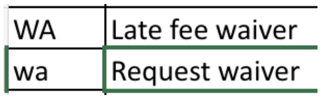
Data cleaning pipeline

Kelsey

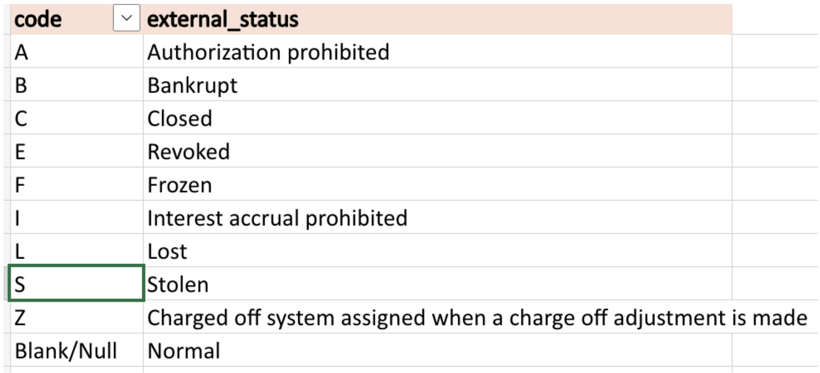
- **MOS**: lowercase and uppercase letters should be kept. There’s no need to convert to anything else.



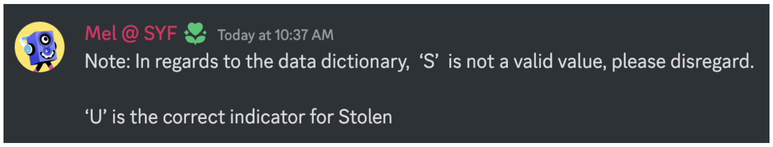
WA	Late fee waiver
wa	Request waiver

This doesn’t need to be changed

- **external_status**: S should be converted to U



code	external_status
A	Authorization prohibited
B	Bankrupt
C	Closed
E	Revoked
F	Frozen
I	Interest accrual prohibited
L	Lost
S	Stolen
Z	Charged off system assigned when a charge off adjustment is made
Blank/Null	Normal



Mel @ SYF Today at 10:37 AM

Note: In regards to the data dictionary, ‘S’ is not a valid value, please disregard.

‘U’ is the correct indicator for Stolen

This doesn’t need to be changed, because we don’t have S in our data.

- **auto_pay_enrolled_status**: make sure 0 and 1 in this column are all integers
  0 represents the absence, and 1 represents the presence of that category

- **account_balance**:
  - Do not remove null rows
  - make sure they are all numbers, not words
- **resolved**:
  - convert *resolved* to be 1, *floor* to be 0 (number) (this is for simpler calculations)
- **no_of_account_with_syf**: make sure they are numbers
- **account_open_date_13_march** and **account_open_date_18_march:**
  - there are 2 rows with different open dates (previously asked in discord). Use the older date.
- card_activation_status:
  - make sure they are all numbers
  - check if there are numbers other than 0, 7, 8, 9

Chage “ ” to 6

- **account_status:**
  - do not remove NULL since it means no restrictions

    - replace B, C, E, F, I, Z as letter C (closed)
    - replace blank as N (no restriction)
- **serial**:
    - probably there are duplicate serial number -> same person call more than once
- **delinquency_history**:
    - compare **delinquency_history_13_March** and **delinquency_history_18_March**, see if they are equal
    - separate e.g., [01] to two columns like:
        - delinquency_history_18_March_current: 0
        - delinquency_history_18_March_past: 1
    - create a new attribute **delinquency_compare_13_March**
        - M: [32] more delinquency in current than the past
            - There can only be 1 more comparing current and past
        - N: [00] no delinquency at all
        - E: [22] current = past, but with delinquency
        - P: [03] or [23] paid delinquency currently
        - NA: current - past > 1, bad data
- **e_bill_enrolled_status**:
    - create a new attribute **e_bill_enrolled_status_combined:**
        - replace blank to be P (paper)
        - replace B, D, L to be B (both paper and electronic)
        - keep E (electronic)

5. Zheer

Delinquency_history:

First number-current, second num-past

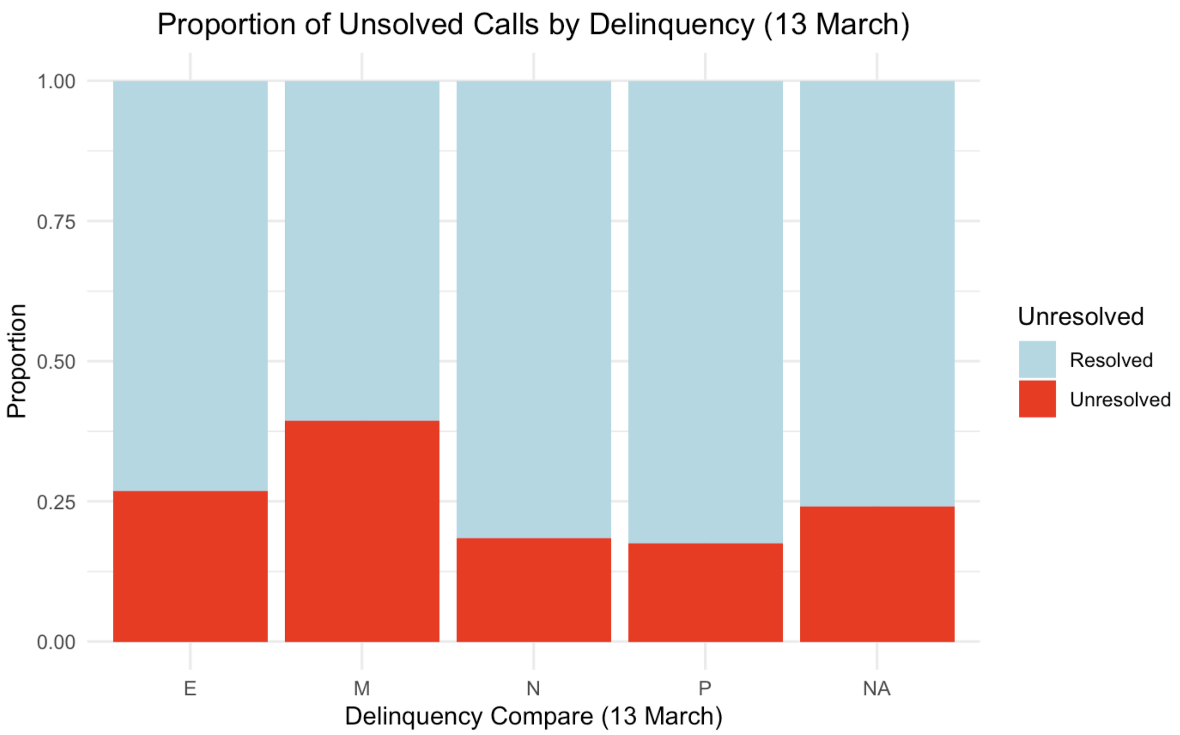
Find the relationship between Delinquency_history and unresolved percentage.

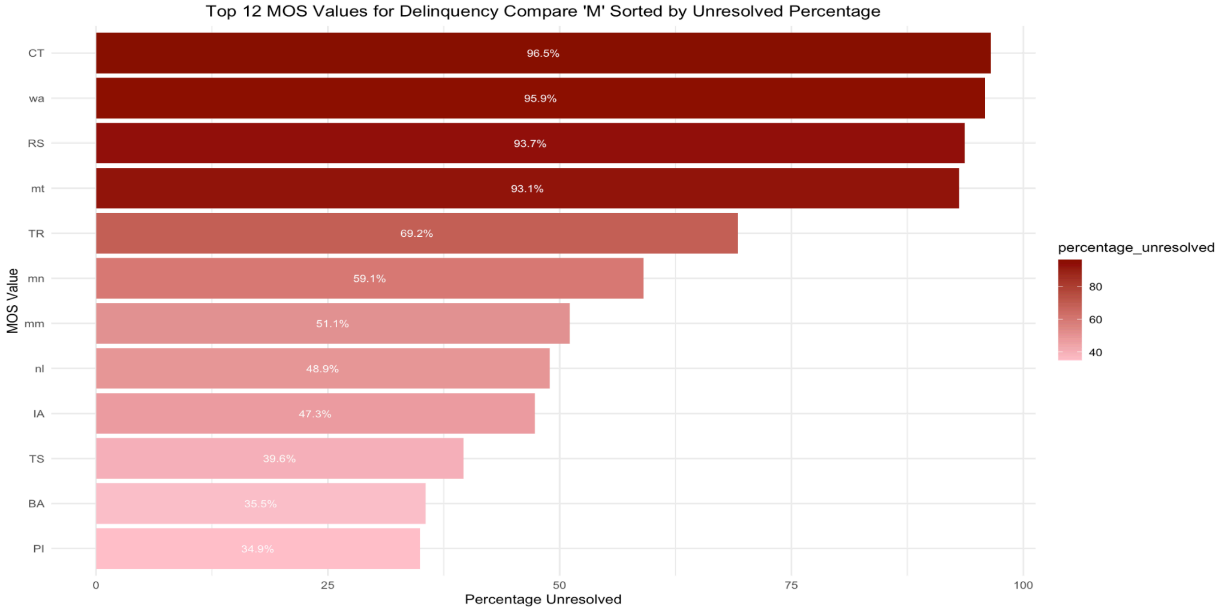
Categorize :

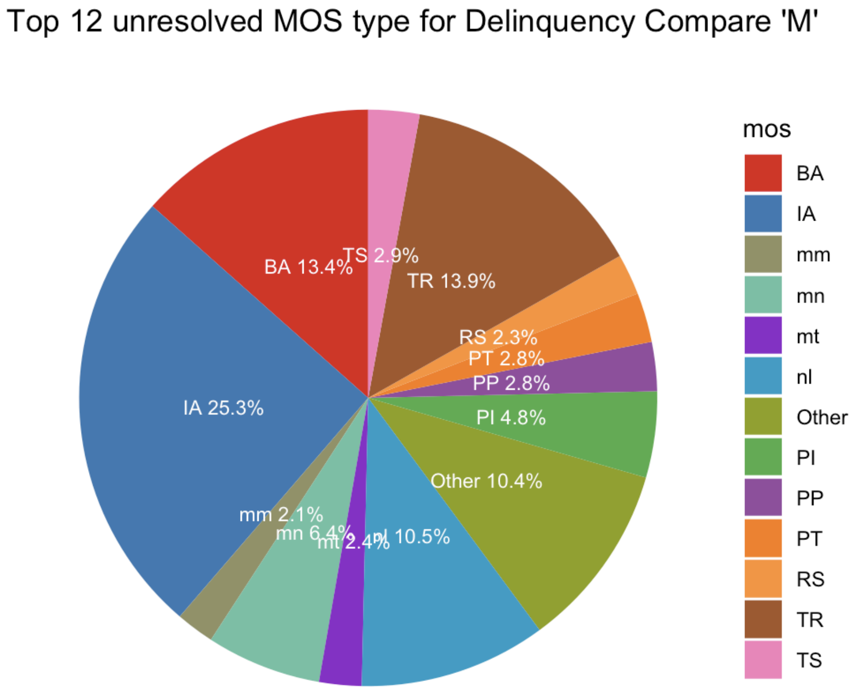
1. Current - past < 0 (paid)

2. Current - past >= 0 (unpaid)

```
# Plot for delinquency_compare_13_March
ggplot(data, aes(x = delinquency_compare_13_March, fill = Unresolved)) +
  geom_bar(position = "fill") +
  labs(x = "Delinquency Compare (13 March)", y = "Proportion") +
  theme_minimal() +
  scale_fill_manual(values = c("Resolved" = "lightblue", "Unresolved" = "red")) +
  ggtitle("Proportion of Unsolved Calls by Delinquency (13 March)") +
  theme(plot.title = element_text(hjust = 0.5))
```


Proportion of Unsolved Calls by Delinquency (13 March)
Proportion
1.00
0.75
0.50
0.25
0.00
E
M
N
P
NA
Delinquency Compare (13 March)
Unresolved
Resolved
Unresolved


Top 12 MOS Values for Delinquency Compare 'M' Sorted by Unresolved Percentage
MOS Value
CT
wa
RS
mt
TR
mn
mm
nl
IA
TS
BA
PI
96.5%
95.9%
93.7%
93.1%
69.2%
59.1%
51.1%
48.9%
47.3%
39.6%
35.5%
34.9%
0
25
50
75
100
Percentage Unresolved
percentage_unresolved
80
60
40



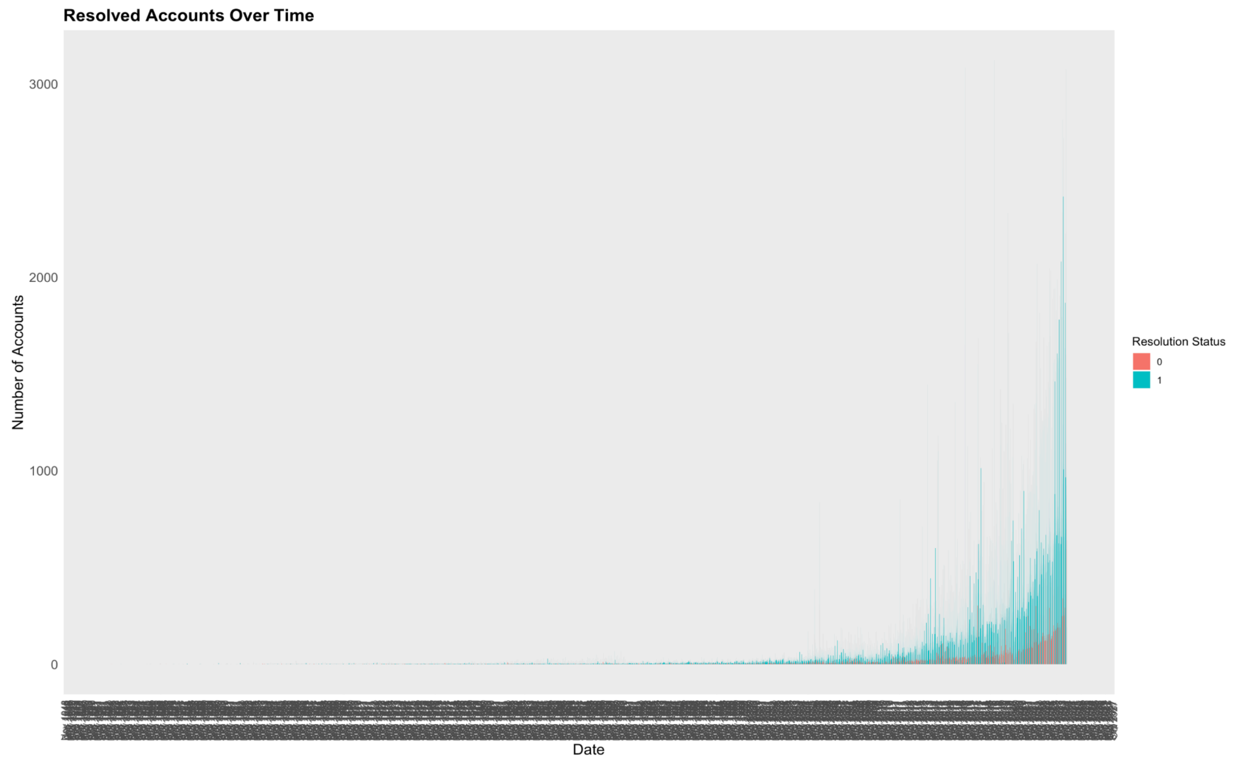
Top 12 unresolved MOS type for Delinquency Compare 'M'

6. Ziqi

Find if there’s the relationship between account_open_date_13_march and the unresolved percentage. (consider customer loyalty)

X-axis: time, y-axis: 0,1 (column resolve)

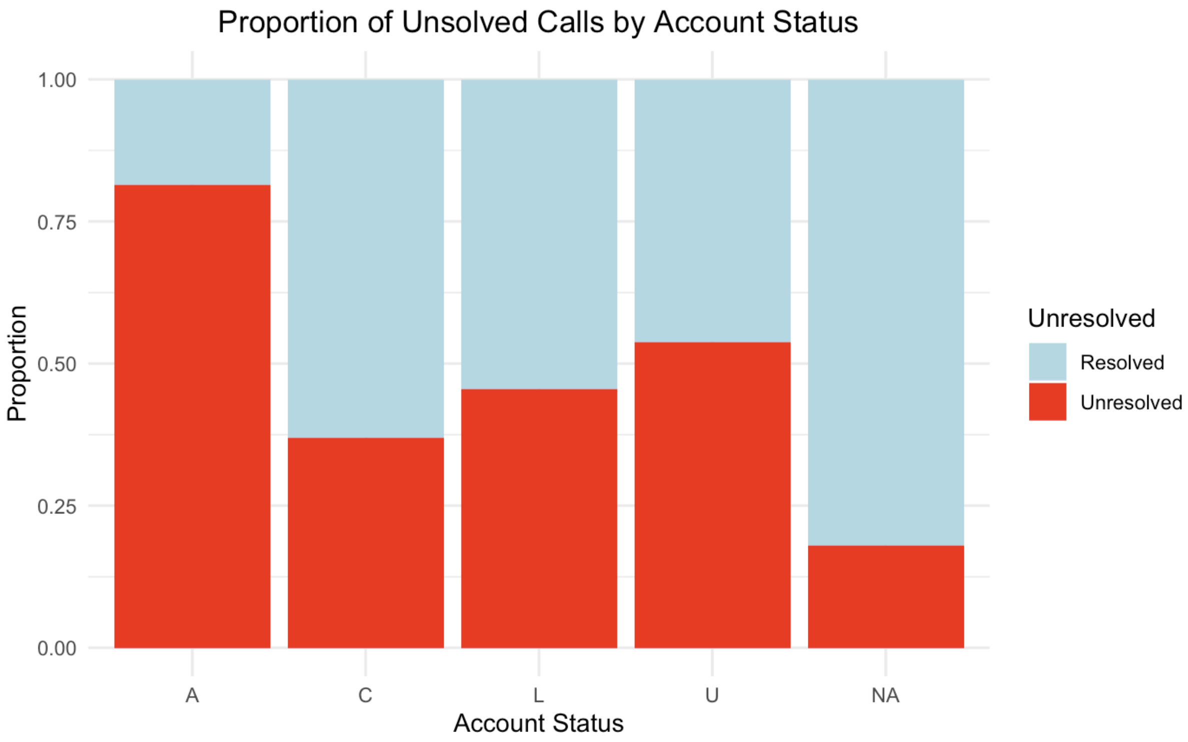
[redacted]



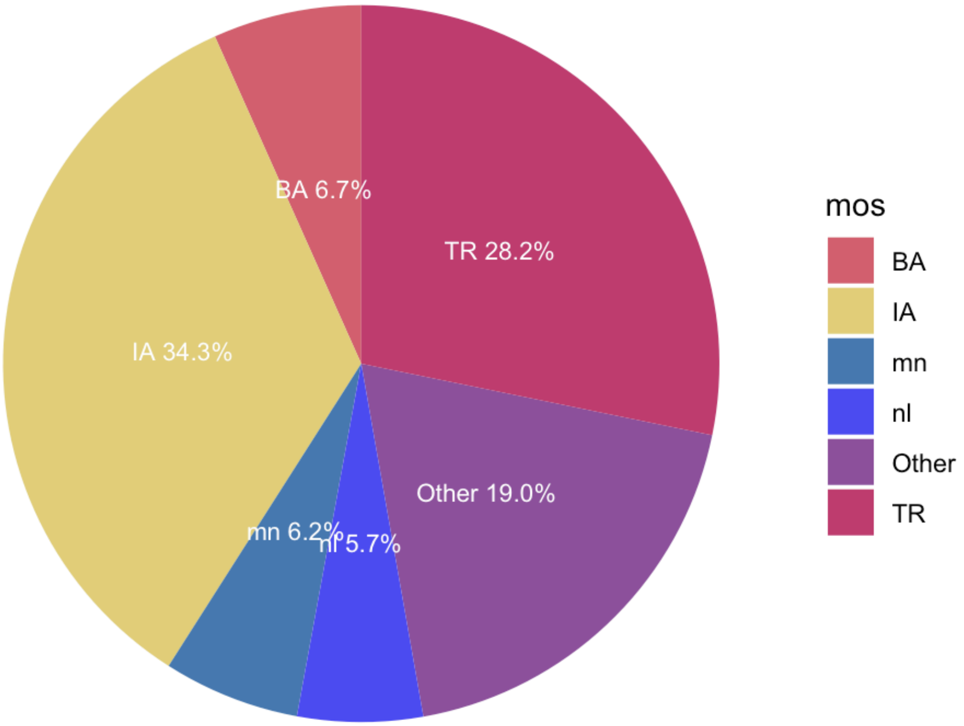
7. Zheer

Find the relationship between account_status/card_activation_status_13_march/ebill_enrolled_status_13_march and unresolved percentage.

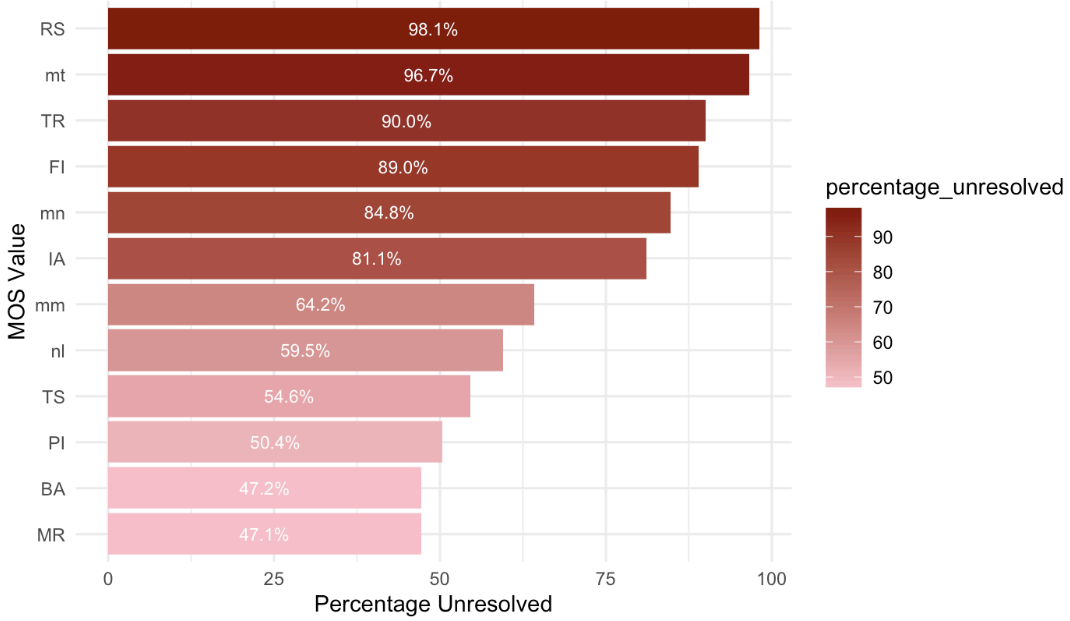
```
library(readxl)
library(dplyr)
library(ggplot2)

data = read_excel("/Users/zheerwang/Desktop/cleaned_data_latest.xlsx")
library(dplyr)

data$Unresolved <- ifelse(data$resolved == 0, "Unresolved", "Resolved")

# Plot for Account Status
data$account_status_13_march <- factor(data$account_status_13_march)
ggplot(data, aes(x = account_status_13_march, fill = Unresolved)) +
  geom_bar(position = "fill") +
  labs(x = "Account Status", y = "Proportion") +
  theme_minimal() +
  scale_fill_manual(values = c("Resolved" = "lightblue", "Unresolved" = "red")) +
  ggtitle("Proportion of Unsolved Calls by Account Status") +
  theme(plot.title = element_text(hjust = 0.5))

# Plot for Card Activation Status as of 13 March
ggplot(data, aes(x = as.factor(card_activation_status_13_march), fill = Unresolved)) +
  geom_bar(position = "fill") +
  scale_x_discrete(limits =
levels(droplevels(as.factor(data$card_activation_status_13_march)))) +
  scale_fill_manual(values = c("Resolved" = "lightblue", "Unresolved" = "red")) +
  labs(x = "Card Activation Status (13 March)", y = "Proportion") +
  ggtitle("Proportion of Unsolved Calls by Card Activation Status (13 March)") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))

# Plot for E-bill Enrolled Status as of 13 March
ggplot(data, aes(x = ebill_enrolled_status_13_march, fill = Unresolved)) +
  geom_bar(position = "fill") +
  labs(x = "E-bill Enrolled Status (13 March)", y = "Proportion") +
  theme_minimal() +
  scale_fill_manual(values = c("Resolved" = "lightblue", "Unresolved" = "red")) +
  ggtitle("Proportion of Unsolved Calls by E-bill Enrolled Status (13 March)") +
  theme(plot.title = element_text(hjust = 0.5))
```



Top 5 Unresolved MOS Types for Account Status A



Top 12 MOS Values for Account Status 'A' Sorted by Unresolved Percentage



IA, mn, TR have high unresolved percentages (> 80%) when Account Status = A and they occupy 68.7% of the total unresolved cases. Therefore we should pay close attention to these MOS types.

Separated data:

```
separated_data <- converted_data %>%
  separate_rows(mos, sep = " ")
```

Pie chart:

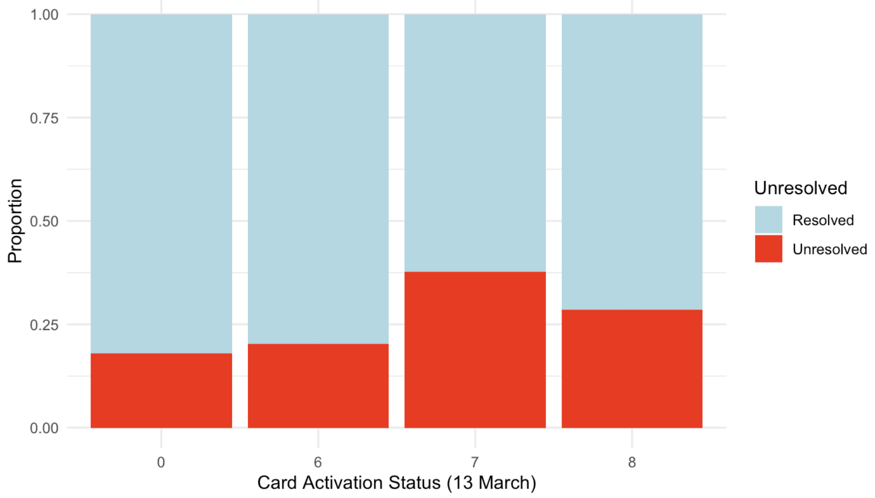
```
# Step 1: Filter the data for account_status_13_march 'A' and calculate counts
filtered_data <- separated_data %>%
  filter(account_status_13_march %in% c('A')) %>%
  count(mos)

# Step 2: Identify the top 5 mos types based on count
top_5_mos <- filtered_data %>%
  top_n(5, wt = n)

# Step 3: Create a dataset with an "Other" category for MOS types outside the top 5
pie_data <- filtered_data %>%
  mutate(mos = if_else(mos %in% top_5_mos$mos, as.character(mos), "Other")) %>%
  group_by(mos) %>%
  summarise(count = sum(n)) %>%
  mutate(percentage = count / sum(count) * 100) %>%
  ungroup()  # Ensure that the data is no longer grouped for plotting

# Define colors for the top 5 MOS types plus "Other", adjust the number of colors
accordingly
pie_chart_colors <- c("#E45A6C", "#E4cf6c", "#377EB8", "#4f4FfA", "#984EA3",
"#cF2c6f")

# Step 4: Create the pie chart
ggplot(pie_data, aes(x = "", y = percentage, fill = mos)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start = 0) +
  scale_fill_manual(values = pie_chart_colors) +
  theme_void() +
  geom_text(aes(label = paste(mos, sprintf("%.1f%%", percentage))),
            position = position_stack(vjust = 0.5),
            color = "white", size = 3) +
  labs(title = "Top 5 Unresolved MOS Types for Account Status A") +
  theme(legend.position = "right")
```

Box chart:

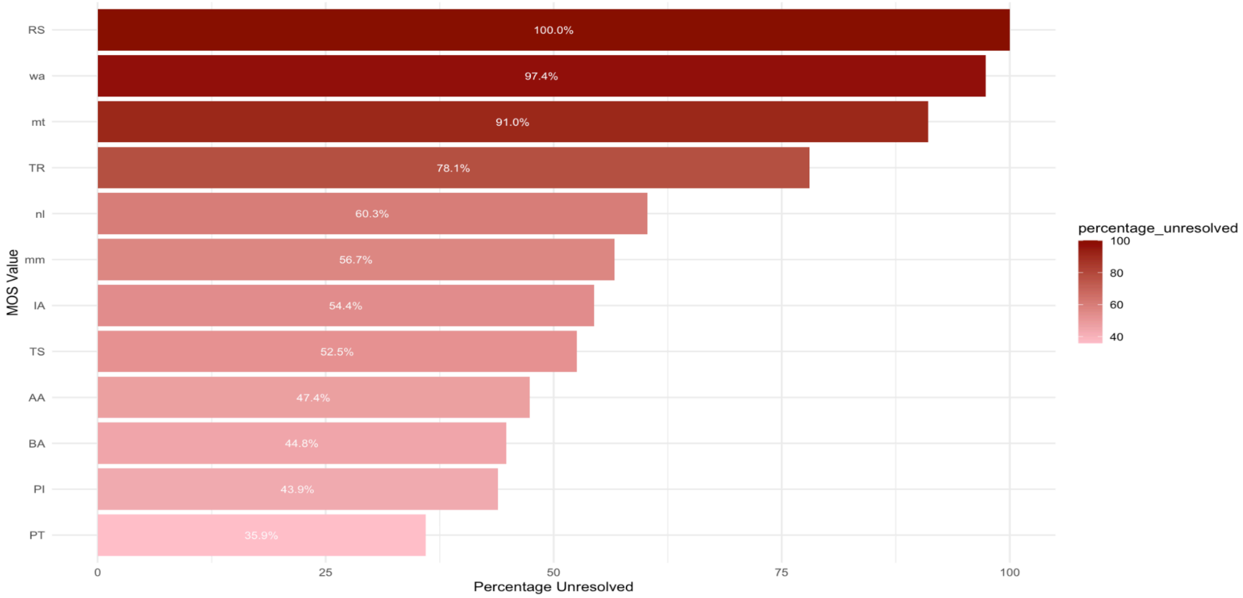
```
percentage_resolved <- separated_data %>%
  filter(account_status_13_march %in% c('A')) %>%
  group_by(mos) %>%
  summarise(
    total_count = n(),  # Total number of cases for this card_activation_status
    solved_count = sum(resolved == 1, na.rm = TRUE),  # Number of resolved cases
    unsolved_count = total_count - solved_count,  # Number of unresolved cases
    percentage_resolved = (solved_count / total_count) * 100,  # Calculate the
percentage resolved
```

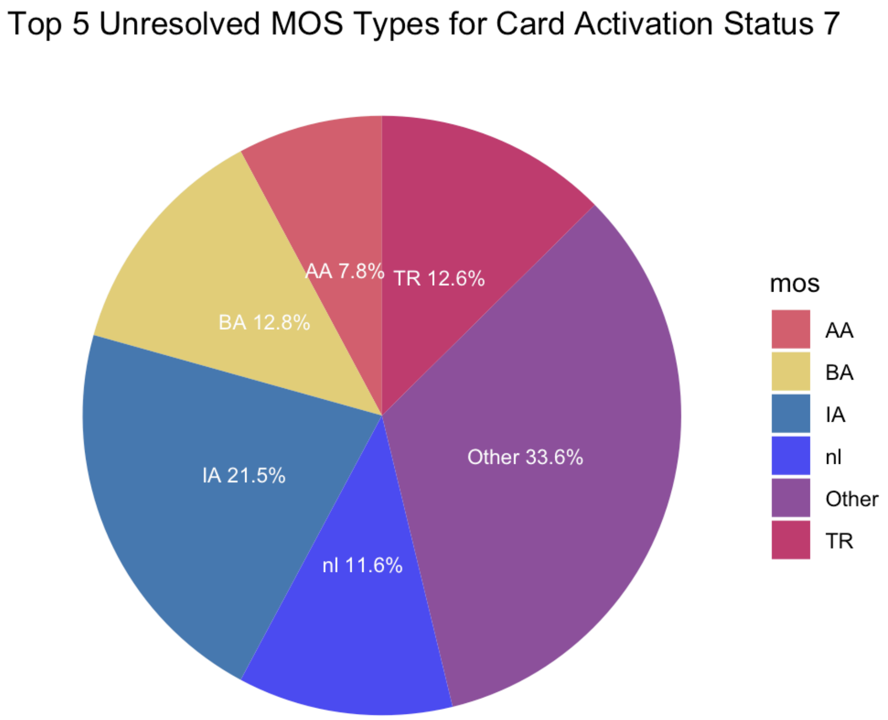
```
    percentage_unresolved = (unsolved_count / total_count) * 100  # Calculate the
percentage unresolved
  ) %>%
  ungroup()  # Remove the grouping

# Debug: Check the intermediate values for a specific 'mos' type
cat("Debug Info for a specific MOS type:\n")
print(percentage_resolved)

# Filter the top 12 unresolved MOS types
top_unresolved <- percentage_resolved %>%
  slice_max(order_by = unsolved_count, n = 12) %>%
  arrange(desc(percentage_unresolved))

# Plot the unresolved percentages for the top 12 MOS types
ggplot(top_unresolved, aes(x = reorder(mos, percentage_unresolved), y =
percentage_unresolved, fill = percentage_unresolved)) +
  geom_col() +
  geom_text(aes(label = sprintf("%.1f%%", percentage_unresolved)), position =
position_stack(vjust = 0.5), color = "white", size = 3) +
  labs(
    x = "MOS Value",
    y = "Percentage Unresolved",
    title = "           Top 12 MOS Values for Account Status 'A' Sorted by Unresolved
Percentage"
  ) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5)) +
  coord_flip() +
  scale_fill_gradient(low = "pink", high = "darkred")
```

Proportion of Unsolved Calls by Card Activation Status (13 March)



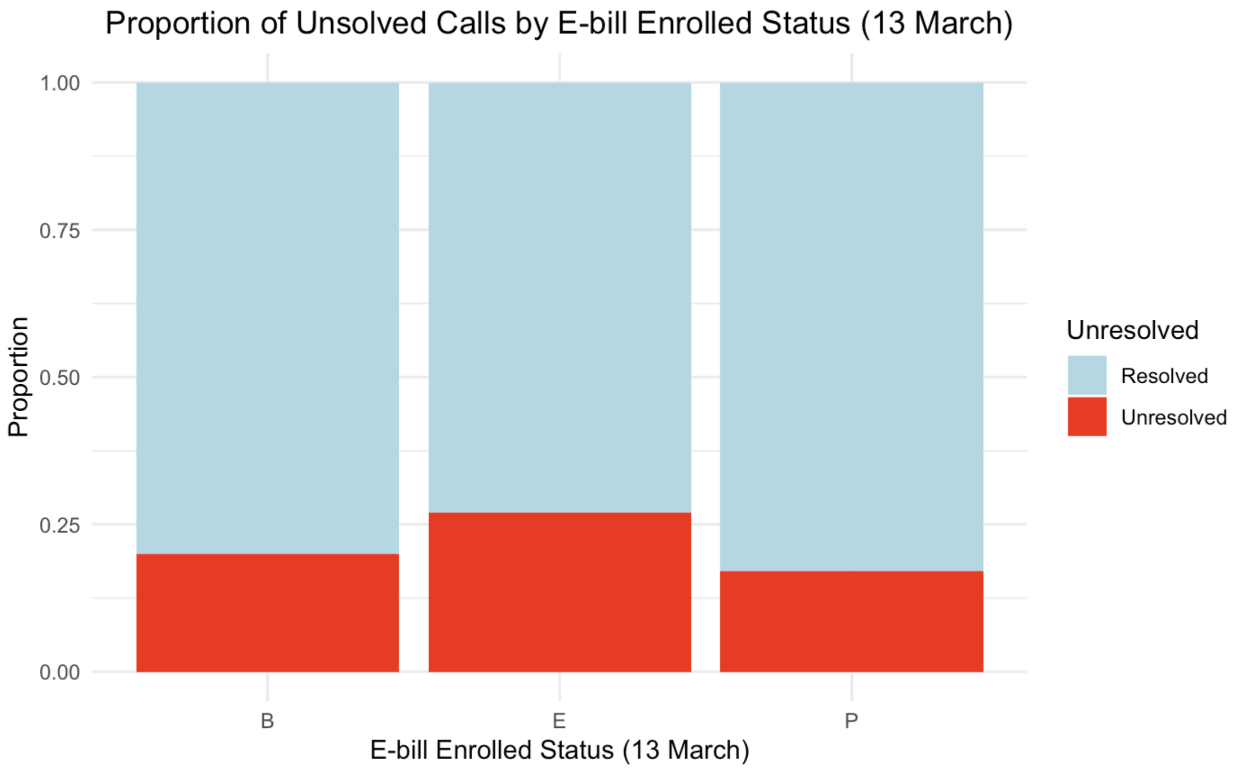
Top 12 MOS Values for Card Activation Status 7 Sorted by Unresolved Percentage





```
# Step 1: Filter the data for account_status_13_march 'A' and calculate counts
filtered_data <- separated_data %>%
  filter(card_activation_status_13_march %in% c(7)) %>%
  count(mos)

# Step 2: Identify the top 5 mos types based on count
top_5_mos <- filtered_data %>%
  top_n(5, wt = n)

# Step 3: Create a dataset with an "Other" category for MOS types outside the top 5
pie_data <- filtered_data %>%
  mutate(mos = if_else(mos %in% top_5_mos$mos, as.character(mos), "Other")) %>%
  group_by(mos) %>%
  summarise(count = sum(n)) %>%
  mutate(percentage = count / sum(count) * 100) %>%
  ungroup()  # Ensure that the data is no longer grouped for plotting

# Define colors for the top 5 MOS types plus "Other", adjust the number of colors
accordingly
pie_chart_colors <- c("#E45A6C", "#E4cf6c", "#377EB8", "#4f4FfA", "#984EA3",
"#cF2c6f", "#cF2ccf", "#af2c6f")

# Step 4: Create the pie chart
ggplot(pie_data, aes(x = "", y = percentage, fill = mos)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start = 0) +
  scale_fill_manual(values = pie_chart_colors) +
  theme_void() +
  geom_text(aes(label = paste(mos, sprintf("%.1f%%", percentage))),
            position = position_stack(vjust = 0.5),
            color = "white", size = 3) +
  labs(title = "Top 5 Unresolved MOS Types for Card Activation Status 7") +
  theme(legend.position = "right")
```

Continue: try to analyze the percentage of L and U when card activation status is 7. Conclusion: No relation :(



The percentages of B, E, and P don't have a large difference. We can conclude that E-bill Enrolled Status doesn't influence the resolved rate a lot.

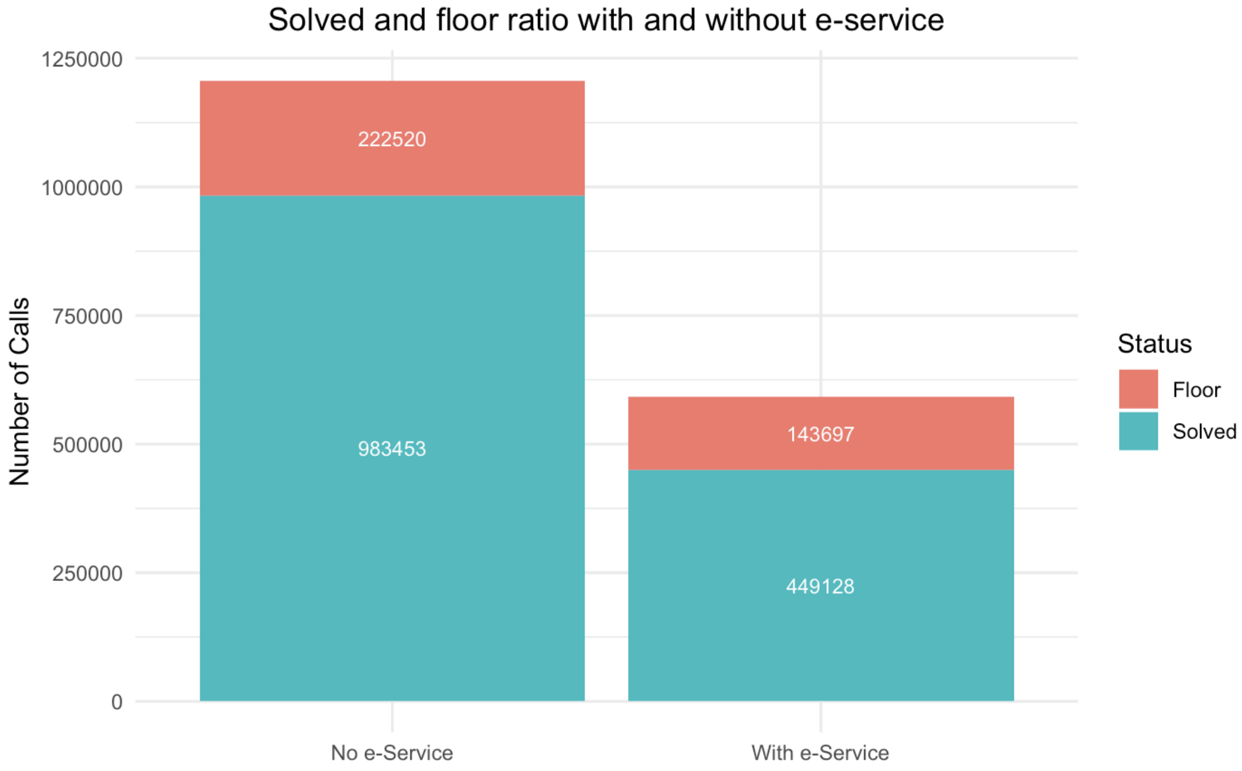
8. Kelsey

eservice_ind_13_march/auto_pay_enrolled_status_13_march; floor, resolved percentage.

```
result <- data %>%
  mutate(combination = interaction(resolved, eservice_ind_13_march)) %>%
  count(combination) %>%
  spread(key = combination, value = n, fill = 0) %>%
  mutate(
    percent_resolved_eservice = `1.1` / (`1.1` + `0.1`) * 100,
    percent_resolved_no_eservice = `1.0` / (`1.0` + `0.0`) * 100,
    percent_floor_eservice = `0.1` / (`1.1` + `0.1`) * 100,
    percent_floor_no_eservice = `0.0` / (`1.0` + `0.0`) * 100
  )
print(result)
```

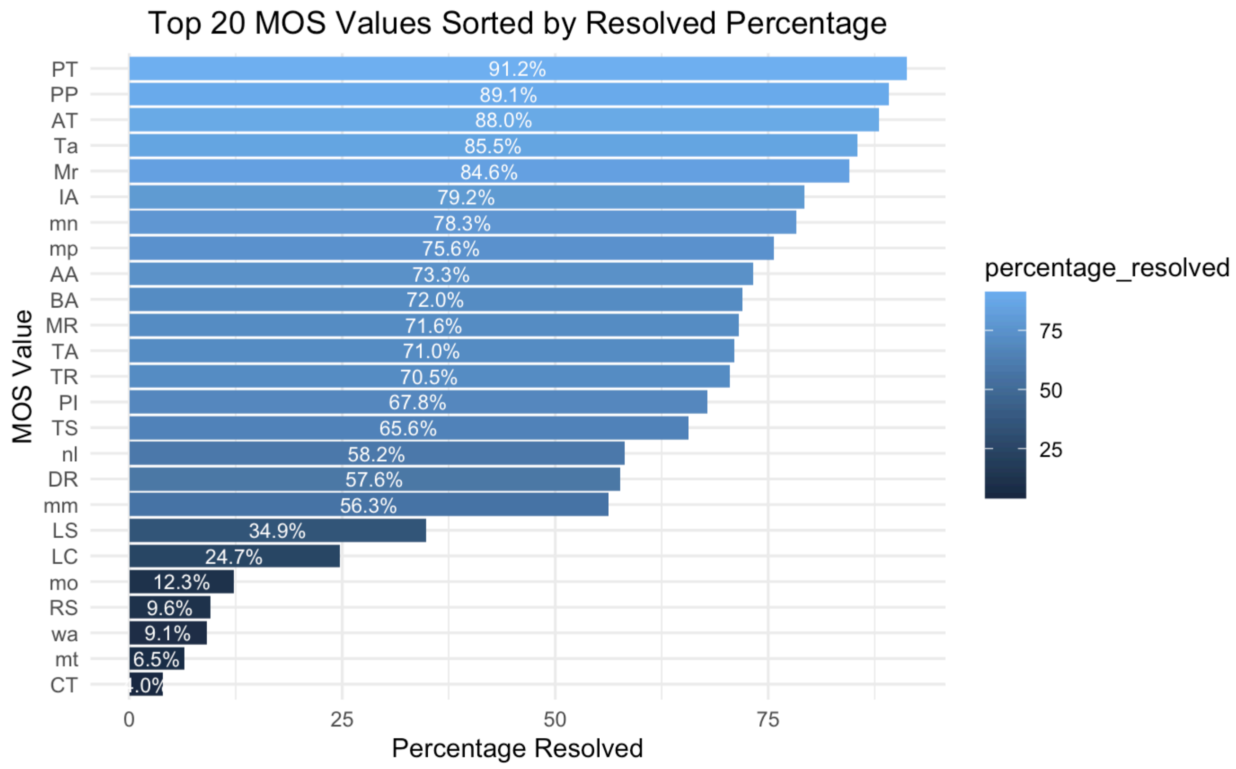
```
# Create a data frame with your data
data <- data.frame(
  e_service = rep(c("With e-Service", "No e-Service"), each = 2),
  status = rep(c("Floor", "Solved"), times = 2),
  calls = c(143697, 449128, 222520, 983453)  # Reordered to match the new status order
)

# Arrange the data to make 'Floor' come on top of 'Solved' in the plot
data$status <- factor(data$status, levels = c("Floor", "Solved"))

# Plot
ggplot(data, aes(x = e_service, y = calls, fill = status)) +
  geom_col(position = "stack") +
  geom_text(aes(label = calls), position = position_stack(vjust = 0.5), color =
"white", size = 3) +
  labs(
    x = NULL,
    y = "Number of Calls",
    fill = "Status",
    title = "Number of Calls With and Without e-Service",
    subtitle = "Comparing Floor and Solved Cases"
  ) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```

percent_resolved_eservice = 0.7576064
percent_resolved_no_eservice = 0.8154851



9. Kelsey
   1. calculate the percentage of resolved calls for each MOS type

2. sort them based on the total number of calls (pick the MOS with top 25 calls)
3. sort them based on the percentage of resolved calls (highest to lowest)



code

```
separated_data <- converted_data %>%
  separate_rows(mos, sep = " ")

percentage_resolved_by_mos <- separated_data %>%
  group_by(mos) %>%
  summarise(
    total_count = n(),  # Total occurrences of each 'mos'
    resolved_count = sum(resolved, na.rm = TRUE),  # Count of resolved cases
for each 'mos'
    floor_count = total_count - resolved_count,
    percentage_resolved = (resolved_count / total_count) * 100  # Calculate the
percentage resolved
  ) %>%
  ungroup()


top_resolved <- percentage_resolved_by_mos %>%
  slice_max(order_by = total_count, n = 25) %>%
  arrange(percentage_resolved)

ggplot(top_resolved, aes(x = reorder(mos, percentage_resolved), y =
percentage_resolved, fill = percentage_resolved)) +
  geom_col() +
  geom_text(aes(label = sprintf("%.1f%%", percentage_resolved)), position =
position_stack(vjust = 0.5), color = "white", size = 3) +
  labs(
    x = "MOS Value",
    y = "Percentage Resolved",
    title = "Top 20 MOS Values Sorted by Resolved Percentage"
  ) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5)) +
  coord_flip()
```

Conclusion:

- CT, mt, wa, RS, mo, LC, LS has much lower resolved percentage than others (<50%) This means we should train the employees to to increase the resolved rate.
    - CT: CIT Change in Terms
    - mt: pre transfer menu
    - wa: Request waiver
    - RS: Global router
    - mo: more menu options
    - LC: Live chat
    - LS: Report lost stolen
- mm, DR, nl, TS, PI, TR, TA, MR, BA, AA have resolved percentage < 75%
    - mm: main menu
    -
- mp, mn, IA, Mr, Ta, AT, PP, PT have resolved percentage > 75%. This means currently for these MOS reasons we have enough knowledge and experience to deal with.