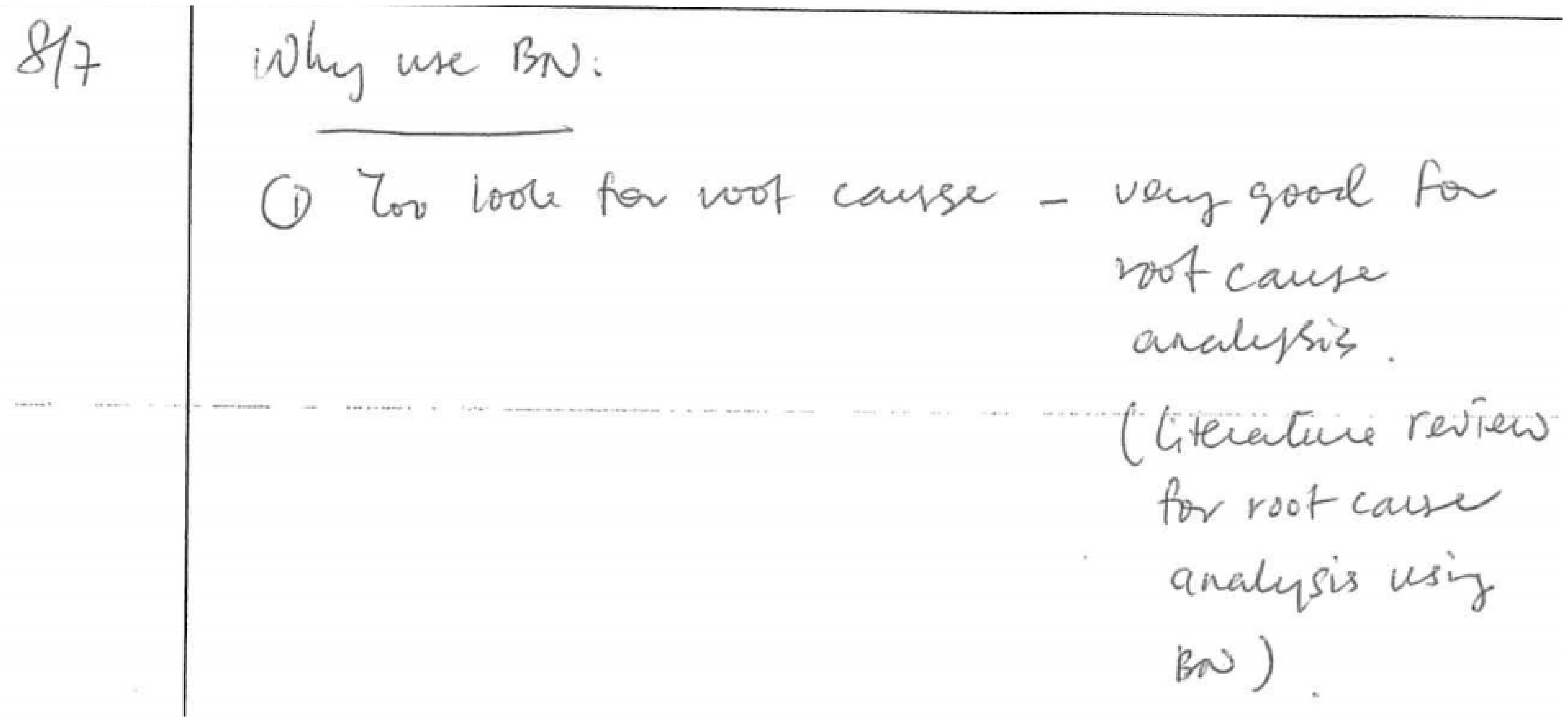
PHD - Progress update 14/10/2015

Fauzy Bin Che Yayah

October 14, 2015

- Why use Bayesian Net ?

1. To look for the root cause - the dataset is in factor type which is status not a number.
2. To find out the probabilistic relationship between the symptom error code and the resolution
3.



8/7

Why use BN:

① Too look for root cause - very good for root cause analysis.

(Literature review for root cause analysis using BN).

1

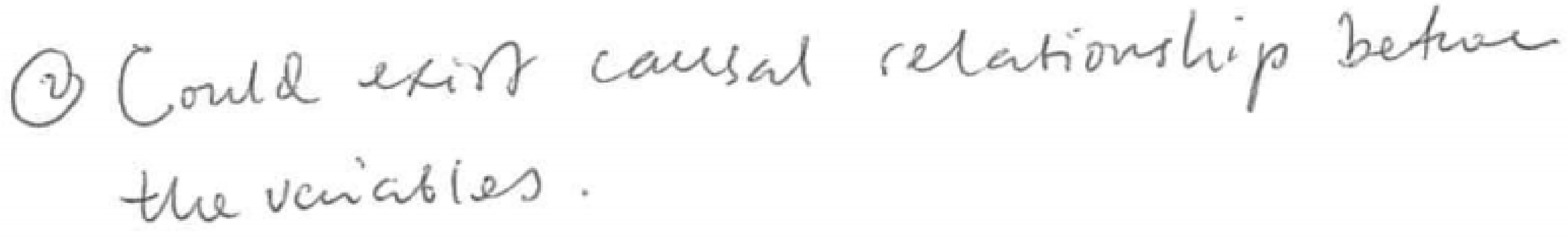
Citation	
TroubleMiner: Mining network trouble tickets Medem, A. ; Akodjenou, M.-I ; Teixeira, R. 20091	
Knowledge Discovery from Trouble Ticketing Reports in a Large Telecommunication Company Temprado, Y. ; Garcia, C. ; Molinero, F.J. 2009	Data Mining , Text Mining an
A Bayesian Approach To Stochastic Root Finding 2011	
A Fully Bayesian Approach For Unit Root Testing 2011	
Online Root-Cause Analysis Of Alarms In Discrete Bayesian 2014	
Documents Categorization Based On Bayesian Spanning Tree 2006	
Benefits of a Bayesian Approach to Anomaly and Failure 2009	

List of literature review regarding Bayesian Net :-

1.A real-life application of multi-agent systems for fault diagnosis in the provision of an Internet business service

2.A Bayesian Network approach to diagnosing the root cause of failure

3.sss



② Could exist causal relationship betwee the variables.

Process on gathering the dataset

- Acquiring dataset for 100 records, for each zone , randomize , selective year ; ie . 2015
- Rules :-

Rules	Description
status = 'Closed'	Dataset must be closed for complete information
network_tt_id is NULL	Dataset must be not related to Network Trouble Ticket
trouble ticket type <> PASSIVE	Trouble Ticket must related to the Active elements such as routers, switches , modem , etc
installed_date is NOT NULL	This field must have value
created_date is NOT NULL	This field must have value

Rules	Description
closed_date is NOT NULL	This field must have value
closed_date is NOT NULL	This field must have value
product is NOT NULL	This field must have value
sub_product is NOT NULL	This field must have value
length description > 10	This field is useful for text analysis
rand()	Record selection is in random mode
zone	Should selective from different zone , sparse

For sample purpose - selecting dataset from `ZONE KEPONG` for the analysis due to this zone has the highest records inside the Trouble Ticket dataset.

- Using Impala for the data retrieval :-

Documentation - https://github.com/piersharding/dplyrimpaladb

- Data processing using DplyrImpalaDb
- Package installation manual below :-

```
install.packages(c("RJDBC", "devtools", "dplyr"))
devtools::install_github("jwills/dplyrimpaladb")
install.packages("dplyrimpaladb")
```

- Basic notes why choosing Impala.

1. Cloudera 'Impala', which is a massively parallel processing (MPP) SQL query engine runs natively in Apache Hadoop
2. Impala's Place in the Big Data Ecosystem
3. Flexibility for Big Data Workflow
4. High-Performance Analytics

Connection to Impala

Basic Impala drivers can be downloaded from https://github.com/Mu-Sigma/RImpala/blob/master/impala-jdbc-cdh5.zip

Below is the components required and how to set the class path for the Impala drivers , RJava , RJDBC and dplyr

```
suppressWarnings(suppressMessages(library("rJava")))
suppressWarnings(suppressMessages(library("RJDBC")))
suppressWarnings(suppressMessages(library("dplyr")))
suppressWarnings(suppressMessages(library("caret")))
suppressWarnings(suppressMessages(library("corrplot")))
suppressWarnings(suppressMessages(library("lazy")))
suppressWarnings(suppressMessages(library("dplyrimpaladb")))
suppressWarnings(suppressMessages(library("rpart")))


.jaddClassPath(c(list.files(paste(getwd(),"/lib",sep = ''),pattern="jar$",full.names=T)))

.jinit(classpath = c(list.files(paste(getwd(),"/lib",sep = ''),pattern="jar$",full.names=T)))

dplyr.jdbc.classpath = c(list.files(paste(getwd(),"/lib",sep = ''),pattern="jar$",full.names=T))

conn <- src_impaladb(dbname='nova', host='10.54.1.151')
```

```
## Loading required package: testthat
```

```
## [1] "here:"
## [1] FALSE
```

- Zone list

```
result <-  tbl(conn, sql("select zone from nova.nova_trouble_ticket where zone <> 'null' group by zone order by zone limit 1000"))
as.data.frame(result)
```

```
##                       zone
## 1             ZONE AIR ITAM
## 2                ZONE BANGI
## 3              ZONE BANGSAR
## 4              ZONE BANTING
## 5                 ZONE BATU
## 6           ZONE BATU PAHAT
## 7           ZONE BAYAN BARU
## 8             ZONE BINTULU
## 9       ZONE BUKIT ANGGERIK
## 10      ZONE BUKIT MERTAJAM
## 11          ZONE BUKIT RAJA
## 12         ZONE BUTTERWORTH
## 13           ZONE CYBERJAYA
## 14              ZONE GOMBAK
## 15                ZONE IPOH
## 16              ZONE KAJANG
## 17              ZONE KEPONG
## 18               ZONE KERAMAT
## 19               ZONE KINRARA
## 20            ZONE KL CENTRAL
## 21                 ZONE KLANG
## 22 ZONE KOTA KINABALU SELATAN
## 23   ZONE KOTA KINABALU UTARA
## 24               ZONE KUCHING
## 25                 ZONE KULIM
## 26              ZONE LANGKAWI
## 27                ZONE MALURI
## 28          ZONE MELAKA UTARA
## 29                  ZONE MIRI
## 30     ZONE N. SEMBILAN UTARA
## 31                ZONE PANDAN
## 32               ZONE PELANGI
## 33                ZONE PERLIS
## 34         ZONE PETALING JAYA
## 35               ZONE PUCHONG
## 36         ZONE SEBERANG JAYA
## 37                 ZONE SENAI
## 38             ZONE SG PETANI
## 39             ZONE SHAH ALAM
## 40                  ZONE SIBU
## 41        ZONE SKUDAI PONTIAN
## 42               ZONE STAMPIN
## 43           ZONE SUBANG JAYA
## 44        ZONE TAMAN PETALING
## 45                ZONE TAMPOI
## 46                   ZONE TAR
## 47                 ZONE TASEK
## 48          ZONE TASIK AMPANG
## 49                   ZONE TDI
## 50           ZONE TELUK INTAN
## 51    ZONE TERENGGANU SELATAN
## 52              ZONE TERUNTUM
```

- Trouble Ticket Data Dictionary

```
result <-  tbl(conn, sql("select * from nova_trouble_ticket where zone <> 'null' limit 1"))
as.data.frame(apply(as.data.frame(result),2,class))
```

```
##                         apply(as.data.frame(result), 2, class)
## tt_row_id                                           character
## tt_num                                              character
## tt_type                                             character
## tt_sub_type                                         character
## status                                              character
## severity                                            character
## important_message                                   character
## appointment_flag                                    character
## nova_account_name                                   character
## nova_subscriber_num                                 character
## nova_account_num                                    character
## package_row_id                                      character
## created_by                                          character
## category                                            character
## symptom_error_code                                  character
## priority                                            character
## product                                             character
## sub_product                                         character
## package_name                                        character
## network_tt_id                                       character
## swap_order_num                                      character
## cause_category                                      character
## cause_code                                          character
## resolution_code                                     character
## closure_category                                    character
## resolution_team                                     character
## service_affected                                    character
## service_order_num                                   character
## btu_type                                            character
## owner                                               character
## owner_name                                          character
## group_owner                                         character
## owner_position                                      character
## btu_platform                                        character
## dp_location                                         character
## created_date                                        character
## pending_verify_date                                 character
## closed_by                                           character
## closed_date                                         character
## source                                              character
## installed_date                                      character
## description                                         character
## repeat_ticket_count                                 character
## follow_up_ticket_count                              character
## fdp_device_name                                     character
## fdp_site_name                                      character
## olt_site_name                                      character
## exchange                                           character
## timestamp                                          character
## contact_id                                         character
## contact_name                                       character
## contact_office_phone                               character
## contact_mobile_phone                               character
## contact_home_phone                                 character
## contact_email_addr                                 character
## due_date                                           character
## part_num                                           character
## network_layer                                      character
## network_row_id                                     character
## asset_id                                           character
## ptt                                                character
## zone                                               character
## service_point_id                                   character
```

Getting the dataset from Impala

Sample dataset - Selection trouble tickets only from Zone Kepong. The SQL is define by :-

- Why Kepong zone ?

`Zone Kepong` contains very rich information especially for the textual analysis and also one of the largest composition of the cause code & the resolution code which is good for the supervised learning.

Rules	
a.status like '%Closed%'	
network_tt_id = 'null'	Data
trouble ticket type <> PASSIVE	Trouble Ticket must related to the Active elements such as routers, switches , modem , etc. Excluding for now if related to the `3rd party` c
installed_date is NOT NULL	
created_date is NOT NULL	
closed_date is NOT NULL	
closed_date is NOT NULL	
product is NOT NULL	
sub_product is NOT NULL	
length description > 10	
rand()	
zone	S

Generated SQL :-

```
select * from nova_trouble_ticket a join active_code b on (trim(a.cause_code) = trim(b.cause_code)) join exchange_zone c ON (trim(a.exchange)=trim(
```

Datset filtering

Removing non-related fields such as trouble ticket key , trouble ticket number , trouble ticket date etc.

```
conn <- src_impaladb(dbname='nova', host='10.54.1.151')
```

```
## [1] "here:"
## [1] FALSE
```

```
result <-  tbl(conn, sql("select a.tt_row_id,a.tt_num,a.tt_type,a.tt_sub_type,a.status,a.severity,a.important_message,a.appointment_flag,a.nova_acc
a.fdp_site_name,a.olt_site_name,a.exchange,a.`timestamp`,a.contact_id,a.contact_name,a.contact_office_phone,a.contact_mobile_phone,a.contact_home_p

result <- as.data.frame(result)
```

Close the connection from Impala

```
x <- conn$con
class(x) <- c('JDBCConnection')
dbDisconnect(x)
```

```
## [1] TRUE
```

Save the class as the data.frame

```
df <- as.data.frame(result)
df$contact_name <- NULL
df$contact_home_phone <- NULL
df$contact_email_addr <- NULL
df$contact_office_phone <- NULL
df$contact_mobile_phone <- NULL
df$`tt_row_id` <- NULL
df$`tt_num` <- NULL
df$tt_type <- NULL
```

```
df$`created_date` <- NULL
df$`closed_date` <- NULL
df$`installed_date` <- NULL
df$timestamp <- NULL
df$service_point_id <- NULL
df$contact_id <- NULL
df$owner_position <- NULL
df$tt_sub_type <- NULL
df$severity <- NULL
df$status <- NULL
df$important_message <- NULL
df$network_tt_id <- NULL
df$swap_order_num <- NULL
df$appointment_flag <- NULL
df$nova_account_name <- NULL
df$nova_subscriber_num <- NULL
df$nova_account_num <- NULL
df$repeat_ticket_count <- NULL
df$follow_up_ticket_count <- NULL
df$service_order_num <- NULL
df$source <- NULL
df$owner_name <- NULL
df$description <- NULL
df$due_date <- NULL
df$part_num <- NULL
df$zone <- NULL
df$ptt <- NULL
df$asset_id <- NULL
df$network_layer <- NULL
df$network_row_id <- NULL
df$pending_verify_date <- NULL
df$package_row_id <- NULL
df$priority <- NULL
summary(df)
```

```
##   created_by          category         symptom_error_code
##  Length:100         Length:100         Length:100        
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##    product          sub_product        package_name      
##  Length:100         Length:100         Length:100        
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##  cause_category      cause_code        resolution_code   
##  Length:100         Length:100         Length:100        
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##  closure_category   resolution_team    service_affected  
##  Length:100         Length:100         Length:100        
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##    btu_type            owner           group_owner       
##  Length:100         Length:100         Length:100        
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##  btu_platform       dp_location         closed_by        
##  Length:100         Length:100         Length:100        
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##  fdp_device_name    fdp_site_name      olt_site_name     
##  Length:100         Length:100         Length:100        
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##    exchange          zone_name           district        
##  Length:100         Length:100         Length:100        
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##     state              region         
##  Length:100         Length:100        
##  Class :character   Class :character  
##  Mode  :character   Mode  :character  
```

Looping the columns name and rename it to [column name]+1 as the factor name

```
for(i in names(df)){

  num <- as.numeric(as.factor(df[,i]))-1
  df <- cbind(df,num)
  names(df)[names(df)=="num"] <- paste(names(df[i]),"_factor",sep = "")
  print(paste(names(df[i]),"1",sep = ""))
}
```

```
## [1] "created_by1"
## [1] "category1"
## [1] "symptom_error_code1"
## [1] "product1"
## [1] "sub_product1"
## [1] "package_name1"
## [1] "cause_category1"
## [1] "cause_code1"
## [1] "resolution_code1"
## [1] "closure_category1"
## [1] "resolution_team1"
## [1] "service_affected1"
## [1] "btu_type1"
## [1] "owner1"
## [1] "group_owner1"
## [1] "btu_platform1"
## [1] "dp_location1"
## [1] "closed_by1"
## [1] "fdp_device_name1"
## [1] "fdp_site_name1"
## [1] "olt_site_name1"
## [1] "exchange1"
## [1] "zone_name1"
## [1] "district1"
## [1] "state1"
## [1] "region1"
```

```
df <- df[27:52]
names(df)
```

```
##  [1] "created_by_factor"          "category_factor"
##  [3] "symptom_error_code_factor" "product_factor"
##  [5] "sub_product_factor"        "package_name_factor"
##  [7] "cause_category_factor"     "cause_code_factor"
##  [9] "resolution_code_factor"    "closure_category_factor"
## [11] "resolution_team_factor"    "service_affected_factor"
## [13] "btu_type_factor"           "owner_factor"
## [15] "group_owner_factor"        "btu_platform_factor"
## [17] "dp_location_factor"        "closed_by_factor"
## [19] "fdp_device_name_factor"    "fdp_site_name_factor"
## [21] "olt_site_name_factor"      "exchange_factor"
## [23] "zone_name_factor"          "district_factor"
## [25] "state_factor"              "region_factor"
```

Remove the predictors column which might have one unique value which can leads to zero variance result

The list below is non-zero variance variables

```
df <- df[,-nearZeroVar(df)]
names(df)
```

```
##  [1] "created_by_factor"          "category_factor"
##  [3] "symptom_error_code_factor" "product_factor"
##  [5] "sub_product_factor"        "package_name_factor"
##  [7] "cause_category_factor"     "cause_code_factor"
##  [9] "resolution_code_factor"    "closure_category_factor"
## [11] "resolution_team_factor"    "service_affected_factor"
## [13] "btu_type_factor"           "owner_factor"
## [15] "group_owner_factor"        "btu_platform_factor"
## [17] "dp_location_factor"        "closed_by_factor"
## [19] "fdp_device_name_factor"    "fdp_site_name_factor"
## [21] "olt_site_name_factor"      "exchange_factor"
```

Find the correlation between the variables using Pearson.

```
correlations <- cor(df, use="pairwise.complete.obs", method="pearson")
print(correlations)
```

```
##                           created_by_factor category_factor
## created_by_factor                1.00000000    -0.072387248
## category_factor                 -0.07238725     1.000000000
## symptom_error_code_factor       -0.08267702     0.198715902
## product_factor                   0.08388301    -0.044170248
## sub_product_factor               0.16617625    -0.087225808
## package_name_factor              0.06271467    -0.104937159
## cause_category_factor            0.08414186     0.053469506
## cause_code_factor                0.06781792     0.072915126
## resolution_code_factor          -0.01756382    -0.040664360
## closure_category_factor          0.05410537     0.103440224
## resolution_team_factor           0.09611271     0.214571696
## service_affected_factor         -0.10211774    -0.294634711
## btu_type_factor                 -0.08708122    -0.014813596
## owner_factor                     0.43534980     0.046706493
## group_owner_factor               0.01388123     0.049131139
## btu_platform_factor             -0.03920239    -0.095018399
## dp_location_factor               0.01464646     0.045696501
## closed_by_factor                 0.43534980     0.046706493
## fdp_device_name_factor          -0.03085109     0.034875988
## fdp_site_name_factor            -0.03085109     0.034875988
## olt_site_name_factor            -0.01589588     0.006246004
## exchange_factor                  0.03843539     0.025839067
##                           symptom_error_code_factor product_factor
## created_by_factor                      -0.082677023     0.08388301
## category_factor                         0.198715902    -0.04417025
## symptom_error_code_factor               1.000000000    -0.10651881
## product_factor                         -0.106518810     1.00000000
## sub_product_factor                     -0.048073165     0.48755359
## package_name_factor                     0.015390256     0.77553621
## cause_category_factor                   0.158916426     0.24381425
## cause_code_factor                       0.136987911     0.06411406
## resolution_code_factor                  0.062227127     0.28855270
## closure_category_factor                 0.005429765    -0.03349907
## resolution_team_factor                  0.101226596    -0.37627774
## service_affected_factor                -0.317962939     0.23751093
## btu_type_factor                        -0.079434062     0.17131098
## owner_factor                            0.005034200    -0.02332783
## group_owner_factor                      0.106279328    -0.04467218
## btu_platform_factor                    -0.031844469     0.10889355
## dp_location_factor                      0.139383400    -0.03037303
## closed_by_factor                        0.005034200    -0.02332783
## fdp_device_name_factor                 -0.073277163    -0.02430041
## fdp_site_name_factor                   -0.073277163    -0.02430041
## olt_site_name_factor                   -0.094243710    -0.04138103
## exchange_factor                        -0.051399544    -0.01354695
##                           sub_product_factor package_name_factor
## created_by_factor                 0.16617625         0.062714675
## category_factor                  -0.08722581        -0.104937159
## symptom_error_code_factor        -0.04807316         0.015390256
## product_factor                    0.48755359         0.775536214
## sub_product_factor                1.00000000         0.555410598
## package_name_factor               0.55541060         1.000000000
## cause_category_factor             0.09546730         0.146596285
## cause_code_factor                 0.21221574         0.093513229
## resolution_code_factor            0.11958663         0.168443826
## closure_category_factor          -0.01441844        -0.061428430
## resolution_team_factor           -0.18675861        -0.295632612
## service_affected_factor           0.24294257         0.238479552
## btu_type_factor                   0.08918906         0.014639816
## owner_factor                      0.03674080         0.011394674
## group_owner_factor               -0.03216436         0.059038196
## btu_platform_factor               0.08599936        -0.003912656
## dp_location_factor                0.04786905         0.068812495
## closed_by_factor                  0.03674080         0.011394674
## fdp_device_name_factor           -0.09447411        -0.164135482
## fdp_site_name_factor             -0.09447411        -0.164135482
## olt_site_name_factor             -0.10305364        -0.202252716
## exchange_factor                  -0.10932264        -0.122840263
##                           cause_category_factor cause_code_factor
## created_by_factor                    0.08414186        0.06781792
## category_factor                      0.05346951        0.07291513
## symptom_error_code_factor            0.15891643        0.13698791
## product_factor                       0.24381425        0.06411406
## sub_product_factor                   0.09546730        0.21221574
## package_name_factor                  0.14659629        0.09351323
## cause_category_factor                1.00000000        0.25932650
## cause_code_factor                    0.25932650        1.00000000
## resolution_code_factor               0.46757026        0.15279525
## closure_category_factor              0.13686841       -0.11892871
## resolution_team_factor              -0.27080474        0.14141374
## service_affected_factor             -0.01872455        0.01142710
## btu_type_factor                      0.06395550       -0.05762476
## owner_factor                        -0.16523259       -0.09099995
## group_owner_factor                  -0.13980399       -0.12887935
## btu_platform_factor                  0.06837131       -0.07330032
## dp_location_factor                   0.02200759        0.17676213
## closed_by_factor                    -0.16523259       -0.09099995
## fdp_device_name_factor              -0.02086297       -0.28877825
## fdp_site_name_factor                -0.02086297       -0.28877825
## olt_site_name_factor                 0.01073973       -0.30538266
## exchange_factor                     -0.01147051       -0.32079033
##                           resolution_code_factor closure_category_factor
## created_by_factor                    -0.01756382             0.054105367
## category_factor                      -0.04066436             0.103440224
## symptom_error_code_factor             0.06222713             0.005429765
## product_factor                        0.28855270            -0.033499075
## sub_product_factor                    0.11958663            -0.014418442
## package_name_factor                   0.16844383            -0.061428430
## cause_category_factor                 0.46757026             0.136868408
## cause_code_factor                     0.15279525            -0.118928712
## resolution_code_factor                1.00000000             0.144926338
## closure_category_factor               0.14492634             1.000000000
## resolution_team_factor               -0.41490201            -0.054166643
## service_affected_factor              -0.01306866             0.104292947
## btu_type_factor                       0.33161471             0.054083570
## owner_factor                         -0.18849713            -0.108159548
## group_owner_factor                   -0.24600919            -0.097840859
## btu_platform_factor                   0.32934291             0.073878259
## dp_location_factor                   -0.24313781            -0.078841974
## closed_by_factor                     -0.18849713            -0.108159548
## fdp_device_name_factor                0.22182602             0.027996173
## fdp_site_name_factor                  0.22182602             0.027996173
## olt_site_name_factor                  0.25645181             0.054411002
## exchange_factor                       0.15430579             0.015553757
##                           resolution_team_factor service_affected_factor
## created_by_factor                     0.09611271             -0.10211774
## category_factor                       0.21457170             -0.29463471
## symptom_error_code_factor             0.10122660             -0.31796294
## product_factor                       -0.37627774              0.23751093
## sub_product_factor                   -0.18675861              0.24294257
## package_name_factor                  -0.29563261              0.23847955
## cause_category_factor                -0.27080474             -0.01872455
## cause_code_factor                     0.14141374              0.01142710
## resolution_code_factor               -0.41490201             -0.01306866
## closure_category_factor              -0.05416664              0.10429295
## resolution_team_factor                1.00000000             -0.13246490
## service_affected_factor              -0.13246490              1.00000000
## btu_type_factor                      -0.25822644              0.06582288
## owner_factor                          0.27837402             -0.19032833
## group_owner_factor                   -0.01349542             -0.28777200
## btu_platform_factor                  -0.21900417              0.11727936
## dp_location_factor                    0.16254032             -0.09467019
## closed_by_factor                      0.27837402             -0.19032833
## fdp_device_name_factor               -0.11517543              0.05452589
## fdp_site_name_factor                 -0.11517543              0.05452589
## olt_site_name_factor                 -0.10949318              0.04547937
## exchange_factor                      -0.03080287              0.11172718
##                           btu_type_factor owner_factor group_owner_factor
## created_by_factor              -0.08708122   0.43534980         0.01388123
## category_factor                -0.01481360   0.04670649         0.04913114
## symptom_error_code_factor      -0.07943406   0.00503420         0.10627933
## product_factor                  0.17131098  -0.02332783        -0.04467218
## sub_product_factor              0.08918906   0.03674080        -0.03216436
## package_name_factor             0.01463982   0.01139467         0.05903820
## cause_category_factor           0.06395550  -0.16523259        -0.13980399
## cause_code_factor              -0.05762476  -0.09099995        -0.12887935
## resolution_code_factor          0.33161471  -0.18849713        -0.24600919
## closure_category_factor         0.05408357  -0.10815955        -0.09784086
## resolution_team_factor         -0.25822644   0.27837402        -0.01349542
## service_affected_factor         0.06582288  -0.19032833        -0.28777200
## btu_type_factor                 1.00000000  -0.15504445        -0.12921915
## owner_factor                   -0.15504445   1.00000000         0.33030464
## group_owner_factor             -0.12921915   0.33030464         1.00000000
## btu_platform_factor             0.91632426  -0.18329232        -0.14965284
## dp_location_factor             -0.77537561   0.12603192         0.14968938
## closed_by_factor               -0.15504445   1.00000000         0.33030464
## fdp_device_name_factor          0.68689996  -0.12146031        -0.07982665
## fdp_site_name_factor            0.68689996  -0.12146031        -0.07982665
## olt_site_name_factor            0.67620677  -0.11011242        -0.09849533
## exchange_factor                 0.39386981  -0.07144686        -0.09452025
##                           btu_platform_factor dp_location_factor
## created_by_factor                 -0.039202385         0.01464646
## category_factor                   -0.095018399         0.04569650
## symptom_error_code_factor         -0.031844469         0.13938340
## product_factor                     0.108893546        -0.03037303
## sub_product_factor                 0.085999357         0.04786905
## package_name_factor               -0.003912656         0.06881250
## cause_category_factor              0.068371308         0.02200759
## cause_code_factor                 -0.073300324         0.17676213
## resolution_code_factor             0.329342909        -0.24313781
## closure_category_factor            0.073878259        -0.07884197
## resolution_team_factor            -0.219004166         0.16254032
## service_affected_factor            0.117279362        -0.09467019
## btu_type_factor                    0.916324258        -0.77537561
## owner_factor                      -0.183292319         0.12603192
## group_owner_factor                -0.149652836         0.14968938
## btu_platform_factor                1.000000000        -0.79164019
## dp_location_factor                -0.791640195         1.00000000
## closed_by_factor                  -0.183292319         0.12603192
## fdp_device_name_factor             0.668271342        -0.65681126
## fdp_site_name_factor               0.668271342        -0.65681126
## olt_site_name_factor               0.670017850        -0.62280188
## exchange_factor                    0.421064534        -0.37695976
##                           closed_by_factor fdp_device_name_factor
## created_by_factor               0.43534980            -0.03085109
## category_factor                 0.04670649             0.03487599
## symptom_error_code_factor       0.00503420            -0.07327716
## product_factor                 -0.02332783            -0.02430041
## sub_product_factor              0.03674080            -0.09447411
## package_name_factor             0.01139467            -0.16413548
## cause_category_factor          -0.16523259            -0.02086297
## cause_code_factor              -0.09099995            -0.28877825
## resolution_code_factor         -0.18849713             0.22182602
## closure_category_factor        -0.10815955             0.02799617
## resolution_team_factor          0.27837402            -0.11517543
## service_affected_factor        -0.19032833             0.05452589
## btu_type_factor                -0.15504445             0.68689996
## owner_factor                    1.00000000            -0.12146031
## group_owner_factor              0.33030464            -0.07982665
## btu_platform_factor            -0.18329232             0.66827134
## dp_location_factor              0.12603192            -0.65681126
## closed_by_factor                1.00000000            -0.12146031
## fdp_device_name_factor         -0.12146031             1.00000000
## fdp_site_name_factor           -0.12146031             1.00000000
## olt_site_name_factor           -0.11011242             0.97273379
## exchange_factor                -0.07144686             0.88203830
##                           fdp_site_name_factor olt_site_name_factor
## created_by_factor                  -0.03085109         -0.015895876
## category_factor                     0.03487599          0.006246004
## symptom_error_code_factor          -0.07327716         -0.094243710
## product_factor                     -0.02430041         -0.041381032
## sub_product_factor                 -0.09447411         -0.103053644
## package_name_factor                -0.16413548         -0.202252716
## cause_category_factor              -0.02086297          0.010739731
## cause_code_factor                  -0.28877825         -0.305382658
## resolution_code_factor              0.22182602          0.256451813
## closure_category_factor             0.02799617          0.054411002
## resolution_team_factor             -0.11517543         -0.109493184
## service_affected_factor             0.05452589          0.045479369
## btu_type_factor                     0.68689996          0.676206768
## owner_factor                       -0.12146031         -0.110112425
## group_owner_factor                 -0.07982665         -0.098495332
## btu_platform_factor                 0.66827134          0.670017850
## dp_location_factor                 -0.65681126         -0.622801882
## closed_by_factor                   -0.12146031         -0.110112425
## fdp_device_name_factor              1.00000000          0.972733794
## fdp_site_name_factor                1.00000000          0.972733794
## olt_site_name_factor                0.97273379          1.000000000
## exchange_factor                     0.88203830          0.868256058
##                           exchange_factor
## created_by_factor              0.03843539
## category_factor                0.02583907
## symptom_error_code_factor     -0.05139954
## product_factor                -0.01354695
## sub_product_factor            -0.10932264
## package_name_factor           -0.12284026
## cause_category_factor         -0.01147051
## cause_code_factor             -0.32079033
## resolution_code_factor         0.15430579
## closure_category_factor        0.01555376
## resolution_team_factor        -0.03080287
## service_affected_factor        0.11172718
## btu_type_factor                0.39386981
## owner_factor                  -0.07144686
## group_owner_factor            -0.09452025
## btu_platform_factor            0.42106453
## dp_location_factor            -0.37695976
## closed_by_factor              -0.07144686
## fdp_device_name_factor         0.88203830
## fdp_site_name_factor           0.88203830
## olt_site_name_factor           0.86825606
## exchange_factor                1.00000000
```

Find the highest correlated variables.

Rules	Description
- +.70 or higher	Very strong relationship
- +.40 to +.69	Strong positive relationship
- +.30 to +.39	Moderate relationship
- +.20 to +.29	weak relationship
- +.01 to +.19	No or negligible relationship

```
# Choose 0.7 Very strong relationship
highlyCorrelated <- findCorrelation(correlations, 0.7 ,verbose = FALSE,names = TRUE)
highlyCorrelated
```

```
## [1] "olt_site_name_factor"   "fdp_device_name_factor"
## [3] "fdp_site_name_factor"   "btu_platform_factor"
## [5] "btu_type_factor"        "owner_factor"
## [7] "package_name_factor"
```

Summary of the correlated variables.

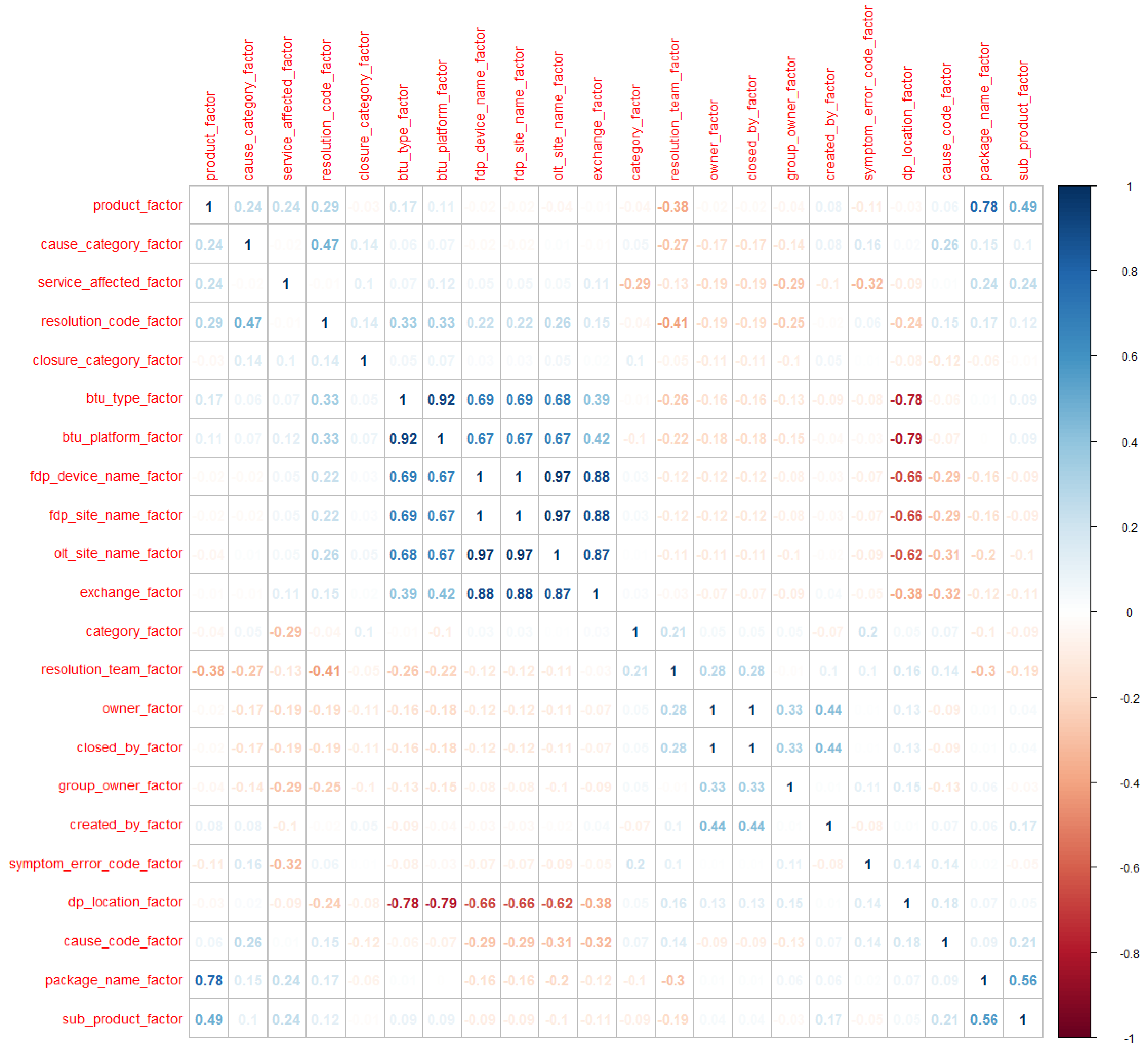
```
summary(correlations)
```

```
##  created_by_factor  category_factor    symptom_error_code_factor
##  Min.   :-0.10212   Min.   :-0.29463   Min.   :-0.317963
##  1st Qu.:-0.03085   1st Qu.:-0.04329   1st Qu.:-0.073277
##  Median : 0.02654   Median : 0.03488   Median : 0.005034
##  Mean   : 0.09427   Mean   : 0.05361   Mean   : 0.044360
##  3rd Qu.: 0.08408   3rd Qu.: 0.05238   3rd Qu.: 0.105016
##  Max.   : 1.00000   Max.   : 1.00000   Max.   : 1.000000
##  product_factor     sub_product_factor package_name_factor
##  Min.   :-0.37628   Min.   :-0.18676   Min.   :-0.29563
##  1st Qu.:-0.03272   1st Qu.:-0.07744   1st Qu.:-0.09406
##  Median :-0.01844   Median : 0.04230   Median : 0.01502
##  Mean   : 0.12161   Mean   : 0.10936   Mean   : 0.09555
##  3rd Qu.: 0.22096   3rd Qu.: 0.15453   3rd Qu.: 0.13333
##  Max.   : 1.00000   Max.   : 1.00000   Max.   : 1.00000
##  cause_category_factor cause_code_factor  resolution_code_factor
##  Min.   :-0.27080      Min.   :-0.32079   Min.   :-0.41490
##  1st Qu.:-0.02033      1st Qu.:-0.11195   1st Qu.:-0.03489
##  Median : 0.05871      Median : 0.03777   Median : 0.14886
##  Mean   : 0.09083      Mean   : 0.02840   Mean   : 0.11669
##  3rd Qu.: 0.14416      3rd Qu.: 0.14031   3rd Qu.: 0.24780
##  Max.   : 1.00000      Max.   : 1.00000   Max.   : 1.00000
##  closure_category_factor resolution_team_factor service_affected_factor
##  Min.   :-0.11893        Min.   :-0.41490       Min.   :-0.31796
##  1st Qu.:-0.05961        1st Qu.:-0.21094       1st Qu.:-0.12488
##  Median : 0.02177        Median :-0.08183       Median : 0.02845
##  Mean   : 0.05125        Mean   :-0.01453       Mean   : 0.02918
##  3rd Qu.: 0.06901        3rd Qu.: 0.13137       3rd Qu.: 0.10987
##  Max.   : 1.00000        Max.   : 1.00000       Max.   : 1.00000
##  btu_type_factor     owner_factor      group_owner_factor
##  Min.   :-0.77538   Min.   :-0.19033   Min.   :-0.28777
##  1st Qu.:-0.08517   1st Qu.:-0.12146   1st Qu.:-0.12128
##  Median : 0.05902   Median :-0.04739   Median :-0.06225
##  Mean   : 0.15632   Mean   : 0.07912   Mean   : 0.01893
##  3rd Qu.: 0.37831   3rd Qu.: 0.10620   3rd Qu.: 0.05656
##  Max.   : 1.00000   Max.   : 1.00000   Max.   : 1.00000
##  btu_platform_factor dp_location_factor closed_by_factor
##  Min.   :-0.79164    Min.   :-0.79164   Min.   :-0.19033
##  1st Qu.:-0.08959    1st Qu.:-0.34350   1st Qu.:-0.12146
##  Median : 0.07112    Median : 0.01833   Median :-0.04739
##  Mean   : 0.15262    Mean   :-0.10218   Mean   : 0.07912
##  3rd Qu.: 0.39813    3rd Qu.: 0.12603   3rd Qu.: 0.10620
##  Max.   : 1.00000    Max.   : 1.00000   Max.   : 1.00000
##  fdp_device_name_factor fdp_site_name_factor olt_site_name_factor
##  Min.   :-0.65681       Min.   :-0.65681     Min.   :-0.622802
##  1st Qu.:-0.11000       1st Qu.:-0.11000     1st Qu.:-0.107883
##  Median :-0.02258       Median :-0.02258     Median :-0.004825
##  Mean   : 0.17081       Mean   : 0.17081     Mean   : 0.169093
##  3rd Qu.: 0.55666       3rd Qu.: 0.55666     3rd Qu.: 0.566626
##  Max.   : 1.00000       Max.   : 1.00000     Max.   : 1.000000
##  exchange_factor
##  Min.   :-0.376960
##  1st Qu.:-0.071447
##  Median : 0.002042
##  Mean   : 0.159936
##  3rd Qu.: 0.333979
##  Max.   : 1.000000
```

Plot correlated variables.

```
png(height=1200, width=1200, pointsize=15, file="corrplot.png")
corrplot(correlations, method = "number",tl.cex = 0.9 ,addCoef.col="grey", order = "AOE")
dev.off()
```

```
## pdf
##   2
```



Feature selection process to confirm which variable does become the independent and resolution code is the dependent variable via GBM (Stochastic Gradient Boosting).

List of other available model - http://topepo.github.io/caret/modelList.html

```
set.seed(777)
suppressWarnings(suppressMessages(library(mlbench)))
control <- trainControl(method = "repeatedcv", number = 10, repeats = 3)
model <-
train(
resolution_code_factor ~ ., data = df, method = "gbm", preProcess = "scale", trControl =
control , verbose = FALSE
)
```

```
## Loading required package: gbm
```

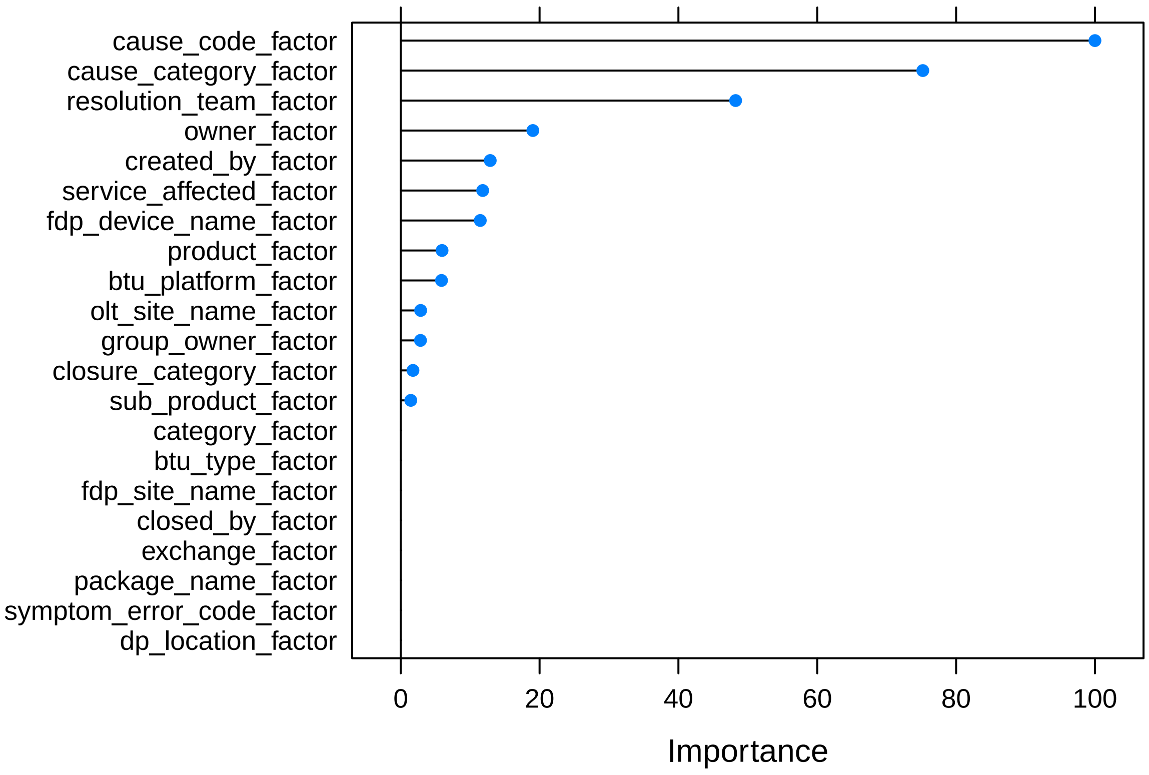
```
## Warning: package 'gbm' was built under R version 3.2.2
```

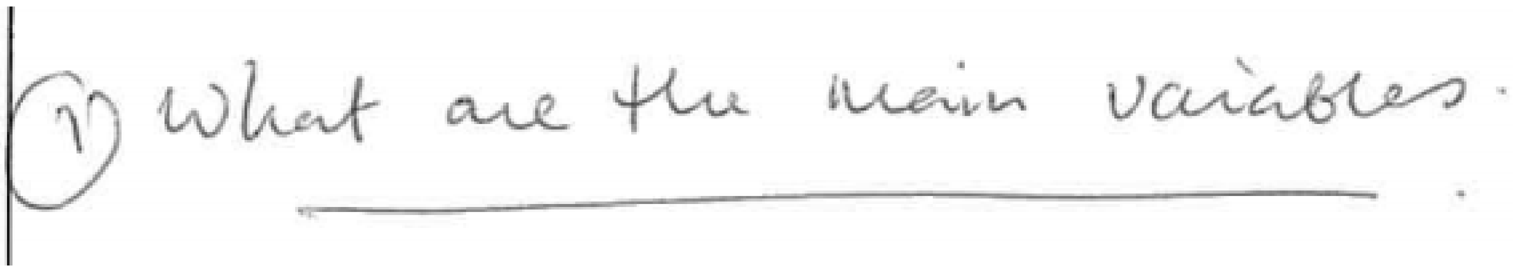
```
## Loading required package: survival
##
## Attaching package: 'survival'
##
## The following object is masked from 'package:caret':
##
##     cluster
##
## Loading required package: splines
## Loading required package: parallel
## Loaded gbm 2.1.1
## Loading required package: plyr
## -------------------------------------------------------------------------
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
## -------------------------------------------------------------------------
##
## Attaching package: 'plyr'
##
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
importance <- varImp(model, scale = TRUE)
print(importance)
```

```
## gbm variable importance
##
##   only 20 most important variables shown (out of 21)
##
##                          Overall
## cause_code_factor        100.000
## cause_category_factor     75.203
## resolution_team_factor    48.239
## owner_factor              19.032
## created_by_factor         12.890
## service_affected_factor   11.806
## fdp_device_name_factor    11.458
## product_factor             5.951
## btu_platform_factor        5.872
## olt_site_name_factor       2.868
## group_owner_factor         2.845
## closure_category_factor    1.762
## sub_product_factor         1.444
## btu_type_factor            0.000
## closed_by_factor           0.000
## exchange_factor            0.000
## dp_location_factor         0.000
## category_factor            0.000
## package_name_factor        0.000
## fdp_site_name_factor       0.000
```

```
plot(importance)
```




(v) What are the main variables.

The main variables / factors are :-

- cause_code_factor

- resolution_team_factor
- cause_category_factor

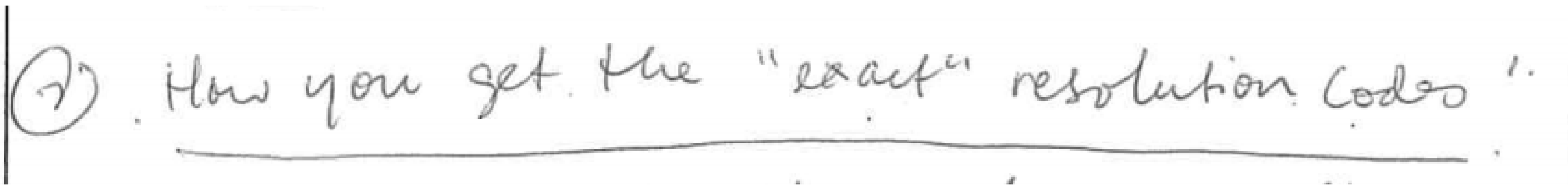
- fdp_device_name_factor
- owner_factor

- created_by_factor

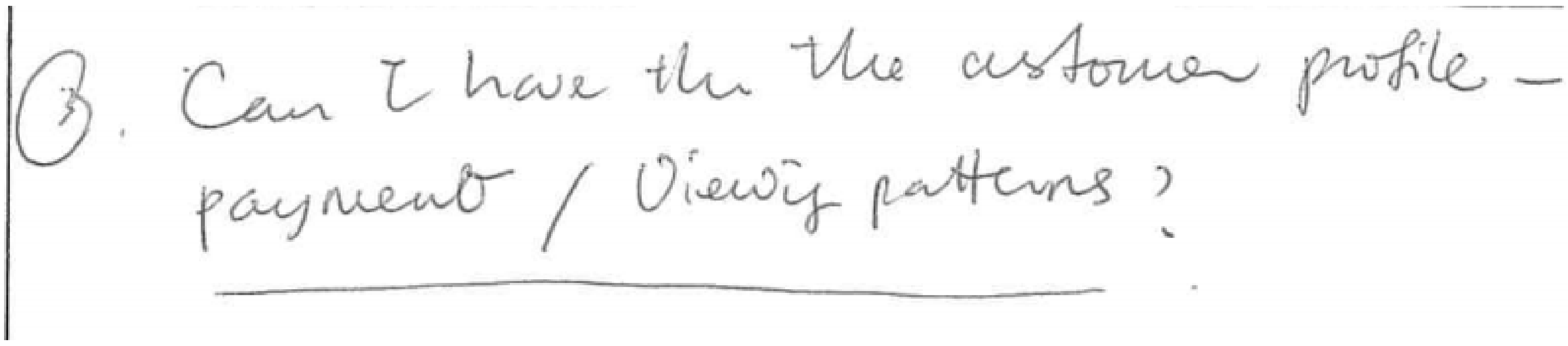
- service_affected_factor
- dp_location_factor

- btu_type_factor

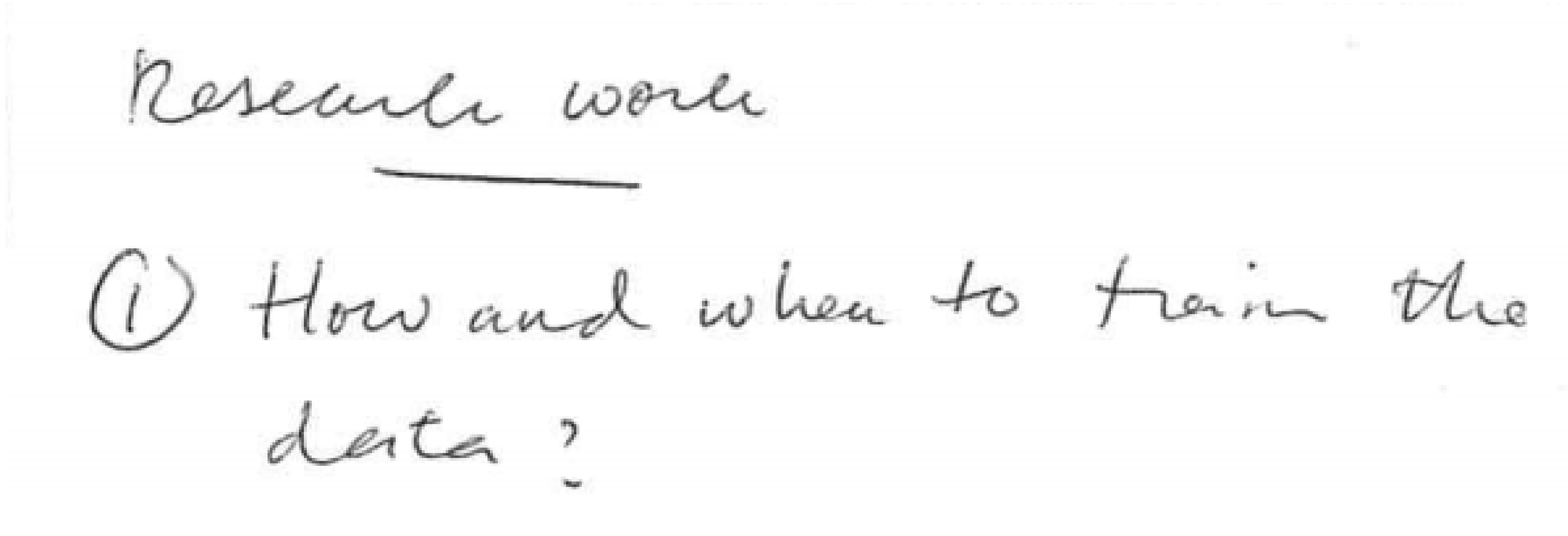
as listed from the importance plot


(2). How you get the "exact" resolution Codes !.

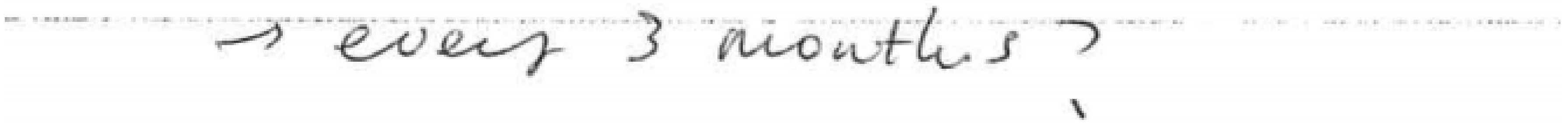
sss


(3). Can I have the the customer profile –
payment / Viewing patterns ?

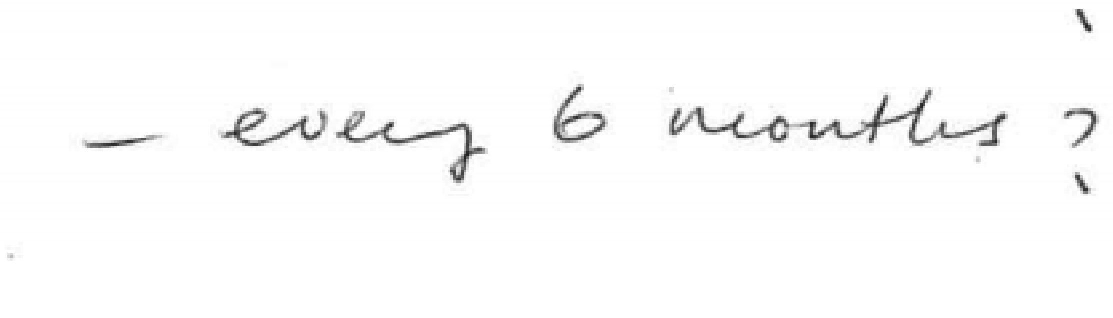
sss


Research work

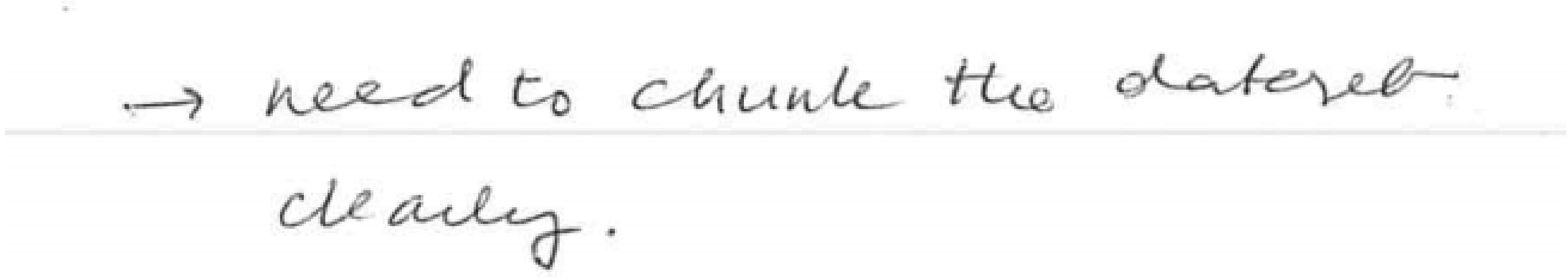
(1) How and when to train the
data ?


→ every 3 months ?

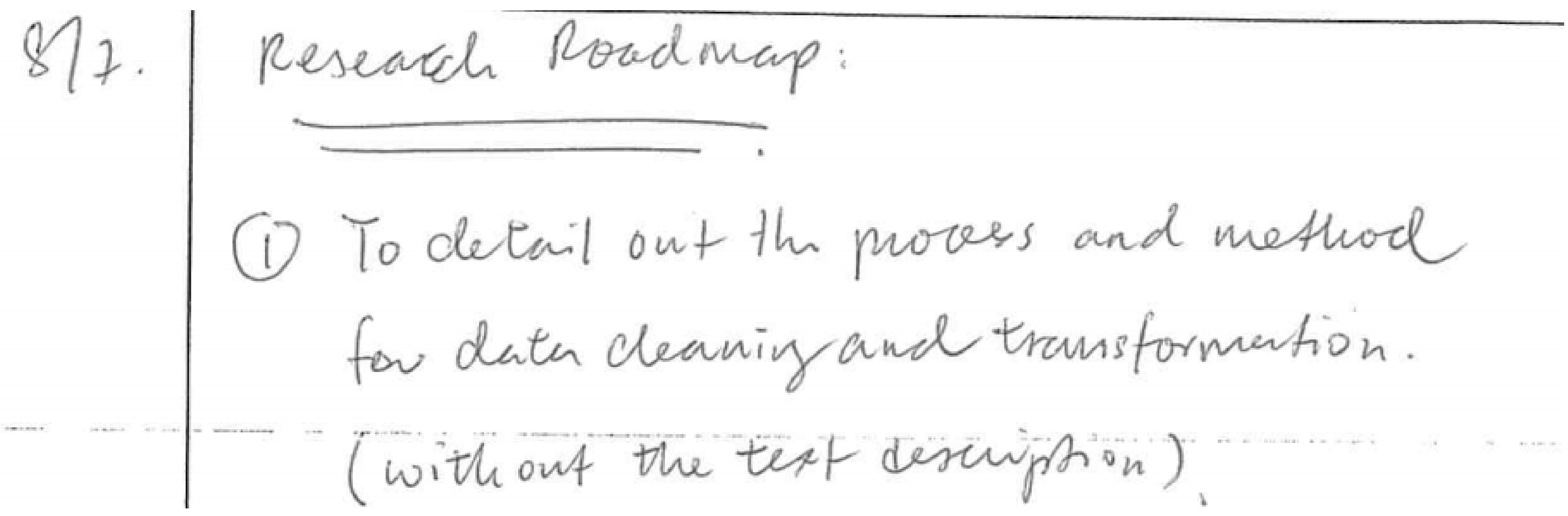
sss


– every 6 months ?

sss


→ need to chunk the dataset
clearly.

sss



8/7. Research Roadmap:

① To detail out the process and method for data cleaning and transformation. (without the text description).

1