practice 1

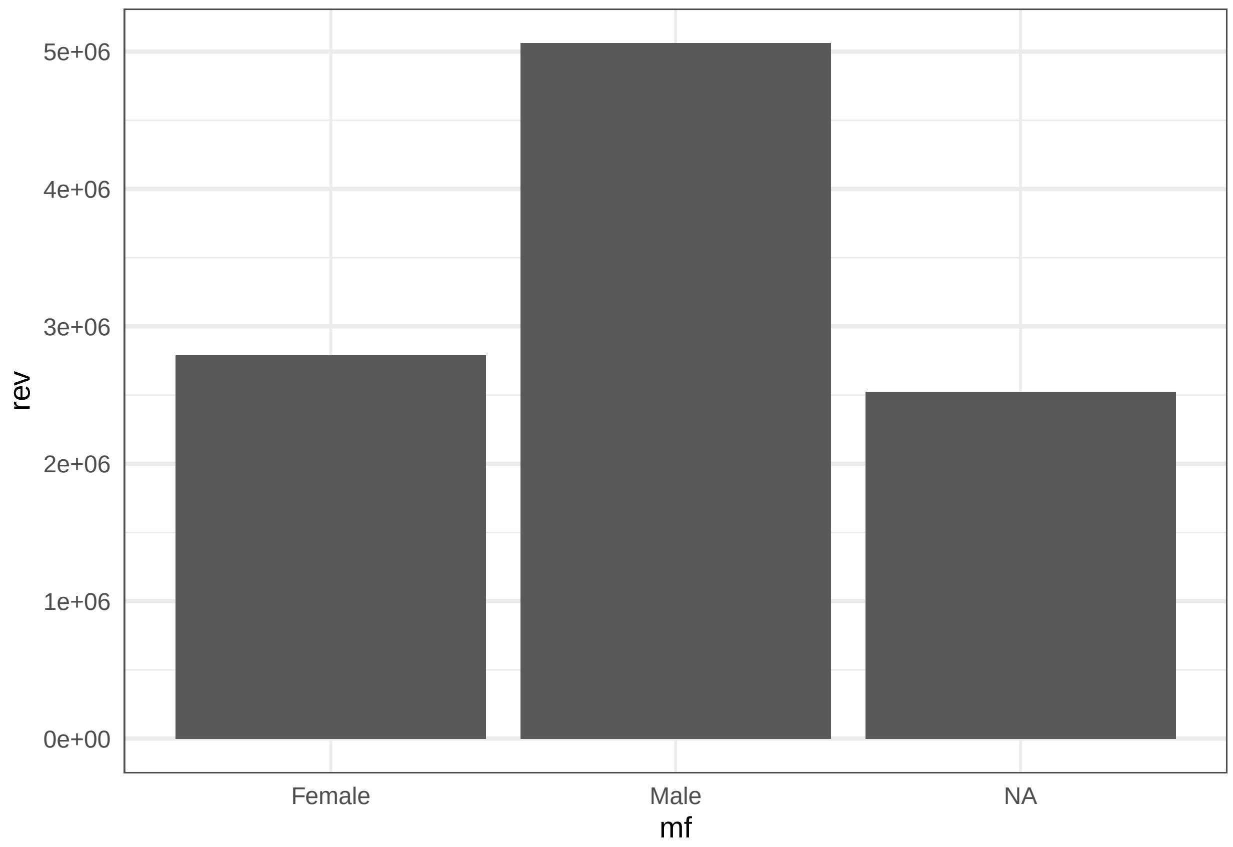
```
library(tidyverse)

df <- read_csv("/Users/Josiah/Downloads/customertxndata.csv",
               col_names = c("nv","nt","os","mf","rev"))

# calculates na's means and sds
summary(df)
```

```
##        nv              nt              os                 mf           
##  Min.   : 0.00   Min.   :0.000   Length:22800       Length:22800      
##  1st Qu.: 6.00   1st Qu.:1.000   Class :character   Class :character  
##  Median :12.00   Median :1.000   Mode  :character   Mode  :character  
##  Mean   :12.49   Mean   :0.993                                        
##  3rd Qu.:19.00   3rd Qu.:1.000                                        
##  Max.   :25.00   Max.   :2.000                                        
##                  NA's   :1800                                         
##       rev        
##  Min.   :   0.0  
##  1st Qu.: 170.0  
##  Median : 344.7  
##  Mean   : 454.9  
##  3rd Qu.: 576.9  
##  Max.   :2000.0  
## 
```

```
ggplot(df, aes(mf, rev)) +
  geom_bar(stat = "identity")
```



```
# (5 pts) What is the Pearson Moment of Correlation between number of visits and revenue? Comment on th
cor(df$rev, df$nv)
```

```
## [1] 0.7388448
```

```
# (10 pts) Which columns have missing data? How did you recognize them? How would you impute missing va
# look at the NA counts
summary(df)
```

```
##        nv              nt              os                  mf
##  Min.   : 0.00   Min.   :0.000   Length:22800       Length:22800
##  1st Qu.: 6.00   1st Qu.:1.000   Class :character   Class :character
##  Median :12.00   Median :1.000   Mode  :character   Mode  :character
##  Mean   :12.49   Mean   :0.993
##  3rd Qu.:19.00   3rd Qu.:1.000
##  Max.   :25.00   Max.   :2.000
##                  NA's   :1800
##       rev
##  Min.   :   0.0
##  1st Qu.: 170.0
##  Median : 344.7
##  Mean   : 454.9
##  3rd Qu.: 576.9
##  Max.   :2000.0
##
```

```r
#   (15 pts) Impute missing transaction and gender values. Use the mean for transaction (rounded to the
df <- mutate(df,
       nt = ifelse(is.na(nt),
                   yes = round(mean(nt, na.rm = TRUE), 0),
                   no = nt),
       mf = ifelse(is.na(mf), "Male", mf)
)


# (20 pts) Split the data set into two equally sized data sets where one can be used for training a mod
index <- 1:nrow(df)
evens <- index %% 2 == 0
odds <- index %% 2 == 1

training <- slice(df, index[odds])
testing <- slice(df, index[evens])
# (10 pts) Calculate the mean revenue for the training and the validation data sets and compare them. C
mean(training$rev)
```

```
## [1] 449.6105
```

```r
mean(testing$rev)
```

```
## [1] 460.26
```

```r
# (15 pts) For many data mining and machine learning tasks, there are packages in R. Use the sample() f
set.seed(0)

df <- mutate(df, id = row_number()) # add id for referencing

analysis <- sample_frac(df, 0.6)
assessment <- sample_frac(anti_join(df, analysis), .2)

validation <- anti_join(df, analysis) %>%
  anti_join(assessment)
```