EDA_bank

William HU ZIHAO

2024-10-04

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
##
##
##
```

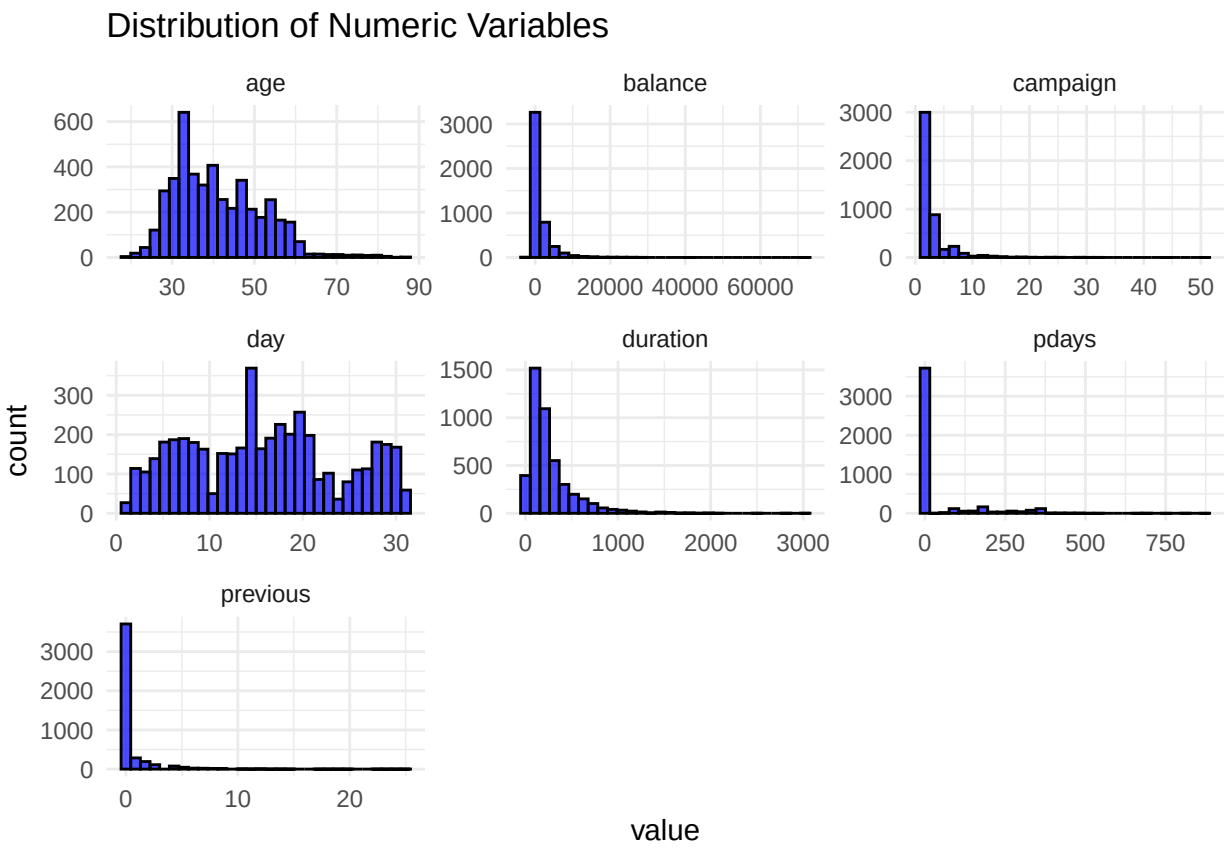
2.2 Distribution of Categorical Variables

```
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")
```

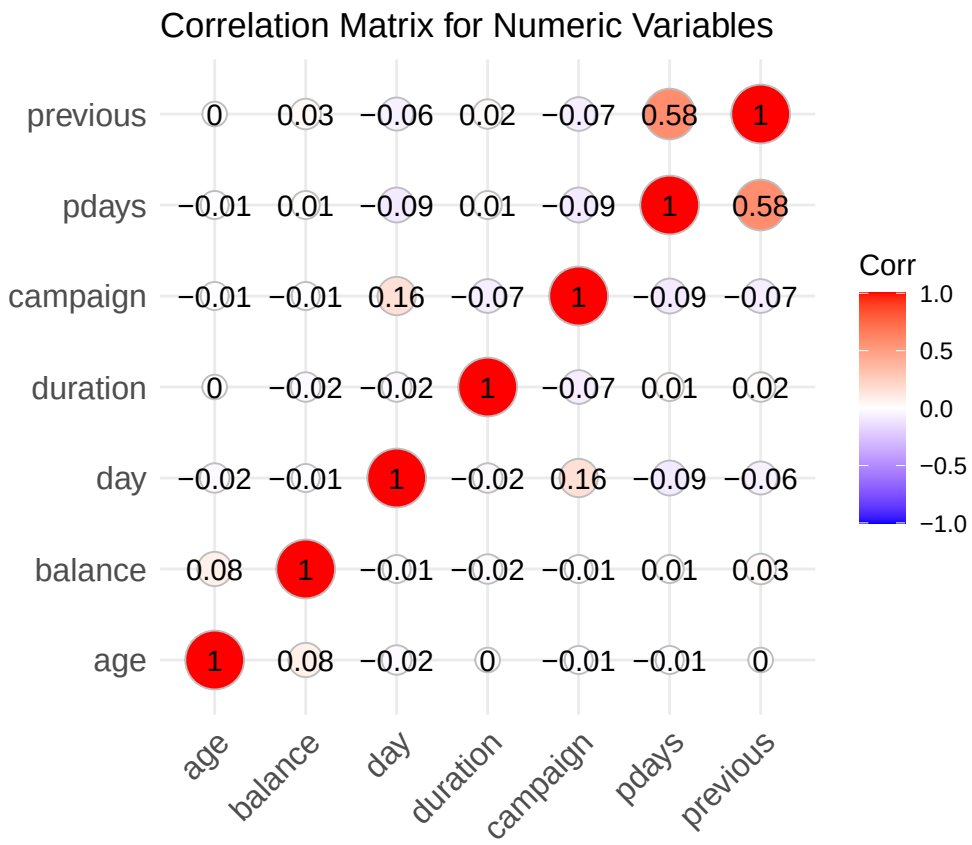


2.4 Correlations between Numerical Variables

```
numeric_vars <- bank %>% select_if(is.numeric)
cor_matrix <- cor(numeric_vars, use = "complete.obs")
cor_matrix
##                  age      balance          day     duration     campaign
## age       1.000000000  0.083820142 -0.017852632 -0.002366889 -0.005147905
## balance   0.083820142  1.000000000 -0.008677052 -0.015949918 -0.009976166
## day      -0.017852632 -0.008677052  1.000000000 -0.024629306  0.160706069
## duration -0.002366889 -0.015949918 -0.024629306  1.000000000 -0.068382000
## campaign -0.005147905 -0.009976166  0.160706069 -0.068382000  1.000000000
## pdays    -0.008893530  0.009436676 -0.094351520  0.010380242 -0.093136818
## previous -0.003510917  0.026196357 -0.059114394  0.018080317 -0.067832630
##                 pdays     previous
## age      -0.008893530 -0.003510917
## balance   0.009436676  0.026196357
```

```
## day      -0.094351520 -0.059114394
## duration  0.010380242  0.018080317
## campaign -0.093136818 -0.067832630
## pdays     1.000000000  0.577561827
## previous  0.577561827  1.000000000
```

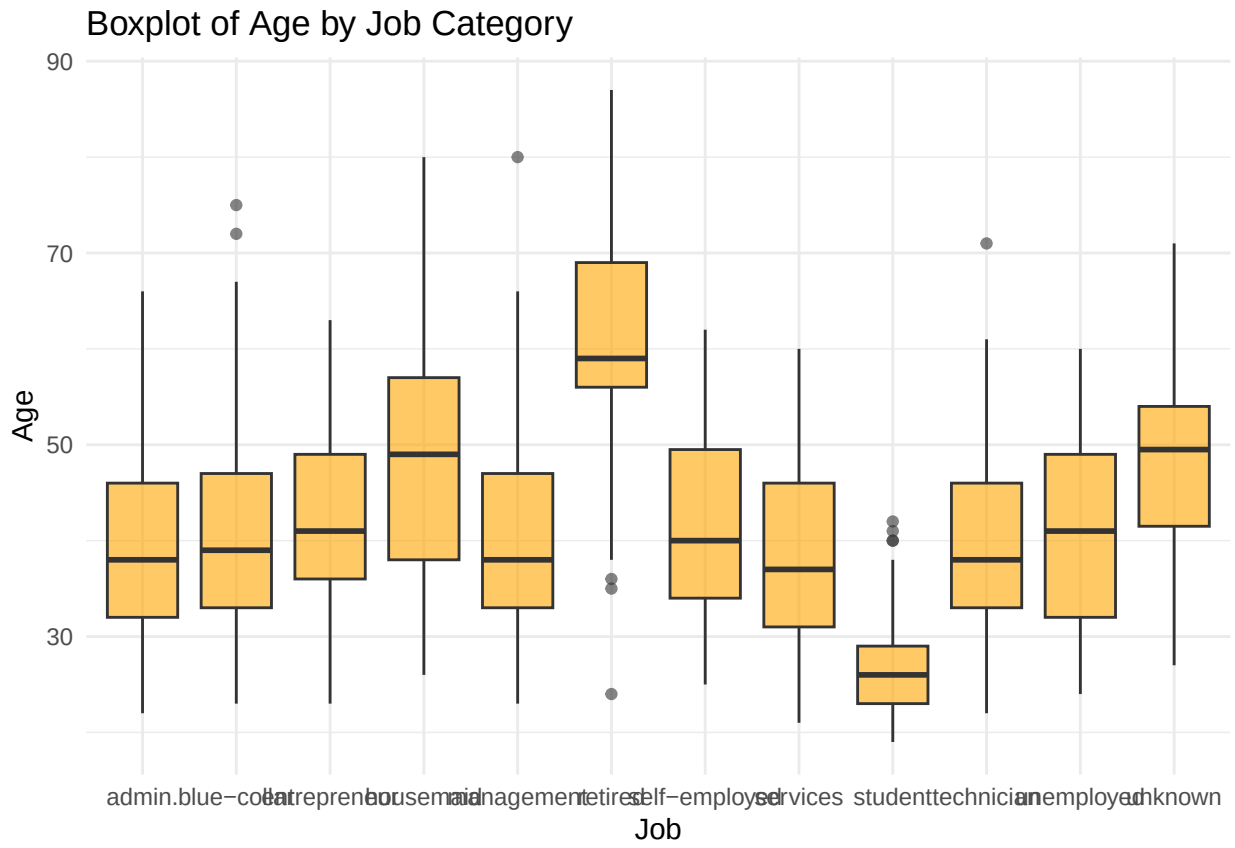
```
# Plot a heatmap of the correlation matrix
library(ggcorrplot)
ggcorrplot(cor_matrix, method = "circle", lab = TRUE) +
  labs(title = "Correlation Matrix for Numeric Variables")
```



2.5 Potential Relationship between Variables

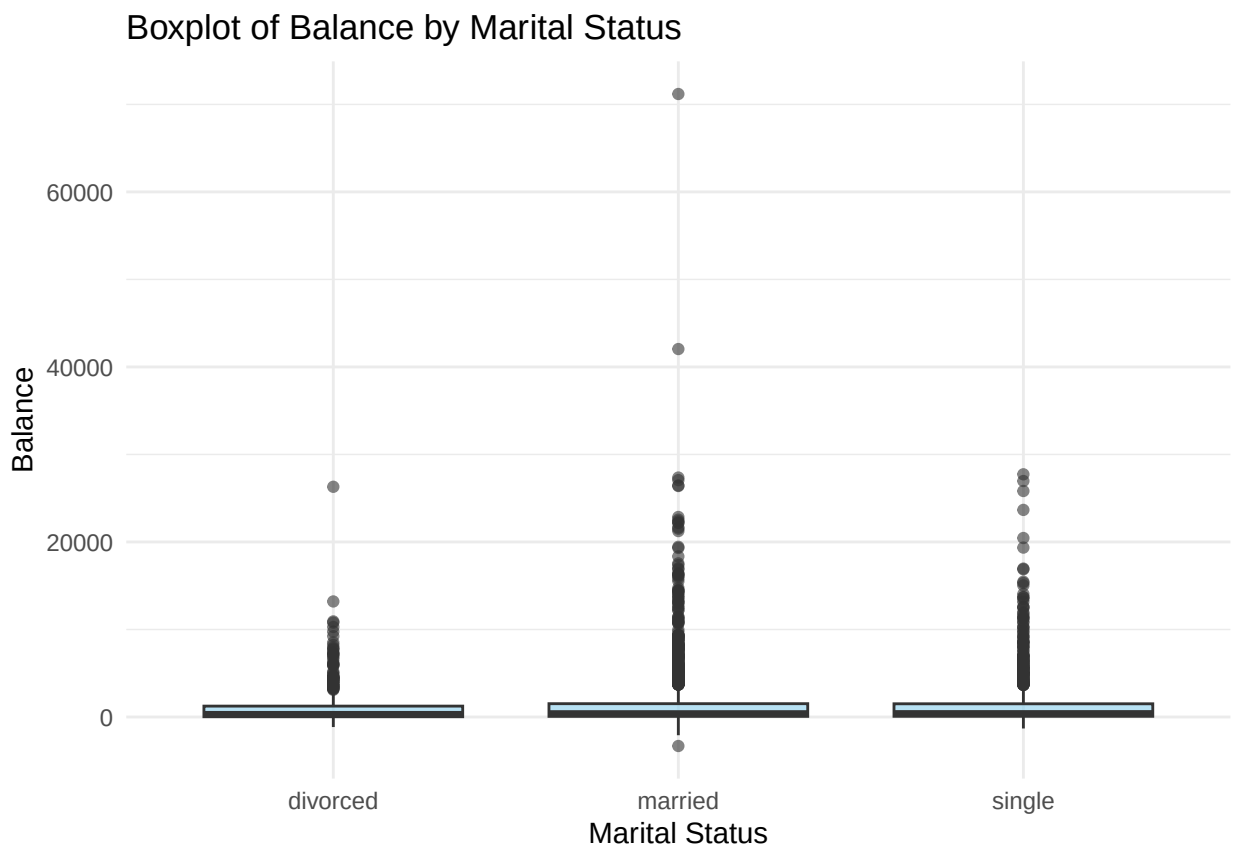
```
ggplot(bank, aes(x = job, y = age)) +
  geom_boxplot(fill = "orange", alpha = 0.6) +
  theme_minimal() +
  labs(title = "Boxplot of Age by Job Category", x = "Job", y = "Age")
```

2.5.1 Job vs. Age



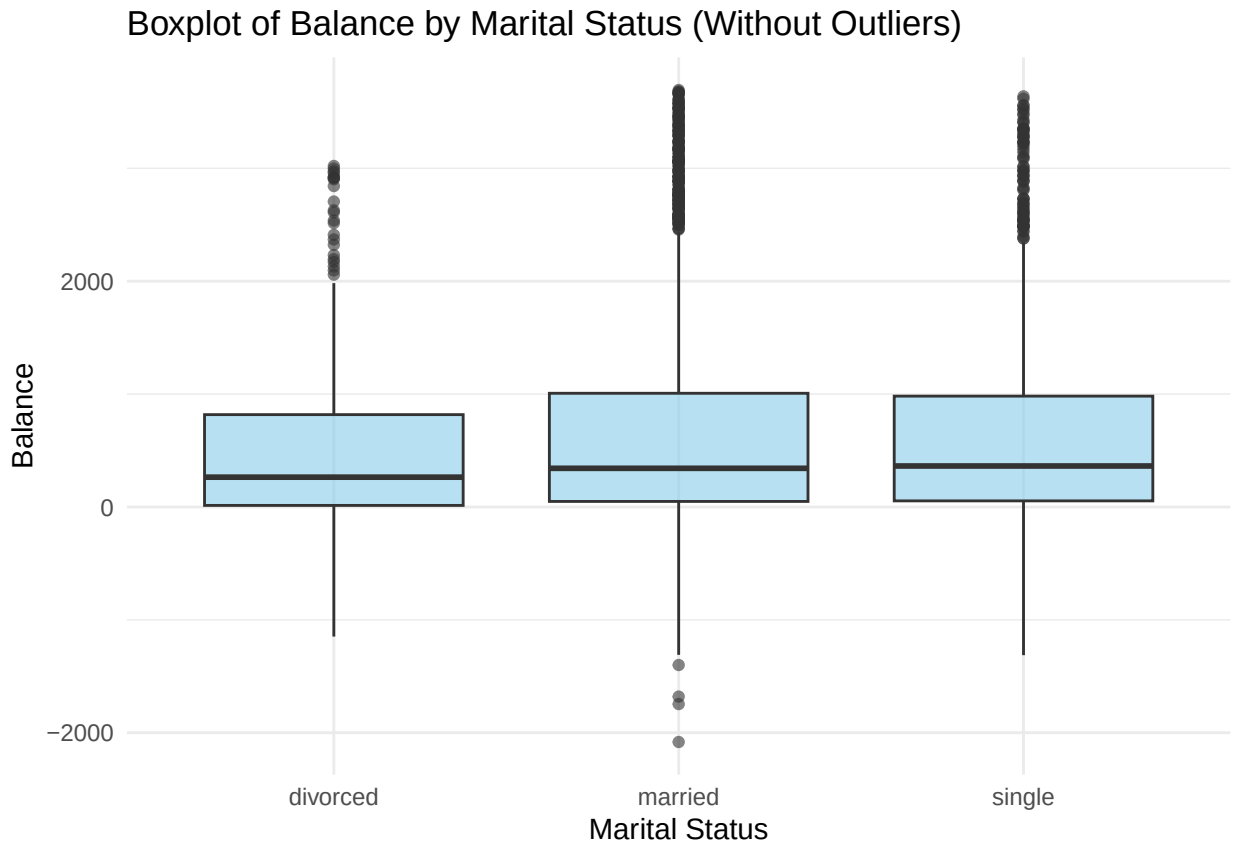

```
ggplot(bank, aes(x = marital, y = balance)) +
  geom_boxplot(fill = "skyblue", alpha = 0.6) +
  theme_minimal() +
  labs(title = "Boxplot of Balance by Marital Status", x = "Marital Status", y = "Balance")
```

2.5.2 Martial vs. Balance



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

```
# Identify outliers by marital status using IQR
bank_no_outliers <- bank %>%
  group_by(marital) %>%
  mutate(IQR = IQR(balance, na.rm = TRUE),
         Q1 = quantile(balance, 0.25, na.rm = TRUE),
         Q3 = quantile(balance, 0.75, na.rm = TRUE)) %>%
  filter(balance >= (Q1 - 1.5 * IQR) & balance <= (Q3 + 1.5 * IQR)) %>%
  ungroup()  # Remove grouping to get the cleaned dataset


# Boxplot of balance by marital status after removing outliers
ggplot(bank_no_outliers, aes(x = marital, y = balance)) +
  geom_boxplot(fill = "skyblue", alpha = 0.6) +
  theme_minimal() +
  labs(title = "Boxplot of Balance by Marital Status (Without Outliers)",
       x = "Marital Status", y = "Balance")
```



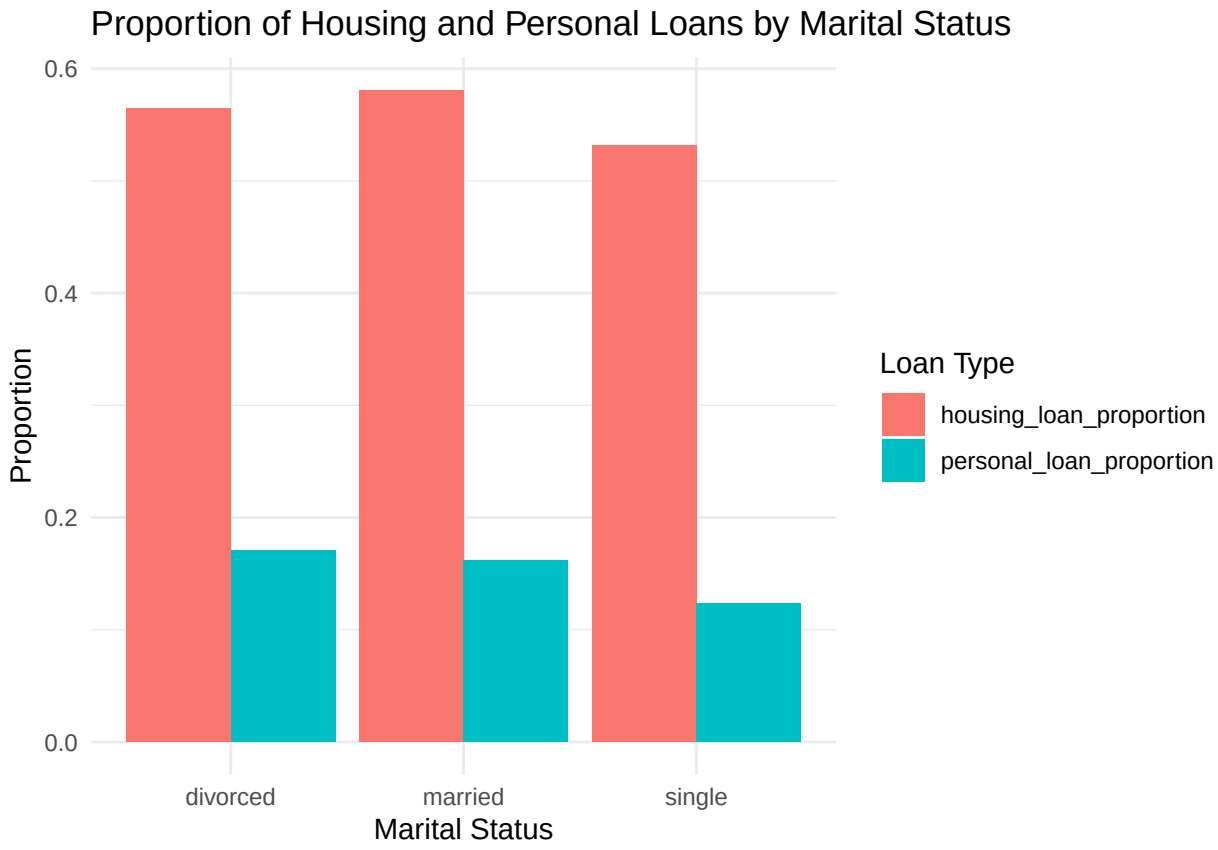

```
# Summarize average and median balance by marital status
summary_stats <- bank_no_outliers %>%
  group_by(marital) %>%
  summarise(
    avg_balance = mean(balance, na.rm = TRUE),
    median_balance = median(balance, na.rm = TRUE)
  )

print(summary_stats)
## # A tibble: 3 x 3
##   marital  avg_balance median_balance
##   <chr>          <dbl>          <dbl>
## 1 divorced        510.           264
## 2 married         664.           342.
## 3 single          670.           363
```

```
housing_loan_stats <- bank %>%
  group_by(marital) %>%
  summarise(housing_loan_yes = sum(housing == "yes", na.rm = TRUE),
            housing_loan_no = sum(housing == "no", na.rm = TRUE),
            total = n()) %>%
  mutate(housing_loan_proportion = housing_loan_yes / total)

personal_loan_stats <- bank %>%
  group_by(marital) %>%
  summarise(personal_loan_yes = sum(loan == "yes", na.rm = TRUE),
```

```
            personal_loan_no = sum(loan == "no", na.rm = TRUE),
            total = n()) %>%
  mutate(personal_loan_proportion = personal_loan_yes / total)

combined_stats <- housing_loan_stats %>%
  select(marital, housing_loan_proportion) %>%
  left_join(personal_loan_stats %>% select(marital, personal_loan_proportion), by = "marital")

print(combined_stats)
## # A tibble: 3 x 3
##   marital  housing_loan_proportion personal_loan_proportion
##   <chr>                      <dbl>                    <dbl>
## 1 divorced                   0.564                    0.170
## 2 married                    0.581                    0.162
## 3 single                     0.532                    0.124
```

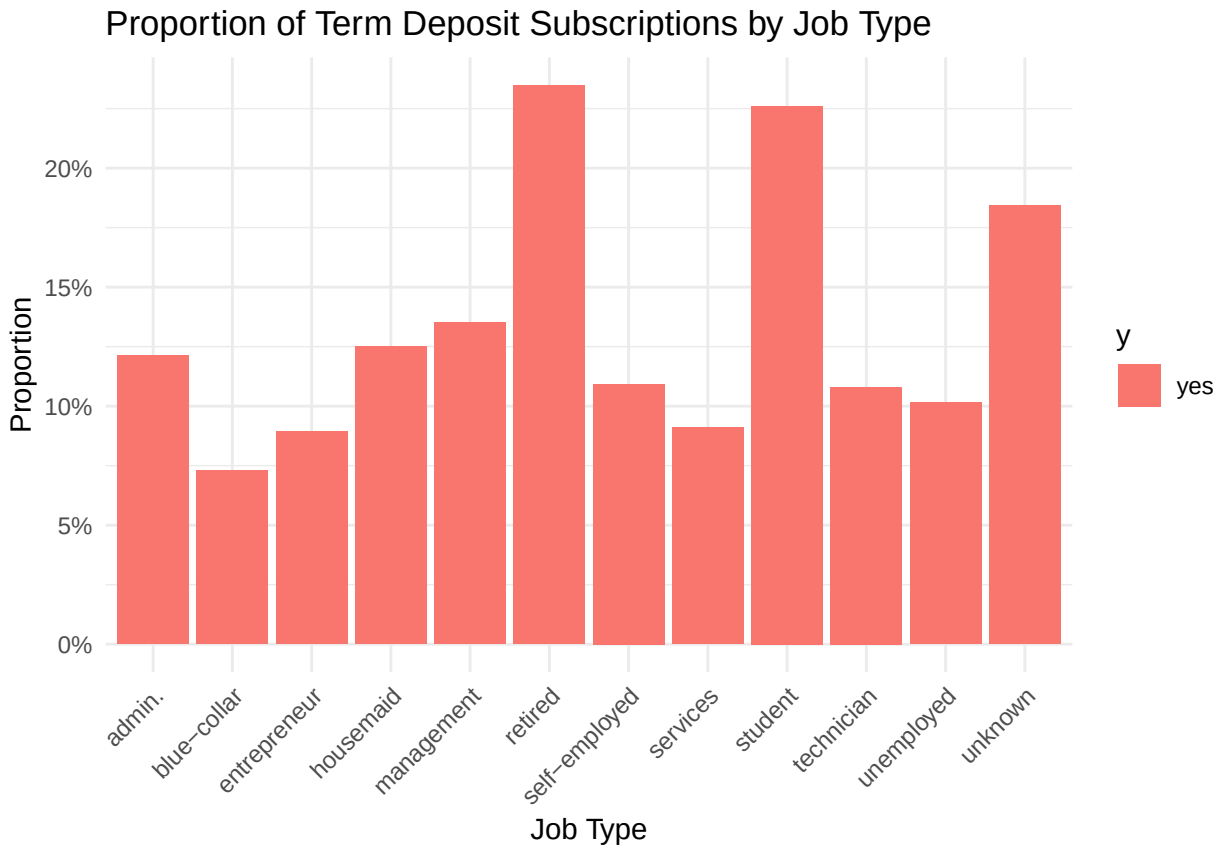
```
# Reshape data for visualization
combined_long <- combined_stats %>%
  pivot_longer(cols = c(housing_loan_proportion, personal_loan_proportion),
               names_to = "loan_type", values_to = "proportion")

# Bar plot for housing and personal loans by marital status
ggplot(combined_long, aes(x = marital, y = proportion, fill = loan_type)) +
  geom_bar(stat = "identity", position = "dodge") +
  theme_minimal() +
  labs(title = "Proportion of Housing and Personal Loans by Marital Status",
       x = "Marital Status",
       y = "Proportion",
       fill = "Loan Type")
```

2.5.3 Martial vs. Housing & Personal Loan




```
job_proportions <- bank %>%
  group_by(job, y) %>%
  summarise(count = n()) %>%
  mutate(ratio = count / sum(count)) %>%
  filter(y == "yes")
## `summarise()` has grouped output by 'job'. You can override using the `.groups`
## argument.

# Create a bar plot for job type proportions
ggplot(job_proportions, aes(x = job, y = ratio, fill = y)) +
  geom_bar(stat = "identity") +  # Use stat = "identity" since we provide y values
  theme_minimal() +
  labs(title = "Proportion of Term Deposit Subscriptions by Job Type", x = "Job Type", y = "Proportion"
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_y_continuous(labels = scales::percent)  # Format y-axis as percentages
```

2.5.4 Job vs. Subscription (y)




```
# Exclude job type 'unknown'
job_subscription_rates<-job_proportions %>% filter(job!='unknown') %>% select(job,ratio) %>% arrange(des

job_subscription_rates
## # A tibble: 11 x 2
## # Groups:   job [11]
##    job           ratio
##    <chr>         <dbl>
##  1 retired       0.235
##  2 student       0.226
##  3 management    0.135
##  4 housemaid     0.125
##  5 admin.        0.121
##  6 self-employed 0.109
##  7 technician    0.108
##  8 unemployed    0.102
##  9 services      0.0911
## 10 entrepreneur  0.0893
## 11 blue-collar   0.0729
```

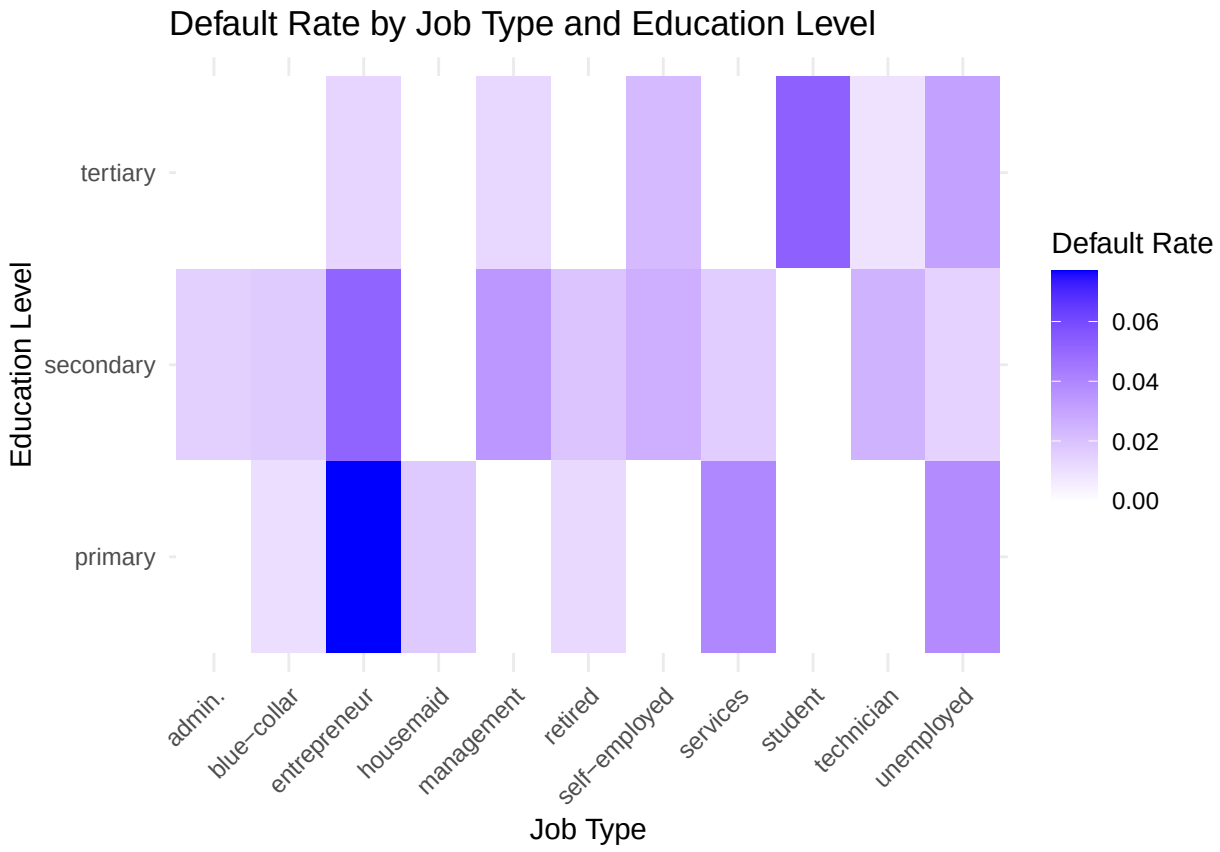
```
head(job_subscription_rates,3)
## # A tibble: 3 x 2
## # Groups:   job [3]
##   job        ratio
##   <chr>      <dbl>
## 1 retired    0.235
## 2 student    0.226
```

```
## 3 management 0.135
tail(job_subscription_rates,3)
## # A tibble: 3 x 2
## # Groups:   job [3]
##   job           ratio
##   <chr>         <dbl>
## 1 services     0.0911
## 2 entrepreneur 0.0893
## 3 blue-collar  0.0729
```

```
combined_insight <- bank %>%
  filter(job != "unknown", education != "unknown") %>%  # Exclude unknowns
  group_by(job, education) %>%
  summarise(
    total = n(),
    default_count = sum(default == "yes"),
    default_rate = default_count / total
  ) %>%
  arrange(desc(default_rate))
## `summarise()` has grouped output by 'job'. You can override using the `.groups`
## argument.

# Print the combined insight
print(combined_insight)
## # A tibble: 33 x 5
## # Groups:   job [11]
##    job           education total default_count default_rate
##    <chr>         <chr>     <int>         <int>        <dbl>
##  1 entrepreneur  primary      26             2       0.0769
##  2 student       tertiary     19             1       0.0526
##  3 entrepreneur  secondary    58             3       0.0517
##  4 services      primary      25             1       0.04
##  5 unemployed    primary      26             1       0.0385
##  6 management    secondary   116             4       0.0345
##  7 unemployed    tertiary     32             1       0.0312
##  8 self-employed secondary    76             2       0.0263
##  9 technician    secondary   520            13       0.025
## 10 self-employed tertiary     88             2       0.0227
## # i 23 more rows
```

```
# Visualize the subscription rates by job type and education level
ggplot(combined_insight, aes(x = job, y = education, fill = default_rate)) +
  geom_tile() +
  scale_fill_gradient(low = "white", high = "blue") +
  theme_minimal() +
  labs(title = "Default Rate by Job Type and Education Level",
       x = "Job Type", y = "Education Level", fill = "Default Rate") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

2.5.5 Combined effect of Job and Education on Default Rate



Step 3 Conclusions

3.1 Average Yearly Balance by Martial Group

Code & Results are in 2.5.2 Martial vs. Balance

- **Single and Married individuals** have significantly **higher average and median balances** compared to Divorced individuals.
  - The **average balance** for Singles is 669.83 and for Married individuals is 663.57, whereas Divorced individuals have a much lower average of 509.89.
  - Similarly, the **median balance** follows a similar trend, with Divorced individuals having a median balance of 264.00, compared to 363.00 and 342.50 for Singles and Married individuals, respectively.
- The **disparity between the median and average balances** across all groups indicates **right skewness** in the balance distributions, potentially due to a few individuals with exceptionally high balances.
- It could be a bit surprising that the mean and median balance of single individuals are higher than that of married individuals. Thus, it might be useful to investigate whether housing loans or personal loans are more common among married individuals, impacting their balances and this is illustrated in 3.2 Proportion of Housing & Personal Loan by Martial Group

3.2 Proportion of Housing & Personal Loan by Martial Group

Code & Results are in 2.5.3 Martial vs. Housing & Personal Loan

- Based on the analysis of housing and personal loans across different marital groups, **married individuals** have the highest proportion of clients with housing loans (58.10%), closely followed by **divorced**

individuals (56.44%). **Single individuals** have the lowest proportion of housing loans (53.18%).

- In terms of personal loans, **divorced individuals** lead with 17.05%, while **single individuals** have the lowest proportion at 12.37%.
- These findings are particularly interesting in light of the previous insight, where it was noted that **single individuals** have higher average and median balances compared to **married individuals**. This indicates that single individuals may manage their finances differently, potentially saving more and incurring fewer debts, thus contributing to their higher average and median balances.

3.3 Subscription Rate by Job Type

Code & Results are in 2.5.4 Job vs. Subscription (y)

- The subscription rates indicate that **retired individuals** have the highest subscription rate at **23.48%**, closely followed by **students** at **22.62%**. This suggests a strong inclination towards saving and investing among these demographics.
- The low subscription rate of **blue-collar** reveals that they have the lowest subscription rate at **7.29%**. This might suggest that individuals in blue-collar jobs may prioritize immediate cash flow needs over long-term savings options such as term deposits.
- Additionally, **entrepreneurs** (8.93%) and **service workers** (9.11%) also exhibit low subscription rates, indicating a possible preference for more flexible financial instruments due to the variability in their incomes.

3.4 Combined Effect of Job and Education Level on Default Rate

Code & Results are in 2.5.5 Combined effect of Job and Education on Default Rate

- **Entrepreneurs** with primary and secondary education have the **highest default rates**, with **7.69%** for those with primary education and **5.17%** for those with secondary education. This suggests that entrepreneurs, especially those with lower education levels, may face financial challenges or inconsistent cash flows that contribute to a higher risk of default.
- **Retired individuals** and **blue-collar workers** exhibit some of the lowest default rates, especially for those with secondary or tertiary education. This may indicate that these groups have more stable financial management, possibly due to pensions or steady income streams for blue-collar workers.