EDA Playground

Nicholas Mobley

```
raw_data = read_csv("prac_data.csv") %>%
  mutate(
    price = unlist(read_csv("price.csv")),
    job = as.factor(replace_na(job,"Unknown")),
    education = replace_na(education,"Unknown"),
    education = replace(education,which(education == "unknown"),"Unknown"),
    contact = replace_na(contact,"Unknown"),
    contact = replace(contact,which(contact == "unknown"),"Unknown"),
    pdays = replace_na(pdays,-1),
    default = replace_na(default,"Unknown"),
    maritial = replace_na(default,"Unkown"),
    housing = replace_na(housing,"Unkown"),
    loan = replace_na(loan,"Unkown"),
    previous = replace_na(previous,0),
    poutcome = replace_na(poutcome,"Unknown"),
    inPrevious = if_else(previous > 0,TRUE,FALSE),
    numMissing = rowSums(across(everything(), ~is.na(.)))
  )
```

```
## Parsed with column specification:
## cols(
##   age = col_double(),
##   job = col_character(),
##   marital = col_character(),
##   education = col_character(),
##   default = col_character(),
##   balance = col_double(),
##   housing = col_character(),
##   loan = col_character(),
##   contact = col_character(),
##   day = col_double(),
##   month = col_character(),
##   duration = col_double(),
##   campaign = col_double(),
##   pdays = col_double(),
##   previous = col_double(),
##   poutcome = col_character(),
##   y = col_character()
## )

## Parsed with column specification:
## cols(
##   price = col_double()
## )
```

```
numericColumns = c(1,6,10,12,13,14,15,18)
categoricalColumns = c(2,3,4,5,7,8,9)
df_colnames = colnames(raw_data)
```

```
proportion_df = raw_data %>%
  group_by(job) %>%
  summarize(n = n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
mean_df = raw_data %>%
  summarize(across(numericColumns,function(x) round(mean(x, na.rm = TRUE),2))) %>%
  pivot_longer(everything(), names_to = "Variable",values_to = "mean") %>%
  na.omit()
```

```
## Note: Using an external vector in selections is ambiguous.
## i Use `all_of(numericColumns)` instead of `numericColumns` to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```

```
sd_df = raw_data %>%
  summarize(across(numericColumns,function(x) round(sd(x, na.rm = TRUE),2))) %>%
  pivot_longer(everything(),names_to = "Variable",values_to = "sd") %>%
  na.omit()
median_df = raw_data %>%
  summarize(across(numericColumns,function(x) median(x,na.rm = TRUE))) %>%
  pivot_longer(everything(),names_to = "Variable", values_to = "median") %>%
  na.omit()
pecent_missing = raw_data %>%
  summarize(across(numericColumns,function(x) round((sum(is.na(x))/nrow(raw_data))*100,2))) %>%
  pivot_longer(everything(),names_to = "Variable", values_to = "Missing Number") %>%
  na.omit()
display = inner_join(mean_df,sd_df) %>%
  inner_join(median_df) %>%
  inner_join(pecent_missing)
```

```
## Joining, by = "Variable"
```

```
## Joining, by = "Variable"
## Joining, by = "Variable"
```

```
ddf = transpose_df(display)
ddf
```

```
## # A tibble: 5 x 9
##   rowname        `1`   `2`     `3`   `4`      `5`      `6`    `7`      `8`
##   <chr>          <chr> <chr>   <chr> <chr>    <chr>    <chr>  <chr>    <chr>
## 1 Variable       age   balance day   duration campaign pdays  previous price
## 2 mean           40.94 1365.28 15.81 257.58   2.76     40.07  0.56     52.04
## 3 sd             10.61 3054.52 8.32  256.45   3.1      100.03 2.28     27.65
## 4 median         39    450     16    180      2        -1     0        50
## 5 Missing Number 5.68  1.76    0     6.91     3.26     0      0        0
```

```
\begin{tabular}{|c |c |c |c |c |c |c |c |c|}
\hline
Variable & Age & Balance & Day & Duration & Campaign & Prev Days & Previous & Price \\
\hline\hline
Mean & 40.94 & 1365.28 &15.81 & 257.58 & 2.76 & 40.07 & 0.56 & 52.04 \\
Std Dev & 10.61 & 3054.52 &8.32 & 256.45 & 3.1 & 100.03 & 2.28 & 27.65 \\
Median & 39 & 450 &16 & 180 & 2 & -1 & 0 & 50 \\
Percent Missing & 5.68 & 1.76 &0 & 6.91 & 3.26 & 0 & 0 & 0 \\
\hline
\end{tabular}
```

```
bar_graphs <- function(variable) {
  ggplot(raw_data, aes(x = fct_reorder(!!sym(variable),!!sym(variable),.fun='length'))) +
    geom_bar(stat = 'count') +
    xlab(variable)
}                                                  #remove job
graphs = lapply(df_colnames[categoricalColumns[2:length(categoricalColumns)]],bar_graphs)
ggarrange(plotlist = graphs)
```