Appointment Capstone Milestone Report

Derek Samsom

4/29/2018

Missed medical appointments are a major problem in the medical industry, resulting in lost revenue. Medical providers can over-book appointments to try to minimize the lost revenue, but without any way to predict the probability of an appointment being missed, there will be times where more or fewer patients show up at a given time than expected. The result will be that lost revenue will be reduced but not eliminated, as there will still be times that more appointments are missed than expected. There will aslo be times more apppointments show up than expected, which can overwhelm staff and resources and affect the level of patient care.

This project is a classification problem that will explore the prediction of whether a medical appointment will be missed, and its probability of being kept or missed. The goal of this project is to minimize the prediction error of missed appointments, as the error results in times where there are too many or too few patients at a given time. This is important to the client because it reduces the revenue loss caused by missed appontments in a way that reduces the undesired consequences of overbooking.

There are countless reasons and circumstances that can lead someone to miss an appointment, such as a last minute work meeting or a family emergency, that aren't directly captured in the data and are impossible to know in advance of future appointments. Missed appointments can only be predicted based on indirect factors that are known, such as patient history and demographics. Because of this, there will be a level of error that cannot be eliminated, however, any reduction in error compared to having no predictive model at all is still beneficial as it allows more overbooking to be done with fewer negative consequences.

Medical providers can use the missed appointment predictions by incorporating them into their booking methods and systems. The methods used in booking will have to consider the implications of the inherent prediction errors and balance the risk the errors represent: too many patients leading to staff/resource shortage, and too few patients leading to lost revenue. The methods of implementing the use of missed appointentment predictions into a booking system are client-specific and not included in the scope of this project, which is limited to minimizing the error while predicting the classification and associated probability that an appointment will be missed or kept.

I will start off by loading the required packages and the data.

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

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

The raw data, which has been named `appointments`, contains information on 342862 past appointments, pre-sorted by the date and time of appointment. The dependended variable, `kept_status`, shows whether the appintment was kept or missed.

There is no field that can be used to identify a specific patient in the data set. A patient may have had more than one appointment during the time-period represented in the data, meaning that one individual patient may make up one or multiple observations. If there was a patient ID field, it would allow the data to be grouped by patient and give the option of organizing the data by patient rather than by appointment.

A secondary data set, `zipcodes`, has information about the county the offices are located in. This will be used to see if the location can help predict whether an appointment will be missed. The county names are converted to a 2-letter code for confidentiality.

Data Dictionary

```
var_descriptions <- c(
    "Dependent variable: kept or missed",
    "Appointment date",
    "Appointment time",
    "Appointment length in minutes",
    "Date appointment was scheduled",
    "Patient age",
    "Patient gender",
    "Billing type",
    "Number of prior missed appointments",
    "Number of prior kept appointments",
    "Patient distance from office in miles",
    "Office Zip Code - Anonymized",
    "Provider primary specialty code",
    "Reminder Call result")
var <- colnames(appointments)
var_type <- unlist(map(appointments, class))
var_type <- var_type[-4]
as_data_frame(cbind(c(1:length(var)), var, var_type, var_descriptions))
```

```
## # A tibble: 14 x 4
##    V1    var               var_type  var_descriptions
##    <chr> <chr>             <chr>     <chr>
##  1 1     kept_status       character Dependent variable: kept or missed
##  2 2     appt_date         character Appointment date
##  3 3     appt_time         hms       Appointment time
##  4 4     appt_length       integer   Appointment length in minutes
##  5 5     date_scheduled    character Date appointment was scheduled
##  6 6     patient_age       integer   Patient age
##  7 7     patient_gender    character Patient gender
##  8 8     billing_type      character Billing type
##  9 9     prior_missed      integer   Number of prior missed appointments
## 10 10    prior_kept        integer   Number of prior kept appointments
## 11 11    patient_distance  integer   Patient distance from office in mil~
## 12 12    office_zip        character Office Zip Code - Anonymized
## 13 13    provider_specialty character Provider primary specialty code
## 14 14    remind_call_result character Reminder Call result
```

The `appt_date` and `appt_time` variables can be combined into one variable, `appt_datetime`.

```
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")
```

Data Exploration

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 to see if there are any missing values that could indicate reduced data integrity or adversely affect the modelling.

```
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. This is fairly minor considering the size of the data set and will be evaluated later on when exploring the variable further.

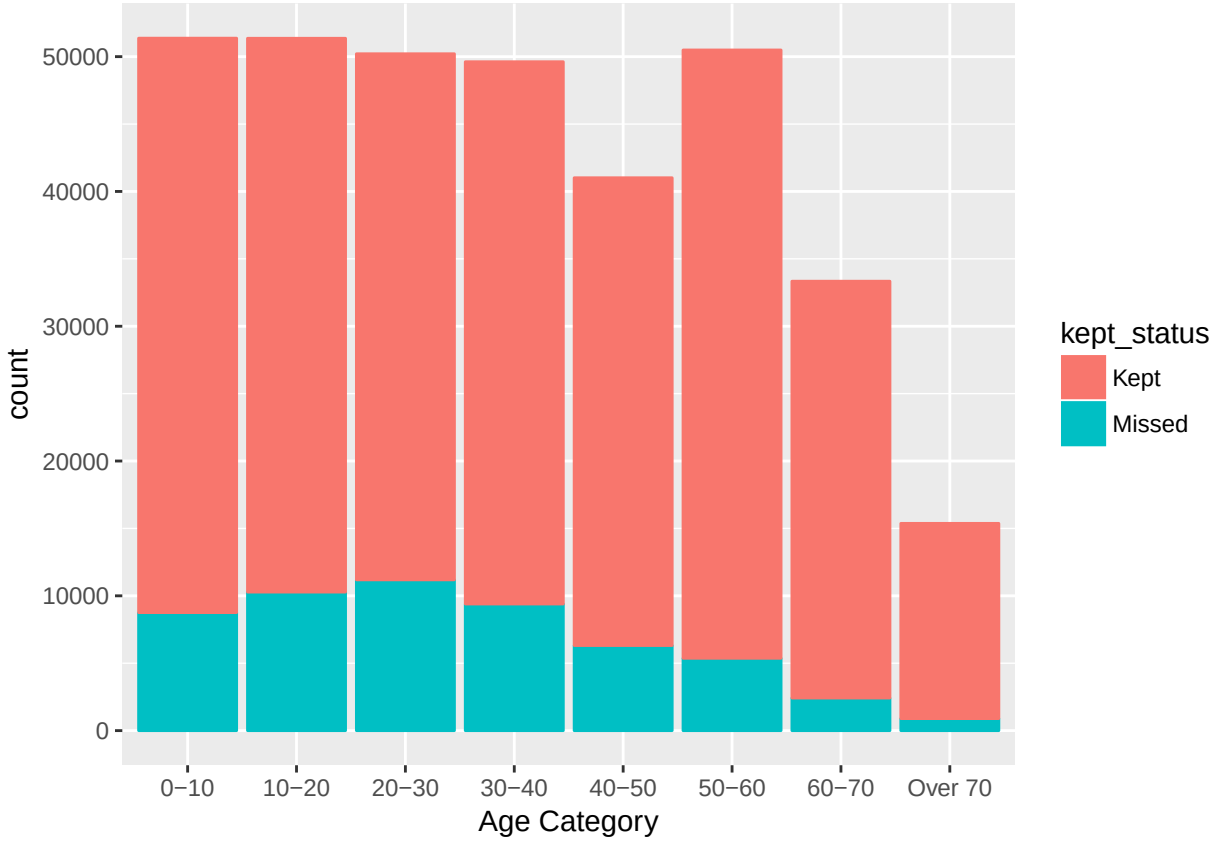
Patient Age

I expect missed appointments to vary across age ranges. Perhaps older patients have fewer commitments with children or work, and make their appointments more regularly, or perhaps younger adults might skip more appointments because they aren't as critical.

There are a small number of observations where the age is higher than plausible. Therefore, the observations greater than age 110 will be removed from the data. To make the plot easier to read, I will group the ages. The final model will still use the original continuous variable.

```
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.

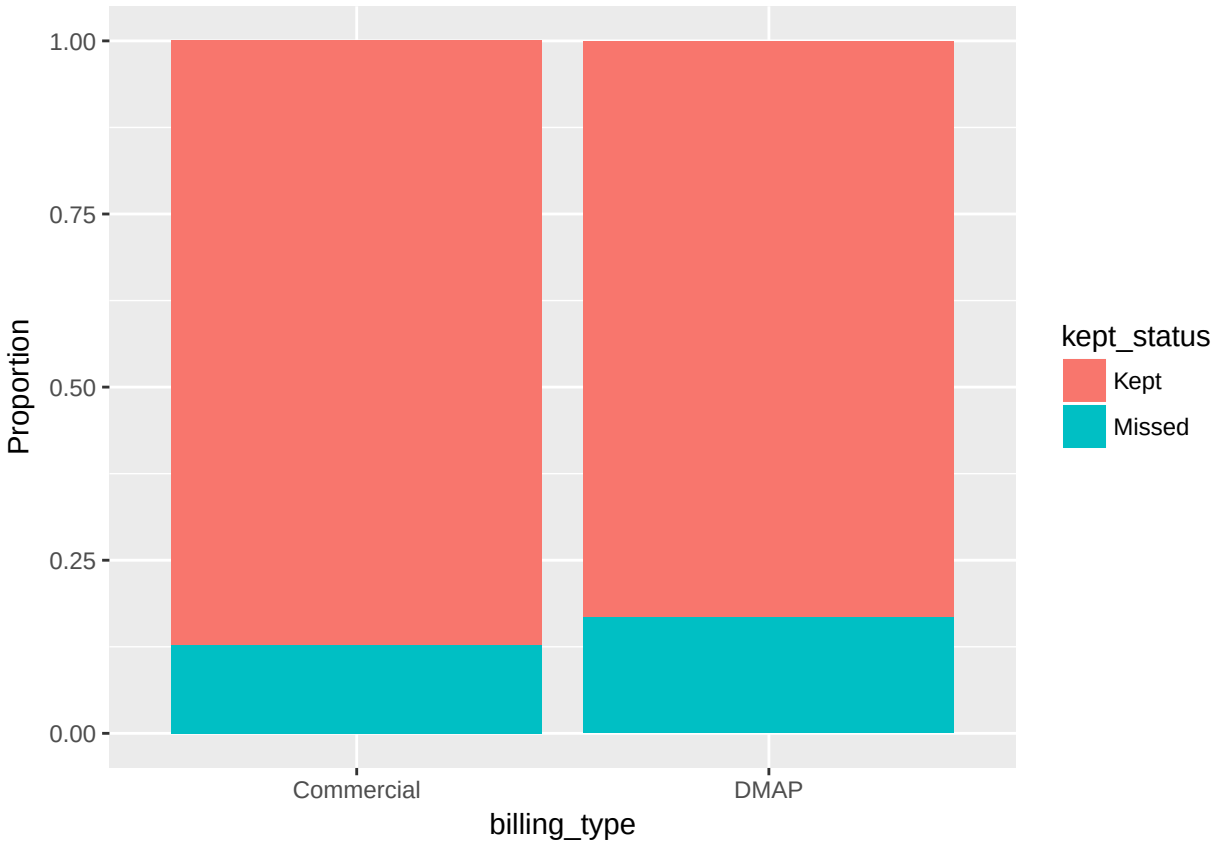
Billing Type

```
table(appointments$billing_type)
```

```
##
##     Commercial           DMAP To Be Assigned
##          78282         264500              1
```

There is only one observation of "To Be Assigned", therefore it will be removed from the data.

```
appointments <- subset(appointments, billing_type != "To Be Assigned")

ggplot(
    appointments,
    aes(x = billing_type, fill = kept_status)
) +
    geom_bar(position = "fill") +
    labs(y = "Proportion")
```



There is a fairly small yet significant difference between billing types. DMAP has a higher proportion of missed appointments than commercial.

Appointment Datetime

For the variable `appt_datetime`, I will create an `hour` variable to see the variation in missed appointments by hour of day. There are many ways the time of day can have an effect, such as rush hour traffic in the morning and afternoon causing more missed appointments, whereas mid-day appointments could be more likely to be missed by work factors.

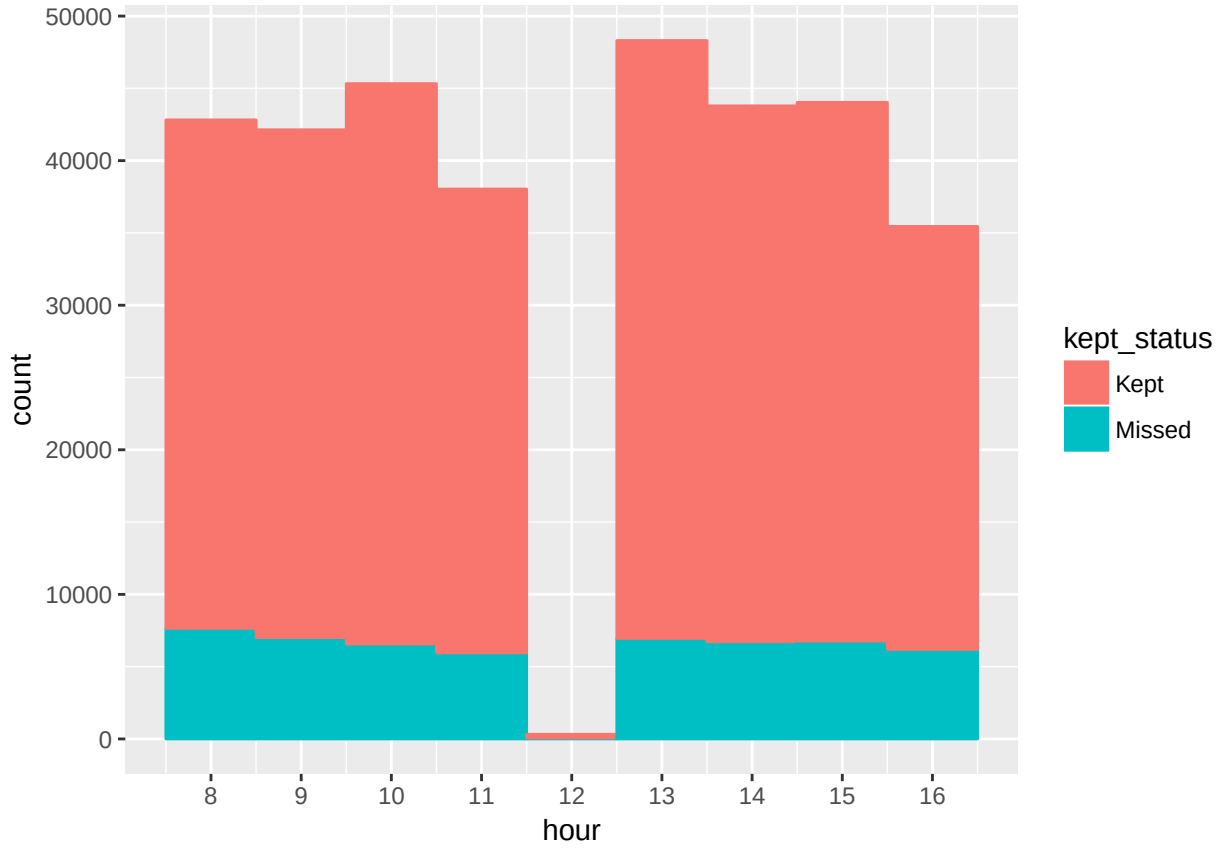
```
appointments <- appointments %>%
    mutate(hour = lubridate::hour(appointments$appt_datetime))

table(appointments$hour)
```

```
##
##     0     5     6     7     8     9    10    11    12    13    14    15
##     7    24    25    98 42816 42133 45326 38033   321 48307 43787 44033
##    16    17    18    19    20    21
## 35449  2180   205    33     3     2
```

Most appointments are scheduled between 8:00 AM and 5:00 PM, with a one hour gap starting at 12:00.

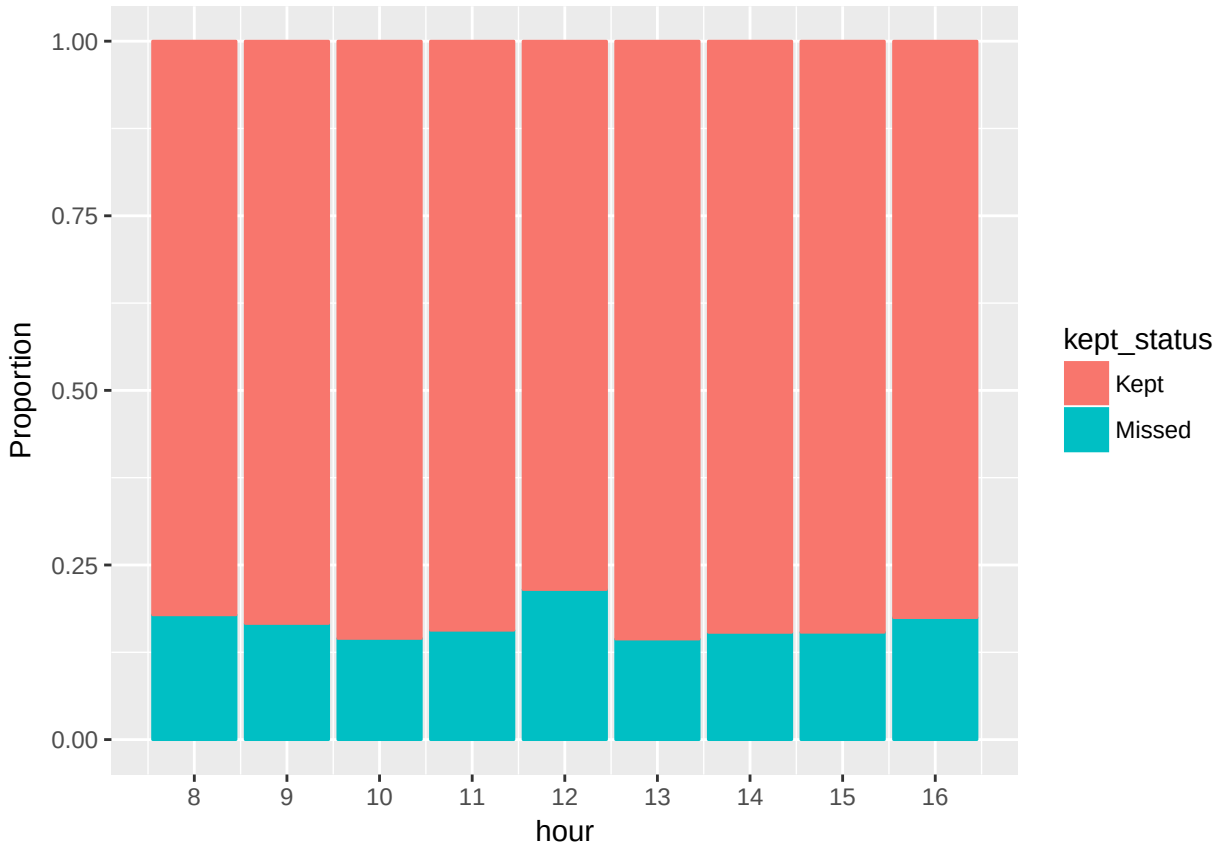
```
appointments_hour <- appointments %>%
    select(kept_status, hour) %>%
    filter(hour >= 8 & hour <= 16)

ggplot(
```

```
    appointments_hour,
    aes(x = hour, col = kept_status, fill = kept_status)
) +
    geom_histogram(binwidth = 1) +
    scale_x_continuous(breaks = seq(8, 17, 1))
```



There is a decline in the total number of missed appointments as both the morning and afternoon period progress, however, there are fewer appointments towards the end of the two periods. The ratio of missed appointments is hard to read, so I will create a proportional plot to see if it shows any trends.

```
ggplot(
    appointments_hour,
    aes(x = hour, col = kept_status, fill = kept_status)
) +
    geom_bar(position = "fill") +
    scale_x_continuous(breaks = seq(8, 17, 1)) +
    labs(y = "Proportion")
```



Proportionally more appointments are missed at the beginning and end of the typical scheduling hours, and during the few noon appointments.

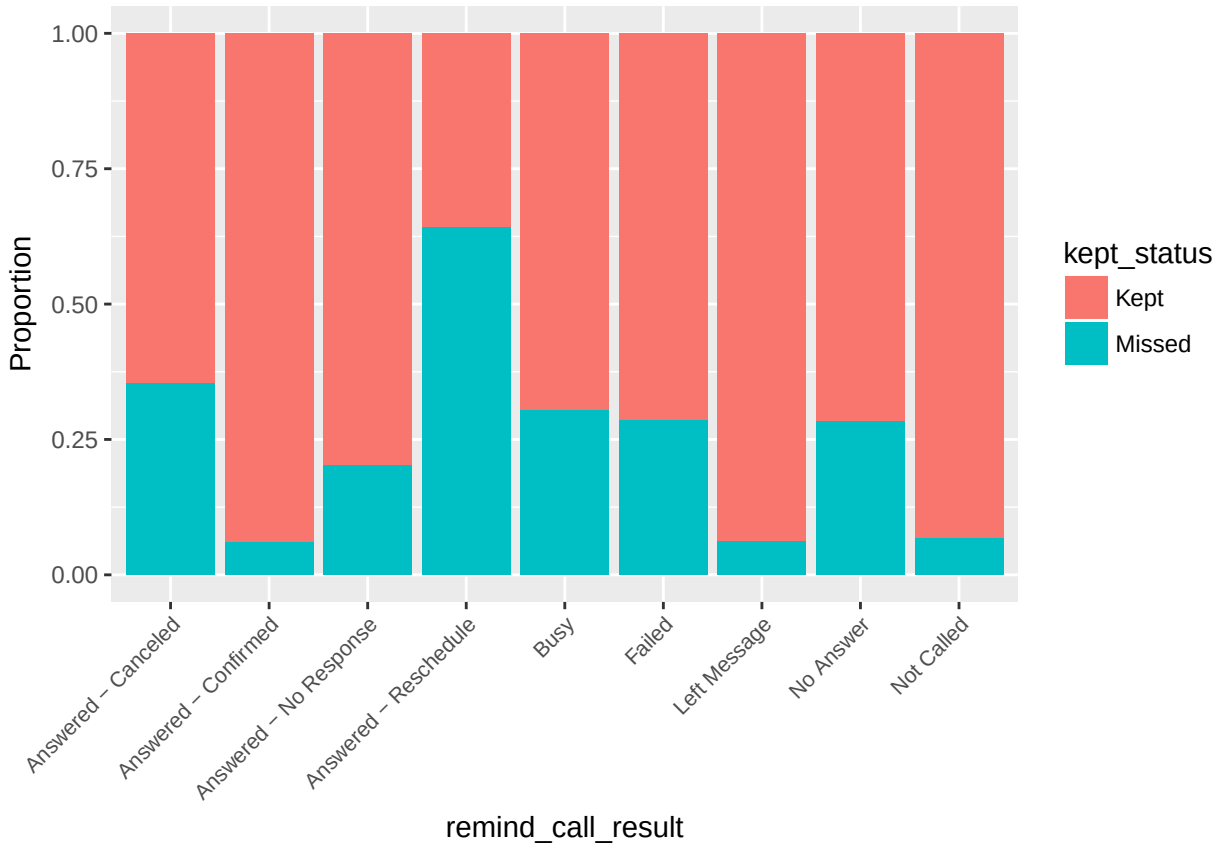
Reminder Call Result

```
table(appointments$remind_call_result)
```

```
##
##    Answered - Canceled   Answered - Confirmed Answered - No Response
##                    152                  49108                 180869
##  Answered - Reschedule                   Busy                 Failed
##                   1369                   1104                  27944
##           Left Message              No Answer             Not Called
##                  18430                    377                  63429
```

There are relatively few instances of “Answered - Cancelled”, “Answered - Reschedule”, “Busy”, and “No Answer”

```
ggplot(
    appointments,
    aes(x = remind_call_result, fill = kept_status)
) +
    geom_bar(position = "fill") +
    theme(axis.text.x = element_text(size = 8, angle = 45,hjust = 1, vjust = 1)) +
    labs(y = "Proportion")
```



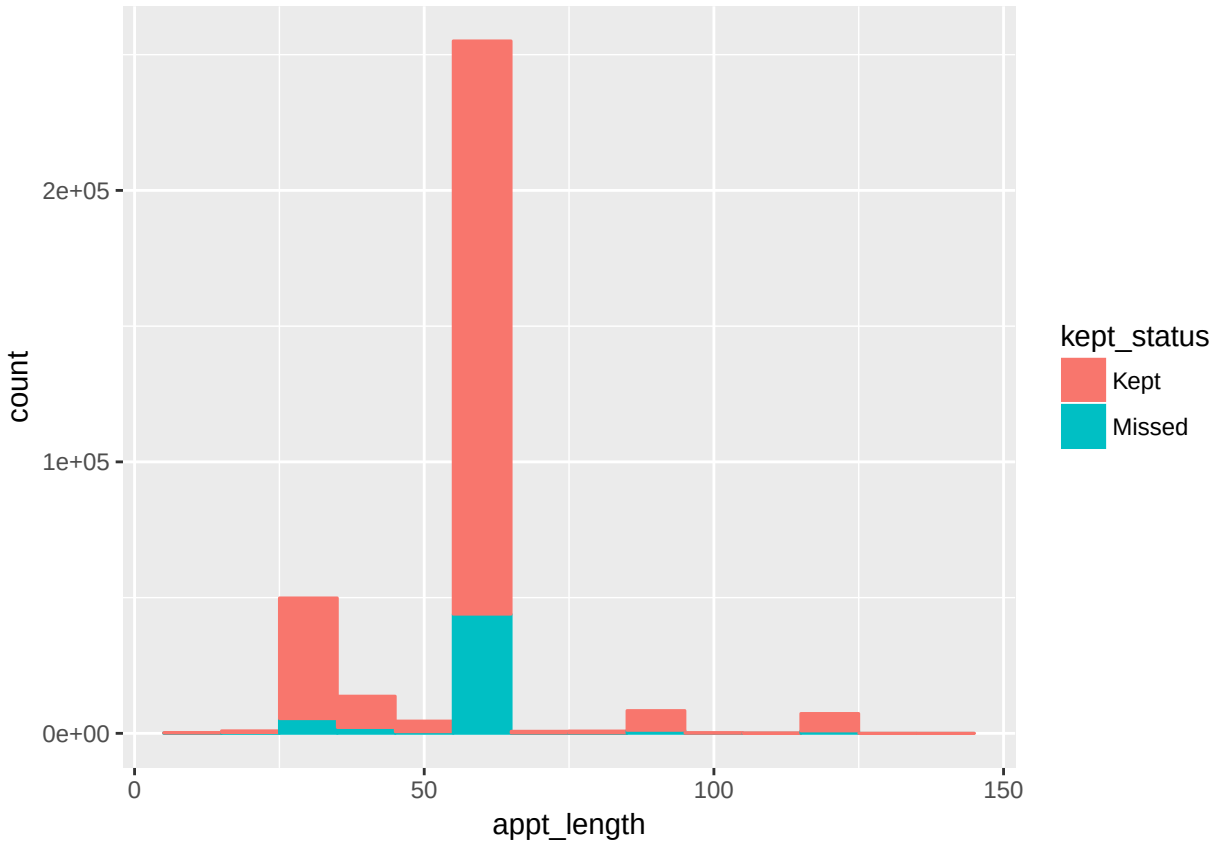
Reminder call responses of "Answered - Confirmed", "Left Message", and "Not Called" have the lowest ratios of missed appointments. This seems logical although I would not expect "Not Called" to have as low of a rate, since I would rate it a more neutral response. The best explanation for the low miss rate is a bias in choosing who doesn't need a reminder call.

Responses of "Answered - No Response", "Busy", "Failed", and "No Answer" have average or higher rates of missed appointments, which makes sense as I would rate these responses as neutral to slightly negative.

Responses "Answered - Cancelled" and "Answered - Reschedule" have the highest missed rates of about 35% and 65%, respectively. It's not surprising that these have the highest missed rates, but I would expect them to nearly 100%, particularly for "Answered - Cancelled". Perhaps the patients wanted to cancel or reschdule, but circumstances changed and they ultimately decided to come. Since there are so few cases with these responses, I won't look into this further, but it would be interesting to learn the reasons for these results.

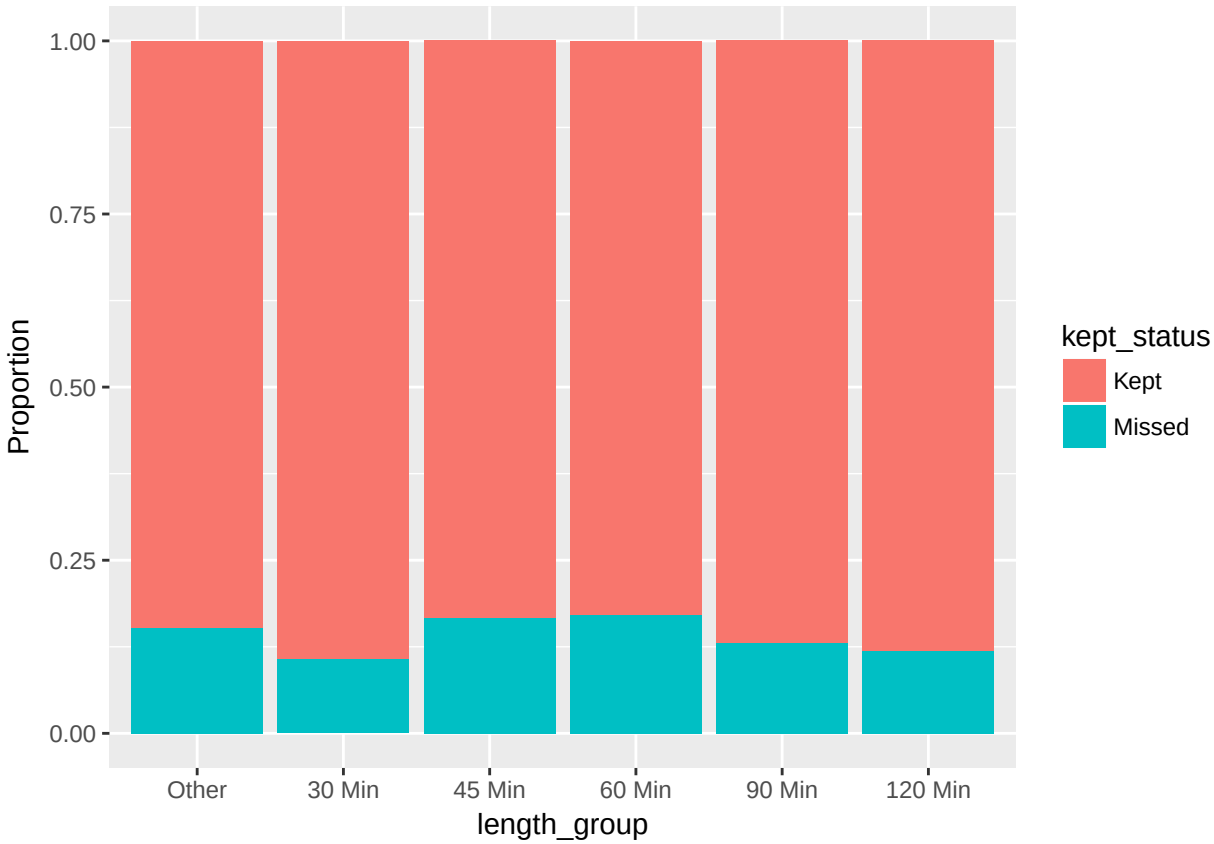
Appointment Length

```
appointments %>%
    filter(appt_length < 150) %>%
    ggplot(
        aes(x = appt_length, color = kept_status, fill = kept_status)
    ) +
        geom_histogram(binwidth = 10)
```



Most Appointments are 60 minutes long. 30-minute appointments are the next most common, followed by 45-minute, 90-minute, and 120-minute. I'll group them as 30, 45, 60, 90, or 120 minutes, and the rest as "Other", and take a look at the differences.

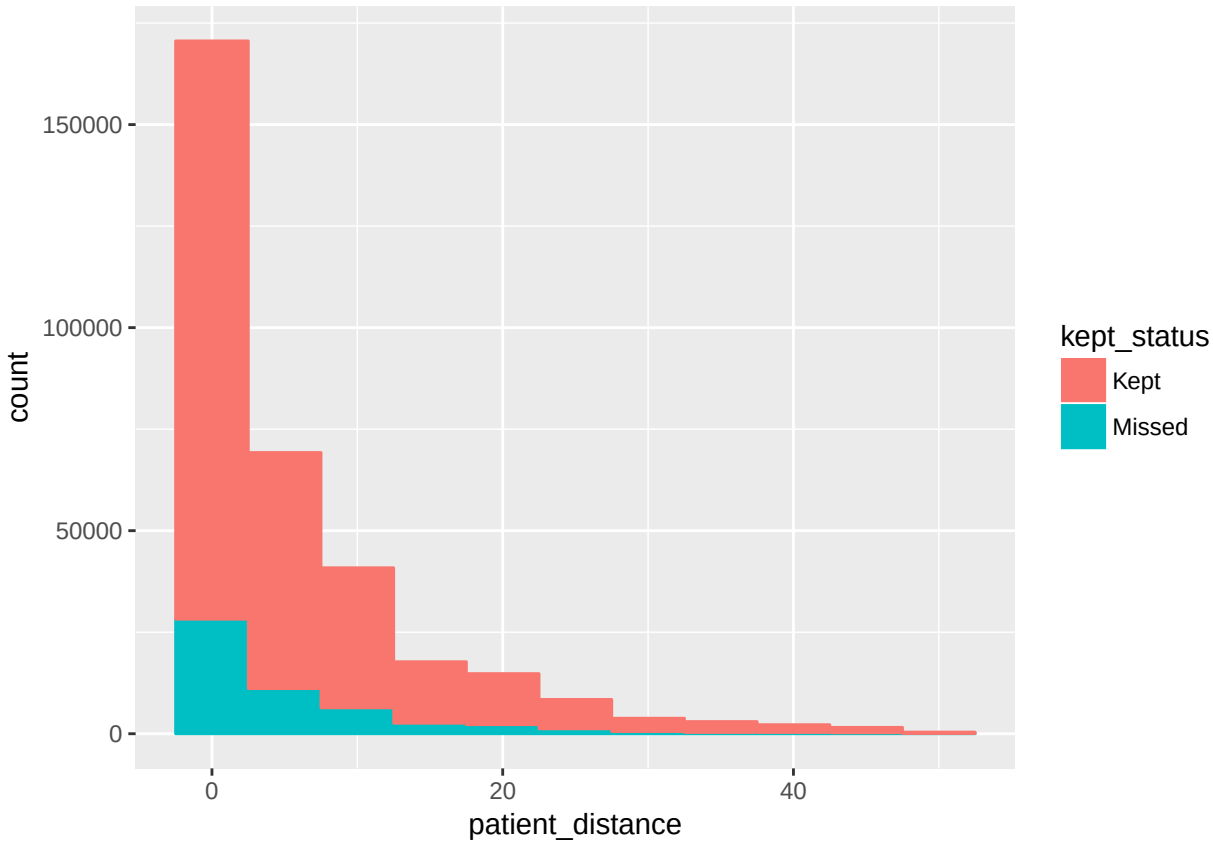
```
length_breaks <- c(-1, 29, 30, 44, 45, 59, 60, 89, 90, 119, 120, 1000)

length_labels <- c(
    "Other1", "30 Min", "Other2", "45 Min", "Other3", "60 Min", "Other4",
    "90 Min", "Other5", "120 Min", "Other6")

appointments <- appointments %>%
    mutate(
        length_group = cut(
            appt_length, breaks = length_breaks, labels = length_labels))

appointments$length_group <- appointments$length_group %>%
    fct_collapse(Other = c("Other1", "Other2", "Other3", "Other4", "Other5",
                           "Other6"))

ggplot(
    data = appointments,
    mapping = aes(x = length_group, fill = kept_status)
) +
    geom_bar(position = "fill") +
    labs(y = "Proportion")
```



Of the most common lengths, 30 and 60 minutes, the longer appointments are more likely to be missed. This could be because a patient is more inclined to go to a shorter appointment, or because shorter appointments are less likely to be impacted by scheduling conflicts. 90 and 120 minute appointments also have lower miss rates, which might be because they are more likely to be for a more important procedure. There could be some interaction between appointment length and provider specialty.

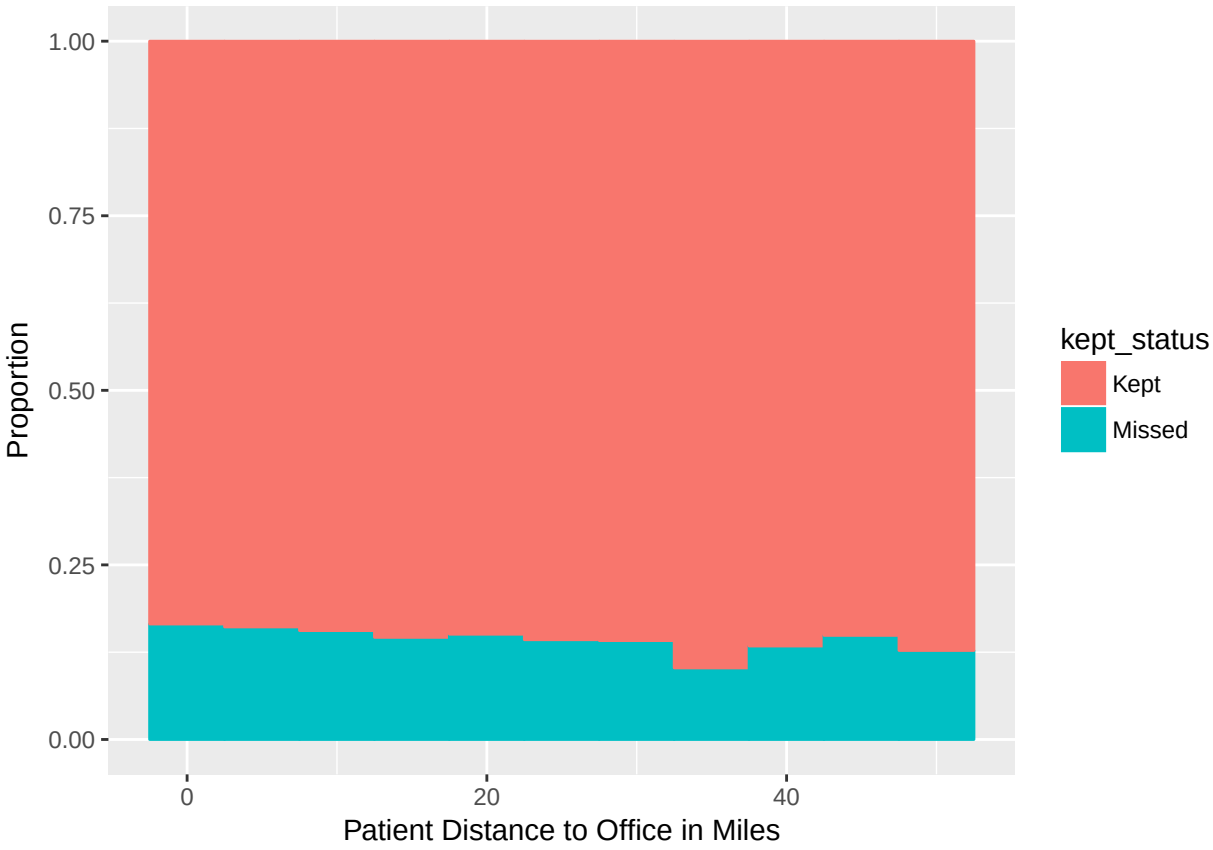
Patient Distance

```
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 the only one with missing values, which will be replaced with median rather than mean since it is a right-skewed distribution as would be expected. 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.

Milestone Summary

So far, the data hasn't required a lot of wrangling. There were just a few missing values that were replaced with the median for `patient_distance`, and some groupings that have been created. The `billing_type` variable had only one instance of "unkown", which was removed from the data.

Based on the initial findings, I havent had to alter my approach. The approach will be to create some new features, and then divide the data into a training, validation and test set. I will use the train data to train using glm and random forest, then run predictions with the models on the validation set to see which one performs better. I will select the best one and use the test data to evaluate the estimated performance.