Appointment_Capstone_Statistics

Derek Samsom

4/9/2018

The goal of the Appointment_Capstone project is to predict missed medical appointments based on historical appointment data.

This mini-project will take a look at some statistics on the data. First I'll load the packages and data.

```
library(tidyverse)
library(lubridate)
library(caret)

appointments <- read_csv("Final_Data.csv")
appointments_original <- appointments
zipcodes <- read_csv("zipcodes.csv")

appointments <- appointments %>%
    mutate(appt_datetime = lubridate::mdy_hms(paste(appt_date, appt_time)))

appointments$date_scheduled <- lubridate::as_date(
    appointments$date_scheduled, format = "%m/%d/%y", tz = "UTC")
```

First I want to calculate the percent of missed appointments overall by creating a logical variable `missed`, where 1 represents a missed appointment and 0 represents a kept appointment. This will determine the degree of class imbalance.

```
appointments <- appointments %>%
    mutate(missed = ifelse(appointments$kept_status == "Missed", 1, 0))
missed_rate <- mean(appointments$missed)
missed_rate
```

```
## [1] 0.1592944
```

15.93 % of the total appointments are missed. This is an imbalanced classification, which will have implications in the modeling. For example, the model could predict all of the appointments will be kept and be correct 84.07 % of the time. This results in a high accuracy without providing any useful prediction of which appointments will be missed.

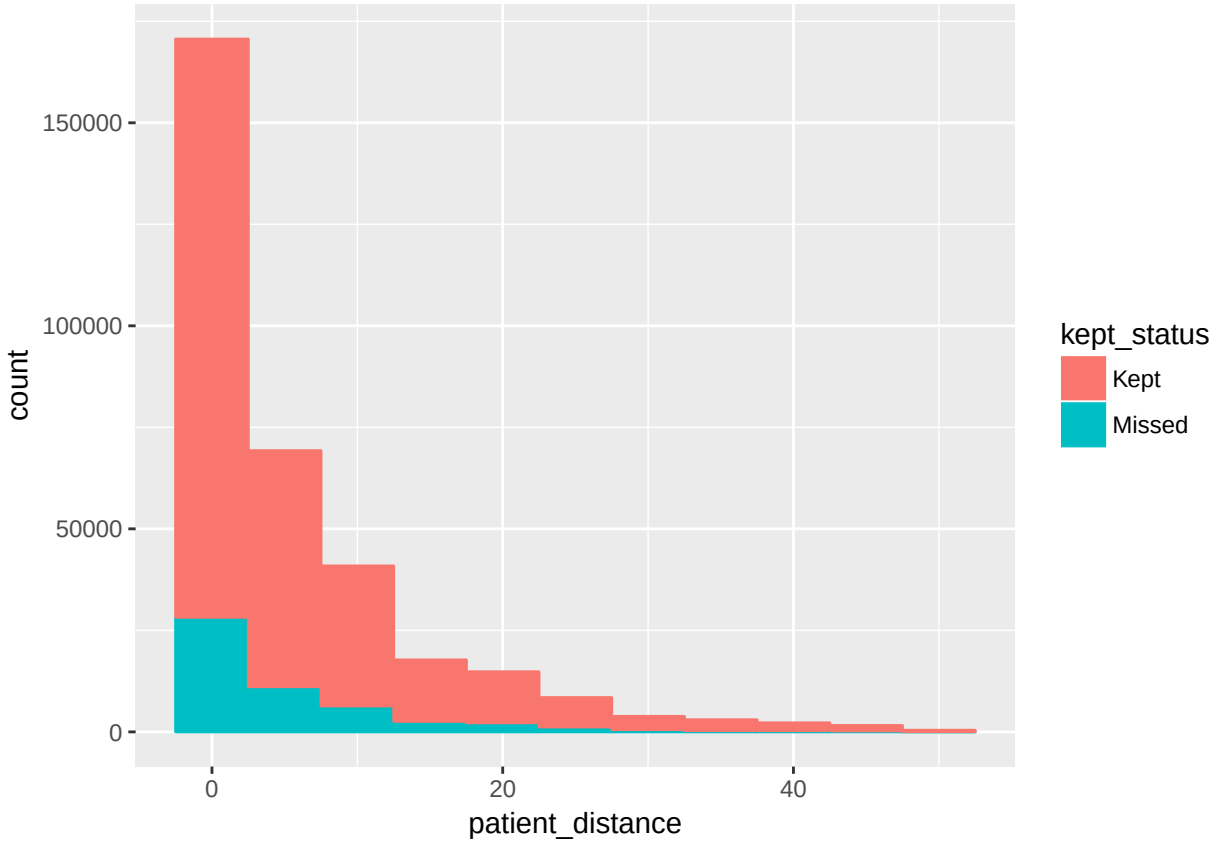
Next I want to check the data integrity by counting the missing values for each variable.

```
map_dbl(appointments, ~sum(is.na(.)))
```

```
##        kept_status          appt_date          appt_time
##                  0                  0                  0
##        appt_length     date_scheduled        patient_age
##                  0                  0                  0
##     patient_gender       billing_type       prior_missed
##                  0                  0                  0
##         prior_kept   patient_distance         office_zip
##                  0                974                  0
## provider_specialty remind_call_result     appt_datetime
##                  0                  0                  0
##             missed
##                  0
```

One variable, `patient_distance` has 974 missing value. I expect most people to live fairly close to the office, with fewer and fewer as the distance increases. This would give a right skewed distribution. I'll take a closer look at the histogram.

```
appointments %>%
    filter(patient_distance < 50) %>%
    ggplot(
        aes(x = patient_distance, color = kept_status, fill = kept_status)
    ) +
            geom_histogram(binwidth = 5)
```

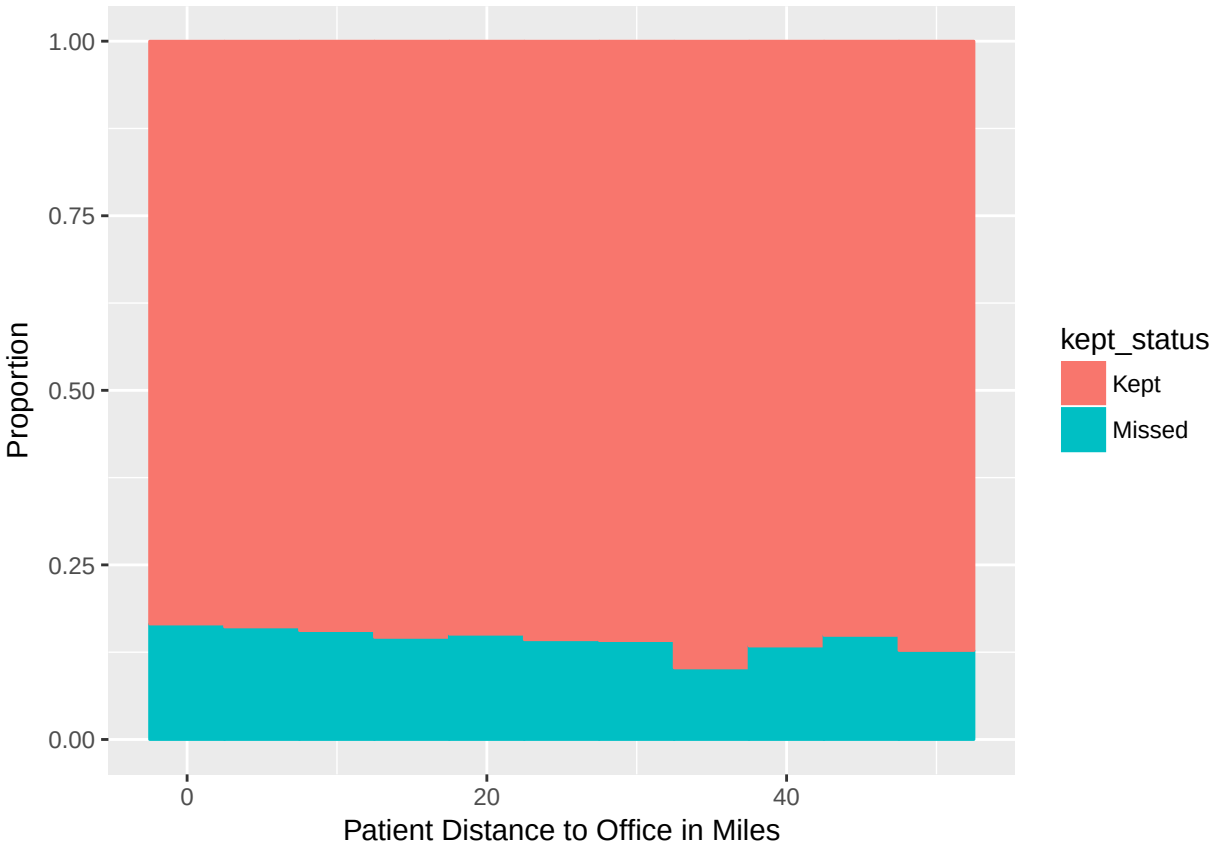


The variable `patient_distance` is a right-skewed distribution as would be expected. Because it is right skewed, I will replace the NA values with the median instead of the mean.

Next I'll add a proportional plot to make it easier to see the change in the miss rate based on distance.

```
appointments$patient_distance <- appointments$patient_distance %>%
    replace_na(median(appointments$patient_distance, na.rm = TRUE))

appointments %>%
    filter(patient_distance < 50) %>%
    ggplot(
        aes(x = patient_distance, color = kept_status, fill = kept_status)
    ) +
        geom_histogram(position = "fill", binwidth = 5) +
        labs(x = "Patient Distance to Office in Miles", y = "Proportion")
```



In general, the closer a patient lives to the office, the more likely they are to miss their appointment. This seems a little counter-intuitive, since the greater the distance from the office, the more potential for there to be hurdles to getting to the appointment. Those who are further away live most likely live in a more rural area, where perhaps trips to town are better planned out and prepared for.

Next I'll take a look at `patient_age`. I'll start by seeing if there are any unreasonable values by looking at the summary for age.

```
summary(appointments_original$patient_age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   17.00   34.00   35.56   54.00  264.00
```

There are no negative values, which is good. The maximum age is 264, which is not possible. I'll create a table that counts the observations of each age over 100.

```
age_over_100 <- appointments_original %>%
    filter(patient_age > 100)

table(age_over_100$patient_age)
```

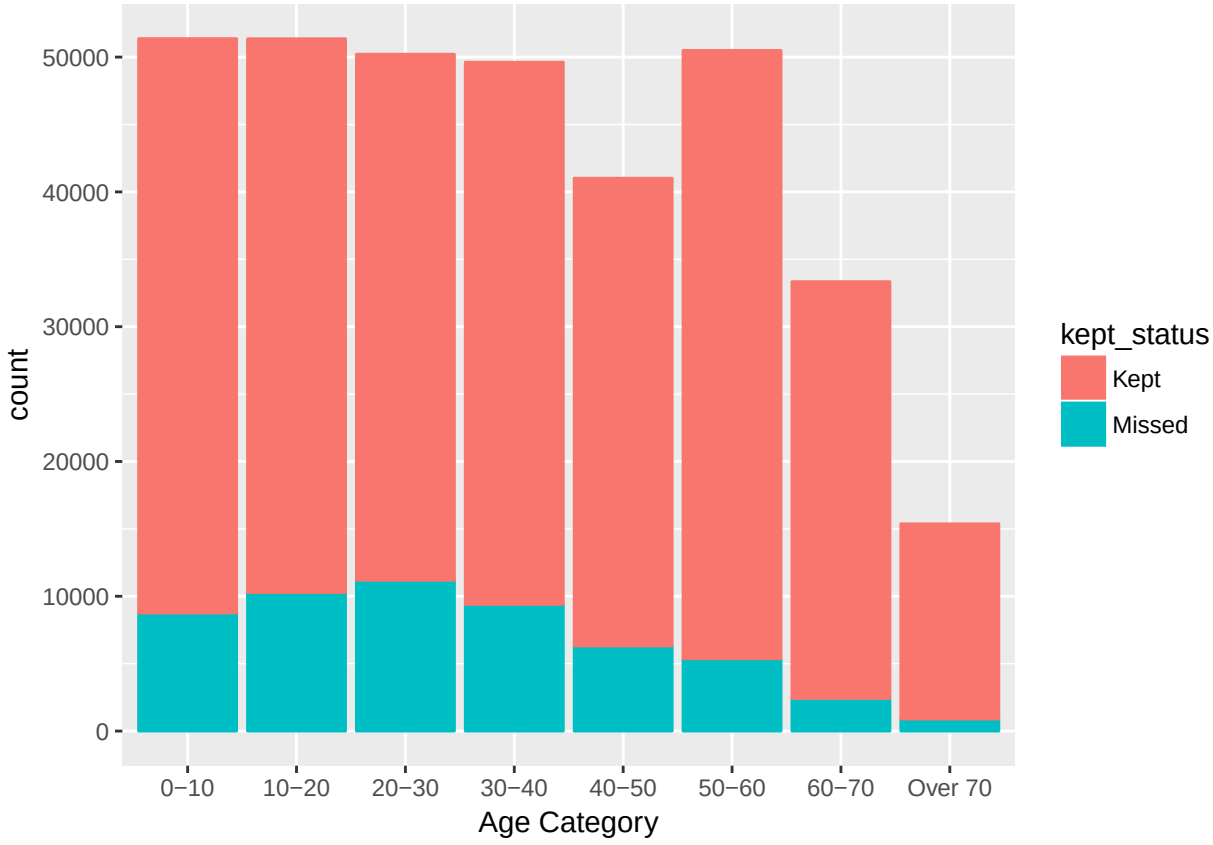
```
##
## 101 102 103 104 107 116 117 263 264
##   3   5   6   3   1  22  41  10   6
```

There are a definitely some observations where the age is higher than plausible. Therefore, the observations greater than age 110 will be removed from the data. Next I will create age groups and create a bar chart to see if the ratio of missed appointment variaes across age groups.

```r
age_labels <- c("0-10", "10-20","20-30", "30-40", "40-50", "50-60", "60-70",
                "Over 70")
age_breaks <- c(-1, 10, 20, 30, 40, 50, 60, 70, 111)

appointments <- appointments %>%
    filter(patient_age <= 110) %>%
    mutate(
        age_cat = cut(patient_age, breaks = age_breaks, labels = age_labels))


ggplot(
    appointments,
    aes(x = age_cat, color = kept_status, fill = kept_status)
) +
    stat_count() +
    labs(x = "Age Category")
```



There is a significant difference in missed appointments across the age groups. Missed appointments are highest with young adults, and decrease with older patients. This follows a similar pattern to what I expected. Older patients are more likely to have retired and therefore may miss fewer appointments due to gettingtied up at work. Younger patients may not have as critical of medical needs, making it easier to skip an appointments. Younger patients are also more likely to miss an appointment due to the responsibilities of having young children.