040	0.8296984	1.0000000	0.6000000
'DEPARTMENT OF EDUCATION	ADVERTISING SERVICES	201308	0.0749676	0.0749676	0.0916677	0.2000000
'DEPARTMENT OF EDUCATION	ADVERTISING SERVICES	201401	2.4359373	2.4359373	2.9785740	0.2000000
'DEPARTMENT OF EDUCATION	AMERICAN AIRLINES	201307	1.7274251	1.4453020	1.1627457	0.1348315
'DEPARTMENT OF EDUCATION	AMERICAN AIRLINES	201308	1.4637270	1.0760503	1.1627457	0.0561798
'DEPARTMENT OF EDUCATION	AMERICAN AIRLINES	201312	1.0986067	-0.1045326	0.1456655	0.0674157

```r
#create new DF, add Max to year_month column
max_per_agency_category <- per_agency_category %>%
mutate(Year_Month = paste("Max", Year_Month, sep = "_")) %>%
select(-mean_amount_ratio,-mean_count_ratio,-median_amount_ratio)

#view head of max_per_agency_category
kable(head(max_per_agency_category)) %>% kable_styling(latex_options = "scale_down")
```

Agency_Name	Merchant_Category	Year_Month	max_amount_ratio
'DEPARTMENT OF EDUCATION	ADVERTISING SERVICES	Max_201307	1.4822040
'DEPARTMENT OF EDUCATION	ADVERTISING SERVICES	Max_201308	0.0749676
'DEPARTMENT OF EDUCATION	ADVERTISING SERVICES	Max_201401	2.4359373
'DEPARTMENT OF EDUCATION	AMERICAN AIRLINES	Max_201307	1.7274251
'DEPARTMENT OF EDUCATION	AMERICAN AIRLINES	Max_201308	1.4637270
'DEPARTMENT OF EDUCATION	AMERICAN AIRLINES	Max_201312	1.0986067

```r
#create new DF, add Med to year_month column
med_per_agency_category <- per_agency_category %>%
mutate(Year_Month = paste("Med", Year_Month, sep = "_")) %>%
select(-mean_amount_ratio,-mean_count_ratio,-max_amount_ratio)
```

```
#view head of med_per_agency_category
kable(head(med_per_agency_category)) %>% kable_styling(latex_options = "scale_down")
```

Agency_Name	Merchant_Category	Year_Month	median_amount_ratio
'DEPARTMENT OF EDUCATION	ADVERTISING SERVICES	Med_201307	1.0000000
'DEPARTMENT OF EDUCATION	ADVERTISING SERVICES	Med_201308	0.0916677
'DEPARTMENT OF EDUCATION	ADVERTISING SERVICES	Med_201401	2.9785740
'DEPARTMENT OF EDUCATION	AMERICAN AIRLINES	Med_201307	1.1627457
'DEPARTMENT OF EDUCATION	AMERICAN AIRLINES	Med_201308	1.1627457
'DEPARTMENT OF EDUCATION	AMERICAN AIRLINES	Med_201312	0.1456655

```
#create new DF, add Mean to year_month column
mean_per_agency_category <- per_agency_category %>%
mutate(Year_Month = paste("Mean", Year_Month, sep = "_")) %>%
select(-max_amount_ratio,-mean_count_ratio,-median_amount_ratio)

#view head of mean_per_agency_category
kable(head(mean_per_agency_category)) %>% kable_styling(latex_options = "scale_down")
```

Agency_Name	Merchant_Category	Year_Month	mean_amount_ratio
'DEPARTMENT OF EDUCATION	ADVERTISING SERVICES	Mean_201307	0.8296984
'DEPARTMENT OF EDUCATION	ADVERTISING SERVICES	Mean_201308	0.0749676
'DEPARTMENT OF EDUCATION	ADVERTISING SERVICES	Mean_201401	2.4359373
'DEPARTMENT OF EDUCATION	AMERICAN AIRLINES	Mean_201307	1.4453020
'DEPARTMENT OF EDUCATION	AMERICAN AIRLINES	Mean_201308	1.0760503
'DEPARTMENT OF EDUCATION	AMERICAN AIRLINES	Mean_201312	-0.1045326

```
# Max variable: Use "dcast" in Library "reshape2" to organize the data so each row
#is a merchant category of an agent.
max_wide <-
dcast(max_per_agency_category,
Agency_Name + Merchant_Category ~ Year_Month)
max_wide = as.matrix(max_wide)
max_wide[is.na(max_wide)] <- 0
max_wide = as.data.frame(max_wide)
#view head
kable(head(max_wide)) %>% kable_styling(latex_options = "scale_down")
```

Agency_Name	Merchant_Category	Max_201307	Max_201308	Max_201309	Max_201310	Max_201311	Max_201312	Max_201401	Max_201402	Max_201403	Max_201404	Max_201405	Max_201406
'DEPARTMENT OF EDUCATION	ADVERTISING SERVICES	1.482204e+00	7.496765e-02	0	0	0	0	2.435037e+00	0	0	0	0	0
'DEPARTMENT OF EDUCATION	AMERICAN AIRLINES	1.727425e+00	1.463727e+00	0	0	0	1.098607e+00	1.127816e+00	0	0	0	0	0
'DEPARTMENT OF EDUCATION	BEST WESTERN	0	1.884583e+00	7.383076e-01	0	0	0	8.754219e-01	0	0	0	0	0
'DEPARTMENT OF EDUCATION	BOOK STORES	2.911303e+00	5.317691e-01	3.179236e-01	4.301184964	0	6.221167e-01	0	0	0	0	0	0
'DEPARTMENT OF EDUCATION	BOOKS, PERIODICALS AND NEWSPAPERS	0	0	0	1.000000000	0	0	0	0	0	0	0	0
'DEPARTMENT OF EDUCATION	BUSINESS SERVICES NOT ELSEWHERE CLASSIFIED	4.027460e-02	1.271839e+00	0	0	1.801759e+00	1.839701e+00	2.204505e+00	0	0	0	0	0

```
# Median variable: Use "dcast" in Library "reshape2" to organize the data so each
#row is a merchant category of an agent.
med_wide <-
dcast(med_per_agency_category,
Agency_Name + Merchant_Category ~ Year_Month)
med_wide = as.matrix(med_wide)
med_wide[is.na(med_wide)] <- 0
med_wide = as.data.frame(med_wide)
#view head
kable(head(med_wide)) %>% kable_styling(latex_options = "scale_down")
```

Agency_Name	Merchant_Category	Med_201307	Med_201308	Med_201309	Med_201310	Med_201311	Med_201312	Med_201401	Med_201402	Med_201403	Med_201404	Med_201405	Med_201406
'DEPARTMENT OF EDUCATION	ADVERTISING SERVICES	1.000000e+00	9.166767e-02	0	0	0	0	2.978574e+00	0	0	0	0	0
'DEPARTMENT OF EDUCATION	AMERICAN AIRLINES	1.162746e+00	1.162746e+00	0	0	0	0.14566549	8.959072e-01	0	0	0	0	0
'DEPARTMENT OF EDUCATION	BEST WESTERN	0	2.152771e+00	8.433735e-01	0	0	0	1.000000e+00	0	0	0	0	0
'DEPARTMENT OF EDUCATION	BOOK STORES	8.430449e-01	8.690801e-01	5.195884e-01	1.640218200	0	0.72198116	0	0	0	0	0	0
'DEPARTMENT OF EDUCATION	BOOKS, PERIODICALS AND NEWSPAPERS	0	0	0	1.000000000	0	0	0	0	0	0	0	0
'DEPARTMENT OF EDUCATION	BUSINESS SERVICES NOT ELSEWHERE CLASSIFIED	3.166667e-02	1.881667e-01	0	0	1.416666667	1.38158333	5.000000e-01	0	0	0	0	0

```
# Mean variable: Use "dcast" in Library "reshape2" to organize the data so each
#row is a merchant category of an agent.
mean_wide <-
dcast(mean_per_agency_category,
Agency_Name + Merchant_Category ~ Year_Month)
mean_wide = as.matrix(mean_wide)
mean_wide[is.na(mean_wide)] <- 0
mean_wide = as.data.frame(mean_wide)
#view head
kable(head(mean_wide)) %>% kable_styling(latex_options = "scale_down")
```

Agency_Name	Merchant_Category	Mean_201307	Mean_201308	Mean_201309	Mean_201310	Mean_201311	Mean_201312	Mean_201401	Mean_201402	Mean_201403	Mean_201404	Mean_201405	Mean_201406
'DEPARTMENT OF EDUCATION	ADVERTISING SERVICES	8.296984e-01	7.496765e-02	0	0	0	0	2.435937e+00	0	0	0	0	0
'DEPARTMENT OF EDUCATION	AMERICAN AIRLINES	1.445302e+00	1.076050e+00	0	0	0	-0.1045326132	8.628471e-01	0	0	0	0	0
'DEPARTMENT OF EDUCATION	BEST WESTERN	0	1.884583e+00	7.383076e-01	0	0	0	8.754219e-01	0	0	0	0	0
'DEPARTMENT OF EDUCATION	BOOK STORES	9.944309e-01	5.317691e-01	3.179236e-01	1.403551754	0	0.4417627979	0	0	0	0	0	0
'DEPARTMENT OF EDUCATION	BOOKS, PERIODICALS AND NEWSPAPERS	0	0	0	1.000000000	0	0	0	0	0	0	0	0
'DEPARTMENT OF EDUCATION	BUSINESS SERVICES NOT ELSEWHERE CLASSIFIED	4.027460e-02	5.065414e-01	0	0	1.8017585163	1.7571384966	9.489262e-01	0	0	0	0	0

```
#merge dataframes
model_df_amt <-
left_join(max_wide, mean_wide, by = c('Agency_Name', 'Merchant_Category'))
model_df_amt <-
merge(model_df_amt,
med_wide,
by = c("Agency_Name", "Merchant_Category"))
```

3.2 Recency Feature

```
#Add Recency feature (time since last transaction) by agency_name and merchant category
#Add Max, Average and Median Recency Ratio Features by agency_name and merchant category
#create new DF grouped by agencies and by Merchant_Category,
#with Recency column
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))

#view head of time_by_Merchant_Category
kable(head(time_by_Merchant_Category[, c("Agency_Number",
"Agency_Name",
"Merchant_Category",
"Transaction_Date",
"Recency")])) %>% kable_styling(latex_options = "scale_down")
```

Agency_Number	Agency_Name	Merchant_Category	Transaction_Date	Recency
26500	'DEPARTMENT OF EDUCATION	ADVERTISING SERVICES	2013-06-29	NA
26500	'DEPARTMENT OF EDUCATION	ADVERTISING SERVICES	2013-07-01	2 days
26500	'DEPARTMENT OF EDUCATION	ADVERTISING SERVICES	2013-07-12	11 days
26500	'DEPARTMENT OF EDUCATION	ADVERTISING SERVICES	2013-08-01	20 days
26500	'DEPARTMENT OF EDUCATION	ADVERTISING SERVICES	2014-01-21	173 days
26500	'DEPARTMENT OF EDUCATION	AMERICAN AIRLINES	2013-07-03	NA

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

#Calculate the average and median recency by agency_name and merchant category
avg_recency <-
Recency_cat_sorted %>% group_by(Agency_Name, Merchant_Category) %>%
summarise(
mean_recency_amount = mean(Recency),
median_recency_amount = median(Recency),
mean_count_recency = n()
)
#view head of avg_recency
kable(head(avg_recency)) %>% kable_styling(latex_options = "scale_down")
```

Agency_Name	Merchant_Category	mean_recency_amount	median_recency_amount	mean_count_recency
'DEPARTMENT OF EDUCATION	ADVERTISING SERVICES	51.50000 days	15.5 days	4
'DEPARTMENT OF EDUCATION	AMERICAN AIRLINES	2.37500 days	0.5 days	88
'DEPARTMENT OF EDUCATION	BEST WESTERN	28.33333 days	0.0 days	6
'DEPARTMENT OF EDUCATION	BOOK STORES	6.50000 days	1.0 days	28
'DEPARTMENT OF EDUCATION	BUSINESS SERVICES NOT ELSEWHERE CLASSIFIED	14.64286 days	9.0 days	14
'DEPARTMENT OF EDUCATION	CATALOG MERCHANTS	44.50000 days	44.5 days	2

```
# Append the average and median recency statistics back to the data to derive the ratios.
# Select the most recent 4 transactions
per_agency_category_rec <-
Recency_cat_sorted %>% group_by(Agency_Name, Merchant_Category, Year_Month) %>%
summarise(
max_rec = max(Recency),
mean_rec = mean(Recency),
median_rec = median(Recency),
count_rec = n()
)

per_agency_category_rec <-
left_join(per_agency_category_rec,
avg_recency,
by = c('Agency_Name', 'Merchant_Category'))

#view class
lapply(per_agency_category_rec, class)
```

```
## $Agency_Name
## [1] "factor"
##
## $Merchant_Category
## [1] "factor"
##
## $Year_Month
## [1] "integer"
##
## $max_rec
## [1] "difftime"
##
## $mean_rec
## [1] "difftime"
```

```
##
## $median_rec
## [1] "difftime"
##
## $count_rec
## [1] "integer"
##
## $mean_recency_amount
## [1] "difftime"
##
## $median_recency_amount
## [1] "difftime"
##
## $mean_count_recency
## [1] "integer"
```

```
#change all difftime columns to class numeric
per_agency_category_rec$max_rec <-
as.numeric(per_agency_category_rec$max_rec, units = "days")

per_agency_category_rec$mean_rec <-
as.numeric(per_agency_category_rec$mean_rec, units = "days")

per_agency_category_rec$median_rec <-
as.numeric(per_agency_category_rec$median_rec, units = "days")

per_agency_category_rec$mean_recency_amount <-
as.numeric(per_agency_category_rec$mean_recency_amount, units = "days")

per_agency_category_rec$median_recency_amount <-
as.numeric(per_agency_category_rec$median_recency_amount, units = "days")

#add ratios to per_agency_category_rec
per_agency_category_rec <- per_agency_category_rec %>%
mutate(
max_rec_ratio = max_rec / mean_recency_amount,
mean_rec_ratio = mean_rec / mean_recency_amount,
median_rec_ratio = median_rec / median_recency_amount,
mean_rec_ratio  = count_rec / mean_count_recency
) %>%
select(
-mean_recency_amount,
-median_recency_amount,
-mean_count_recency,-max_rec,
-mean_rec,
-median_rec,
-count_rec
) %>%
top_n(-4)  # Use top_n(xx) to select the top xx rows, and top_n(-xx) for the bottom xx rows

#remove INF from median_rec_ratio
per_agency_category_rec <-
per_agency_category_rec[is.finite(per_agency_category_rec$median_rec_ratio),]
```

```
#create new DF, add MaxR to year_month column
max_per_agency_categor_rec <- per_agency_category_rec %>%
mutate(Year_Month = paste("MaxR", Year_Month, sep = "_")) %>%
select(-mean_rec_ratio,-median_rec_ratio)

#view head of max_per_agency_category_rec
kable(head(max_per_agency_categor_rec)) %>% kable_styling(latex_options = "scale_down")
```

Agency_Name	Merchant_Category	Year_Month	max_rec_ratio
'DEPARTMENT OF EDUCATION	ADVERTISING SERVICES	MaxR_201307	0.2135922
'DEPARTMENT OF EDUCATION	ADVERTISING SERVICES	MaxR_201308	0.3883495
'DEPARTMENT OF EDUCATION	ADVERTISING SERVICES	MaxR_201401	3.3592233
'DEPARTMENT OF EDUCATION	AMERICAN AIRLINES	MaxR_201307	2.9473684
'DEPARTMENT OF EDUCATION	AMERICAN AIRLINES	MaxR_201309	3.3684211
'DEPARTMENT OF EDUCATION	AMERICAN AIRLINES	MaxR_201310	4.6315789

```
#create new DF, add MedR to year_month column
med_per_agency_category_rec <- per_agency_category_rec %>%
mutate(Year_Month = paste("MedR", Year_Month, sep = "_")) %>%
select(-mean_rec_ratio,-max_rec_ratio)

#view head of med_per_agency_category_rec
kable(head(med_per_agency_category_rec)) %>% kable_styling(latex_options = "scale_down")
```

Agency_Name	Merchant_Category	Year_Month	median_rec_ratio
'DEPARTMENT OF EDUCATION	ADVERTISING SERVICES	MedR_201307	0.4193548
'DEPARTMENT OF EDUCATION	ADVERTISING SERVICES	MedR_201308	1.2903226
'DEPARTMENT OF EDUCATION	ADVERTISING SERVICES	MedR_201401	11.1612903
'DEPARTMENT OF EDUCATION	AMERICAN AIRLINES	MedR_201307	2.0000000
'DEPARTMENT OF EDUCATION	AMERICAN AIRLINES	MedR_201309	0.0000000
'DEPARTMENT OF EDUCATION	AMERICAN AIRLINES	MedR_201310	2.0000000

```
#create new DF, add MeanR to year_month column
mean_per_agency_category_rec <- per_agency_category_rec %>%
mutate(Year_Month = paste("MeanR", Year_Month, sep = "_")) %>%
select(-median_rec_ratio,-max_rec_ratio)

#view head of mean_per_agency_category_rec
kable(head(mean_per_agency_category_rec)) %>% kable_styling(latex_options = "scale_down")
```

Agency_Name	Merchant_Category	Year_Month	mean_rec_ratio
'DEPARTMENT OF EDUCATION	ADVERTISING SERVICES	MeanR_201307	0.5000000
'DEPARTMENT OF EDUCATION	ADVERTISING SERVICES	MeanR_201308	0.2500000
'DEPARTMENT OF EDUCATION	ADVERTISING SERVICES	MeanR_201401	0.2500000
'DEPARTMENT OF EDUCATION	AMERICAN AIRLINES	MeanR_201307	0.1250000
'DEPARTMENT OF EDUCATION	AMERICAN AIRLINES	MeanR_201309	0.2500000
'DEPARTMENT OF EDUCATION	AMERICAN AIRLINES	MeanR_201310	0.2727273

```
#Max Recency variable: Use "dcast" in Library "reshape2" to organize the data
#so each row is a merchant category of an agent.
max_wide_rec <-
```

```
dcast(max_per_agency_categor_rec,
Agency_Name + Merchant_Category ~ Year_Month)
max_wide_rec = as.matrix(max_wide_rec)
max_wide_rec[is.na(max_wide_rec)] <- 0
max_wide_rec = as.data.frame(max_wide_rec)
#view head
kable(head(max_wide_rec)) %>% kable_styling(latex_options = "scale_down")
```

Agency_Name	Merchant_Category	MaxR_201307	MaxR_201308	MaxR_201309	MaxR_201310	MaxR_201311	MaxR_201312	MaxR_201401	MaxR_201402	MaxR_201403	MaxR_201404	MaxR_201405	MaxR_201406
'DEPARTMENT OF EDUCATION	ADVERTISING SERVICES	0.21359223	0.38834951	0	0	0	0	3.35922330	0	0	0	0	0
'DEPARTMENT OF EDUCATION	AMERICAN AIRLINES	2.94736842	0	3.368421053	4.631578947	6.31578947	0	0	0	0	0	0	0
'DEPARTMENT OF EDUCATION	BOOK STORES	2.00000000	0.61538462	0	4.769230769	1.23076923	0	0	0	0	0	0	0
'DEPARTMENT OF EDUCATION	BUSINESS SERVICES NOT ELSEWHERE CLASSIFIED	0	2.04878049	1.365853659	0	0	2.66341463	0.95609756	0	0	0	0	0
'DEPARTMENT OF EDUCATION	CATALOG MERCHANTS	0	0	0	1.595505618	0	0	0	0	0	0	0	0
'DEPARTMENT OF EDUCATION	CHARITABLE AND SOCIAL SERVICE ORGANIZATIONS	0	1.95348837	1.395348837	2.372093023	0	5.44186047	0	0	0	0	0	0

```
# Median Recency variable: Use "dcast" in Library "reshape2" to organize the
#data so each row is a merchant category of an agent.
med_wide_rec <-
dcast(med_per_agency_category_rec,
Agency_Name + Merchant_Category ~ Year_Month)
med_wide_rec = as.matrix(med_wide_rec)
med_wide_rec[is.na(med_wide_rec)] <- 0
med_wide_rec = as.data.frame(med_wide_rec)
#view head
kable(head(med_wide_rec)) %>% kable_styling(latex_options = "scale_down")
```

Agency_Name	Merchant_Category	MedR_201307	MedR_201308	MedR_201309	MedR_201310	MedR_201311	MedR_201312	MedR_201401	MedR_201402	MedR_201403	MedR_201404	MedR_201405	MedR_201406
'DEPARTMENT OF EDUCATION	ADVERTISING SERVICES	0.41935484	1.29032258	0	0	0	0	11.16129032	0	0	0	0	0
'DEPARTMENT OF EDUCATION	AMERICAN AIRLINES	2.00000000	0	0.000000000	2.000000000	0.00000000	0	0	0	0	0	0	0
'DEPARTMENT OF EDUCATION	BOOK STORES	7.00000000	4.00000000	0	4.000000000	0.00000000	0	0	0	0	0	0	0
'DEPARTMENT OF EDUCATION	BUSINESS SERVICES NOT ELSEWHERE CLASSIFIED	0	1.77777778	0.555555556	0	0	2.33333333	1.11111111	0	0	0	0	0
'DEPARTMENT OF EDUCATION	CATALOG MERCHANTS	0	0	0	1.000000000	0	0	0	0	0	0	0	0
'DEPARTMENT OF EDUCATION	CHARITABLE AND SOCIAL SERVICE ORGANIZATIONS	0	0.50000000	1.000000000	1.000000000	0	1.25000000	0	0	0	0	0	0

```
# Mean Recency variable: Use "dcast" in Library "reshape2" to organize the
#data so each row is a merchant category of an agent.
mean_wide_rec <-
dcast(mean_per_agency_category_rec,
Agency_Name + Merchant_Category ~ Year_Month)
mean_wide_rec = as.matrix(mean_wide_rec)
mean_wide_rec[is.na(mean_wide_rec)] <- 0
mean_wide_rec = as.data.frame(mean_wide_rec)
#view head
kable(head(mean_wide_rec)) %>% kable_styling(latex_options = "scale_down")
```

Agency_Name	Merchant_Category	MeanR_201307	MeanR_201308	MeanR_201309	MeanR_201310	MeanR_201311	MeanR_201312	MeanR_201401	MeanR_201402	MeanR_201403	MeanR_201404	MeanR_201405	MeanR_201406
'DEPARTMENT OF EDUCATION	ADVERTISING SERVICES	0.50000000	0.25000000	0	0	0	0	0.25000000	0	0	0	0	0
'DEPARTMENT OF EDUCATION	AMERICAN AIRLINES	0.12500000	0	0.25000000	0.27272727	0.15909091	0	0	0	0	0	0	0
'DEPARTMENT OF EDUCATION	BOOK STORES	0.14285714	0.03571429	0	0.17857143	0.53571429	0	0	0	0	0	0	0
'DEPARTMENT OF EDUCATION	BUSINESS SERVICES NOT ELSEWHERE CLASSIFIED	0	0.21428571	0.35714286	0	0	0.14285714	0.21428571	0	0	0	0	0
'DEPARTMENT OF EDUCATION	CATALOG MERCHANTS	0	0	0	1.00000000	0	0	0	0	0	0	0	0
'DEPARTMENT OF EDUCATION	CHARITABLE AND SOCIAL SERVICE ORGANIZATIONS	0	0.33333333	0.25000000	0.25000000	0	0.12500000	0	0	0	0	0	0

```
#merge dataframes
model_df_rec <-
left_join(max_wide_rec,
mean_wide_rec,
by = c('Agency_Name', 'Merchant_Category'))
model_df_rec <-
merge(model_df_rec,
med_wide_rec,
by = c("Agency_Name", "Merchant_Category"))

#merge recency and transaction amount dataframes
```

```r
model_df <-
merge(model_df_amt,
model_df_rec,
by = c("Agency_Name", "Merchant_Category"))

#View in excel, write to CSV
#write_csv(model_df, "model_df.csv")

#remove identifier columns
#model_df$Agency_Name <- NULL
#model_df$Merchant_Category <- NULL

#change ratio columns to numeric
cols = c(3:74)
model_df[, cols] = apply(model_df[, cols], 2, function(x)
as.numeric(as.character(x)))

#scale data
model_df_scale <- as.data.frame(scale(model_df[, cols]))

#remove mean ratio calculations, use median as a feature instead as
#it is not impacted by outliers.
to.remove <-
c(
"Mean_201307",
"Mean_201308",
"Mean_201309",
"Mean_201310",
"Mean_201311",
"Mean_201312",
"Mean_201401",
"Mean_201402",
"Mean_201403",
"Mean_201404",
"Mean_201405",
"Mean_201406",
"MeanR_201307",
"MeanR_201308",
"MeanR_201309",
"MeanR_201310",
"MeanR_201311",
"MeanR_201312",
"MeanR_201401",
"MeanR_201402",
"MeanR_201403",
"MeanR_201404",
"MeanR_201405",
"MeanR_201406"
)
`%ni%` <- Negate(`%in%`)
model_df_scale <-
subset(model_df_scale, select = names(model_df_scale) %ni% to.remove)
```

4 DBSCAN Modeling Technique

4.1 DBCAN Method

DBSCAN is a density based clustering algorithm. Unlike K-means, which makes round clusters, DBSCAN can handle clusters of various shapes and sizes. It is therefore able to find clusters that K-means is unable to discover. For fraud analysis, DBSCAN will group together points that are closely packed together and mark outlier points that lie outside these clusters. These outlier points could be possible fraudulent transactions.

4.2 DBSCAN Model

Hyperparameters tuned include:

minPts - how many neighbors a point should have to be included into a cluster

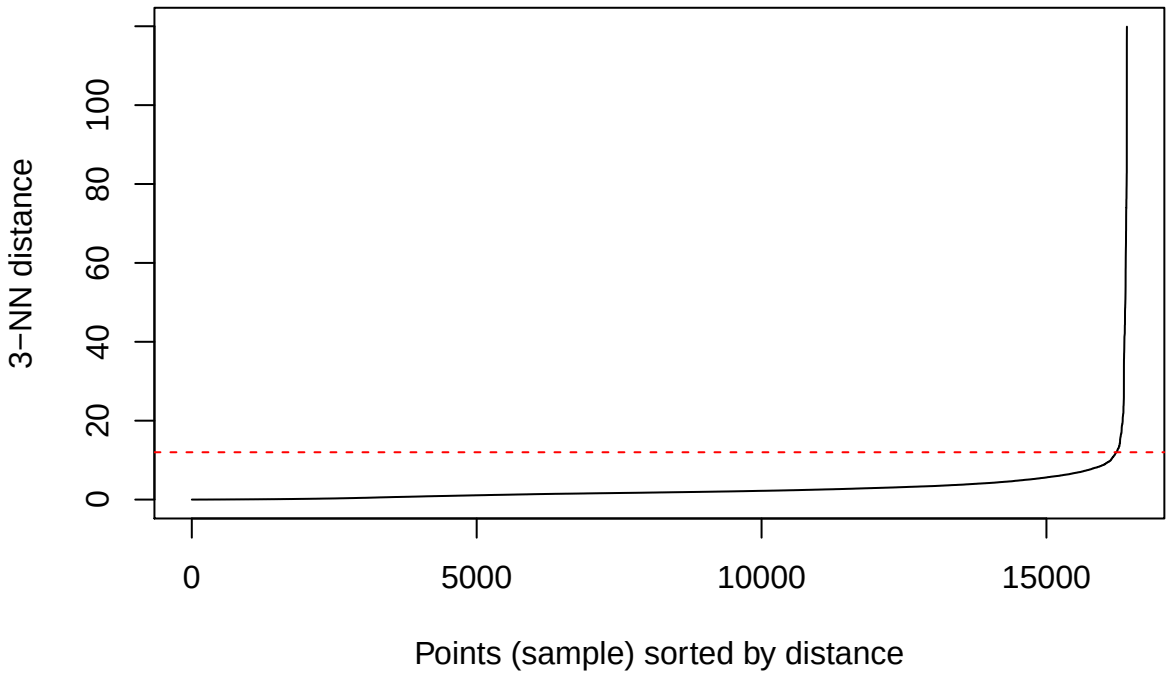
eps (epsilon) - how close points should be to each other to be considered a part of a cluster

(source: https://github.com/alitouka/spark_dbscan/wiki/Choosing-parameters-of-DBSCAN-algorithm)

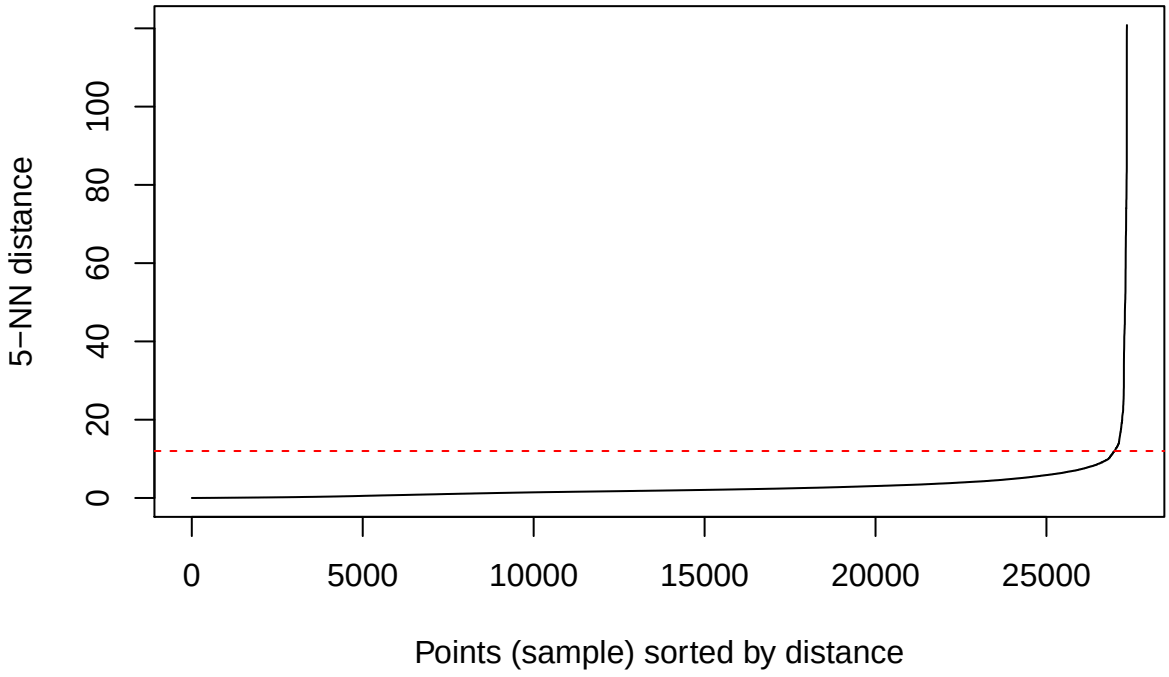
```
#load library
library(dbscan)
library(fpc)
library(factoextra)
library(rattle.data)

#principal component anlaysis to reduce high-dimensional data to two dimensions
fraud_PCA <- prcomp(model_df_scale)
fraud_PCA2 <- fraud_PCA$x[, 1:2]

#view head of ccard
kable(head(fraud_PCA2))
```

PC1	PC2
0.6956969	0.2121071
-4.1154660	0.1033287
-5.1863068	0.3878899
-0.3590404	0.1282656
1.2587429	0.1550933
-1.5495084	0.0768009

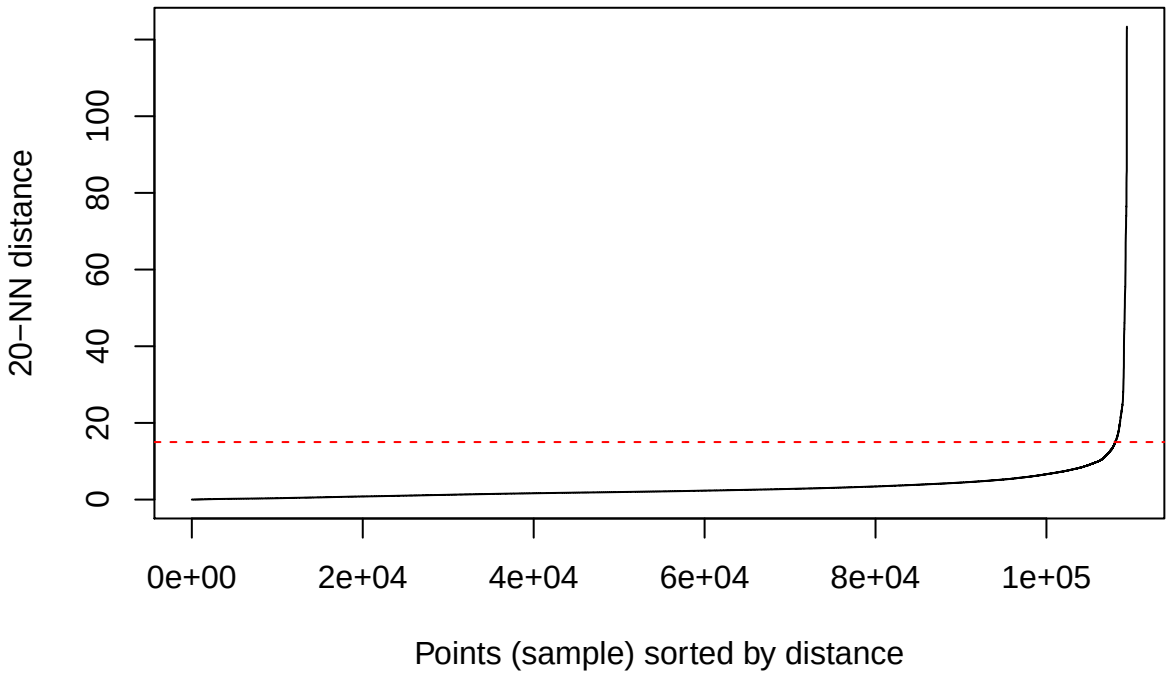
```
#Compute DBSCAN using fpc package
#minPts
#The rule of thumb for minPts is to use at least the number of dimensions of the data set plus one.
#(source: https://cran.r-project.org/web/packages/dbscan/vignettes/dbscan.pdf)
#In our case, this is 3.  However, I tested other MinPts as well.
#I tested 3, 5, 20, 50 and 100.

#eps
#For eps, we can plot the points' kNN distances (i.e., the distance to the kth nearest neighbor)
#in decreasing order and look for a knee in the plot.
#(source: https://cran.r-project.org/web/packages/dbscan/vignettes/dbscan.pdf)

#minPts = 3
kNNdistplot(model_df_scale, k = 3)
abline(h = 12, col = "red", lty = 2) #EPS = 12
```



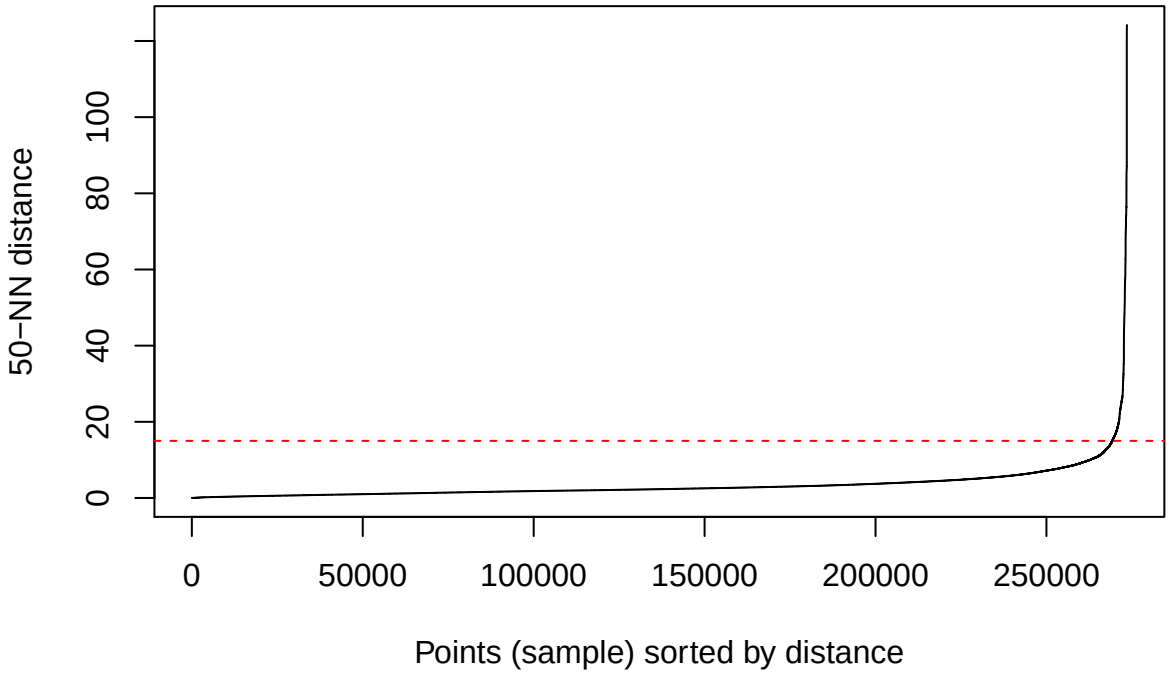

```
#minPts = 5
kNNdistplot(model_df_scale, k = 5)
abline(h = 12, col = "red", lty = 2) #EPS = 12
```



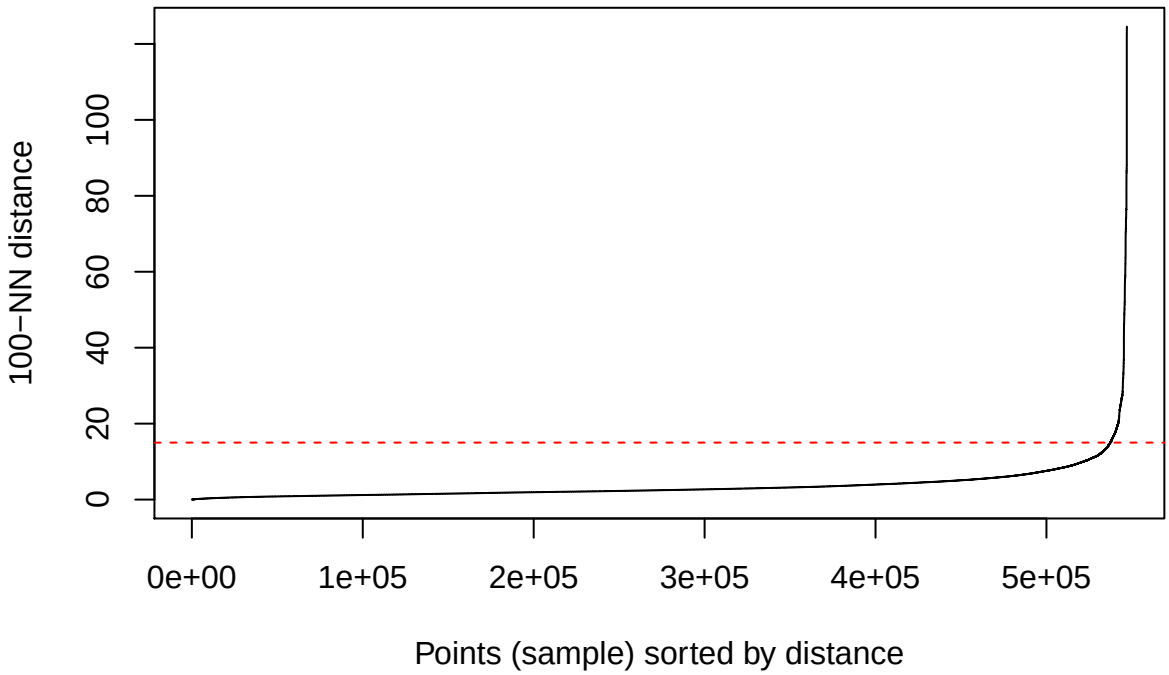

```
#minPts = 20
kNNdistplot(model_df_scale, k = 20)
abline(h = 15, col = "red", lty = 2) #EPS = 15
```




```
#minPts = 50
kNNdistplot(model_df_scale, k = 50)
abline(h = 15, col = "red", lty = 2) #EPS = 15
```



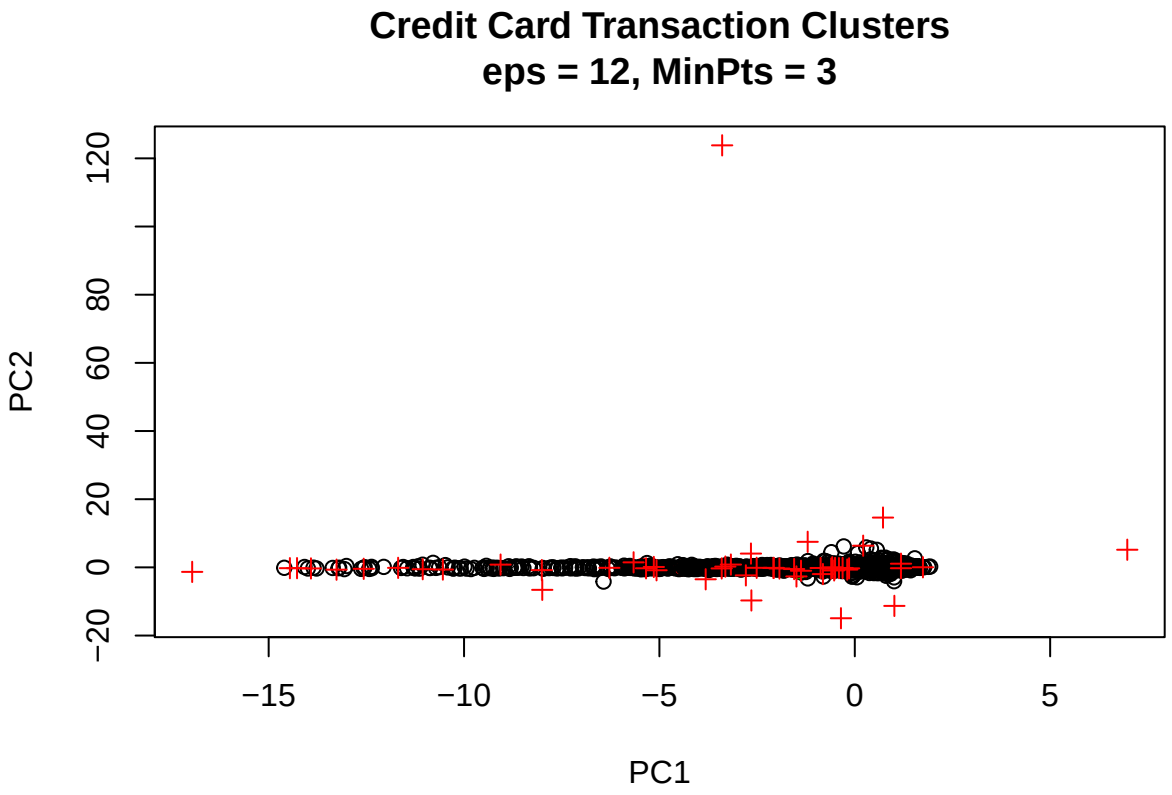

```
#minPts = 100
kNNdistplot(model_df_scale, k = 100)
abline(h = 15, col = "red", lty = 2) #EPS = 15
```




```
#eps = 12, MinPts = 3
set.seed(1)
modl <- fpc::dbscan(model_df_scale, eps = 12, MinPts = 3)
#view table
modl #The clustering contains 3 clusters and 52 noise points.
```

```
## dbscan Pts=5471 MinPts=3 eps=12
##         0    1 2 3
## border 52   10 0 0
## seed    0 5401 3 5
## total  52 5411 3 5
```

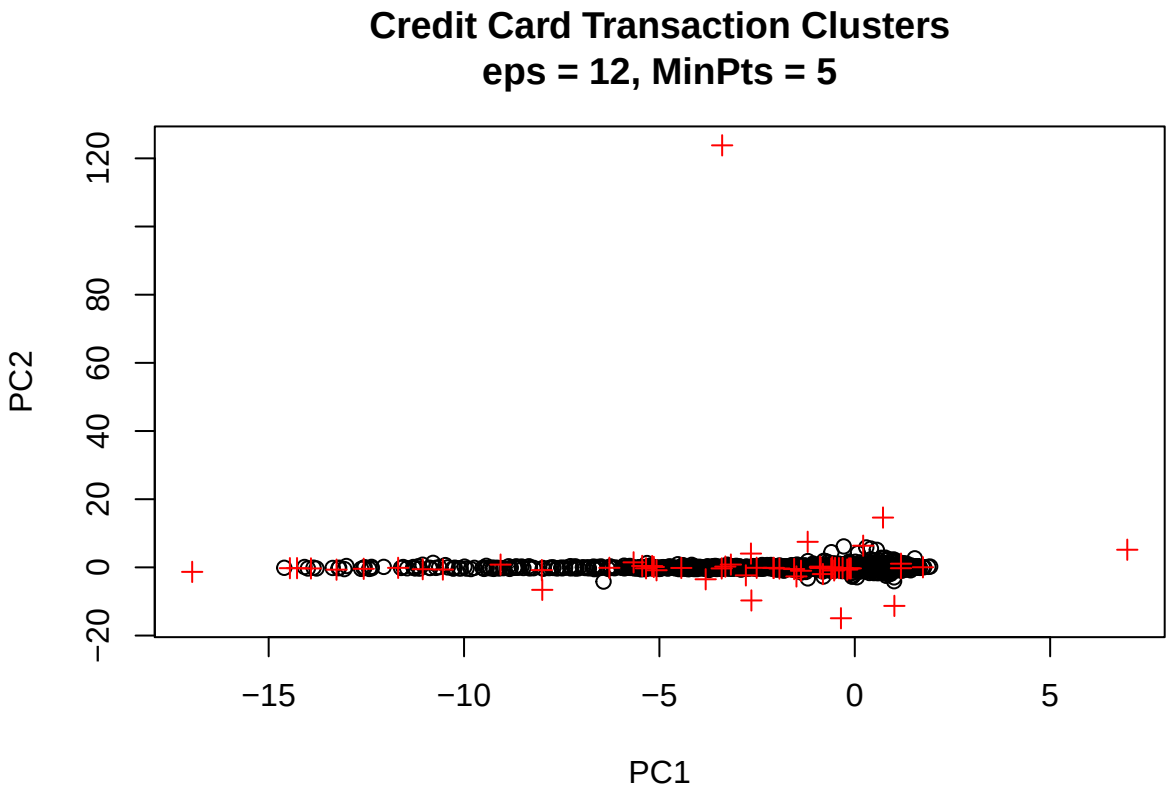
```
#plot clusters and add noise (cluster 0) as crosses.
plot(
fraud_PCA2,
main = "Credit Card Transaction Clusters\neps = 12, MinPts = 3",
sub = "Noise points plotted as crosses",
cex.sub = 0.75,
font.sub = 3,
col.sub = "red",
col =
modl$cluster
)
points(fraud_PCA2[modl$cluster == 0,], pch = 3, col = "red")
```




```
noise <- model_df[modl$cluster == 0,]

#eps = 12, MinPts = 5
set.seed(1)
modl2 <- fpc::dbscan(model_df_scale, eps = 12, MinPts = 5)
#view table
modl2 #The clustering contains 2 clusters and 58 noise points.
```

```
## dbscan Pts=5471 MinPts=5 eps=12
##          0    1 2
## border 58   25 0
## seed    0 5383 5
## total  58 5408 5
```

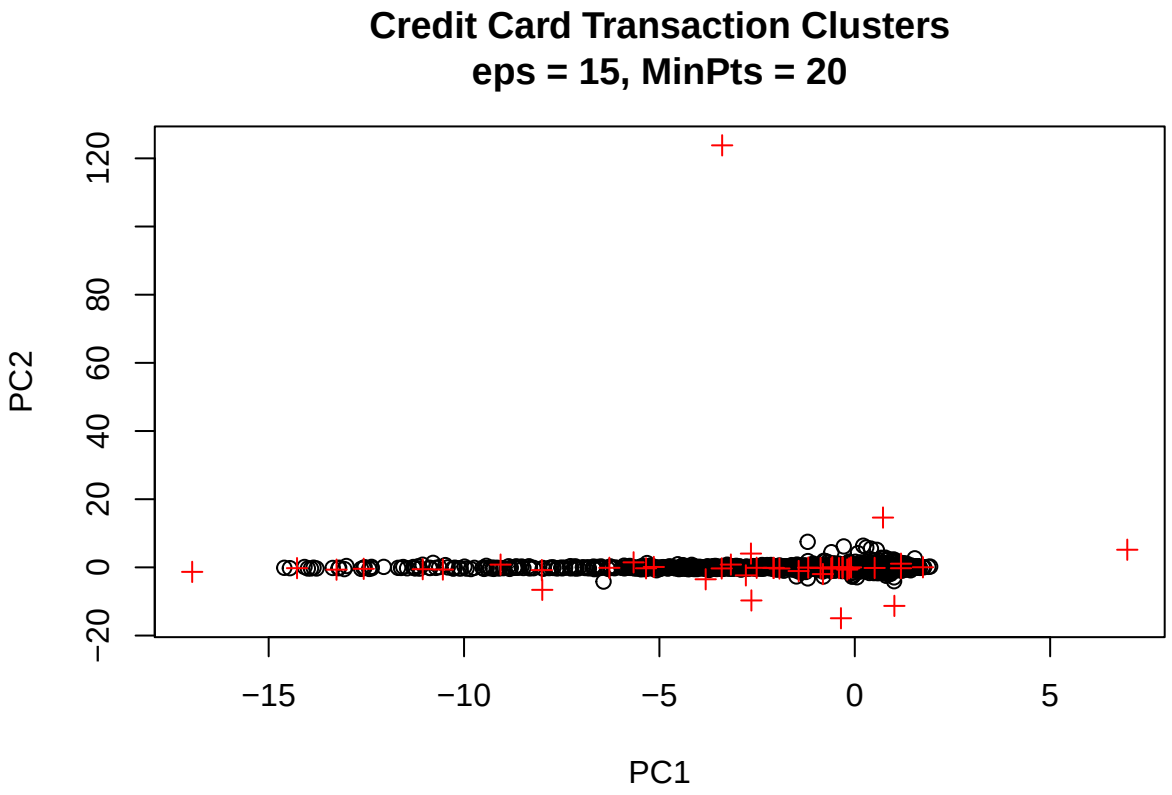
```
#plot clusters and add noise (cluster 0) as crosses.
plot(
fraud_PCA2,
main = "Credit Card Transaction Clusters\neps = 12, MinPts = 5",
sub = "Noise points plotted as crosses",
cex.sub = 0.75,
font.sub = 3,
col.sub = "red",
col =
modl2$cluster
)
points(fraud_PCA2[modl2$cluster == 0,], pch = 3, col = "red")
```




```
noise <- model_df[modl2$cluster == 0,]

#eps = 15, MinPts = 20
set.seed(1)
modl2b <- fpc::dbscan(model_df_scale, eps = 15, MinPts = 20)
#view table
modl2b #The clustering contains 1 cluster and 44 noise points.
```

```
## dbscan Pts=5471 MinPts=20 eps=15
##          0    1
## border 44   44
## seed    0 5383
## total  44 5427
```

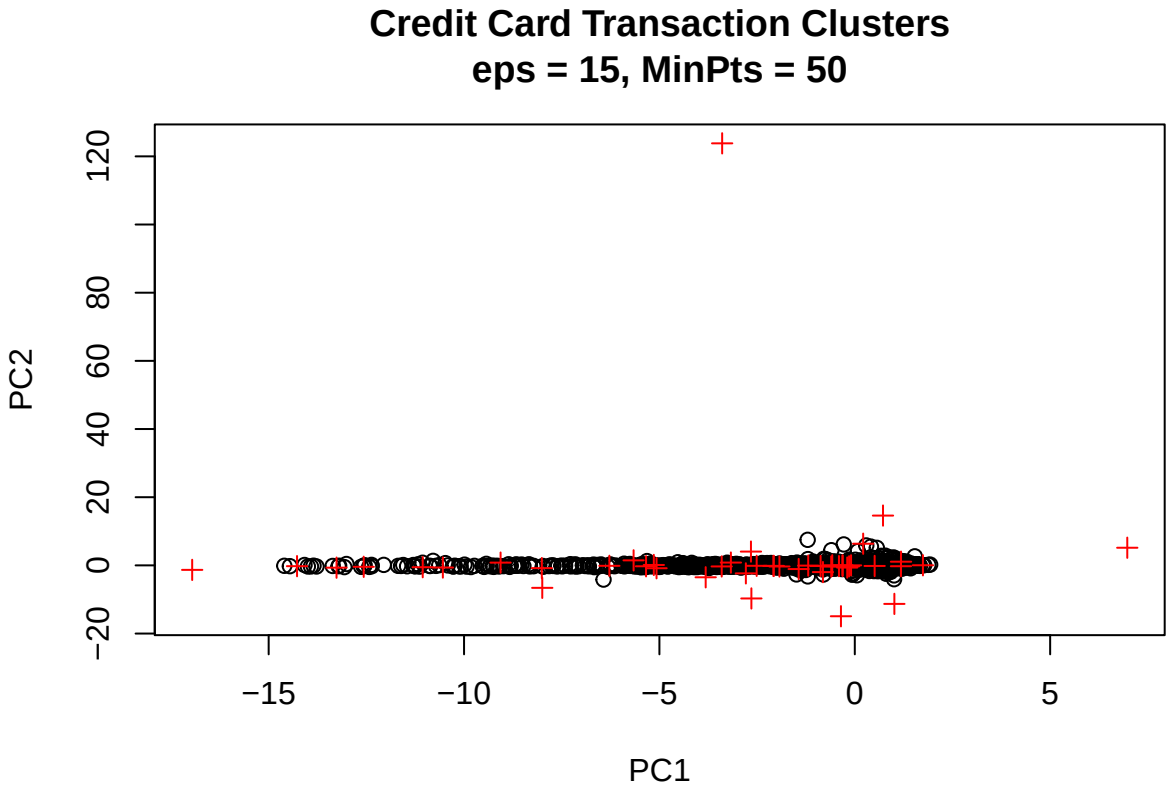
```
#plot clusters and add noise (cluster 0) as crosses.
plot(
fraud_PCA2,
main = "Credit Card Transaction Clusters\neps = 15, MinPts = 20",
sub = "Noise points plotted as crosses",
cex.sub = 0.75,
font.sub = 3,
col.sub = "red",
col =
modl2b$cluster
)
points(fraud_PCA2[modl2b$cluster == 0,], pch = 3, col = "red")
```




```
noise <- model_df[modl2b$cluster == 0,]

#eps = 15, MinPts = 50
set.seed(1)
modl3 <- fpc::dbscan(model_df_scale, eps = 15, MinPts = 50)
#view table
modl3 #The clustering contains 1 cluster and 46 noise points.
```

```
## dbscan Pts=5471 MinPts=50 eps=15
##          0    1
## border 46   57
## seed    0 5368
## total  46 5425
```

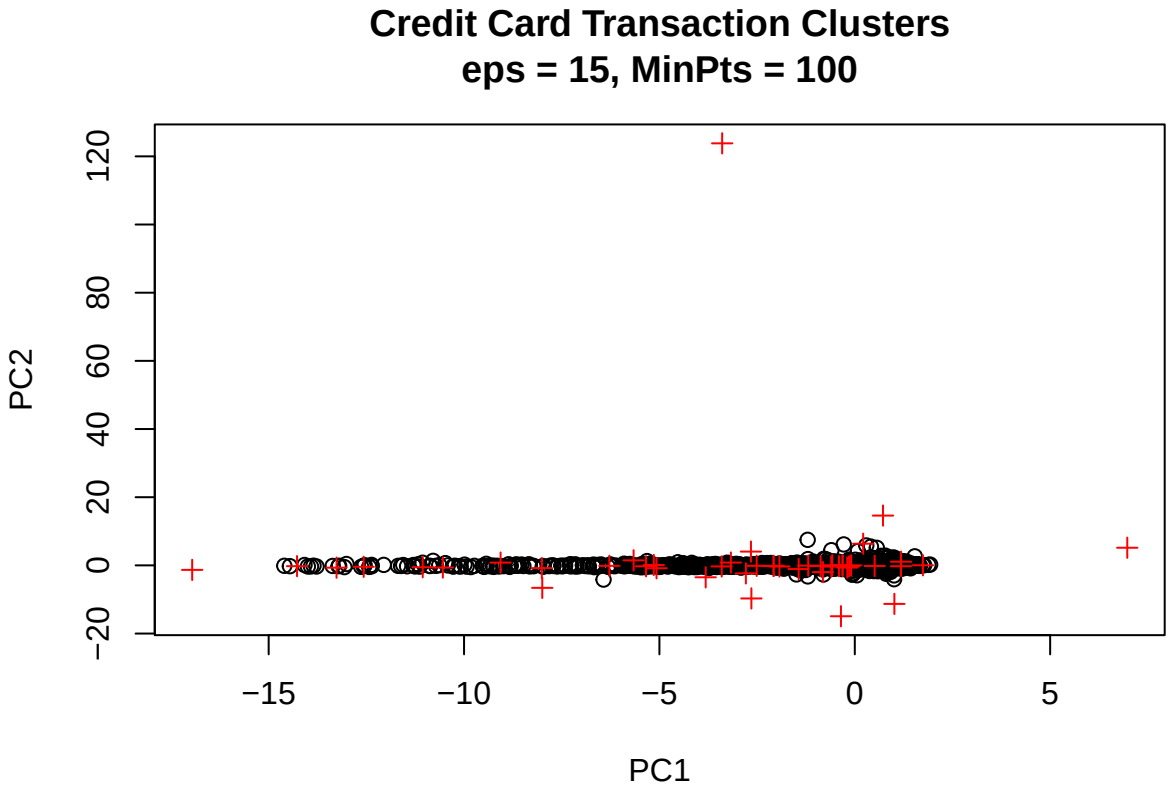
```
#plot clusters and add noise (cluster 0) as crosses.
plot(
fraud_PCA2,
main = "Credit Card Transaction Clusters\neps = 15, MinPts = 50",
sub = "Noise points plotted as crosses",
cex.sub = 0.75,
font.sub = 3,
col.sub = "red",
col =
modl3$cluster
)
points(fraud_PCA2[modl3$cluster == 0,], pch = 3, col = "red")
```




```
noise <- model_df[modl3$cluster == 0,]

#eps = 15, MinPts = 100
set.seed(1)
modl4 <- fpc::dbscan(model_df_scale, eps = 15, MinPts = 100)
#view table
modl4 #The clustering contains 1 cluster and 46 noise points.
```

```
## dbscan Pts=5471 MinPts=100 eps=15
##         0    1
## border 46   67
## seed    0 5358
## total  46 5425
```

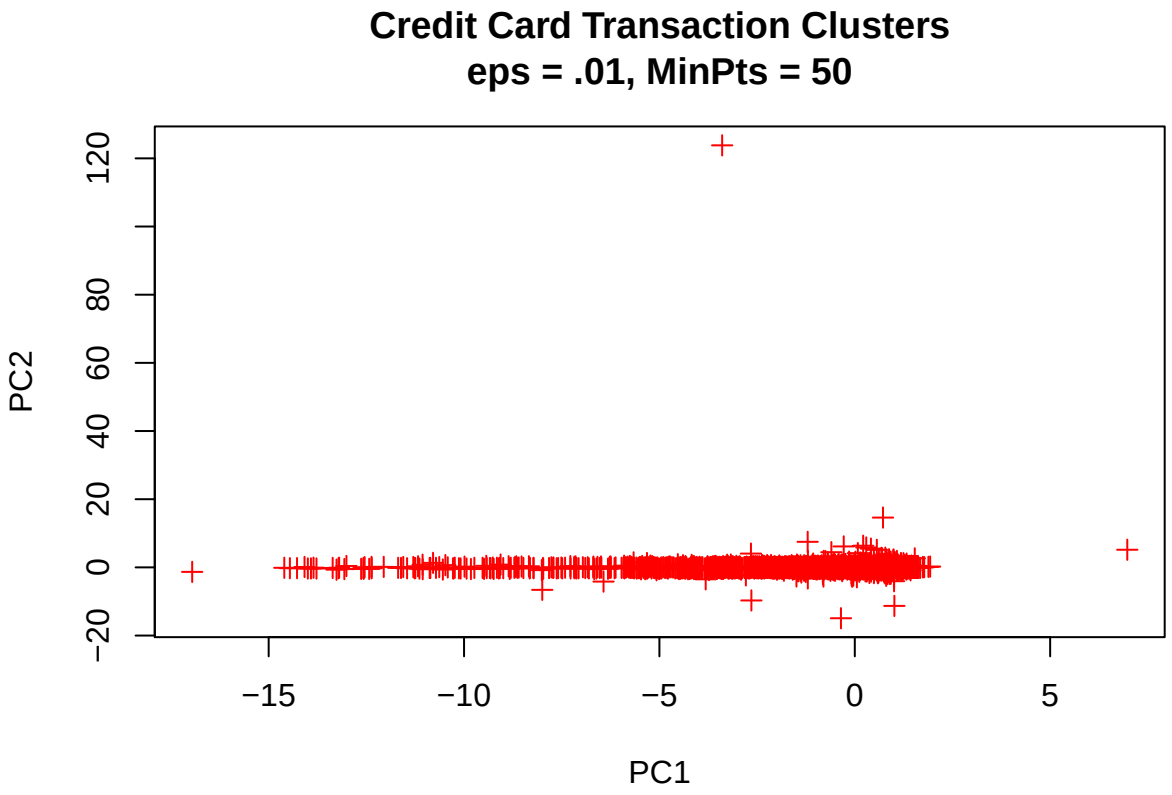
```
#plot clusters and add noise (cluster 0) as crosses.
plot(
fraud_PCA2,
main = "Credit Card Transaction Clusters\neps = 15, MinPts = 100",
sub = "Noise points plotted as crosses",
cex.sub = 0.75,
font.sub = 3,
col.sub = "red",
col =
modl4$cluster
)
points(fraud_PCA2[modl4$cluster == 0,], pch = 3, col = "red")
```




```
noise <- model_df[modl4$cluster == 0,]

#I also tested smaller numbers for EPS - 0.01, 0.15, 0.5, 0.99, 2.0
#eps = .01, MinPts = 50
set.seed(1)
modl5 <- fpc::dbscan(model_df_scale, eps = 0.01, MinPts = 50)
#view table
modl5 #The clustering contains 0 clusters and 5471 noise points, not useful.
```

```
## dbscan Pts=5471 MinPts=50 eps=0.01
##
##    0
## 5471
```

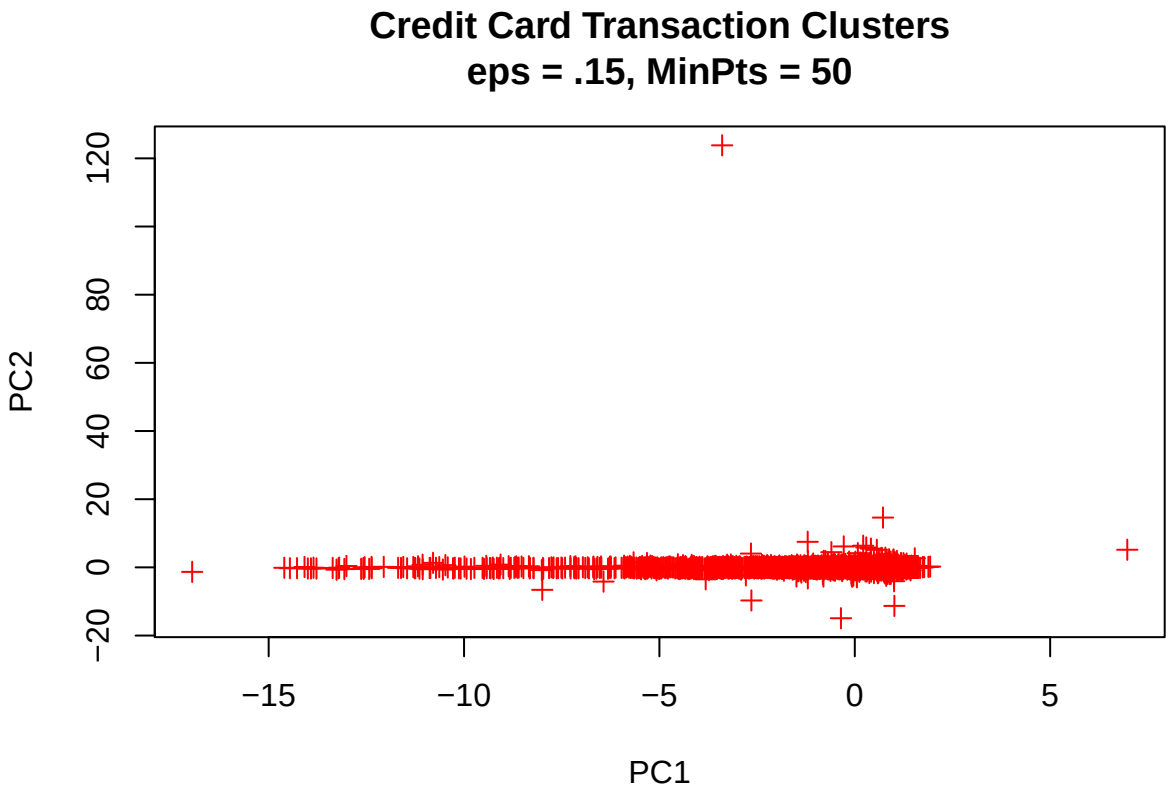
```
#plot clusters and add noise (cluster 0) as crosses.
plot(
fraud_PCA2,
main = "Credit Card Transaction Clusters\neps = .01, MinPts = 50",
sub = "Noise points plotted as crosses",
cex.sub = 0.75,
font.sub = 3,
col.sub = "red",
col =
modl5$cluster
)
points(fraud_PCA2[modl5$cluster == 0,], pch = 3, col = "red")
```




```
noise <- model_df[modl5$cluster == 0,]

#eps = 0.15, MinPts = 50
set.seed(1)
modl6 <- fpc::dbscan(model_df_scale, eps = 0.15, MinPts = 50)
#view table
modl6 #The clustering contains 0 clusters and 5471 noise points, not useful.
```

```
## dbscan Pts=5471 MinPts=50 eps=0.15
##
##    0
## 5471
```

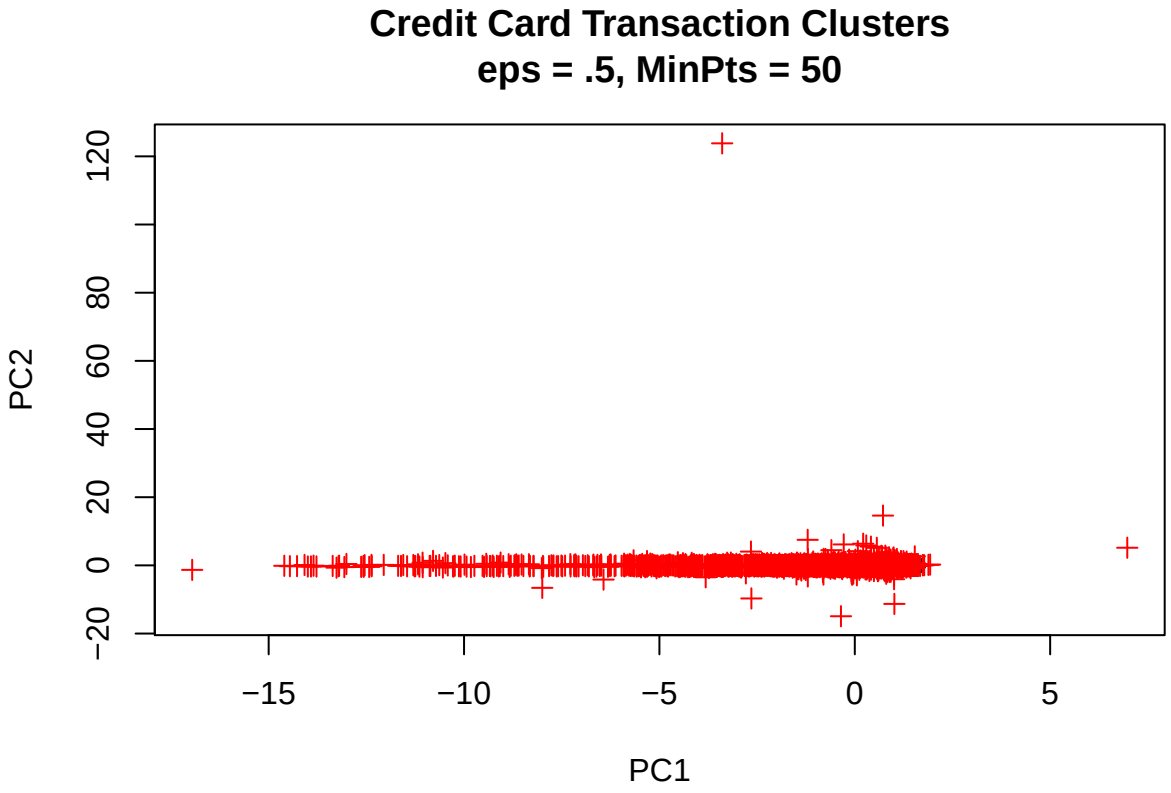
```
#plot clusters and add noise (cluster 0) as crosses.
plot(
fraud_PCA2,
main = "Credit Card Transaction Clusters\neps = .15, MinPts = 50",
sub = "Noise points plotted as crosses",
cex.sub = 0.75,
font.sub = 3,
col.sub = "red",
col =
modl6$cluster
)
points(fraud_PCA2[modl6$cluster == 0,], pch = 3, col = "red")
```




```
noise <- model_df[modl6$cluster == 0,]

#eps = 0.5, MinPts = 50
set.seed(1)
modl7 <- fpc::dbscan(model_df_scale, eps = 0.5, MinPts = 50)
#view table
modl7 #The clustering contains 3 clusters and 5265 noise points, not useful.
```

```
## dbscan Pts=5471 MinPts=50 eps=0.5
##           0  1  2  3
## border 5265 41 60 51
## seed      0 26 16 12
## total  5265 67 76 63
```

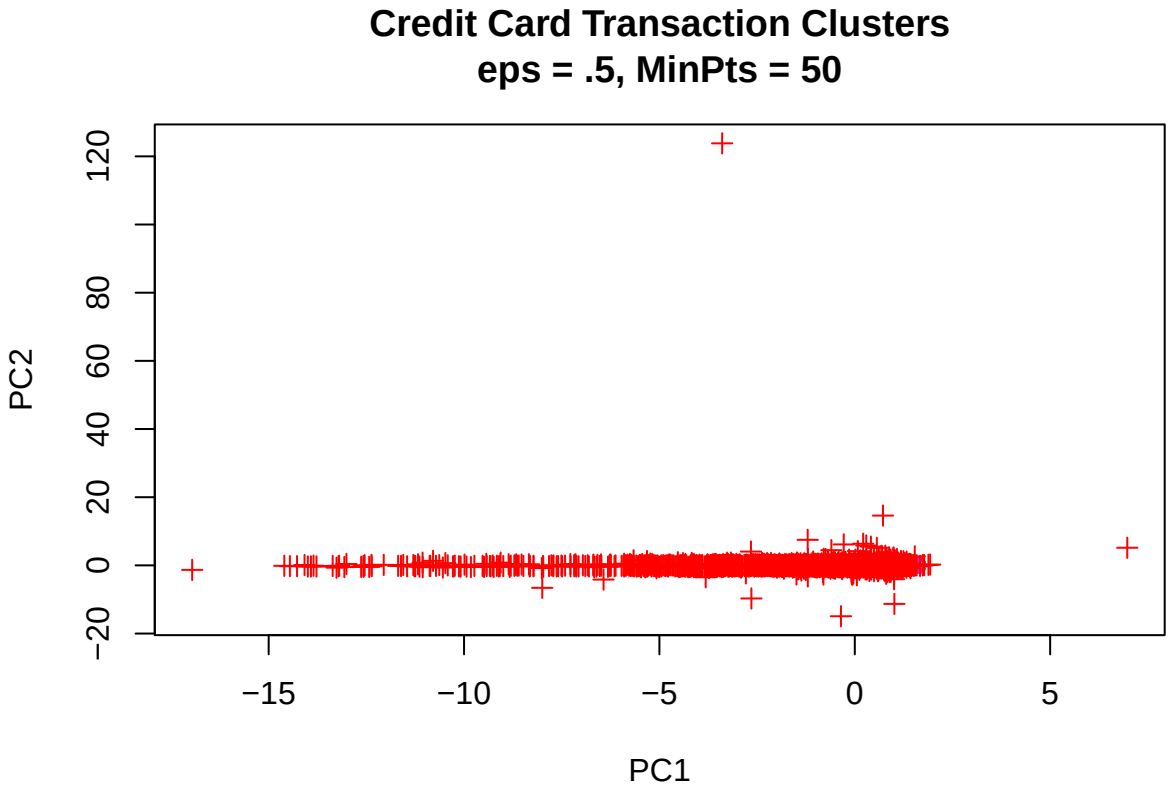
```
#plot clusters and add noise (cluster 0) as crosses.
plot(
fraud_PCA2,
main = "Credit Card Transaction Clusters\neps = .5, MinPts = 50",
sub = "Noise points plotted as crosses",
cex.sub = 0.75,
font.sub = 3,
col.sub = "red",
col =
modl7$cluster
)
points(fraud_PCA2[modl7$cluster == 0,], pch = 3, col = "red")
```




```
noise <- model_df[modl7$cluster == 0,]

#eps = 0.99, MinPts = 50
set.seed(1)
modl8 <- fpc::dbscan(model_df_scale, eps = 0.99, MinPts = 50)
#view table
modl8 #The clustering contains 8 clusters and 4289 noise points, not useful.
```

```
## dbscan Pts=5471 MinPts=50 eps=0.99
##           0  1  2   3  4   5   6  7   8
## border 4289 35 48  42 37 101 122 39  50
## seed      0 43 46  77 27 146 232 57  80
## total  4289 78 94 119 64 247 354 96 130
```

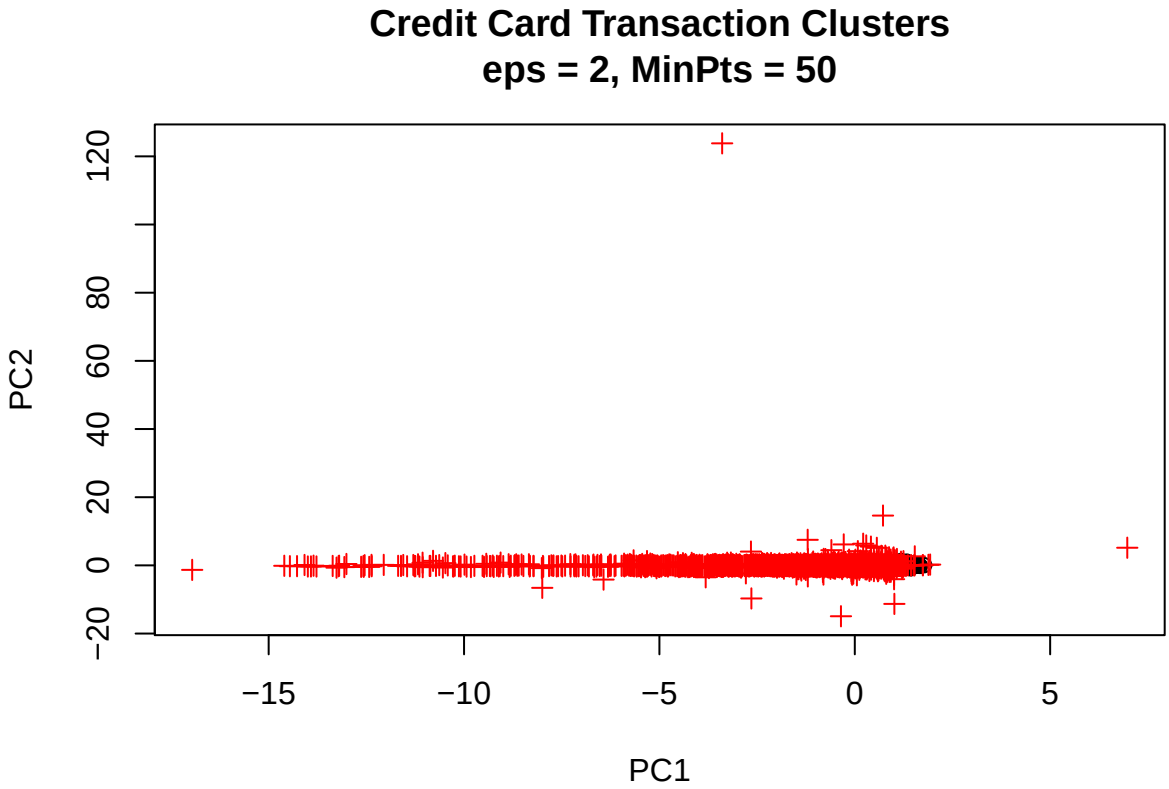
```
#plot clusters and add noise (cluster 0) as crosses.
plot(
fraud_PCA2,
main = "Credit Card Transaction Clusters\neps = .5, MinPts = 50",
sub = "Noise points plotted as crosses",
cex.sub = 0.75,
font.sub = 3,
col.sub = "red",
col =
modl8$cluster
)
points(fraud_PCA2[modl8$cluster == 0,], pch = 3, col = "red")
```




```
noise <- model_df[modl8$cluster == 0,]

#eps = 2, MinPts = 50
set.seed(1)
modl9 <- fpc::dbscan(model_df_scale, eps = 2, MinPts = 50)
#view table
modl9 #The clustering contains 1 cluster1 and 2633 noise points, not useful.
```

```
## dbscan Pts=5471 MinPts=50 eps=2
##           0    1
## border 2633  827
## seed      0 2011
## total  2633 2838
```

```
#plot clusters and add noise (cluster 0) as crosses.
plot(
fraud_PCA2,
main = "Credit Card Transaction Clusters\neps = 2, MinPts = 50",
sub = "Noise points plotted as crosses",
cex.sub = 0.75,
font.sub = 3,
col.sub = "red",
col =
modl9$cluster
)
points(fraud_PCA2[modl9$cluster == 0,], pch = 3, col = "red")
```



Noise points plotted as crosses

```
noise <- model_df[modl9$cluster == 0,]
```

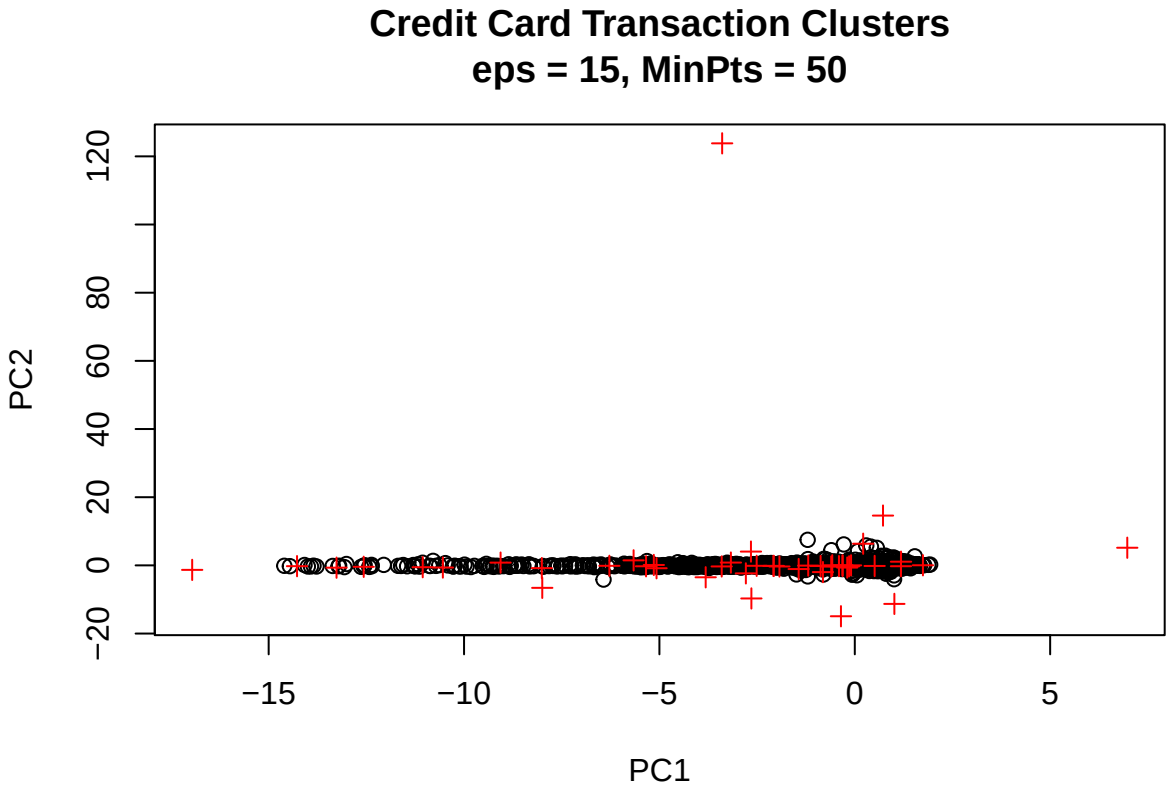
4.3 Best Model with hyperparameters of MinPts=50 and eps=15

It appears that using eps of 15 and MinPts of 50 resulted in a reasonable model. It clustered the data points into 1 cluster with 46 outliers.

```
modl3 #The clustering contains 1 cluster and 46 noise points.
```

```
## dbscan Pts=5471 MinPts=50 eps=15
##         0    1
## border 46   57
## seed    0 5368
## total  46 5425
```

```
#plot clusters and add noise (cluster 0) as crosses.
plot(fraud_PCA2, main = "Credit Card Transaction Clusters\neps = 15, MinPts = 50", sub = "Noise points
modl3$cluster)
points(fraud_PCA2[modl3$cluster == 0, ], pch = 3, col = "red")
```



Noise points plotted as crosses

```
noise <- model_df[modl3$cluster == 0, ]
```

4.4 Outliers

```
#create DF that include possible fraud transactions
fraud <- model_df[modl3$cluster == 0, ]
fraud <- fraud[, 1:2]

#view fraud
kable(fraud) %>% kable_styling(latex_options = "scale_down")
```

	Agency_Name	Merchant_Category
35	ARDMORE HIGHER EDUCATION CENTER	HOUSEHOLD APPLIANCE STORES
279	COMM. ON CONSUMER CREDIT	GOVERNMENT SERVICES–NOT ELSEWHERE CLASSIFIED
293	COMM. ON CONSUMER CREDIT	SHERATON
327	COMPSOURCE OKLAHOMA	DELTA
1301	DEPARTMENT OF REHABILITATION SERVICES	GOVERNMENT SERVICES–NOT ELSEWHERE CLASSIFIED
1328	DEPARTMENT OF REHABILITATION SERVICES	MISCELLANEOUS FOOD STORES-CONV STRS AND SPECIALTY MKTS.
1462	DEPARTMENT OF TOURISM AND RECREATION	MISCELLANEOUS AND SPECIALTY RETAIL STORES
1495	DEPARTMENT OF TOURISM AND RECREATION	SHERATON
1542	DEPARTMENT OF TRANSPORTATION	COMMERCIAL EQUIPMENT, NOT ELSEWHERE CLASSIFIED
1721	DEPARTMENT OF VETERANS AFFAIRS	MARRIOTT
1826	DEPARTMENT OF WILDLIFE CONSERVATION	HILTON HOTELS
2159	GRAND RIVER DAM AUTH.	COMMERCIAL EQUIPMENT, NOT ELSEWHERE CLASSIFIED
2165	GRAND RIVER DAM AUTH.	COMPUTERS, COMPUTER PERIPHERAL EQUIPMENT, SOFTWARE
2174	GRAND RIVER DAM AUTH.	DETECTIVE AGENCIES,PROTECTIVE AGENCIES,AND SECURITY SERVICES
2182	GRAND RIVER DAM AUTH.	EMPLOYMENT AGENCIES AND TEMPORARY HELP SERVICES
2194	GRAND RIVER DAM AUTH.	GOVERNMENT SERVICES–NOT ELSEWHERE CLASSIFIED
2203	GRAND RIVER DAM AUTH.	HOMEWOOD SUITES
2260	GRAND RIVER DAM AUTH.	TELECOMMUNICATION EQUIPMENT AND TELEPHONE SALES
2614	MENTAL HEALTH AND SUBSTANCE ABUSE SERV.	DEPARTMENT STORES
2687	MENTAL HEALTH AND SUBSTANCE ABUSE SERV.	RENAISSANCE HOTELS
2713	N. E. OKLA. A & M COLLEGE	AUTOMATED FUEL DISPENSER
2752	N. E. OKLA. A & M COLLEGE	RECORD STORES
2782	OFFICE OF JUVENILE AFFAIRS	COMFORT HOTEL INTERNATIONAL
2865	OFFICE OF MANAGEMENT AND ENTERPRISE SERV	BUSINESS SERVICES NOT ELSEWHERE CLASSIFIED
2880	OFFICE OF MANAGEMENT AND ENTERPRISE SERV	COMMERCIAL EQUIPMENT, NOT ELSEWHERE CLASSIFIED
2975	OFFICE OF MANAGEMENT AND ENTERPRISE SERV	TOLLS AND BRIDGE FEES
3330	OKLA. CITY COMMUNITY COLLEGE	GROCERY STORES,AND SUPERMARKETS
3374	OKLA. HORSE RACING COMM.	GOVERNMENT SERVICES–NOT ELSEWHERE CLASSIFIED
3471	OKLA. PANHANDLE STATE UNIV.	MOTOR FREIGHT CARRIERS,AND TRUCKING
3485	OKLA. PANHANDLE STATE UNIV.	RECORD STORES
3572	OKLAHOMA AERONAUTICS COMMISSION	COMP PROG,DATA PROCESSING,AND INTEGRATED SYS DES IGN SVCS
3821	OKLAHOMA STATE UNIVERSITY	AIRLINES AND AIR CARRIERS
3937	OKLAHOMA STATE UNIVERSITY	LOCAL AND SUBURBAN COMMUTER PASS TRANS, INCLUDING FEE
3988	OKLAHOMA STATE UNIVERSITY	RECORD STORES
4022	OKLAHOMA STATE UNIVERSITY	VARIETY STORES
4143	REDLANDS COMMUNITY COLLEGE	COLLEGES,UNIVERSITIES,PROFESSIONAL SCHLS AND JR COLLEGES
4165	REDLANDS COMMUNITY COLLEGE	LUXOR HOTEL AND CASINO
4450	SECRETARY OF STATE	TRAVEL AGENCIES
4699	STATE DEPARTMENT OF HEALTH	HYATT PLACE
4713	STATE DEPARTMENT OF HEALTH	MISCELLANEOUS GENERAL MERCHANDISE
4717	STATE DEPARTMENT OF HEALTH	NON-DURABLE GOODS NOT ELSEWHERE CLASSIFIED
4752	STATE ELECTION BOARD	STATIONERY,OFFICE AND SCHOOL SUPPLY STORES
4993	TULSA COMMUNITY COLLEGE	NON-DURABLE GOODS NOT ELSEWHERE CLASSIFIED
5156	UNIV.OF SCIENCE & ARTS OF OK	COMFORT HOTEL INTERNATIONAL
5169	UNIV.OF SCIENCE & ARTS OF OK	FAST FOOD RESTAURANTS
5239	UNIVERSITY OF OKLAHOMA	CAMPER,RECREATIONAL AND UTILITY TRAILER DEALERS

4.5 Business Insight

Agency transactions that occurred within the merchant category listed in the fraud data frame could possibly be fraud based on my DBSCAN analysis. Transactions that occurred within these merchant categories at these agencies require further analysis to determine if fraud actually occurred.