Fraud Analysis_M-KOP@ @SSESSMENT

ERICK@

2024-

Problem Statement: Questions to answer:

1) Which region has the highest Outstanding Loan Balance exposure. Provide reasons and evidence of the analysis. Explain in approximately no less than (80 words)
2) Giving reasons and rank the region that is highly affected by fraud.Show evidence and explain in approximately (70 words)
3) Using the data determine the most affected phone model, and in which region. Show evidence and explain in approximately (70 words).
4) Using the data show the most affected month by fraud. Show evidence and explain in approximately (50 words).
5) How could we potentially improve the fraud identification process? explain in approximately (100 words).
6) What operational improvements should we investigate to improve the fraud investigation process? explain in approximately (100 words).
7) Write an SQL query to replicate the results in Data-Sheet but only getting results for Suwami reg. Use the data on sheet named "Short schema".
8) Show the process you used to clean the data. Show evidence and explain in approximately (80 words)

```
library(tidyverse)
library(janitor)
```

```
df <- read_csv("C:/Users/langa/OneDrive/Desktop/Dataset/M-KOPA Fraud Analyst Intern Technical assessmen
    col_types = cols(`Date of Sale` = col_date(format = "%m/%d/%Y"))) %>%
  clean_names()
head(df)
```

```
## # A tibble: 6 x 8
##   region account_number model     outstanding_loan_balance loan_collection_speed
##   <chr>            <dbl> <chr>     <chr>                                    <dbl>
## 1 Suwami         584025 Nokia C31 23896C31                                  0.8
## 2 Bira           598168 Tecno Ck6 1815C31                                   0.55
## 3 Bira           458938 Tecno Ck6 551573.4542                               0.45
## 4 Bumasi          72228 Tecno Ck6 551345.4231                               0.97
## 5 Bumasi          99694 Tecno Ck6 549234.7974                               1.07
## 6 Bumasi          99246 Tecno Ck6 547262.5153                               1.18
## # i 3 more variables: date_of_sale <date>, investigated <chr>,
## #   investiagtion_outcome <chr>
```

```
str(df)
```

```
## spc_tbl_ [2,343 x 8] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ region                  : chr [1:2343] "Suwami" "Bira" "Bira" "Bumasi" ...
##  $ account_number          : num [1:2343] 584025 598168 458938 72228 99694 ...
##  $ model                   : chr [1:2343] "Nokia C31" "Tecno Ck6" "Tecno Ck6" "Tecno Ck6" ...
##  $ outstanding_loan_balance: chr [1:2343] "23896C31" "1815C31" "551573.4542" "551345.4231" ...
##  $ loan_collection_speed   : num [1:2343] 0.8 0.55 0.45 0.97 1.07 1.18 1.5 1.27 1.65 1.1 ...
##  $ date_of_sale            : Date[1:2343], format: "2022-11-17" "2023-08-05" ...
##  $ investigated            : chr [1:2343] "Uninvestigated" "Uninvestigated" "Investiagted" "Uninvesti
##  $ investiagtion_outcome   : chr [1:2343] "Not investigated" "Not investigated" "Confirmed to be Frau
##  - attr(*, "spec")=
##   .. cols(
##   ..   Region = col_character(),
##   ..   AccountNumber = col_double(),
##   ..   Model = col_character(),
##   ..   ‘Outstanding Loan Balance‘ = col_character(),
##   ..   ‘Loan Collection Speed‘ = col_double(),
##   ..   ‘Date of Sale‘ = col_date(format = "%m/%d/%Y"),
##   ..   Investigated = col_character(),
##   ..   ‘Investiagtion outcome‘ = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
#data cleaning
#remove "C31" Replace with " "
df$outstanding_loan_balance <- str_replace(df$outstanding_loan_balance,
                                           "C31", " ") %>% as.numeric()
#date
df$date_of_sale <- ymd(df$date_of_sale)
```

Which region has the highest Outstanding Loan Balance exposure. Provide reasons and evidence of the analysis. Explain in approximately no less than (80 words)

```
df %>% select(region, outstanding_loan_balance) %>%
      group_by(region) %>%
      summarise(Loan_Balance=sum(outstanding_loan_balance)) %>%
       arrange(desc(Loan_Balance))
```

```
## # A tibble: 4 x 2
##   region  Loan_Balance
##   <chr>          <dbl>
## 1 Suwami      8674531
## 2 Bumasi      8455168.
## 3 Nilmark     5496709
## 4 Bira        1002851.
```

```
# as seen in the calculation,Suwami is leading in the outstanding loan balance by 8674531 followed by,B
# Bira  1002851
```

Giving reasons and rank the region that is highly affected by fraud.Show evidence and explain in approximately (70 words)

```
df %>%  select(region,investiagtion_outcome) %>%
  filter(investiagtion_outcome== "Confirmed to be Fraud") %>%
       group_by(region) %>%
      summarise(Fraud_occurrence=n()) %>% arrange(desc(Fraud_occurrence))
```

```
## # A tibble: 3 x 2
##   region Fraud_occurrence
##   <chr>             <int>
## 1 Suwami               48
## 2 Bumasi                9
## 3 Bira                  5
```

```
# df %>%  pull(investiagtion_outcome) %>% unique()
```

Using the data determine the most affected phone model, and in which region. Show evidence and explain in approximately (70 words).

```
df %>% select(region, model,investiagtion_outcome ) %>%
        filter(investiagtion_outcome=="Confirmed to be Fraud") %>%
        group_by(region, model) %>%
        summarise(Leading_model=n()) %>%
            arrange(desc(Leading_model))
```

```
## # A tibble: 4 x 3
## # Groups:   region [3]
##   region model       Leading_model
##   <chr>  <chr>               <int>
## 1 Suwami Nokia C31              27
## 2 Suwami Samsung A12            21
## 3 Bumasi Nokia C31               9
## 4 Bira   Tecno Ck6               5
```

Using the data show the most affected month by fraud. Show evidence and explain in approximately (50 words).

```
df <- df %>% mutate(month= month(date_of_sale, label=T))# %>% colnames()
# head(df$month)
d <- df %>%  select(month, investiagtion_outcome) %>%
  filter(investiagtion_outcome=="Confirmed to be Fraud") %>%
 group_by(month) %>%
  summarise(NO_Fraud_month=n()) %>% arrange(desc(NO_Fraud_month))

d
```

```
## # A tibble: 11 x 2
##    month NO_Fraud_month
##    <ord>          <int>
##  1 Feb               16
##  2 Dec               12
##  3 Jan               10
##  4 Mar                8
##  5 Apr                3
##  6 Aug                3
##  7 Oct                3
##  8 May                2
##  9 Jul                2
## 10 Nov                2
## 11 Sep                1
```

How could we potentially improve the fraud identification process? explain in approximately (100 words).

What operational improvements should we investigate to improve the fraud investigation process? explain in approximately (100 words).

Write an SQL query to replicate the results in Data-Sheet but only getting results for Suwami reg. Use the data on sheet named "Short schema".

Show the process you used to clean the data. Show evidence and explain in approximately (80 words)