EDA_bank2

William HU ZIHAO

2024-10-08

Step 1 Load Dataset

Step 2 Exploratory Data Analysis

2.1 Fundamental Analysis

```
View(bank)
# Preview the structure of the data
glimpse(bank)
## Rows: 4,521
## Columns: 17
## $ age       <dbl> 30, 33, 35, 30, 59, 35, 36, 39, 41, 43, 39, 43, 36, 20, 31, ~
## $ job       <chr> "unemployed", "services", "management", "management", "blue-~
## $ marital   <chr> "married", "married", "single", "married", "married", "singl~
## $ education <chr> "primary", "secondary", "tertiary", "tertiary", "secondary",~
## $ default   <chr> "no", "no", "no", "no", "no", "no", "no", "no", "no", "no", ~
## $ balance   <dbl> 1787, 4789, 1350, 1476, 0, 747, 307, 147, 221, -88, 9374, 26~
## $ housing   <chr> "no", "yes", "yes", "yes", "yes", "no", "yes", "yes", "yes",~
## $ loan      <chr> "no", "yes", "no", "yes", "no", "no", "no", "no", "no", "yes~
## $ contact   <chr> "cellular", "cellular", "cellular", "unknown", "unknown", "c~
## $ day       <dbl> 19, 11, 16, 3, 5, 23, 14, 6, 14, 17, 20, 17, 13, 30, 29, 29,~
## $ month     <chr> "oct", "may", "apr", "jun", "may", "feb", "may", "may", "may~
## $ duration  <dbl> 79, 220, 185, 199, 226, 141, 341, 151, 57, 313, 273, 113, 32~
## $ campaign  <dbl> 1, 1, 1, 4, 1, 2, 1, 2, 2, 1, 1, 2, 2, 1, 1, 2, 5, 1, 1, 1, ~
## $ pdays     <dbl> -1, 339, 330, -1, -1, 176, 330, -1, -1, 147, -1, -1, -1, -1,~
## $ previous  <dbl> 0, 4, 1, 0, 0, 3, 2, 0, 0, 2, 0, 0, 0, 0, 1, 0, 0, 2, 0, 1, ~
## $ poutcome  <chr> "unknown", "failure", "failure", "unknown", "unknown", "fail~
## $ y         <chr> "no", "no", "no", "no", "no", "no", "no", "no", "no", "no", ~
```

```
# Show summary statistics of the data
summary(bank)
##       age            job              marital           education
##  Min.   :19.00   Length:4521        Length:4521        Length:4521
##  1st Qu.:33.00   Class :character   Class :character   Class :character
##  Median :39.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :41.17
##  3rd Qu.:49.00
##  Max.   :87.00
##    default             balance        housing              loan
##  Length:4521        Min.   :-3313   Length:4521        Length:4521
##  Class :character   1st Qu.:   69   Class :character   Class :character
##  Mode  :character   Median :  444   Mode  :character   Mode  :character
##                     Mean   : 1423
```

```
##                    3rd Qu.: 1480
##                    Max.   :71188
##    contact                day           month              duration
##  Length:4521        Min.   : 1.00   Length:4521        Min.   :   4
##  Class :character   1st Qu.: 9.00   Class :character   1st Qu.: 104
##  Mode  :character   Median :16.00   Mode  :character   Median : 185
##                     Mean   :15.92                      Mean   : 264
##                     3rd Qu.:21.00                      3rd Qu.: 329
##                     Max.   :31.00                      Max.   :3025
##     campaign          pdays           previous         poutcome
##  Min.   : 1.000   Min.   : -1.00   Min.   : 0.0000   Length:4521
##  1st Qu.: 1.000   1st Qu.: -1.00   1st Qu.: 0.0000   Class :character
##  Median : 2.000   Median : -1.00   Median : 0.0000   Mode  :character
##  Mean   : 2.794   Mean   : 39.77   Mean   : 0.5426
##  3rd Qu.: 3.000   3rd Qu.: -1.00   3rd Qu.: 0.0000
##  Max.   :50.000   Max.   :871.00   Max.   :25.0000
##       y
##  Length:4521
##  Class :character
##  Mode  :character
##
##
##
```

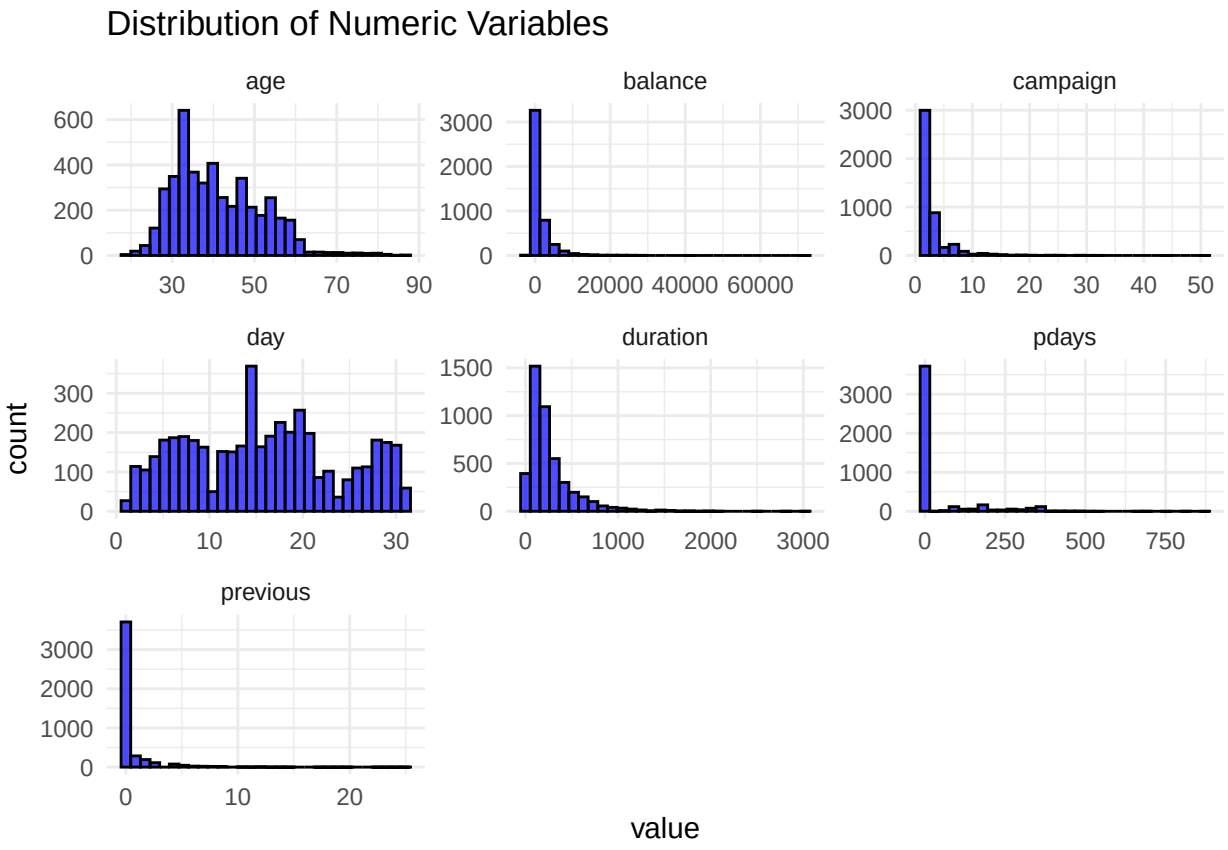
2.2 Distribution of Categorical Variables

```
bank %>%
  select_if(is.character) %>%
  map(~table(.) %>% as.data.frame()) %>%
  imap(~ setNames(.x, c(.y, "Count")))
## $job
##              job Count
## 1         admin.   478
## 2    blue-collar   946
## 3   entrepreneur   168
## 4      housemaid   112
## 5     management   969
## 6        retired   230
## 7  self-employed   183
## 8       services   417
## 9        student    84
## 10    technician   768
## 11    unemployed   128
## 12       unknown    38
##
## $marital
##    marital Count
## 1 divorced   528
## 2  married  2797
## 3   single  1196
##
## $education
##   education Count
```

```
## 1   primary   678
## 2 secondary  2306
## 3  tertiary  1350
## 4   unknown   187
##
## $default
##   default Count
## 1      no  4445
## 2     yes    76
##
## $housing
##   housing Count
## 1      no  1962
## 2     yes  2559
##
## $loan
##   loan Count
## 1   no  3830
## 2  yes   691
##
## $contact
##     contact Count
## 1  cellular  2896
## 2 telephone   301
## 3   unknown  1324
##
## $month
##    month Count
## 1    apr   293
## 2    aug   633
## 3    dec    20
## 4    feb   222
## 5    jan   148
## 6    jul   706
## 7    jun   531
## 8    mar    49
## 9    may  1398
## 10   nov   389
## 11   oct    80
## 12   sep    52
##
## $poutcome
##   poutcome Count
## 1  failure   490
## 2    other   197
## 3  success   129
## 4  unknown  3705
##
## $y
##     y Count
## 1  no  4000
## 2 yes   521
```

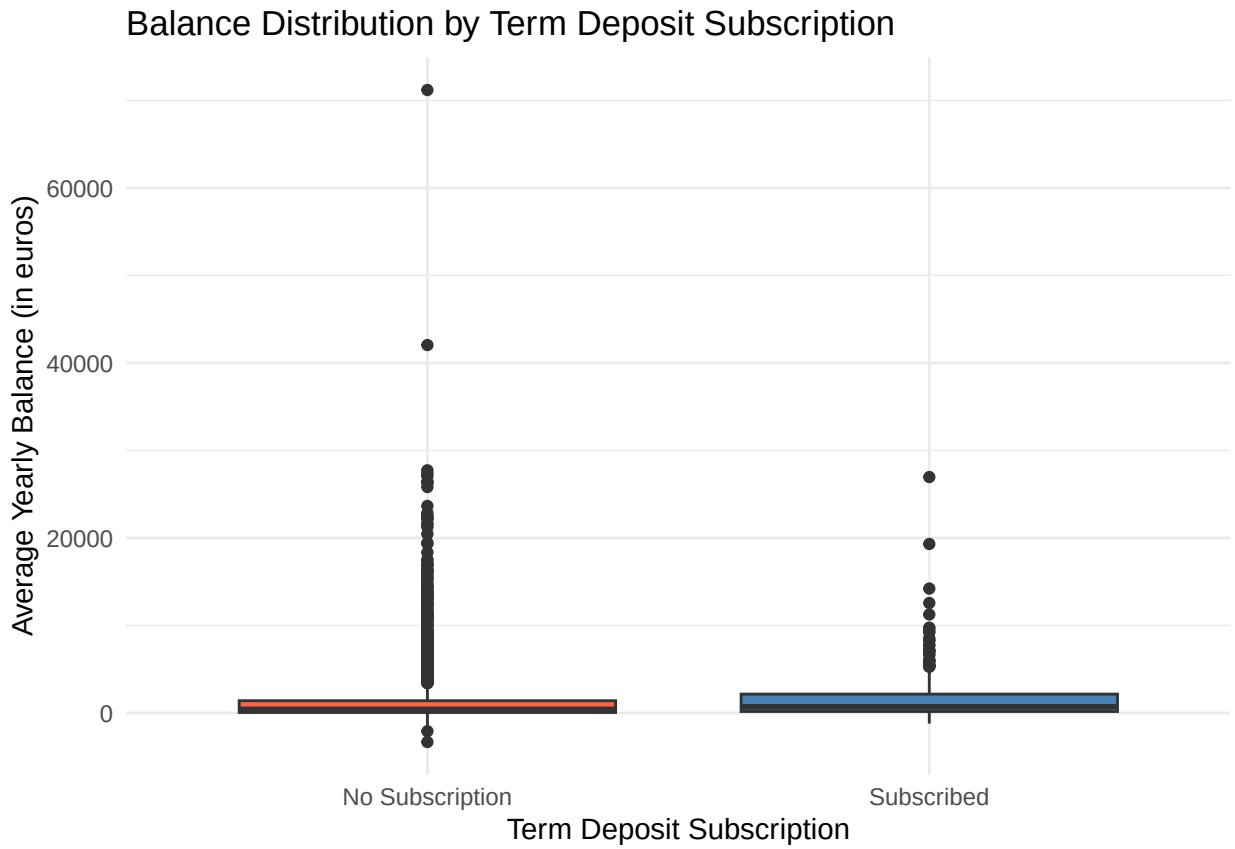
2.3 Distribution of Numerical Variables

```
bank %>%
  select_if(is.numeric) %>%
  gather(key = "variable", value = "value") %>%
  ggplot(aes(x = value)) +
  geom_histogram(bins = 30, fill = "blue", color = "black", alpha = 0.7) +
  facet_wrap(~variable, scales = "free") +
  theme_minimal() +
  labs(title = "Distribution of Numeric Variables")
```



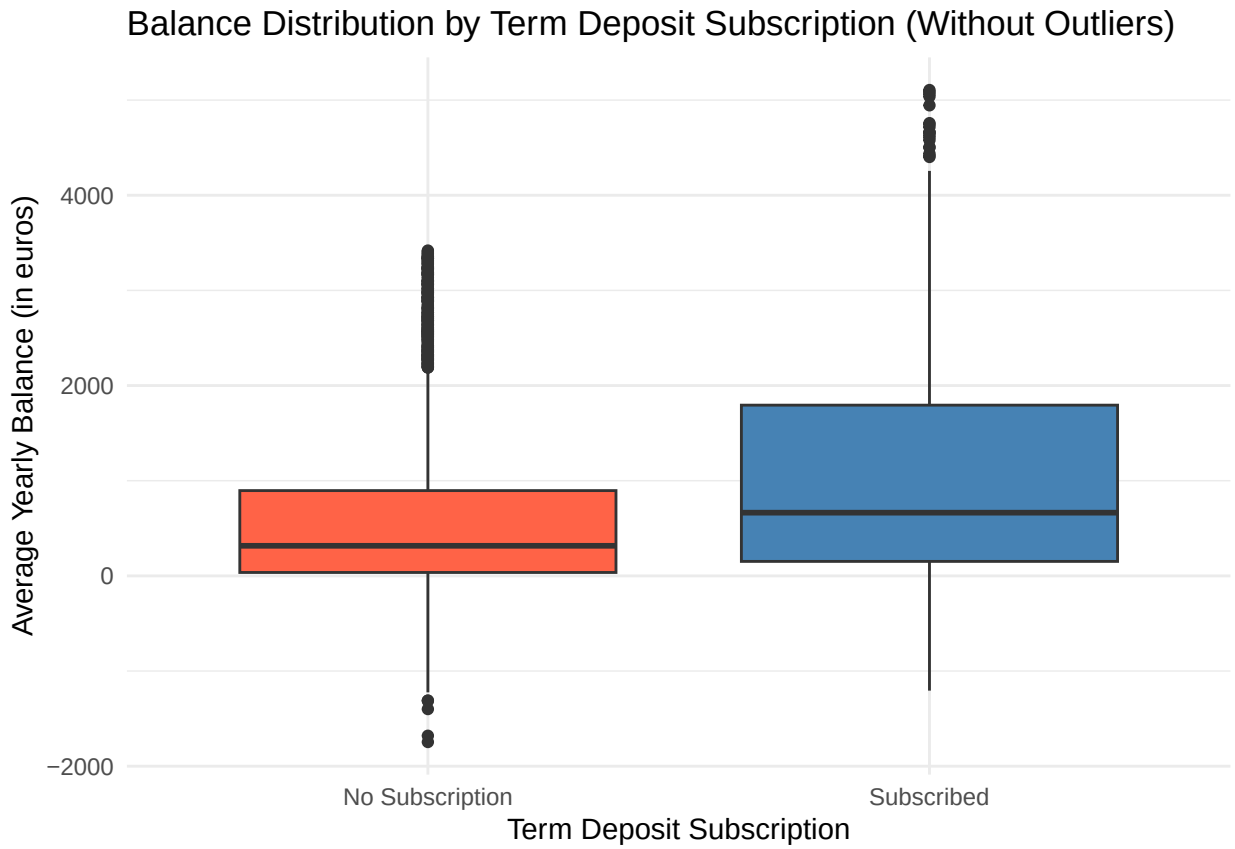
2.4 Balance vs. Subscription

```
# Plot balance distribution using boxplot
bank %>%
  mutate(y = factor(y, levels = c("no", "yes"), labels = c("No Subscription", "Subscribed")))%>% ggplot
  geom_boxplot() +
  labs(title = "Balance Distribution by Term Deposit Subscription",
       x = "Term Deposit Subscription",
       y = "Average Yearly Balance (in euros)") +
  theme_minimal() +
  scale_fill_manual(values = c("No Subscription" = "tomato", "Subscribed" = "steelblue")) +
  theme(legend.position = "none")
```



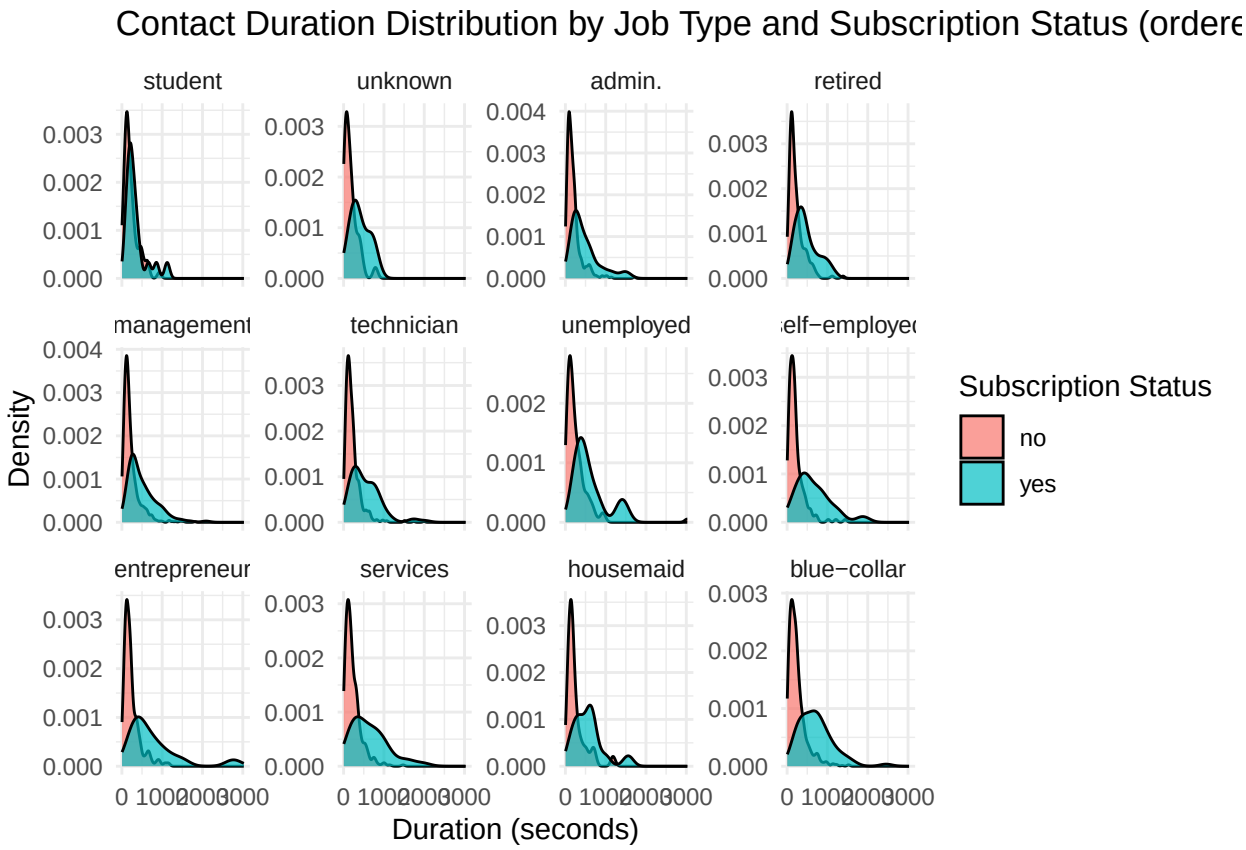
Notice that the visulizations are greatly influenced by the outliers. Thus, we can consider removing outliers in two subscription groups for a more direct visulizations of balance versus two subscription groups.

```
remove_outliers <- function(data) {
  Q1 <- quantile(data$balance, 0.25, na.rm = TRUE)
  Q3 <- quantile(data$balance, 0.75, na.rm = TRUE)
  IQR <- Q3 - Q1
  lower_bound <- Q1 - 1.5 * IQR
  upper_bound <- Q3 + 1.5 * IQR
  data %>%
    filter(balance >= lower_bound & balance <= upper_bound)
}

bank_filtered <- bank %>%
  mutate(y = factor(y, levels = c("no", "yes"), labels = c("No Subscription", "Subscribed"))) %>%
  group_by(y) %>%
  group_modify(~ remove_outliers(.x)) %>%
  ungroup()

ggplot(bank_filtered, aes(x = y, y = balance, fill = y)) +
  geom_boxplot() +
  labs(title = "Balance Distribution by Term Deposit Subscription (Without Outliers)",
       x = "Term Deposit Subscription",
       y = "Average Yearly Balance (in euros)") +
  theme_minimal() +
  scale_fill_manual(values = c("No Subscription" = "tomato", "Subscribed" = "steelblue")) +
  theme(legend.position = "none")
```



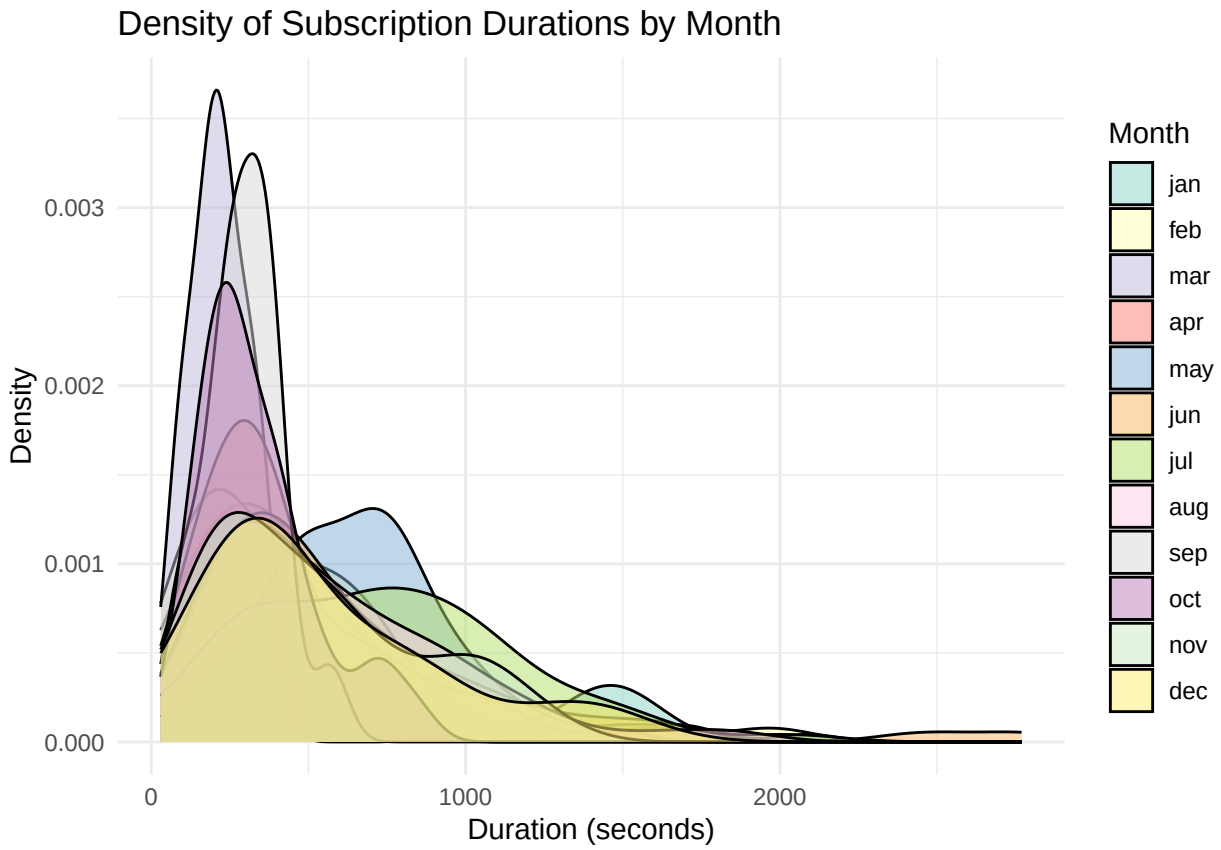
2.5 Density Plot of Contact Duration by subscription behavior of different job types

```
# Step 1: Calculate the median duration for 'yes' subscription by job
medians_sub <- bank %>%
  filter(y == "yes") %>%
  group_by(job) %>%
  summarize(median_duration = median(duration, na.rm = TRUE), .groups = 'drop')

# Step 2: Order jobs based on median duration for 'yes' subscription
ordered_jobs <- medians_sub$job[order(medians_sub$median_duration)]

ggplot(bank, aes(x = duration, fill = y)) +
  geom_density(alpha = 0.7) +
  facet_wrap(~factor(job,levels = ordered_jobs), scales = "free_y") +
  labs(title = "Contact Duration Distribution by Job Type and Subscription Status (ordered by median du
  theme_minimal()
```

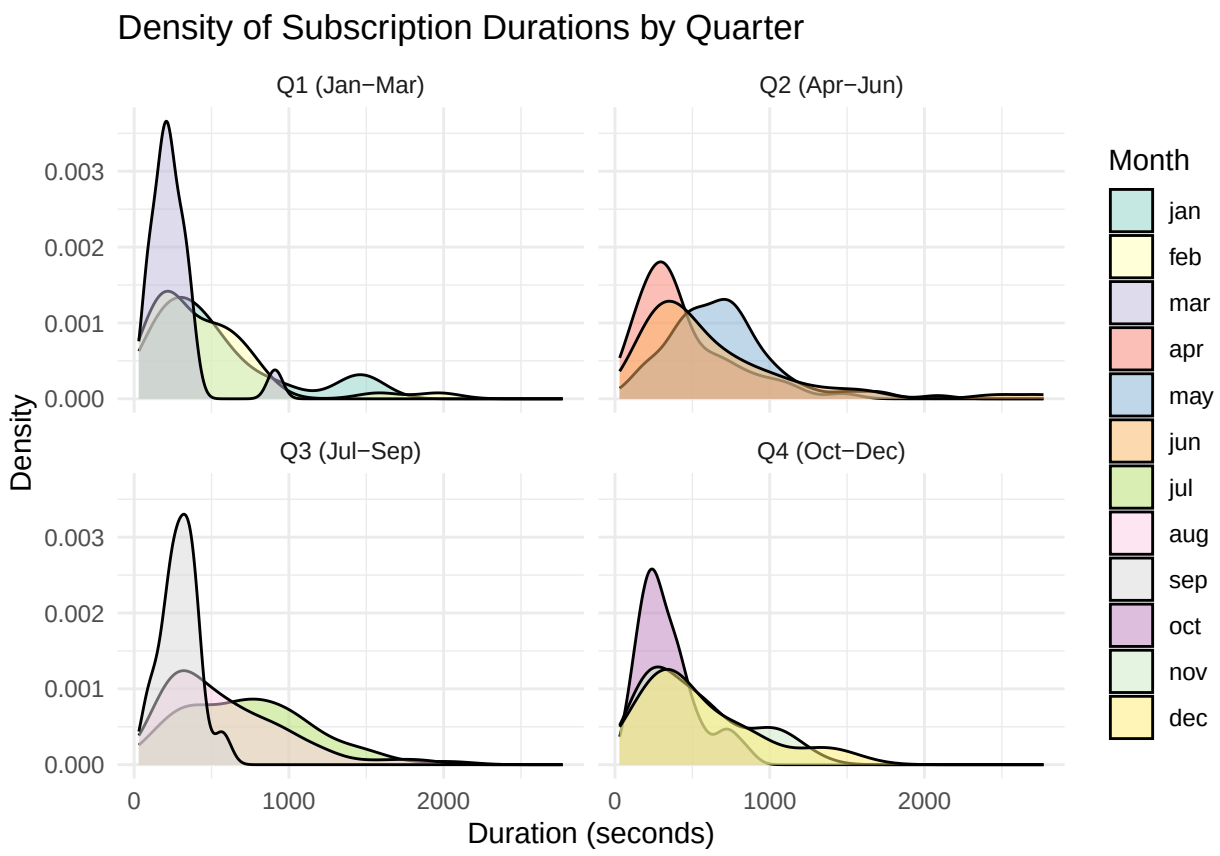


2.6 Density Plot of Duration of Subscribers by Month

```
subscriptions <- bank %>%
    filter(y == "yes") %>%
    mutate(month = factor(month,
                          levels = c("jan", "feb", "mar", "apr", "may", "jun", "jul", "aug", "sep", "oc


# Create the density plot
ggplot(subscriptions, aes(x = duration, fill = month)) +
    geom_density(alpha = 0.5) +  # Density plot with transparency
    labs(title = "Density of Subscription Durations by Month",
         x = "Duration (seconds)",
         y = "Density",
         fill = "Month") +
    theme_minimal() +
    theme(legend.position = "right") +
    scale_fill_brewer(palette = "Set3")
```



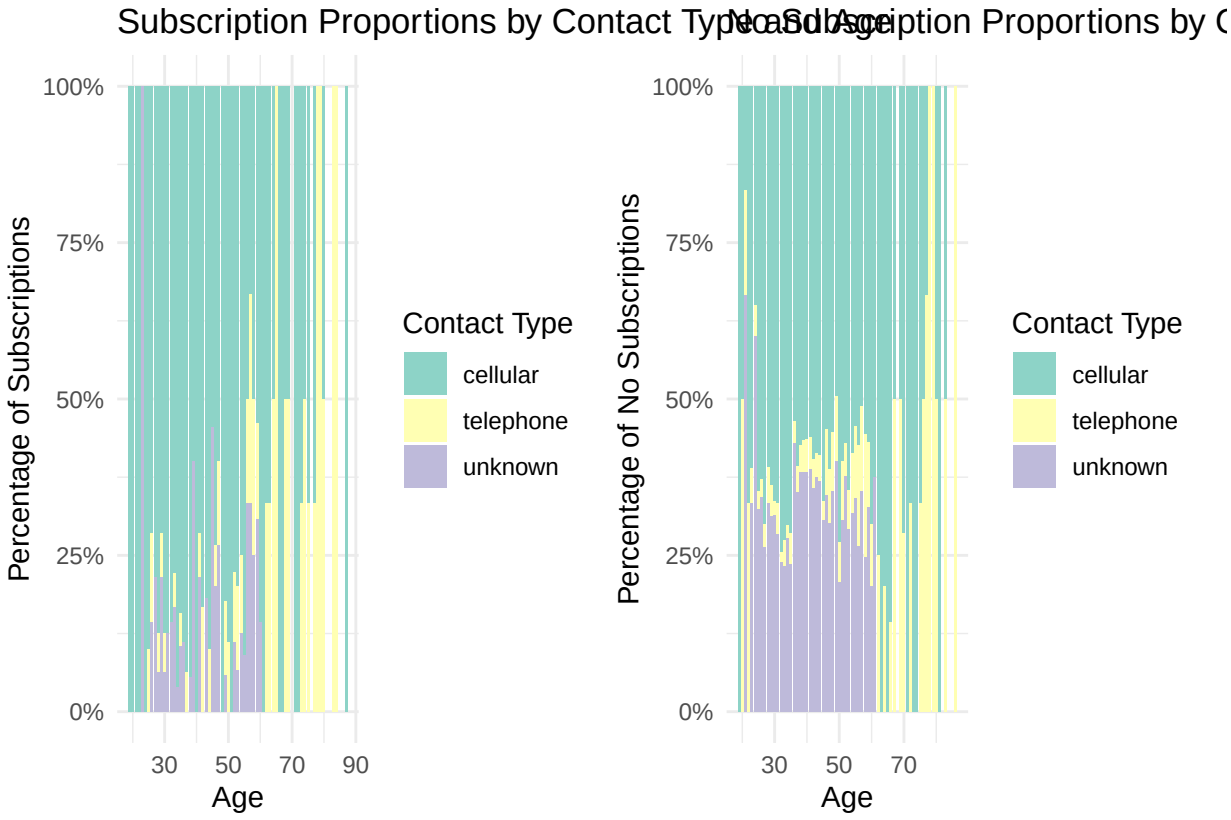

```
subscriptions <- bank %>%
  filter(y == "yes") %>%
  mutate(month = factor(month,
                        levels = c("jan", "feb", "mar", "apr", "may", "jun", "jul", "aug", "sep", "oct"
         quarter = case_when(
           month %in% c("jan", "feb", "mar") ~ "Q1 (Jan-Mar)",
           month %in% c("apr", "may", "jun") ~ "Q2 (Apr-Jun)",
           month %in% c("jul", "aug", "sep") ~ "Q3 (Jul-Sep)",
           month %in% c("oct", "nov", "dec") ~ "Q4 (Oct-Dec)"
         ))

# Create the density plot with facets for each quarter
ggplot(subscriptions, aes(x = duration, fill = month)) +
  geom_density(alpha = 0.5) +
  labs(title = "Density of Subscription Durations by Quarter",
       x = "Duration (seconds)",
       y = "Density",
       fill = "Month") +
  theme_minimal() +
  theme(legend.position = "right") +
  scale_fill_brewer(palette = "Set3") +
  facet_wrap(~quarter)
```



2.7 (No) Subscriptions by Contact and Age

```
library(patchwork)
## Warning: package 'patchwork' was built under R version 4.3.3
sub_percentage_contact<-bank %>%
  filter(y == "yes") %>%
  group_by(age, contact) %>%
  summarise(count = n()) %>%
  ungroup() %>%
  group_by(age) %>%
  mutate(percentage = count / sum(count)) %>%
  ungroup()%>%
# Create the stacked bar plot (3D-like view) with X-axis as percentage, Y-axis as age, and fill by cont
ggplot(aes(x = age, y = percentage, fill = contact)) +
  geom_bar(stat = "identity", position = "fill") +
  scale_y_continuous(labels = scales::percent_format()) +
  labs(title = "Subscription Proportions by Contact Type and Age",
       x = "Age",
       y = "Percentage of Subscriptions",
       fill = "Contact Type") +
  theme_minimal() +
  theme(legend.position = "right") +
  scale_fill_brewer(palette = "Set3")
## `summarise()` has grouped output by 'age'. You can override using the `.groups`
## argument.
```

```
no_sub_percentage_contact<-bank %>%
  filter(y == "no") %>%
  group_by(age, contact) %>%
  summarise(count = n()) %>%
  ungroup() %>%
  group_by(age) %>%
  mutate(percentage = count / sum(count)) %>%
  ungroup()%>%
# Create the stacked bar plot (3D-like view) with X-axis as percentage, Y-axis as age, and fill by cont
ggplot(aes(x = age, y = percentage, fill = contact)) +
  geom_bar(stat = "identity", position = "fill") +
  scale_y_continuous(labels = scales::percent_format()) +
  labs(title = "No Subscription Proportions by Contact Type and Age",
       x = "Age",
       y = "Percentage of No Subscriptions",
       fill = "Contact Type") +
  theme_minimal() +
  theme(legend.position = "right") +
  scale_fill_brewer(palette = "Set3")
## `summarise()` has grouped output by 'age'. You can override using the `.groups`
## argument.

sub_percentage_contact+no_sub_percentage_contact
```



Step 3 Conclusions

3.1 Average Yearly Balance by Subscription Group

Code & Results are in 2.4 Balance vs. Subscription

- **Balance Discrepancy:** Clients who subscribed to a term deposit generally have a higher balance compared to those who did not subscribe. The wider IQR in the subscribed group indicates more variability in the financial status of these clients.
- **Outliers**: While there are extreme cases (outliers) in both groups, the extreme balance values of non-subscribers' group which stretch beyond 20000 euros indicate that while the majority of non-subscribers have low balances, a small number of them have much higher balances than subscribers.
- **Non-subscribers**
  - **Median Balance**: The median is slightly above zero, indicating that for most clients who did not subscribe, their average balance is quite low.
  - **Spread of Balance**: The interquartile range (IQR) is narrow, suggesting that most clients in this group have balances concentrated around 0-1000 euros.
- **Subscribers**
  - **Median Balance**: The median is higher than that of the non-subscribers, around 600 euros, suggesting that clients who subscribed tend to have higher balances.
  - **Spread of Balance**: The IQR is wider compared to the non-subscribers, indicating a more varied distribution of balances in this group.

3.2 Contact Contact Duration by subscription behavior of different job types

Code & Results are in 2.5 Density Plot of Contact Duration by subscription behavior of different job types

- **General Observation**: Across most job types, longer contact durations are associated with a higher likelihood of subscription to the term deposit. This is evident by the increased density of subscribers (in teal) for longer durations, whereas non-subscribers (in red) peak at shorter durations.
- **Job-specific Insights**:
  - **Admin, Blue-collar, Entrepreneur, Housemaid, and Services**: These jobs show a strong trend where short contact durations lead to more non-subscriptions. The density of subscribers rises significantly as the duration increases.
  - **Management, Retired, Self-employed, and Unemployed**: Similarly, there is a noticeable increase in subscription rates as contact duration increases. However, these groups also show some variation, with more balanced densities at certain durations.
- **Short Durations**: For nearly all job types, very short contact durations (under 100 seconds) are predominantly associated with non-subscription.

3.3 Duration of Subscribers by Month

Code & Results are in 2.6 Density Plot of Duration of Subscribers by Month

- **Seasonal Trend of Engagement of Subscribers:**
  - **High Engagement during Q2 and Q3**
    - From density plot of duration of subscribers by quarter, it can be concluded that the campaigns usually experience **high engagements during Q2 and Q3** which could be reflected by higher median and mean of contact duration

    - Understanding that Q2 and Q3 lead to longer subscription durations can guide marketing efforts. Resources can be allocated to intensify campaigns during those times to maximize subscriptions.

  - **Low Engagement during Q1 and Q4**

    - **Q1**: The lower durations in Q1 could suggest that new year resolutions or fresh starts motivate customers to subscribe, but their engagement might not be sustained, leading to shorter subscription durations.

    - **Q4**: In Q4, the lower durations could be influenced by year-end behaviors, holiday distractions, or other seasonal factors that might lead to shorter subscription periods as customers may not prioritize subscription activities during festive seasons.

3.4 (No) Subscriptions by Contact and Age

Code & Results are in 2.7 (No) Subscriptions by Contact and Age

Subscription Proportions by Contact Type and Age

- **Cellular (teal)**:
  - Cellular subscriptions dominate across all age groups, especially for younger people (under 50). The proportion remains high even in older age groups, though it slightly decreases as age increases.
- **Telephone (yellow)**:
  - Telephone subscriptions become more prominent in people aged 50 and above, especially among older individuals, indicating that traditional landlines are more commonly used by older populations.
- **Unknown (purple)**:
  - The "unknown" category shows a slight increase in the middle-aged to older groups (around 50 to 70). However, its overall proportion is low compared to cellular and telephone subscriptions.

No Subscription Proportions by Contact Type and Age

- **Cellular (teal)**:
  - The "No Subscription" plot shows that even among those without subscriptions, cellular contact information remains the most common contact type across all age ranges.
- **Telephone (yellow)**:
  - Landline (telephone) contacts without subscriptions become increasingly prominent as age increases, particularly in those over 50, with a slight peak at around 70 years.
- **Unknown (purple)**:
  - The proportion of "unknown" contact types rises significantly for people without subscriptions in middle-aged (40 to 60) and older groups, but it's relatively rare in younger age groups.