Bank Marketing Analysis by Candace Grant

2025-08-31

Introduction

This data report presents an analysis of a marketing dataset from a Portuguese banking institution's direct marketing campaigns. The dataset focuses on phone-based marketing efforts aimed at promoting term deposits to clients.

The primary objective is to develop a predictive classification model that determines whether a client will subscribe to a term deposit (binary outcome: 'yes' or 'no'). The campaigns often required multiple contacts with the same client to achieve successful conversions, making this a complex customer behavior prediction problem."

Dataset Source: Moro, S., Rita, P., & Cortez, P. (2014). Bank Marketing [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5K306.

```
library(tidyverse)
library(readxl)
library(dplyr)
library(ggplot2)
library(knitr)
```

Loading packages for data analysis

```
df <- read_excel("/Users/candace/Downloads/bank+marketing 2/bank/bank-full.xls")
```

Import Dataset

```
cat("Number of rows:", nrow(df), "\n")
```

Dataset Overview and Structure

```
Number of rows: 16383
```

```
cat("Number of columns:", ncol(df), "\n")
```

```
Number of columns: 17
```

```r
print(names(df))
```

```
 [1] "age"       "job"       "marital"   "education" "default"   "balance"
 [7] "housing"   "loan"      "contact"   "day"       "month"     "duration"
[13] "campaign"  "pdays"     "previous"  "poutcome"  "y"
```

```r
sapply(df, class)
```

```
        age         job     marital   education     default     balance
  "numeric" "character" "character" "character" "character"   "numeric"
    housing        loan     contact         day       month    duration
"character" "character" "character"   "numeric" "character"   "numeric"
   campaign       pdays    previous    poutcome           y
  "numeric"   "numeric"   "numeric" "character" "character"
```

```r
print(head(df))
```

```
# A tibble: 6 x 17
    age job        marital education default balance housing loan  contact   day
  <dbl> <chr>      <chr>   <chr>     <chr>     <dbl> <chr>   <chr> <chr>   <dbl>
1    58 management married tertiary  no         2143 yes     no    unknown     5
2    44 technician single  secondary no           29 yes     no    unknown     5
3    33 entrepren~ married secondary no            2 yes     yes   unknown     5
4    47 blue-coll~ married unknown   no         1506 yes     no    unknown     5
5    33 unknown    single  unknown   no            1 no      no    unknown     5
6    35 management married tertiary  no          231 yes     no    unknown     5
# i 7 more variables: month <chr>, duration <dbl>, campaign <dbl>, pdays <dbl>,
#   previous <dbl>, poutcome <chr>, y <chr>
```

```r
df_clean <- df[, !names(df) %in% c("contact", "poutcome")]
```

```r
str(df_clean)
```

Removing columns, "contact" and "poutcome" where over 70% of values are labeled "unknown."

```
tibble [16,383 x 15] (S3: tbl_df/tbl/data.frame)
 $ age      : num [1:16383] 58 44 33 47 33 35 28 42 58 43 ...
 $ job      : chr [1:16383] "management" "technician" "entrepreneur" "blue-collar" ...
 $ marital  : chr [1:16383] "married" "single" "married" "married" ...
 $ education: chr [1:16383] "tertiary" "secondary" "secondary" "unknown" ...
 $ default  : chr [1:16383] "no" "no" "no" "no" ...
 $ balance  : num [1:16383] 2143 29 2 1506 1 ...
 $ housing  : chr [1:16383] "yes" "yes" "yes" "yes" ...
 $ loan     : chr [1:16383] "no" "no" "yes" "no" ...
 $ day      : num [1:16383] 5 5 5 5 5 5 5 5 5 5 ...
```

```
 $ month    : chr [1:16383] "may" "may" "may" "may" ...
 $ duration : num [1:16383] 261 151 76 92 198 139 217 380 50 55 ...
 $ campaign : num [1:16383] 1 1 1 1 1 1 1 1 1 1 ...
 $ pdays    : num [1:16383] -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
 $ previous : num [1:16383] 0 0 0 0 0 0 0 0 0 0 ...
 $ y        : chr [1:16383] "no" "no" "no" "no" ...
```

```r
library(dplyr)
df_clean <- df_clean %>% rename(
  marital_status = marital,
  education_level = education,
  credit_card_default = default,
  avg_yearly_balance = balance,
  mortgage = housing,
  personal_loan = loan,
  recency = pdays,
  prior_contacts = previous,
  subscribed = y
)
str(df_clean)
```

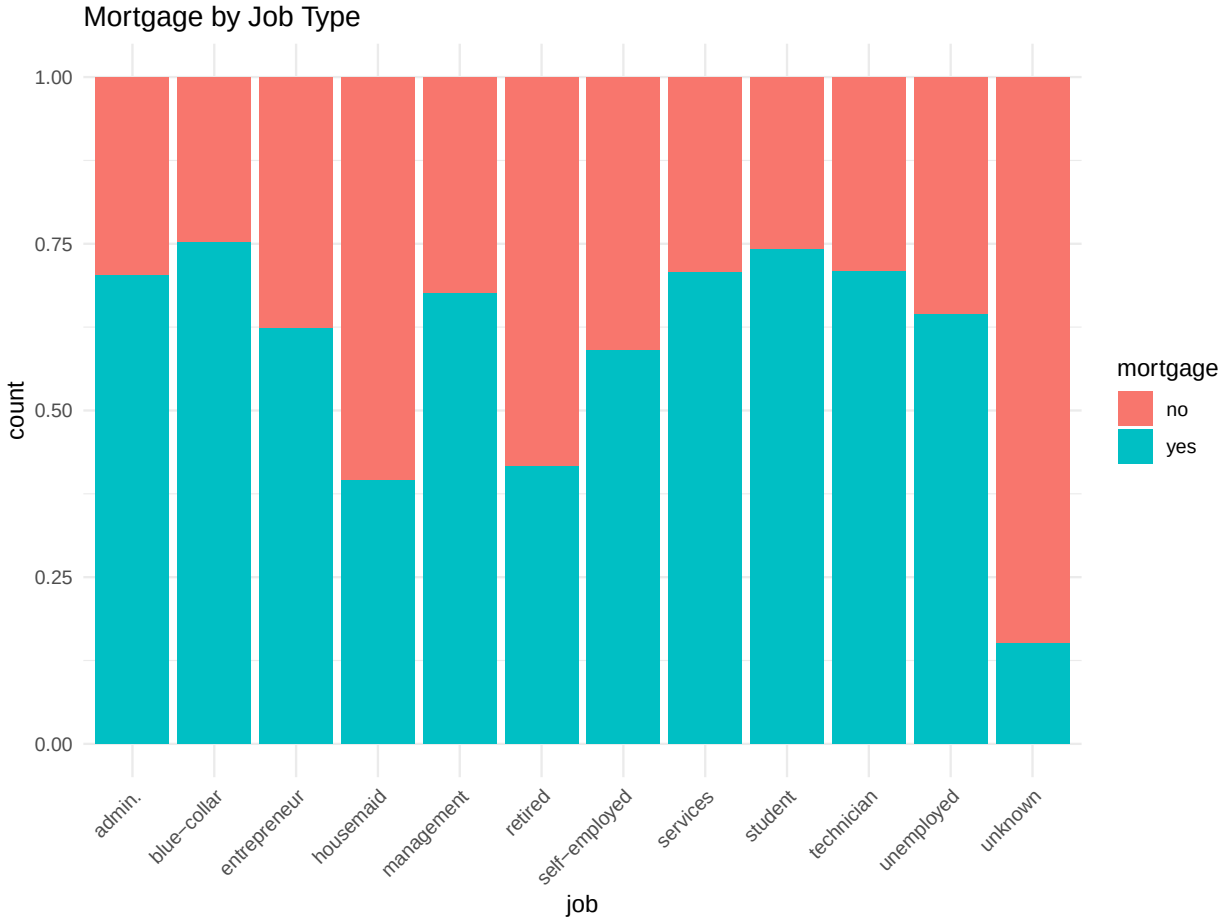
Renaming columns to be more descriptive

```
tibble [16,383 x 15] (S3: tbl_df/tbl/data.frame)
 $ age                : num [1:16383] 58 44 33 47 33 35 28 42 58 43 ...
 $ job                : chr [1:16383] "management" "technician" "entrepreneur" "blue-collar" ...
 $ marital_status     : chr [1:16383] "married" "single" "married" "married" ...
 $ education_level    : chr [1:16383] "tertiary" "secondary" "secondary" "unknown" ...
 $ credit_card_default: chr [1:16383] "no" "no" "no" "no" ...
 $ avg_yearly_balance : num [1:16383] 2143 29 2 1506 1 ...
 $ mortgage           : chr [1:16383] "yes" "yes" "yes" "yes" ...
 $ personal_loan      : chr [1:16383] "no" "no" "yes" "no" ...
 $ day                : num [1:16383] 5 5 5 5 5 5 5 5 5 5 ...
 $ month              : chr [1:16383] "may" "may" "may" "may" ...
 $ duration           : num [1:16383] 261 151 76 92 198 139 217 380 50 55 ...
 $ campaign           : num [1:16383] 1 1 1 1 1 1 1 1 1 1 ...
 $ recency            : num [1:16383] -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
 $ prior_contacts     : num [1:16383] 0 0 0 0 0 0 0 0 0 0 ...
 $ subscribed         : chr [1:16383] "no" "no" "no" "no" ...
```

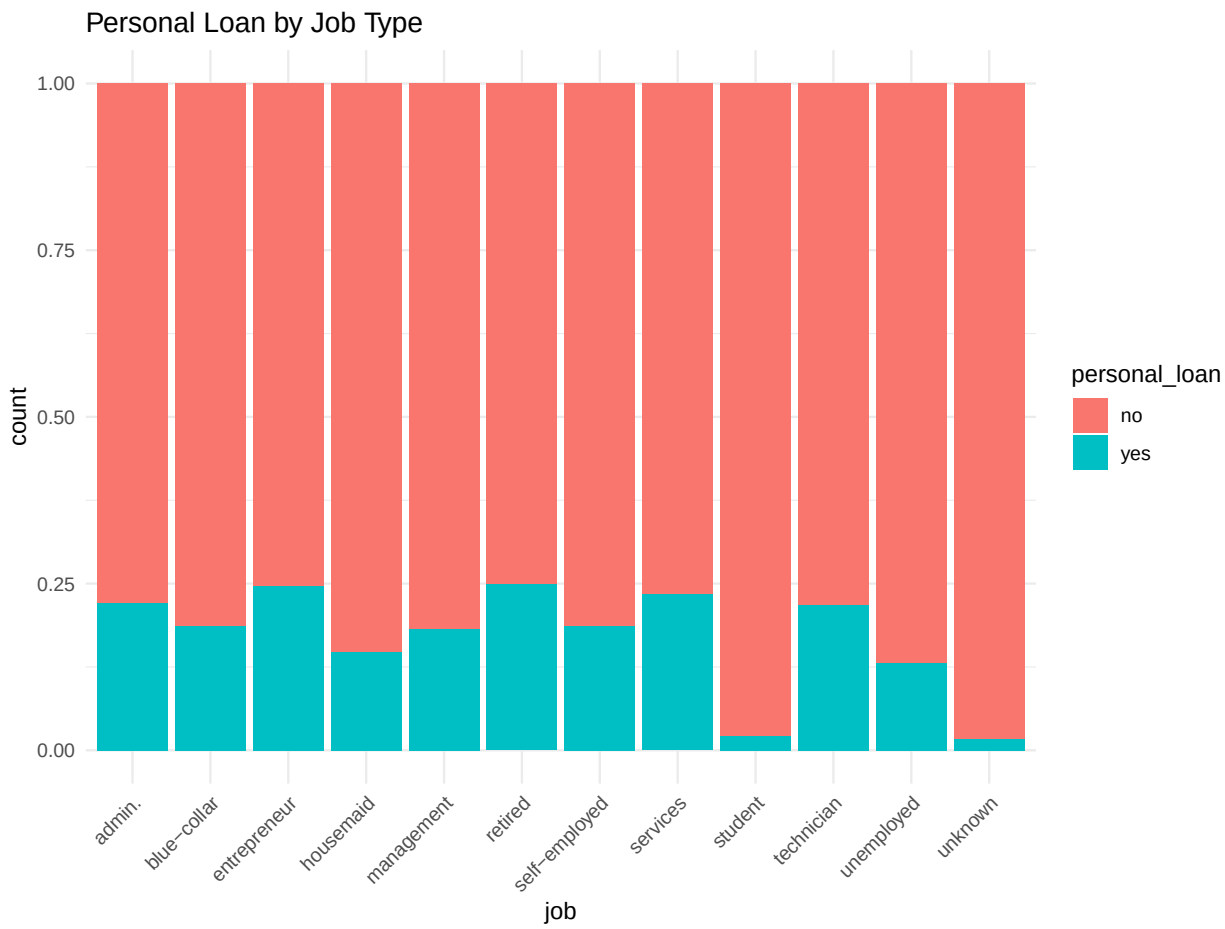
```r
library(ggplot2)

#### Relationship Plots ####

# Job vs Mortgage
ggplot(df_clean, aes(x = job, fill = mortgage)) +
  geom_bar(position = "fill") +
  labs(title = "Mortgage by Job Type") +
```

```
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

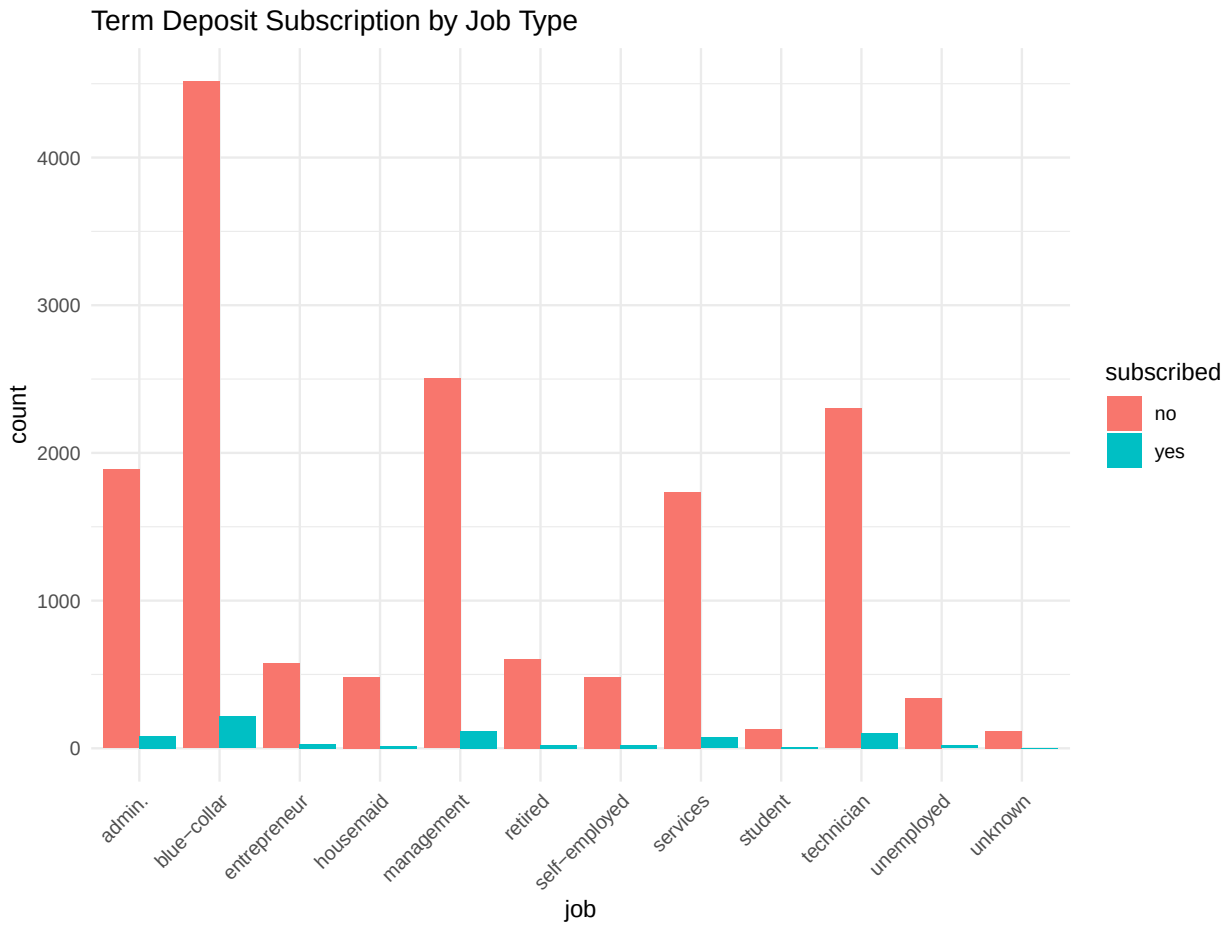
In this section I am creating simple bar plots to give an understanding of the characteristics of potential customers that could shed light on their final decision to subscribe to the term deposit.




```
# Job vs Personal Loan
ggplot(df_clean, aes(x = job, fill = personal_loan)) +
  geom_bar(position = "fill") +
  labs(title = "Personal Loan by Job Type") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
# Side-by-side bars instead of stacked
ggplot(df_clean, aes(x = job, fill = subscribed)) +
  geom_bar(position = "dodge") +
  labs(title = "Term Deposit Subscription by Job Type") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Conclusion

I began this analysis by examining the dataset structure and discovered an important data quality issue: missing values were coded as 'unknown' rather than traditional NA values, which R does not automatically recognize as missing data. To address this, I calculated the percentage of 'unknown' values in each column and removed those with greater than 70% missing data, particularly columns that did not contribute to the primary objective of predicting term deposit subscriptions. Key steps completed:

- Reviewed dataset structure and identified data quality issues
- Renamed columns for better clarity and interpretability
- Removed columns with excessive missing data (>70% 'unknown' values)
- Created visualizations to explore relationships between variables and the target outcome

This preprocessing creates a clean dataset suitable for identifying factors that influence customer decisions to subscribe to term deposits.