SEB task for AML Quantitative Analyst

Warsaw, Poland

Chukwuma Franklin, Ofoegbu

6 March 2022

Contents

Chapter 1

Data Manipulation Task

1.1 Importing the Datasets

The Datasets were be imported, into the software to be used. In this task, R was used. The necessary libraries were also called, after they had been installed. The second sheet is chosen when importing the additional_info.xlsx dataset.

```r
# loading the required libraries
library("readxl")
library(dplyr)
library(ggplot2)
library(VIM)
library(psych)
library(lubridate)
library(ggplot2)
library(plyr)
library(tidyverse)
```

loading the libraries

```r
# importing the data sets

alerts <- read_excel("alerts.xlsx")

additional_info <- read_excel("additional_info.xlsx", sheet = 2)

View(alerts)
View(additional_info)
```

viewing and importing the data

1.2 Cleaning the Datasets

The first column of the alerts dataset was dropped, since it is nameless. Then for the additional_info dataset, the column names were renamed properly.

```r
# Data Cleaning and Manipulation for alerts
alerts1 <- alerts[,-1]
```

dropping the first column

The columns dataset additional_info is renamed, due to the spacing in the name and also to match alerts

```{r}
#renaming columns in additional_info
colnames(additional_info) <- c('IndustryCode','RiskScore','Segment')
```

Renaming the columns in additional_info data

1.3 Merging the Datasets

The datasets alerts and additional_info are joined (left join). They are joined by the IndustryCode.

```{r}
# merging alerts1 with additional_info on Industry Code
alerts2 <- left_join(alerts1, additional_info) %>%  #merge the preliminary data frame with selected regions
  select(AlertState, AlertType, CaseClosed, CaseOpen, CaseReported, CaseState, CusRiskCategory, DateClosed, DateCreated, IndustryCode, intID, PE
RiskScore, Segment)

alerts2
```

Datasets alerts and additional_info are joined by IndustryCode

1.4 Additional Data cleaning - Date and Time; Renaming Categorical Variables

1.4.1 Date and Time

The dates and time columns namely: CaseClosed, CaseOpen, CaseReported, DateClosed, DateCreated will be split into separate time and date columns; and the original columns will be dropped.

```r
# Splitting the CaseClosed, CaseOpen, DateClosed, DateCreated, CaseR
alerts3 <- alerts2

#CaseClosed
alerts3$dCaseClosed <- as.Date(alerts3$CaseClosed)
alerts3$tCaseClosed <- hms(substr(alerts3$CaseClosed, 12, 19))

#CaseOpen
alerts3$dCaseOpen <- as.Date(alerts3$CaseOpen)
alerts3$tCaseOpen <- hms(substr(alerts3$CaseOpen, 12, 19))


#DateClosed
alerts3$dDateClosed <- ymd(substr(alerts3$DateClosed, 1,10))
alerts3$tDateClosed <- hms(substr(alerts3$DateClosed, 12, 19))

#DateCreated
alerts3$dDateCreated <- as.Date(alerts3$DateCreated)
alerts3$tDateCreated <- hms(substr(alerts3$DateCreated, 12, 19))

#CaseReported
alerts3$dCaseReported <- ymd(substr(alerts3$CaseReported, 1,10))
alerts3$tCaseReported <- hms(substr(alerts3$CaseReported, 12, 19))
```

new dates and time columns created

The new columns created for dates will be dCaseClosed, dCaseOpen, dCaseReported, dDateClosed, dDateCreated; while the new columns for the time will be tCaseClosed, tCaseOpen, tCaseReported,t DateClosed, tDateCreated.

```r
#dropping columns CaseClosed, CaseOpen, DateClosed, DateCreated, CaseReported
alerts4 <- alerts3 %>% select(-c(CaseClosed,
                                 CaseOpen, CaseReported, DateClosed, DateCreated))
```

These columns are now dropped: CaseClosed, CaseOpen, CaseReported, DateClosed, DateCreated.

1.4.2 Renaming Categorical Variables

The Categorical Variables are now renamed. They are renamed as follows: CusRiskCategory "Lower Risk" = low, "Medium Risk" = medium, "Higher Risk" = high , "Not Specified" = ns, "NULL" = null

AlertState "Data Created" = created , "Closed - Not Suspicious" = closed_ns , "Closed - Not Investigated" = closed_ni , "Closed - Processed externally" = closed_pe, "Closed - Not Investigated Data Quality" = closed_dq , "Level 2 escalation" = esc_l2 , "Unassigned" = unasgn, "Under Investigation" = u_inv , "Assigned to Investigate" = a_inv

AlertType "New Destinations with high turnover" = new_dest, "Existing Accounts" = exg_a ,

"Unusual behaviour" = unu_b , "Check Countries List" = ccl , "Awakening Account" = awak_a , "Credit Cards" = cc, "repayement of funds" = rpyt_f , "Cash" = cash, "Listed High Risk Banks" = hr_banks, "Close Monitoring" = close_m, "Recurring In-Out scenario" = recur, "International Transfers" = intl_x , "Unusual Cash Behaviour" = uncash_b, "PEP Monitoring" = pep_m

Segment high risk = high, "staff intensive small company" = sisc

CaseState "Closed" = closed , "Report Confirmed" = reported, "NULL" = null

PEP "N" = no, "Y" = yes, "NULL" = NA

1.5 Results - summary tables

The results gotten from the tables will be shown be below:

```
 AlertState         AlertType          CaseState          CusRiskCategory    IndustryCode          intID             PEP                RiskScore
Length:10177       Length:10177       Length:10177       Length:10177       Length:10177       Min.   :     3   Length:10177       Min.   : 20.0
Class :character   Class :character   Class :character   Class :character   Class :character   1st Qu.:  2784   Class :character   1st Qu.: 20.0
Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median : 14320   Mode  :character   Median :100.0
                                                                            Mean   : 31246                      Mean   : 77.6
                                                                            3rd Qu.: 19287                      3rd Qu.:100.0
                                                                            Max.   :203767                      Max.   :100.0
                                                                                                                NA's   :10097

   Segment             Type           dCaseClosed          tCaseClosed                      dCaseOpen           tCaseOpen
Length:10177       Length:10177       Min.   :2010-12-07   Min.   :5H 25M 13S               Min.   :2010-12-07   Min.   :5H 22M 29S
Class :character   Class :character   1st Qu.:2013-12-04   1st Qu.:10H 0M 59S               1st Qu.:2014-05-12   1st Qu.:9H 45M 48S
Mode  :character   Mode  :character   Median :2020-01-27   Median :12H 47M 1S               Median :2017-10-20   Median :12H 20M 1.5S
                                      Mean   :2017-05-27   Mean   :12H 25M 26.62234042552885S Mean   :2017-05-13   Mean   :12H 38M 55.71484375S
                                      3rd Qu.:2021-03-16   3rd Qu.:14H 45M 56S              3rd Qu.:2020-10-23   3rd Qu.:15H 23M 53.5S
                                      Max.   :2021-09-02   Max.   :16H 29M 49S              Max.   :2021-09-06   Max.   :22H 15M 4S
                                      NA's   :9989         NA's   :9989                     NA's   :9921         NA's   :9921
 dDateClosed          tDateClosed                      dDateCreated         tDateCreated dCaseReported        tCaseReported
Min.   :2008-04-29   Min.   :5M 51S                   Min.   :2008-04-17   Min.   :0S   Min.   :2017-10-20   Min.   :7H 40M 39S
1st Qu.:2015-03-02   1st Qu.:8H 59M 35S               1st Qu.:2015-01-01   1st Qu.:0S   1st Qu.:2017-12-12   1st Qu.:11H 35M 35S
Median :2017-05-08   Median :11H 2M 47S               Median :2017-02-04   Median :0S   Median :2018-09-19   Median :14H 46M 25S
Mean   :2016-12-28   Mean   :11H 21M 26.7163170633285S Mean   :2016-12-02   Mean   :0S   Mean   :2019-01-02   Mean   :14H 38M 36.191176470587
3rd Qu.:2019-10-18   3rd Qu.:13H 44M 54.5S            3rd Qu.:2019-10-23   3rd Qu.:0S   3rd Qu.:2019-07-24   3rd Qu.:16H 39M 33.75S
Max.   :2021-10-13   Max.   :23H 49M 13S              Max.   :2021-10-13   Max.   :0S   Max.   :2021-09-06   Max.   :22H 17M 13S
NA's   :261          NA's   :261                                                        NA's   :10109        NA's   :10109
```

new dates and time columns created

CusRiskCategory <chr>	freq <int>
medium	6986
low	1326
high	1183
ns	628
NA	54

5 rows

AlertState <chr>	freq <int>
closed_ns	9078
closed_ni	509
closed_pe	302
created	260
unasgn	12
closed_dq	6
u_inv	6
a_inv	3
esc_l2	1

9 rows

PEP <chr>	freq <int>
NA	7275
no	2784
yes	118

3 rows

Type <chr>	freq <int>
lcfi	6337
pb	3840

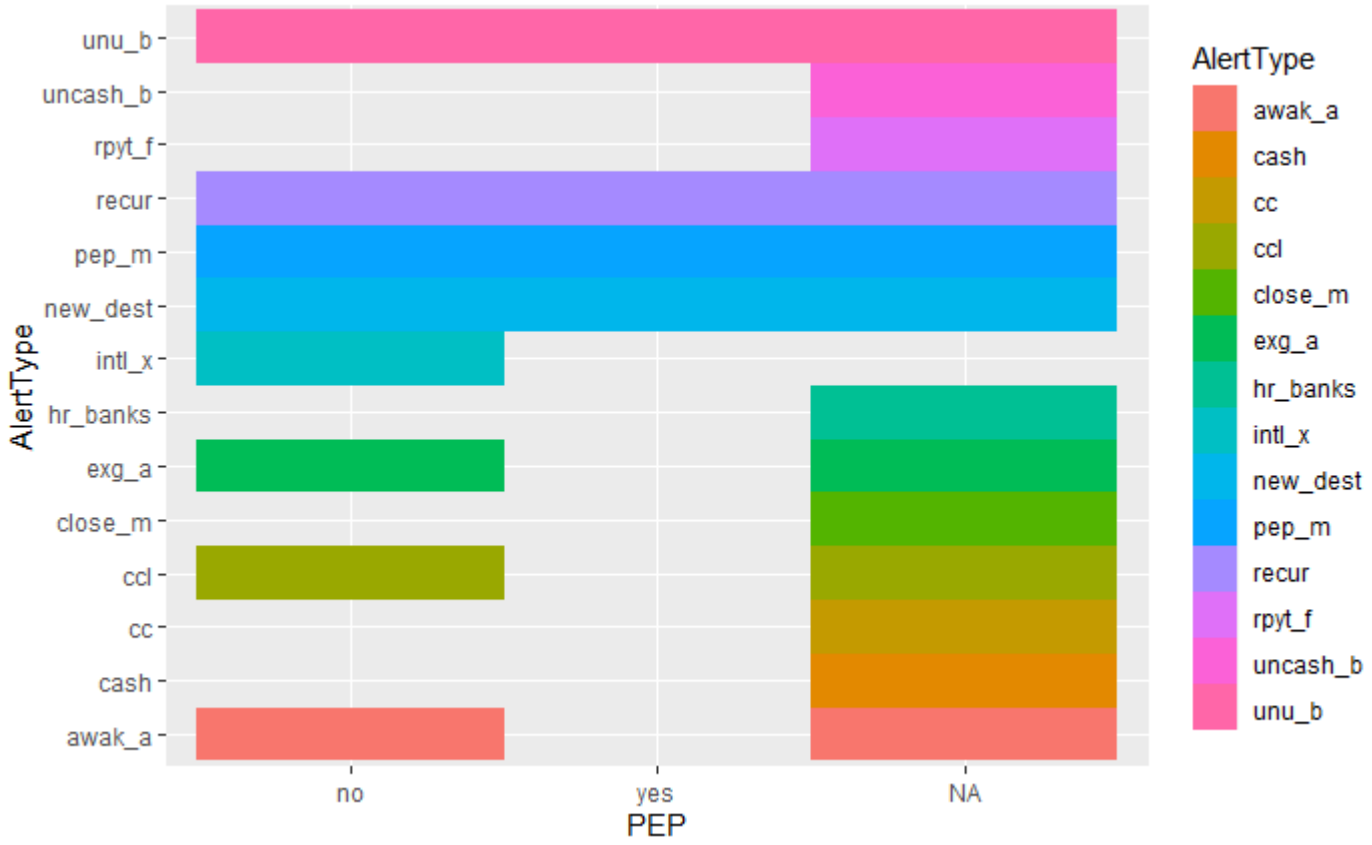
2 rows

Segment <chr>	freq <int>
NA	10097
high	58
sisc	22

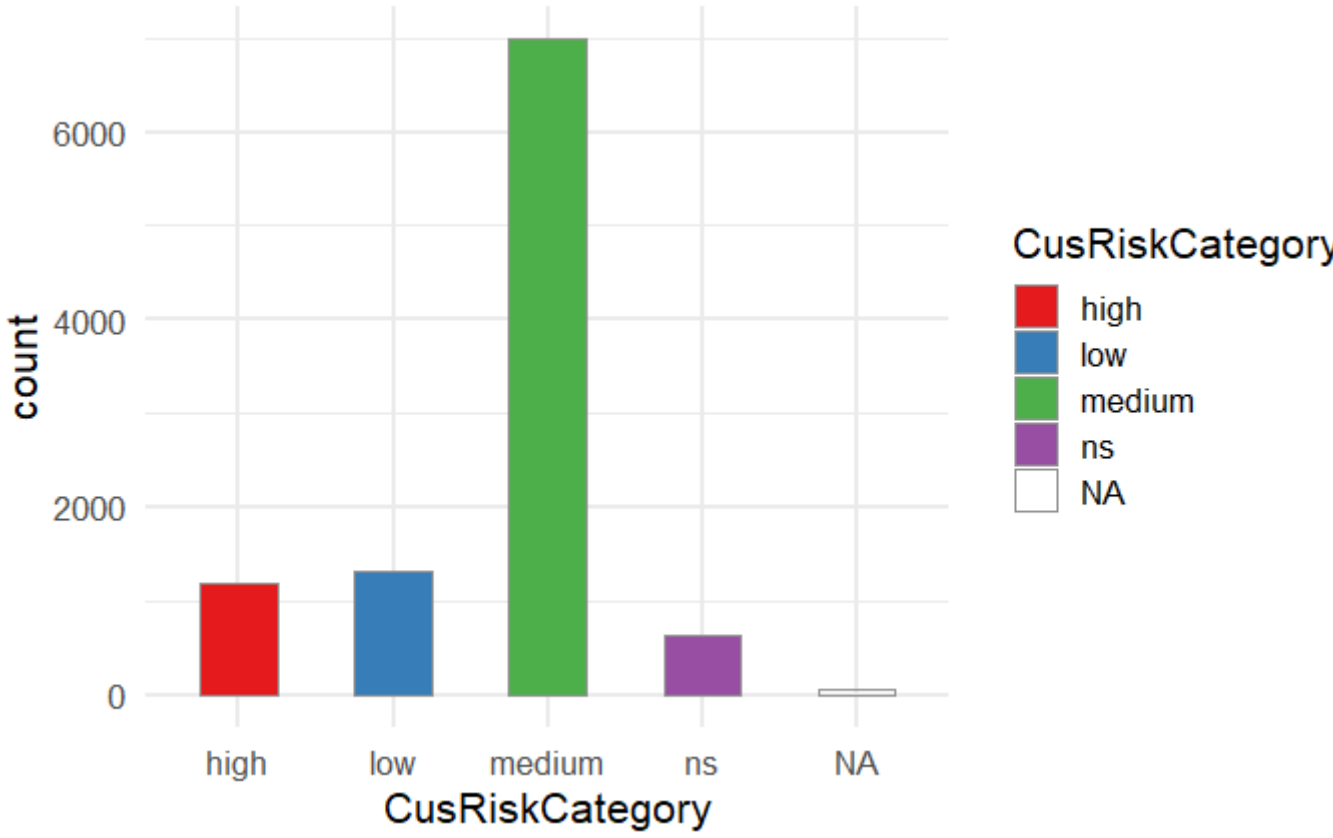
3 rows

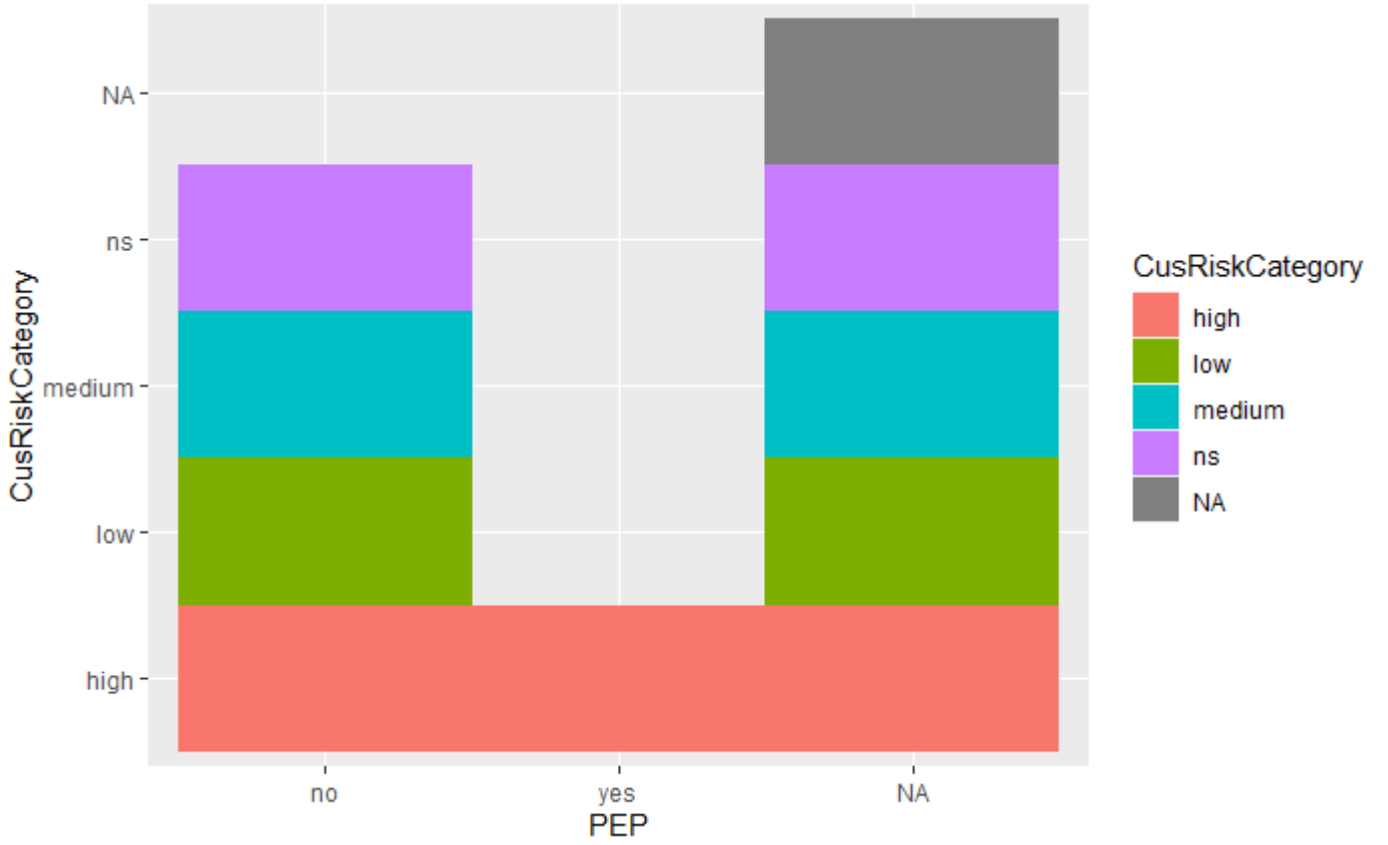
Chapter 2

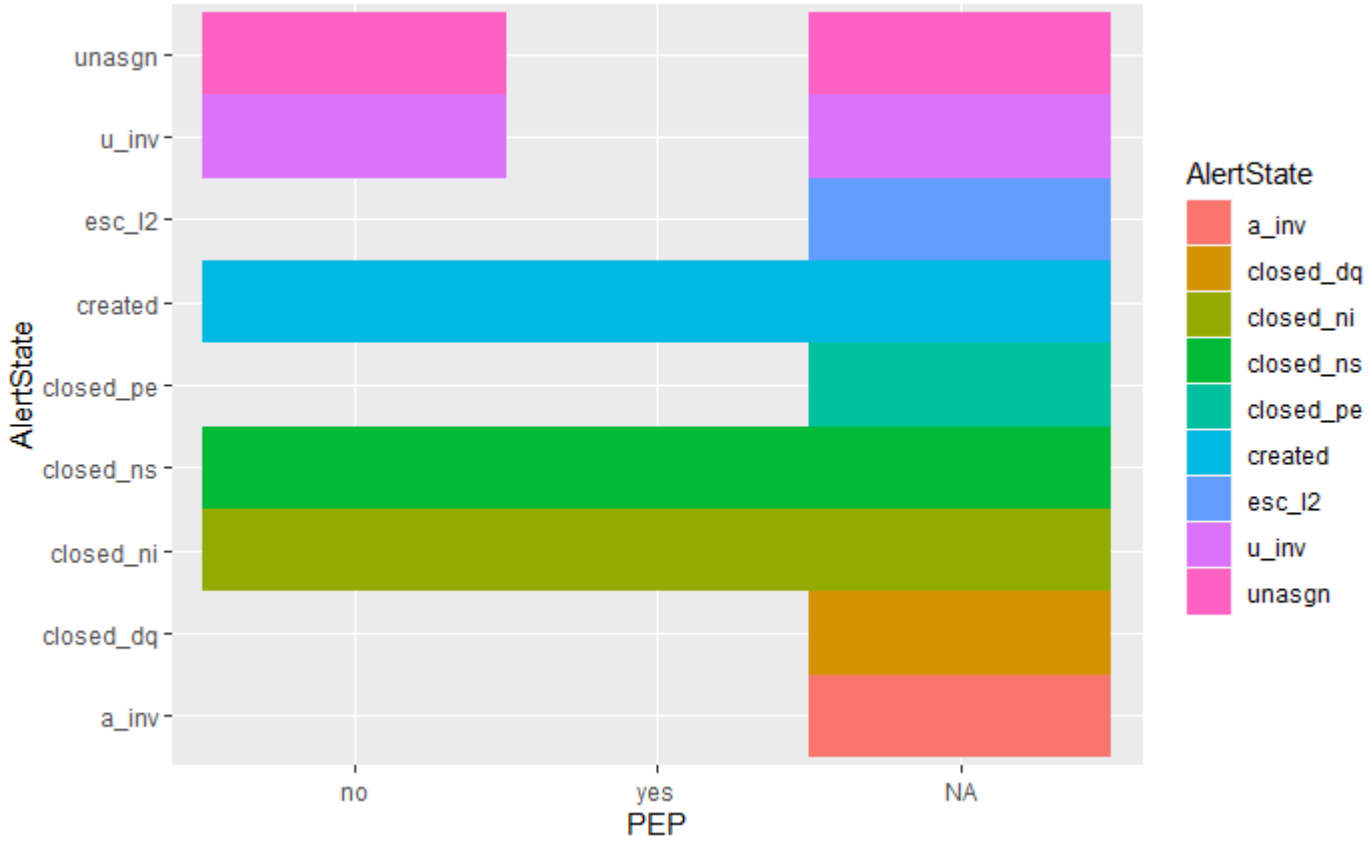
Data Visualisation Task

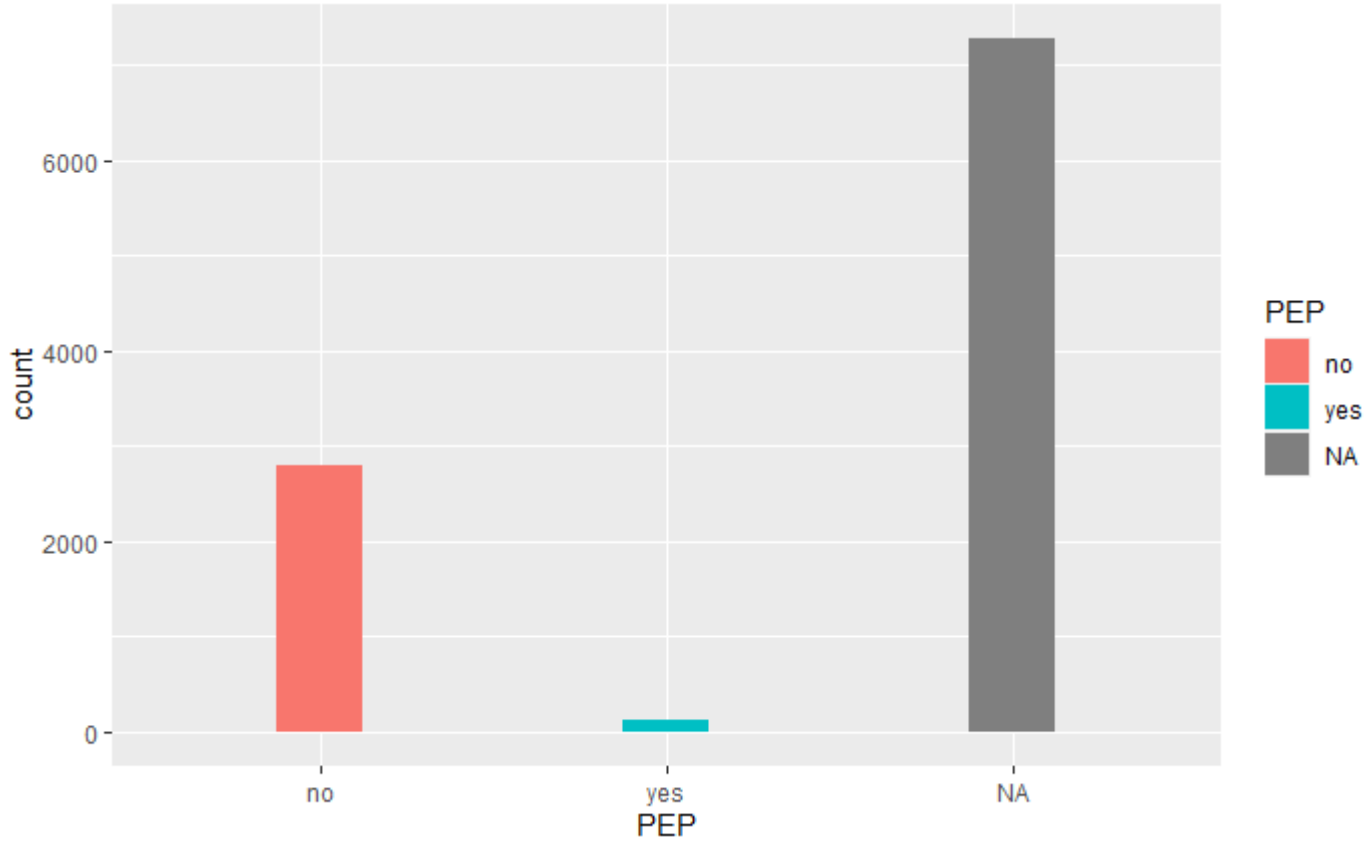


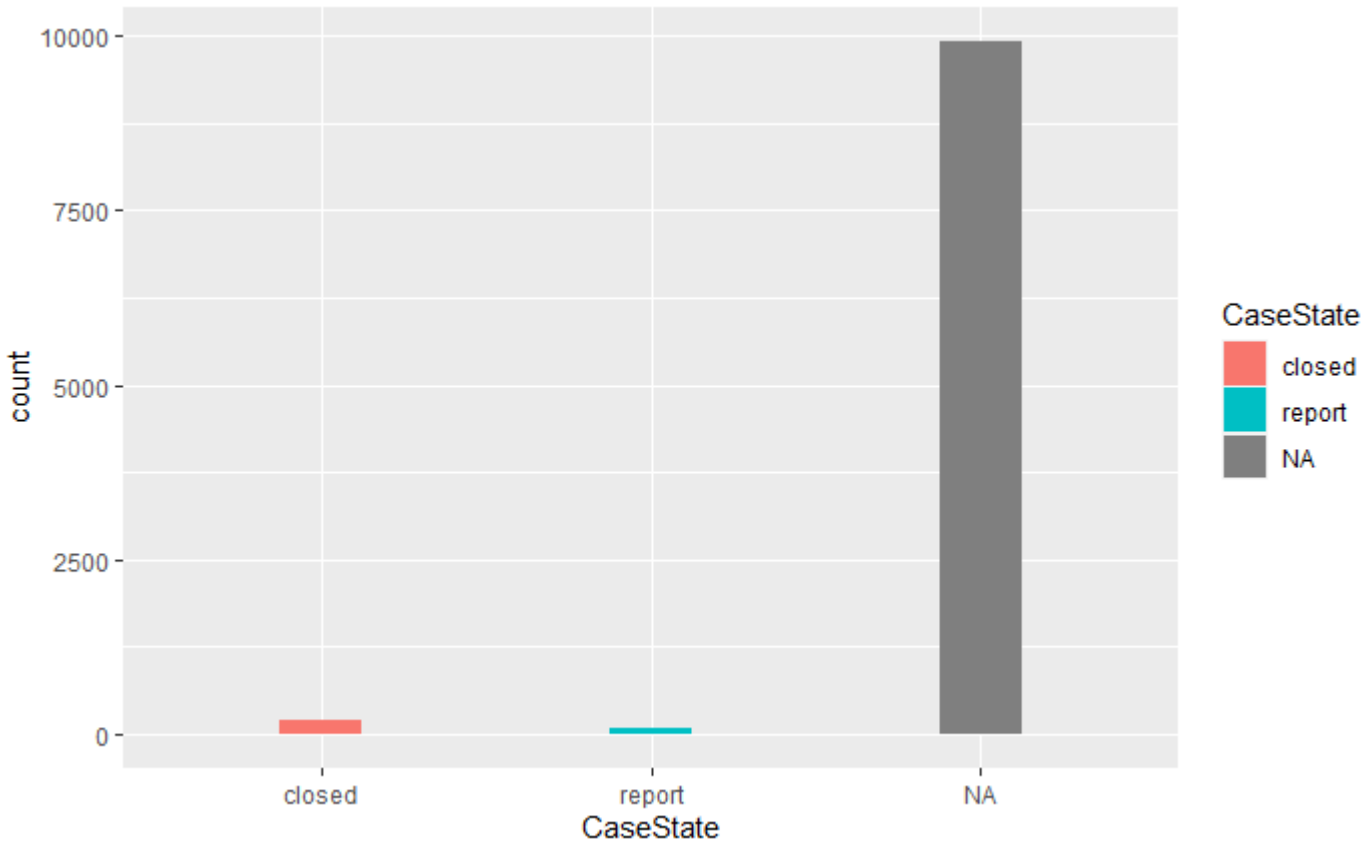
PEP are not involved in cash, unusual cash behaviour, credit cards and repayment funds


6000
4000
2000
0
count
high
low
medium
ns
NA
CusRiskCategory
CusRiskCategory
high
low
medium
ns
NA


NA
ns
medium
low
high
CusRiskCategory
no
yes
NA
PEP
CusRiskCategory
high
low
medium
ns
NA


unasgn
u_inv
esc_l2
created
closed_pe
closed_ns
closed_ni
closed_dq
a_inv
AlertState
no
yes
NA
PEP
AlertState
a_inv
closed_dq
closed_ni
closed_ns
closed_pe
created
esc_l2
u_inv
unasgn


6000
4000
2000
0
count
no
yes
NA
PEP
PEP
no
yes
NA


10000
7500
5000
2500
0
count
closed
report
NA
CaseState
CaseState
closed
report
NA

Chapter 3

Additional Analysis

"All models are wrong, some models are useful".
The analysis that should be included in the qualitative validation part of a Transaction Monitoring Models are:

- Various Scenario Testing: Testing of various scenarios like HRGs, Structuring, MIs etc. should be part of the qualitative validation
- Statistiscal Testing: This can help ensure that the model is properly validated by detecting outliers and using appopriate models.
- Model Should be compared with Standards/Benchmarks/effective models to see its performance.
- Using the opinion of experts/regulators in the AML field
- Outcome analysis and back-testing
- Applying the historical data in the validation analysis.

The areas of Transaction Monitoring models that are essential to analyse are as follows:

- Performance of the Model: ROC, Recall, F1 Score, Precisioon, Accuracy - these could be effective
- Quality of the Data: one must enusre that the data used in the TM models must be of the highest quality and missing/inconsistent data should be dealt with accurately.
- Models should be chosen based on its performance against different metrics .
- Models should strictly adhere to Regulations/Regulatory compliance
- Periodic Validation/Testing
- Trend Review and Defect Analysis: Data is reviewed to identify potential patterns of behaviour/trends
- Hyperparameter Tuning: tuning the hyperparameters of a models can help optimise its performance against benchmarks/starndards.