Analysis of the dataset Alerts in the contex of transaction models

TASK 1.
There are altogether 10177 rows with 13 columns for the raw version of the data, 12 of which are of character type and 1 identified as numeric type. According to the metadata, the data presents status of each transaction which generated alert along with additional information about clients such as industry code or customer risk category.

Type of information	Value
Name	df_raw
Number of rows	10177
Number of columns	13
Number of columns with character type	12
Number of columns with numeric type	1

In case of the columns with character type, there are 6337 missing values identified for the column "PEP" with proportion of non-missing value that equals to around 0.377 and 3840 missing values for the column "Industry Code" with proportion of non-missing value that equals to around 0.622.

Type of column	Variable name	No. of missing values	Proportion of non-missing values
character	AlertType	0	1.0000000
character	AlertState	0	1.0000000
character	DateCreated	0	1.0000000
character	DateClosed	0	1.0000000
character	CaseOpen	0	1.0000000
character	CaseClosed	0	1.0000000
character	CaseReported	0	1.0000000
character	CaseState	0	1.0000000
character	PEP	6337	0.3773214
character	CusRiskCategory	0	1.0000000
character	Type	0	1.0000000
character	IndustryCode	3840	0.6226786

In case of the columns with numeric type, there exists only column "intID" without any missing values. The analysis of its statistics is not necessary since according to metadata, it just determines the ID of a given alert.

Type of column	Variable name	No. of missing values	Proportion of non-missing values
numeric	intID	0	1

To ensure the appropriateness of values in each column and classes of columns, first it need to be analysed whether null values should be treated as missing values. There were identified:
-261 "NULL" values for column "DateClosed"
-9921 "NULL" values for column "CaseOpened"
-9989 "NULL" values for column "CaseClosed"
-10109 "NULL" values for column "CaseReported"
-9921 "NULL" values for column "CaseState"
-938 "NULL" values for column "PEP"
-54 "NULL" values for column "CusRiskCategory"
-4126 "NULL" values for column "IndustryCode"

All columns should be analysed in context of the metadata. There are no “NULL” values for columns “intID”, “Alert type”, “Date created” and “Type” therefore they will not be analysed.
1. Column “PEP”
In case of column “PEP”, classification as a Politically Exposed Person should be analysed in the context of the column “Type” since this term applies generally to individuals and not companies. In this context, “NULL” values should be only considered as missing values for clients classified as “pb” in the column “Type”.
2. Column “IndustryCode”
In case of “NULL” values in the column “IndustryCode”, “NULL” values should not be treated as missing values for client with Type “pb”.With respect to “lcfi” there might be cases where industry code could not be determined such as companies that belong to new, emerging industries which might be difficult to classify. Therefore, “NULL” values for clients classifed to “lcfi” should also not be treated as missing values.
3. Column “CurRiskCategory”
In case of “NULL” values in the column “CusRiskCategory”, there exist value “Not Specified” which determines the lack of information about the client, therefore “NULL” values should be rather treated as missing values.
4. Columns “DataClose”, “CaseOpen”, “CaseClosed”, “CaseReported”, “CaseState”
Since columns “DataClose”, “CaseOpen”, “CaseClosed”, “CaseReported”, “CaseState” are connected to each other, “NULL” values for this columns should be analysed altogether. - In case of column “DataClose”, “NULL” values should be classified as missing data when the “AlertState” is closed or the report was confirmed
- In case of column “CaseOpen”, “NULL” values should be classfied as missing data when the “CasteState” is closed or the report was confirmed or “CaseClosed” is not null
- In case of column “CaseClose”, “NULL” values should be classified as missing data when column “CaseState” is closed
- In case of column “CasteState”, “NULL” values should be classified as missing data when column “CaseClosed” is not null or column “CaseState” is not null

After the analysis of the null values the following change has been made:
-transformation of 6 “NULL” values for column “DateClosed” to missing value
-transformation of 54 “NULL” values for column “CusRiskCategory” to missing value
-transformation of 938 “NULL” values for column “PEP” to missing value
-transformation of 286 “NULL” values for column “IndustryCode” to missing value

Regarding the missing values, through the analysis of “NULL” values, 4 columns with NA values were identified. In the original df_raw, only two columns were indentified as columns with missings values. Therefore the additional analysis of these two columns (“PEP” and “IndustryCode”) need to be conducted.
- With respect to the column “PEP”, missing values for clients classifed as “lcfi”, should be treated as rather “Not Applicable”.
- With respect to the column “IndustryCode”, missing values for the clients classifed as “pb” should not be treated as missing. In case of clients “lcfi”, since there is already a value “NULL”, missing values should not be removed.
- Additionally, there are missing information about customer risk category since there are values “Not Specified” which should be treated as missing values

There exits 6 missing values for the column “DateClosed”, 938 for column “PEP” and 54 for “CusRiskCategory”:
-In case of column “DateClosed”, 6 missing values are assigned since the column “AlertType” suggests in this cases that alert should be closed
-In case of column “PEP”, there exist 938 not assigned categories to private banking clients
-In case of column “CusRiskCategory”, there exist 682 not assigned categories to customers

To deal with the problem of missing values, data imputation is considered:
-In case of column “DateClosed” data imputation is rather not applicable
-In case of column “PEP” some data imputation might be possible since for instance all clients not classifed as PEP that belongs to private banking sector are also classified to MediumRisk category and with AlertType

to “Awakening account” but it would be rather advisable to check each client’s political connections and to fill in missing values

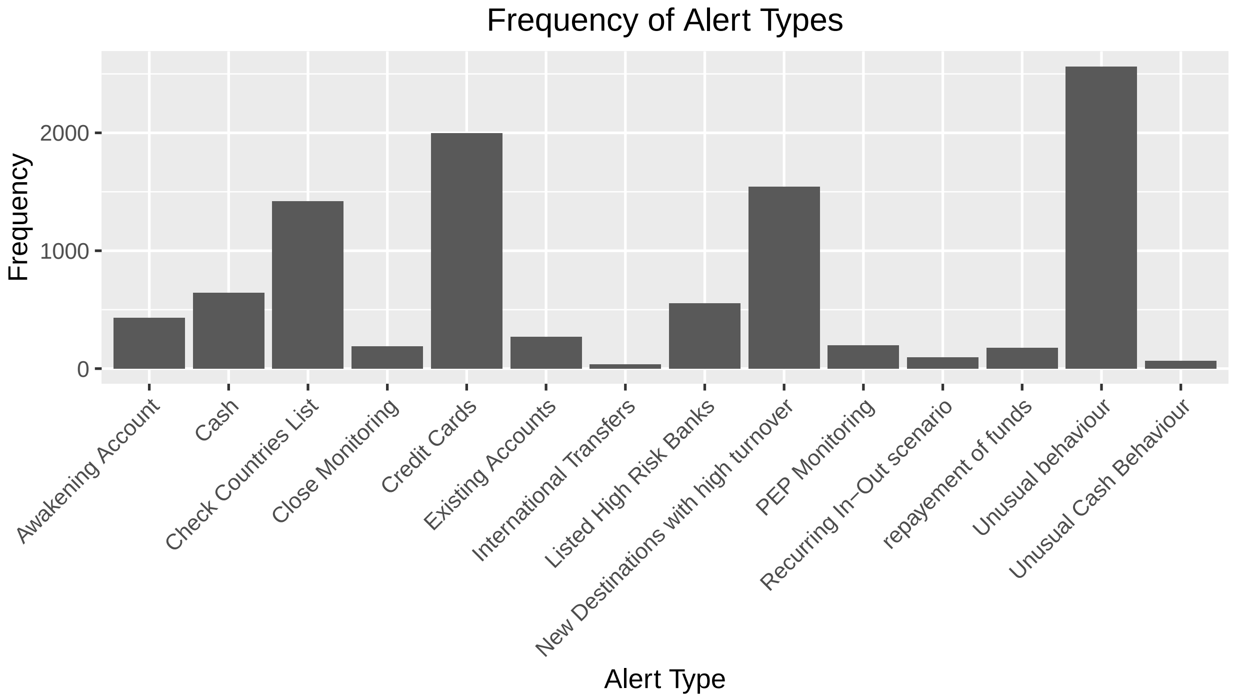
AlertType	CusRiskCategory	Y_count	N_count	NA_count
Awakening Account	Higher Risk	0	2	0
Awakening Account	Lower Risk	0	11	10
Awakening Account	Medium Risk	0	95	24
Awakening Account	NA	0	7	10
Check Countries List	Higher Risk	0	4	0
Check Countries List	Lower Risk	0	4	10
Check Countries List	Medium Risk	0	12	8
Close Monitoring	Higher Risk	0	0	6
Existing Accounts	Higher Risk	0	4	0
Existing Accounts	Lower Risk	0	30	9
Existing Accounts	Medium Risk	0	91	48
Existing Accounts	NA	0	9	29
International Transfers	Lower Risk	0	14	0
International Transfers	Medium Risk	0	20	0
New Destinations with high turnover	Higher Risk	78	106	0
New Destinations with high turnover	Lower Risk	0	155	0
New Destinations with high turnover	Medium Risk	0	1147	5
New Destinations with high turnover	NA	0	52	0
PEP Monitoring	Higher Risk	24	0	0
PEP Monitoring	Medium Risk	0	4	0
Recurring In-Out scenario	Higher Risk	3	3	0
Recurring In-Out scenario	Lower Risk	0	31	2
Recurring In-Out scenario	Medium Risk	0	56	2
Unusual behaviour	Higher Risk	13	32	3
Unusual behaviour	Lower Risk	0	145	255
Unusual behaviour	Medium Risk	0	693	477
Unusual behaviour	NA	0	57	40

-In case of column “CusRiskCategory”, some data imputation might be possible if there were missing values for customers that belong to private banking sector and are indentified as PEP since all such cases are classified to Higher Risk but it would rather be advisable to fill in missing CusRiskCategory

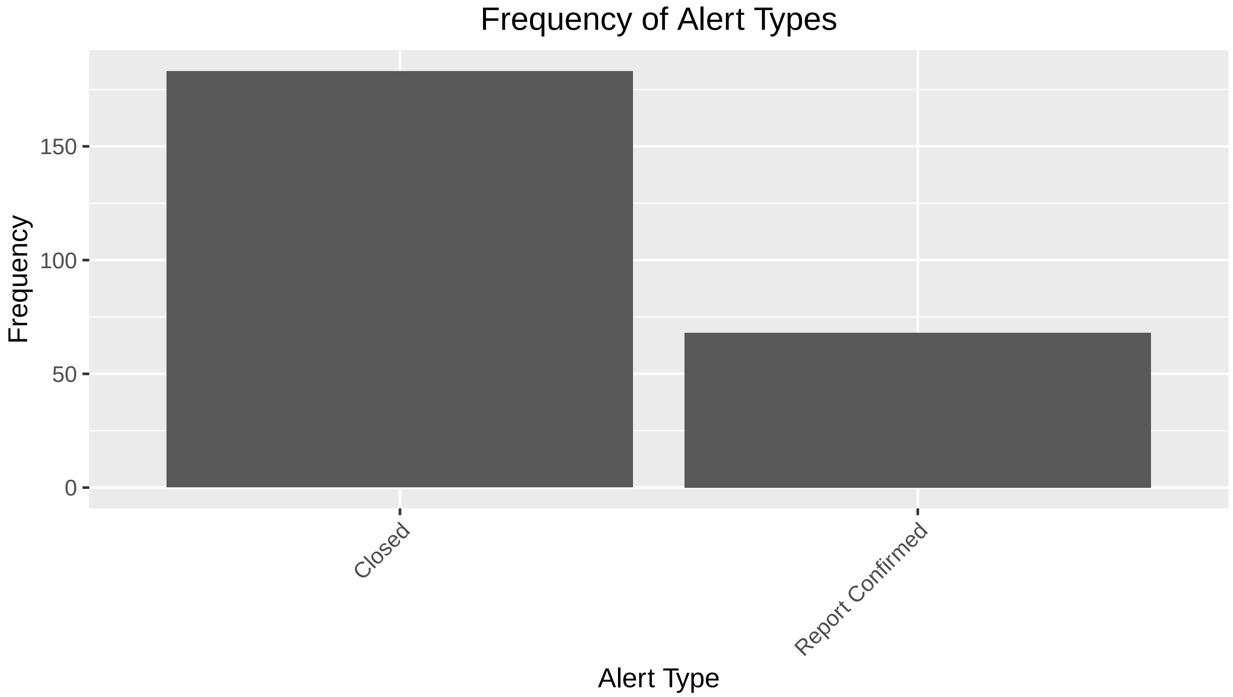
PEP	lower_count	medium_count	higher_count	NA_count
N	390	2118	151	125
Y	0	0	118	0
NA	286	564	9	79

Summarizing the possibility of data imputation in columns “CusRiskCategory”, “DataClosed” and “PEP”, no data imputation processes were conducted since it is more advisable to fill in these data. After the final analysis of the “NULL” values and NA values, there exits missing values for columns “DataClosed” - 6 missing values, “PEP” - 938 missing values, “CusRiskCategory” - 682 missing values. Since the data imputation in all of these cases is rather not recommended due to the context of this missing values, these missing values will be removed from the analysis.
After the cleaning, the dataframe consists of 8361 out of 10177 original rows.
From the context of the data, it would be also advisable to order it chronogically.

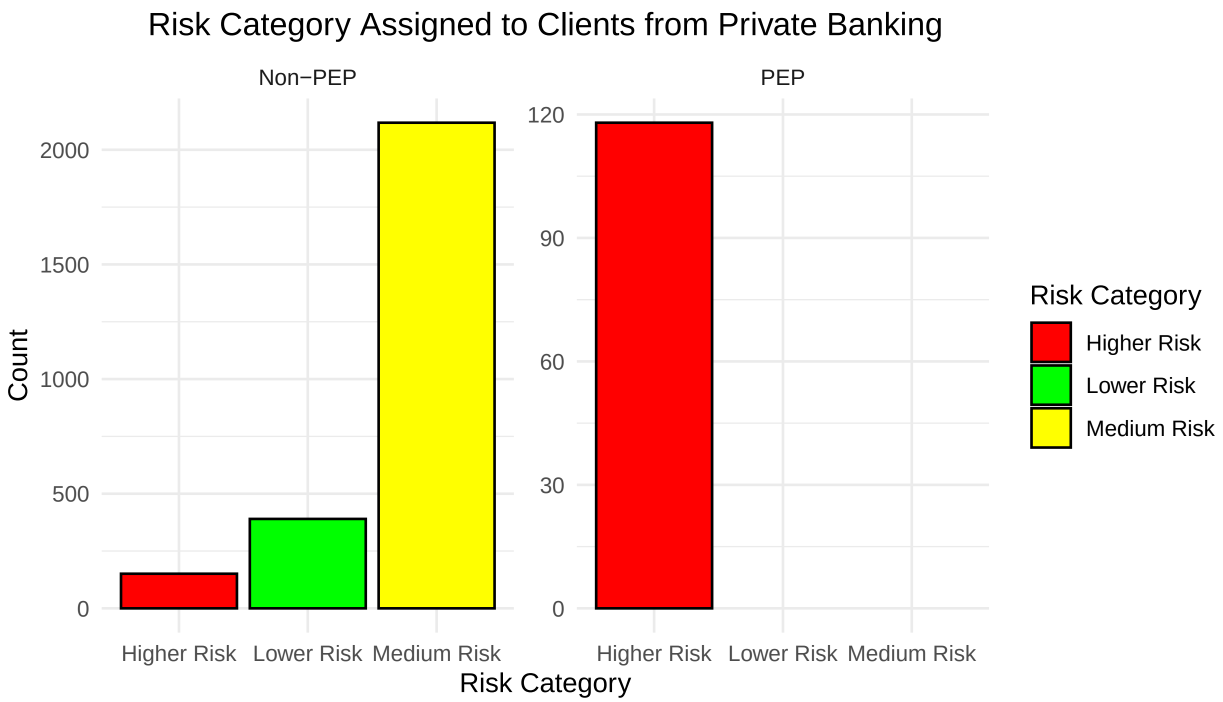
TASK 2.
1. Analysis of column “Alertype”
- The most frequent types of alerts are “Unusual behaviour”, “Credit Cards” and “Check Countries List”



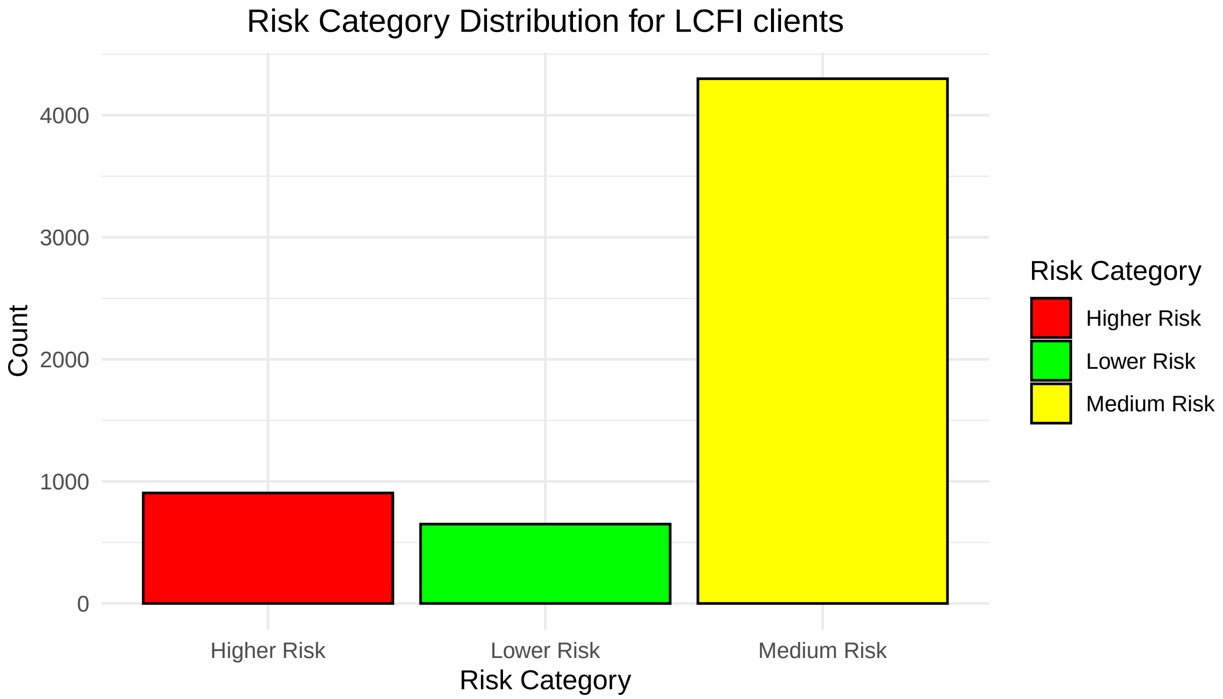
2. Analysis of column "CaseState"
- Excluding "NULL" values from the analysis, when the investigation goes to the second line, the most frequent action is closing the case



3. The analysis to assess whether clients classifed to private banking and as PEP are more likely to be in higher risk category
- All of the clients classifed to private banking and as PEP are at higher risk category while for non-PEP clients the most frequent risk category is medium risk



4.Analysis of customer risk category for lcfi clients
- For the customers classified as lcfi, the most frequent category is medium risk, while the second one is higher risk category.



TASK 3.
In general the most important area in TM models to analyse is it to predict which transactions will be classified as involved with illegal activities in the future and to optimize the reporting to law enforcement according to the risk.
-1. Analysis
-Based on the available data it is possible to analyse whether transactions of clients with higher risk are

more frequently reported to assess whether customer risk category is appropriately assigned
-Based on the analysis, the risk category for clients is assigned appropriately since the percentage of confirmed reports to sum of confirmed reports and closed cases is highest for the customers with the higher risk category (0.4931507).

CusRiskCategory	CaseState	count
Higher Risk	Closed	37
Higher Risk	Report Confirmed	36
Lower Risk	Closed	30
Lower Risk	Report Confirmed	8
Medium Risk	Closed	116
Medium Risk	Report Confirmed	24

High Risk	Medium Risk	Lower Risk
0.4931507	0.2105263	0.1714286

-2. Analysis
Additionally it is also possible to analyse whether PEP clients are more likely to be reported with respect to non-PEP clients.
For the clients classified as PEP there are only 3 cases where state of case is either closed or report was issued, therefore it is not possible to draw any conclusions.

PEP	CaseState	count
N	Closed	35
N	Report Confirmed	2
Y	Closed	3

Since the data can be ordered chronologically, it is also possible to detect whether there is increase in alerts or reported cases based on the year, month or day of the week which could also help to detect some patterns in transactions.