Capstone Project - Predicting Patient No-Shows Using Appointment Data

Derek Samsom

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 still 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 prediction error will result in times where there are too many or too few patients at a given time. The main goal in the prediction will be to minimize the error, as this will reduce instances of having more or fewer patients that desired.

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. Missed appointments can only be predicted based on indirect factors that are known, such as past history as 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.

Medical providers can used the missed appointment predistions 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 an appointment booking system are client-specific ant not included in the scope of this project, which is limited to minimizing the error while predicting the 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)
library(randomForest)

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 be kept or missed.

There is no field that can be used to identify a specific patients 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 Summary and Structure

```r
summary(appointments)
```

```
##   kept_status         appt_date          appt_time         appt_length
##  Length:342862      Length:342862      Length:342862     Min.   : 10
##  Class :character   Class :character   Class1:hms        1st Qu.: 60
##  Mode  :character   Mode  :character   Class2:difftime   Median : 60
##                                        Mode  :numeric    Mean   : 57
##                                                          3rd Qu.: 60
##                                                          Max.   :600
##
##  date_scheduled      patient_age     patient_gender     billing_type
##  Length:342862      Min.   :  0.00   Length:342862      Length:342862
##  Class :character   1st Qu.: 17.00   Class :character   Class :character
##  Mode  :character   Median : 34.00   Mode  :character   Mode  :character
##                     Mean   : 35.56
##                     3rd Qu.: 54.00
##                     Max.   :264.00
##
##   prior_missed       prior_kept     patient_distance  office_zip
##  Min.   :  0.000   Min.   :  0.00   Min.   :   0.0   Length:342862
##  1st Qu.:  1.000   1st Qu.:  2.00   1st Qu.:   0.0   Class :character
##  Median :  2.000   Median :  6.00   Median :   3.0   Mode  :character
##  Mean   :  2.451   Mean   :  8.02   Mean   :  10.8
##  3rd Qu.:  3.000   3rd Qu.: 11.00   3rd Qu.:   9.0
##  Max.   :117.000   Max.   :676.00   Max.   :2688.0
##                                     NA's   :974
##  provider_specialty remind_call_result
##  Length:342862      Length:342862
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
```

```r
str(appointments, give.attr = FALSE)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    342862 obs. of  14 variables:
##  $ kept_status       : chr  "Kept" "Kept" "Kept" "Kept" ...
##  $ appt_date         : chr  "9/1/16" "9/1/16" "9/1/16" "9/1/16" ...
##  $ appt_time         :Classes 'hms', 'difftime'  atomic [1:342862] 19800 28800 28800 28800 28800 288
##  $ appt_length       : int  90 60 120 60 60 60 60 60 60 90 ...
##  $ date_scheduled    : chr  "8/1/16" "1/18/16" "2/3/16" "6/8/16" ...
##  $ patient_age       : int  7 75 31 45 49 71 49 38 36 13 ...
##  $ patient_gender    : chr  "Male" "Female" "Male" "Male" ...
##  $ billing_type      : chr  "DMAP" "Commercial" "DMAP" "DMAP" ...
##  $ prior_missed      : int  1 2 1 6 5 6 8 0 2 3 ...
##  $ prior_kept        : int  3 5 5 15 6 6 20 0 5 12 ...
##  $ patient_distance  : int  41 29 5 5 0 5 0 539 0 4 ...
##  $ office_zip        : chr  "AP" "BL" "BL" "BL" ...
##  $ provider_specialty: chr  "A" "A" "A" "B" ...
##  $ remind_call_result: chr  "Left Message" "Answered - Confirmed" "Left Message" "Answered - No Resp
```

```r
head(appointments[, 1:5])
```

```
## # A tibble: 6 x 5
##   kept_status appt_date appt_time appt_length date_scheduled
##   <chr>       <chr>     <time>          <int> <chr>
## 1 Kept        9/1/16    05:30              90 8/1/16
## 2 Kept        9/1/16    08:00              60 1/18/16
## 3 Kept        9/1/16    08:00             120 2/3/16
## 4 Kept        9/1/16    08:00              60 6/8/16
## 5 Missed      9/1/16    08:00              60 6/28/16
## 6 Kept        9/1/16    08:00              60 7/12/16
```

```r
head(appointments[, 6:10])
```

```
## # A tibble: 6 x 5
##   patient_age patient_gender billing_type prior_missed prior_kept
##         <int> <chr>          <chr>               <int>      <int>
## 1           7 Male           DMAP                    1          3
## 2          75 Female         Commercial              2          5
## 3          31 Male           DMAP                    1          5
## 4          45 Male           DMAP                    6         15
## 5          49 Male           Commercial              5          6
## 6          71 Male           DMAP                    6          6
```

```r
head(appointments[, 11:14])
```

```
## # A tibble: 6 x 4
##   patient_distance office_zip provider_specialty remind_call_result
##              <int> <chr>      <chr>              <chr>
## 1               41 AP         A                  Left Message
## 2               29 BL         A                  Answered - Confirmed
## 3                5 BL         A                  Left Message
## 4                5 BL         B                  Answered - No Response
## 5                0 BL         B                  Answered - No Response
## 6                5 BL         A                  Answered - Confirmed
```

Data Dictionary

```r
variable_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")
variable <- colnames(appointments)
```

```r
variable_type <- unlist(map(appointments, class))
variable_type <- variable_type[-4]
as_data_frame(cbind(c(1:length(variable)), variable, variable_type, variable_descriptions))
```

```
## # A tibble: 14 x 4
##    V1    variable           variable_type variable_descriptions
##    <chr> <chr>              <chr>         <chr>
##  1 1     kept_status        character     Dependent variable: kept or mis~
##  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 appointm~
## 10 10    prior_kept         integer       Number of prior kept appointmen~
## 11 11    patient_distance   integer       Patient distance from office in~
## 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`.

```r
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.

```r
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.

```r
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
```

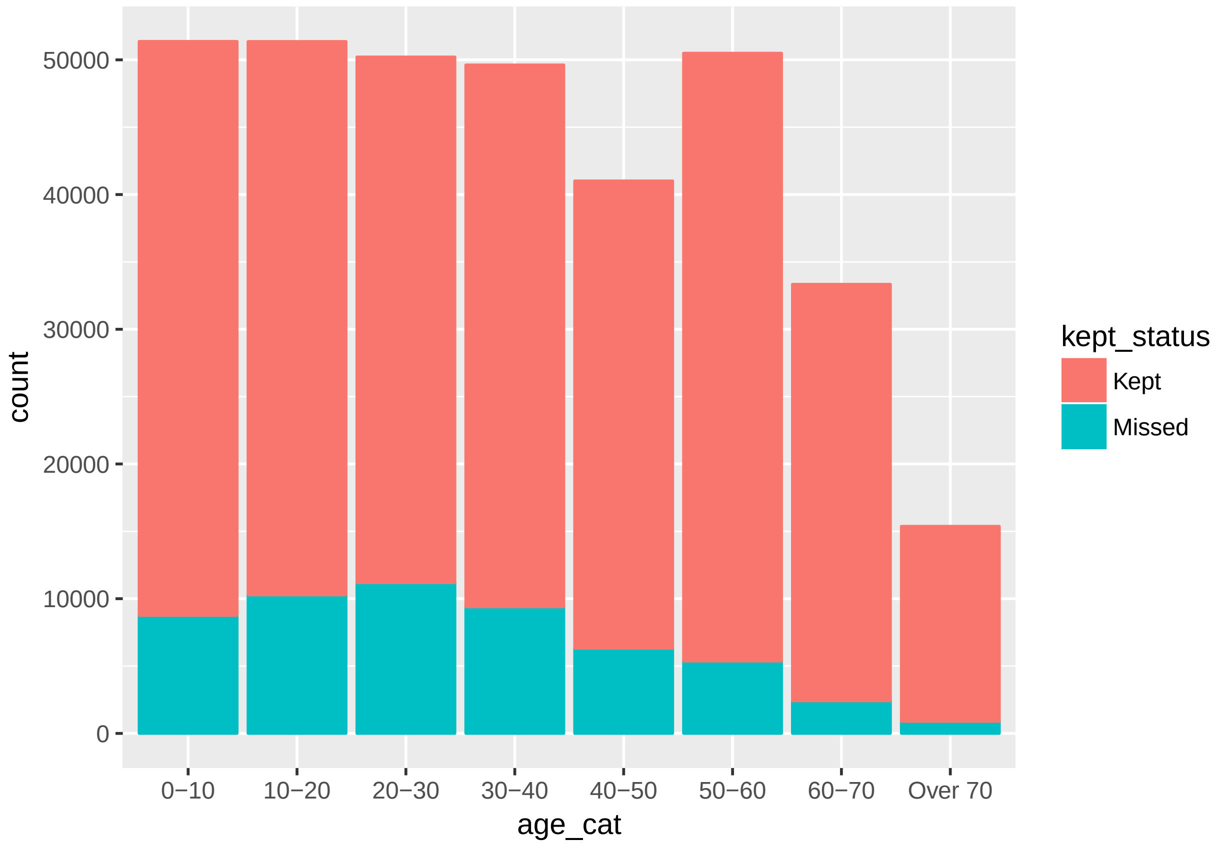
One variable, `patient_distance` has 974 missing value. This is fairly minor and will be evaluated later on when exploring the variable further.

patient_age

I expected missed appointments to have to vary across age ranges. Perhaps older patients have fewer commitments with kids or work, and make their appointments more regularly, or perhaps younger adults might skip more appointments because they aren't as critical? I will break the data into age groups to make the plot simpler to evaluate.

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.

```
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()
```



Missed appointments are highest with young adults, and decrease with older and younger patients.

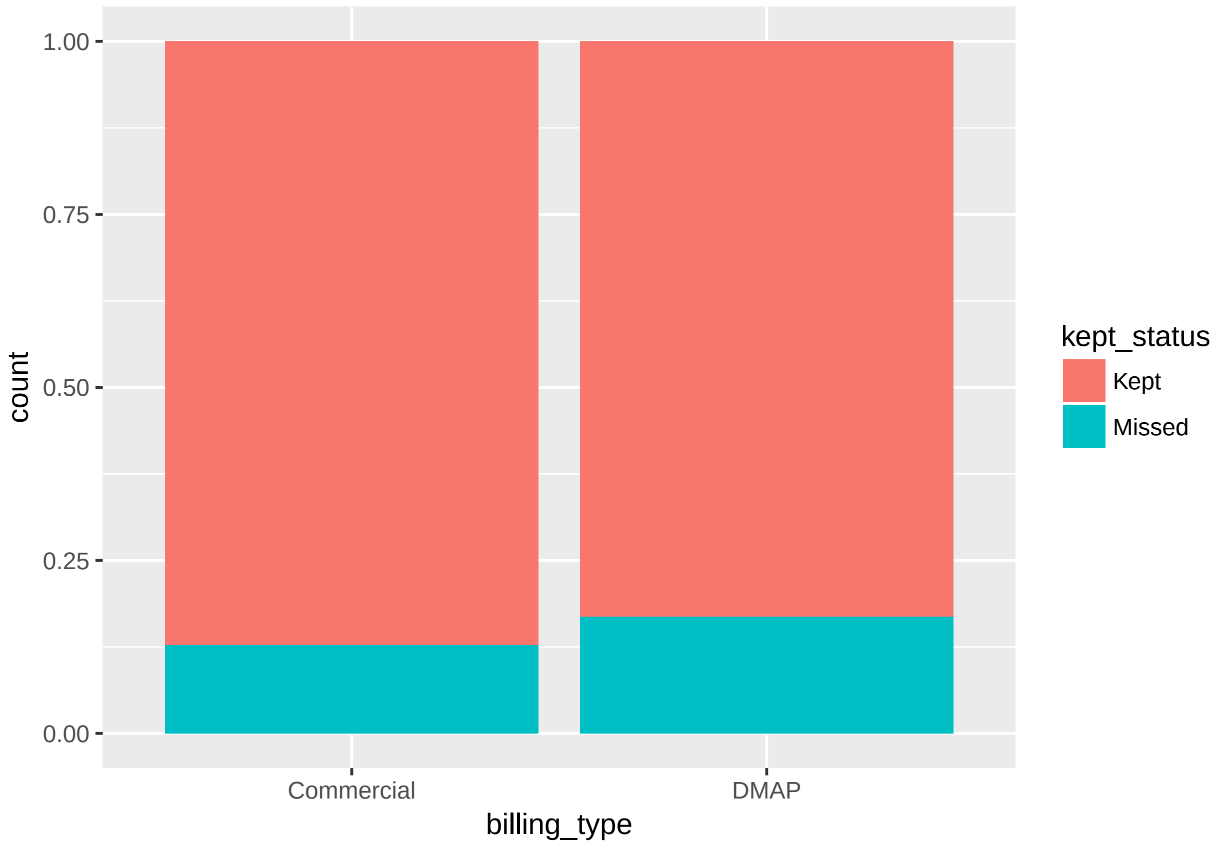
billing_type

```
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")
```



There is a minor difference between billing types. DMAP has a higher proportion of missed appointments than commercial.

appt_datetime

For the variable `appt_datetime`, I will create an `hour` variable to see the variation in missed appointments by hour of day.

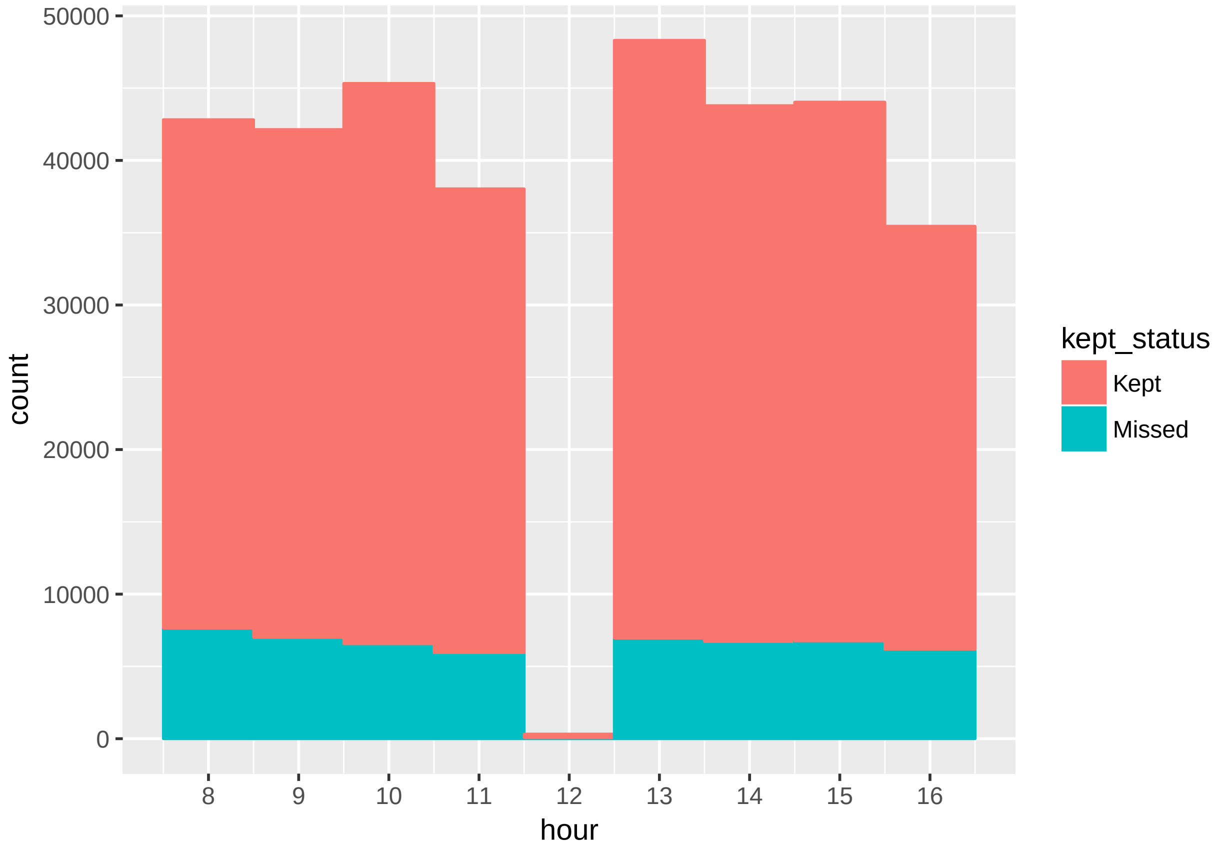
```
appointments <- appointments %>%
    mutate(hour = lubridate::hour(appointments$appt_datetime))

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.

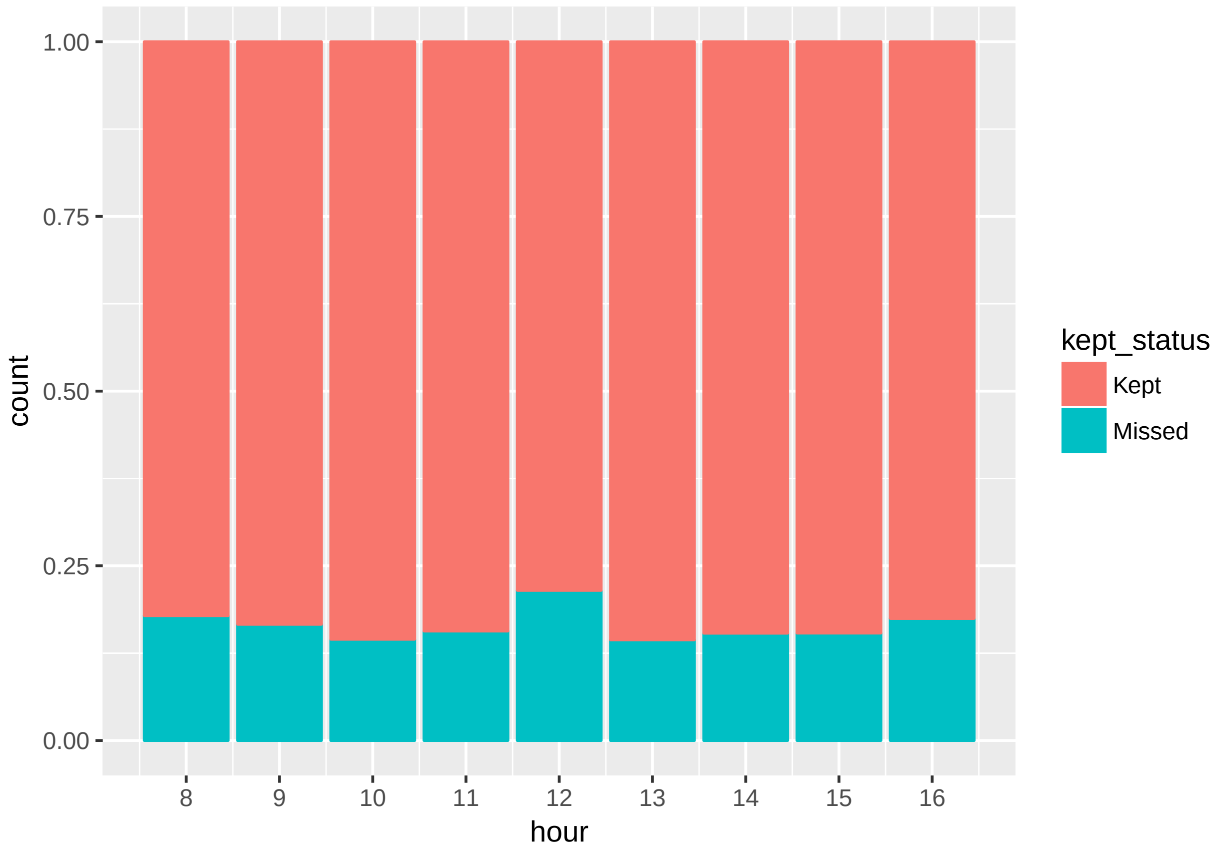
```
appointments_hour <- appointments %>%
    select(kept_status, hour) %>%
    filter(hour >= 8 & hour <= 16)

ggplot(
    appointments_hour,
    aes(x = hour, col = kept_status, fill = kept_status)
```

```r
) +
    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 towars the end of the two periods.

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



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

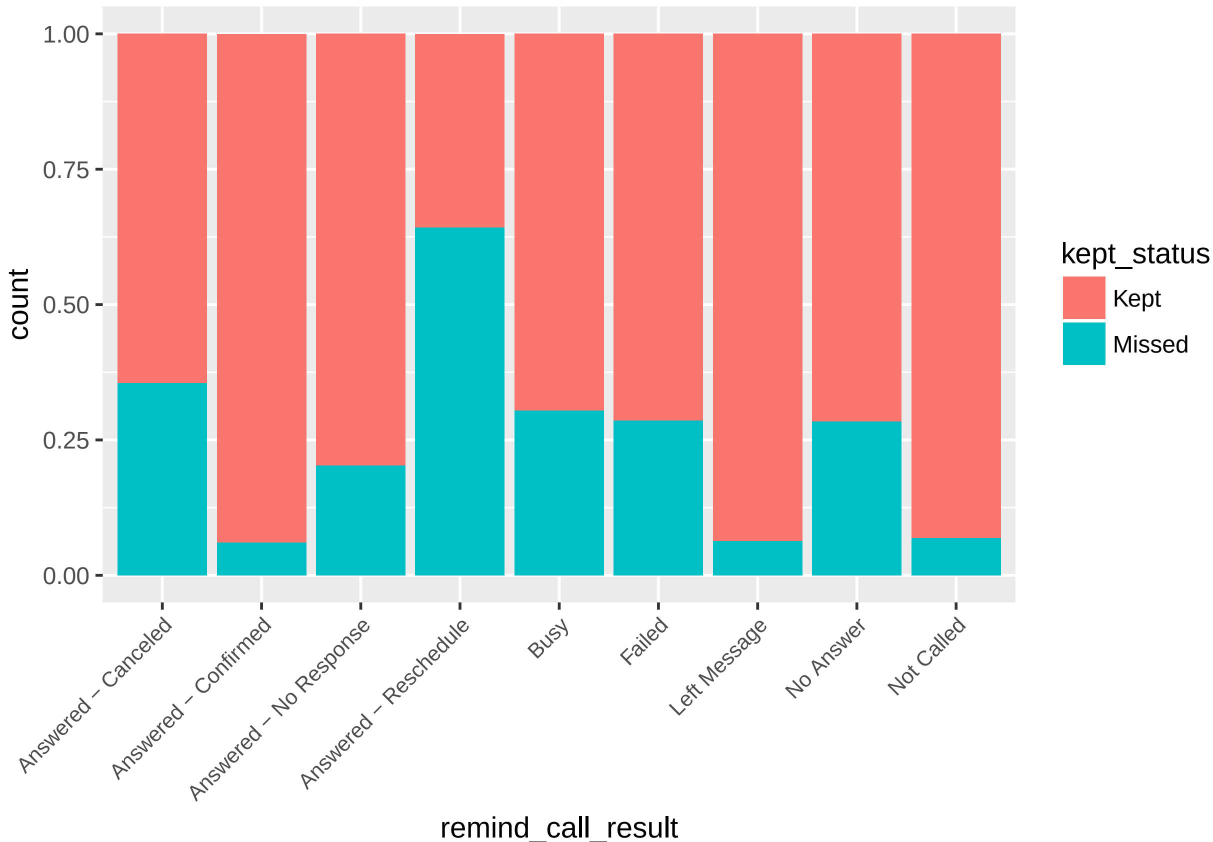
remind_call_result

```r
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
```

Low counts of "Answered - Cancelled", "Answered - Reschedule", "Busy", and "No Answer"

```r
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))
```



~65% of appointments with "Answered - Cancelled" and ~35% with "Answered-Reschedule" still kept their appointments, however, very few observations in these categories.

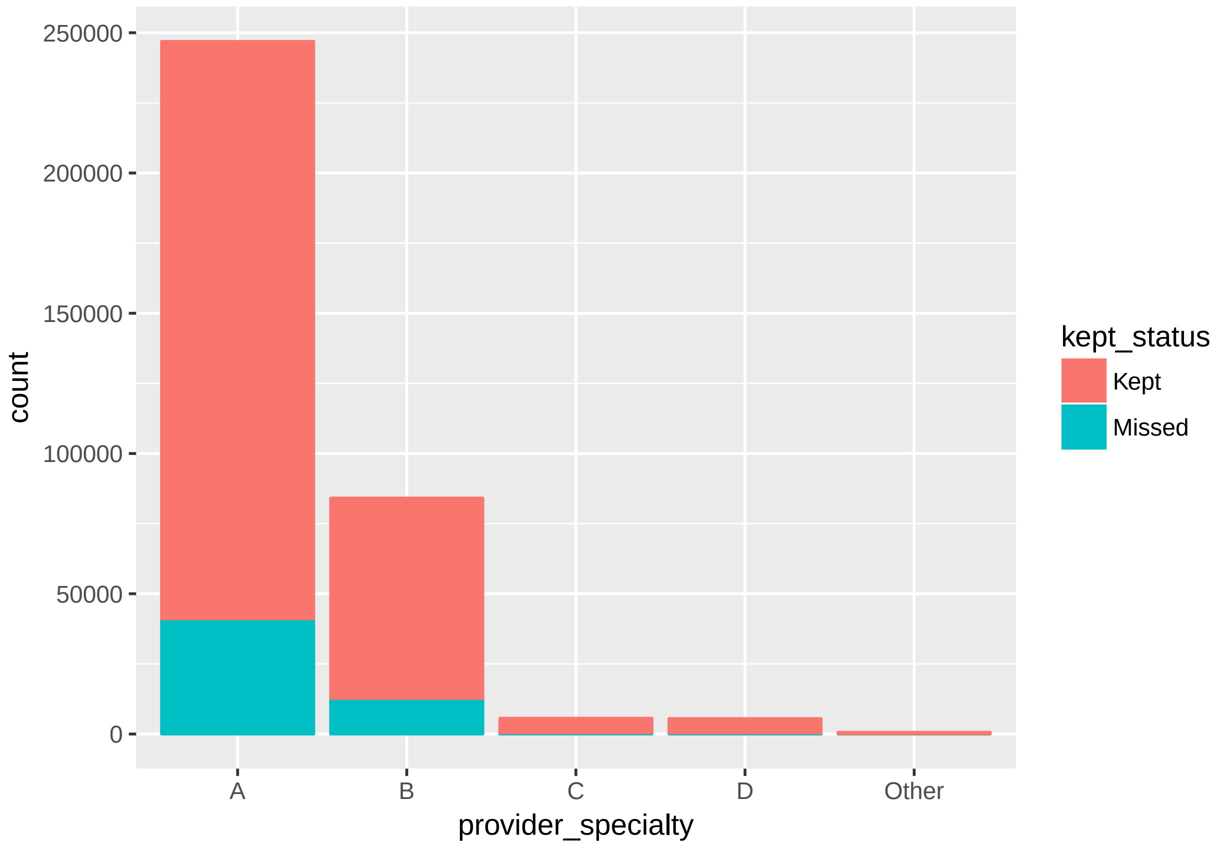
provider_specialty

```
table(appointments$provider_specialty)
```

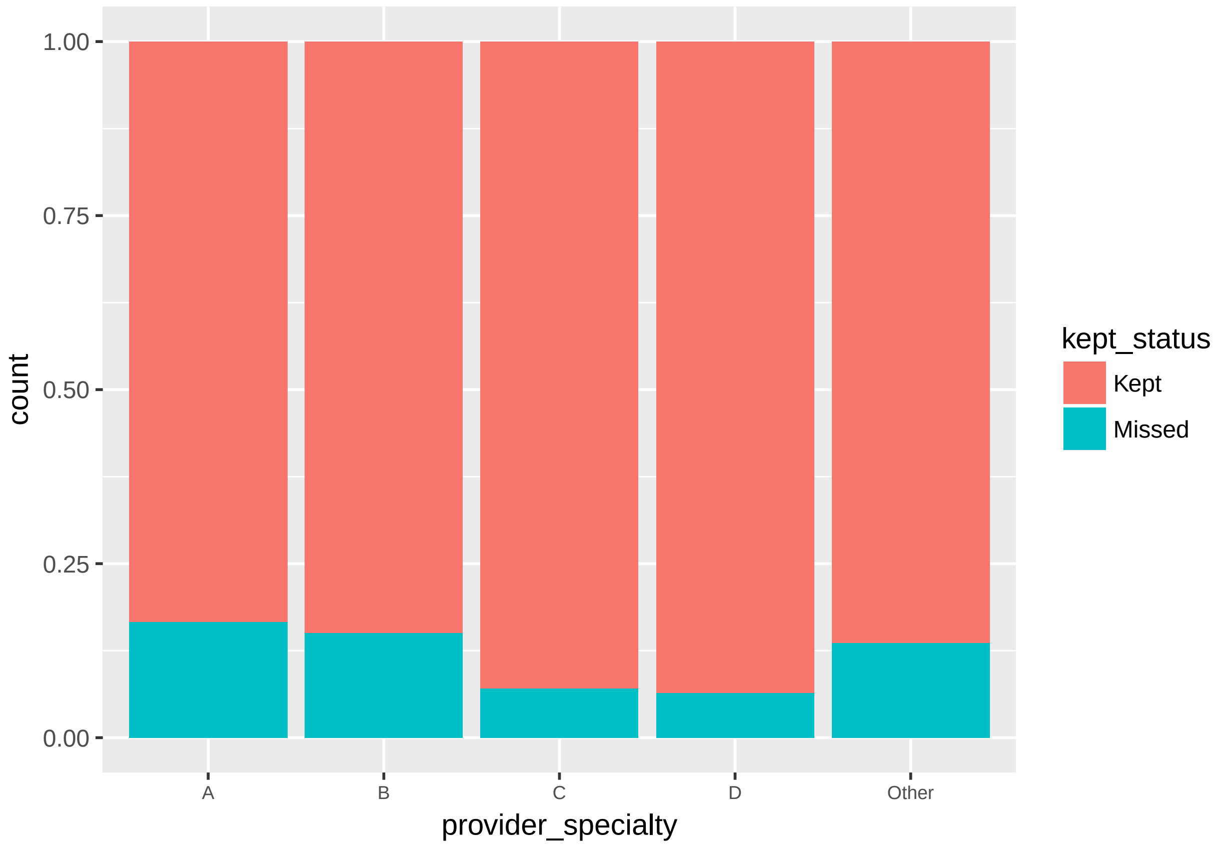
```
##
##       A      B      C      D      E      F      G
## 246917  84115   5623   5512     42    525     48
```

Most observations are specialty A and B. Specialties E, F, and G have very few observations and will be grouped as "Other".

```
appointments$provider_specialty <- appointments$provider_specialty %>%
    fct_collapse(Other = c("E", "F", "G"))


ggplot(
    appointments,
    aes(x = provider_specialty, col = kept_status, fill = kept_status)
) +
    stat_count()
```

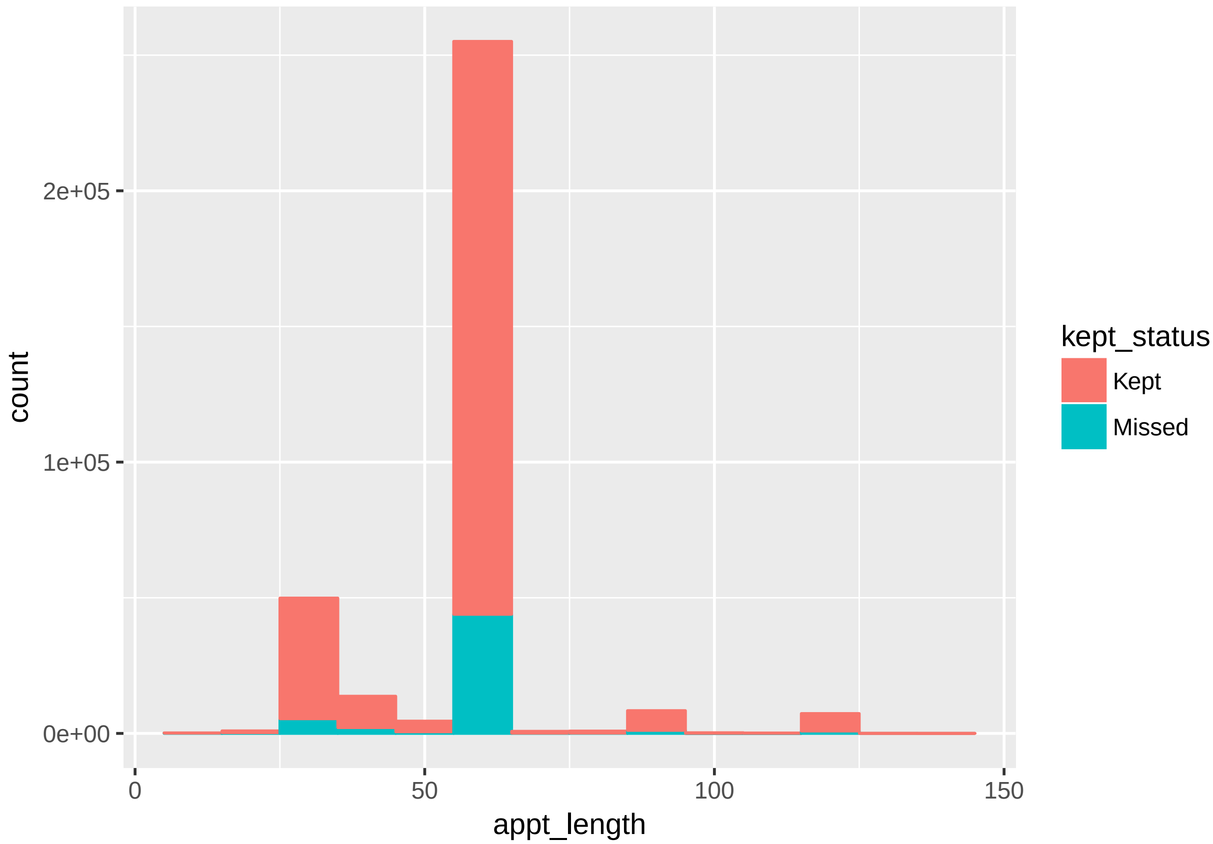



```
ggplot(
    appointments,
    aes(x = provider_specialty, fill = kept_status)
) +
    geom_bar(position = "fill") +
    theme(axis.text.x = element_text(size = 7))
```



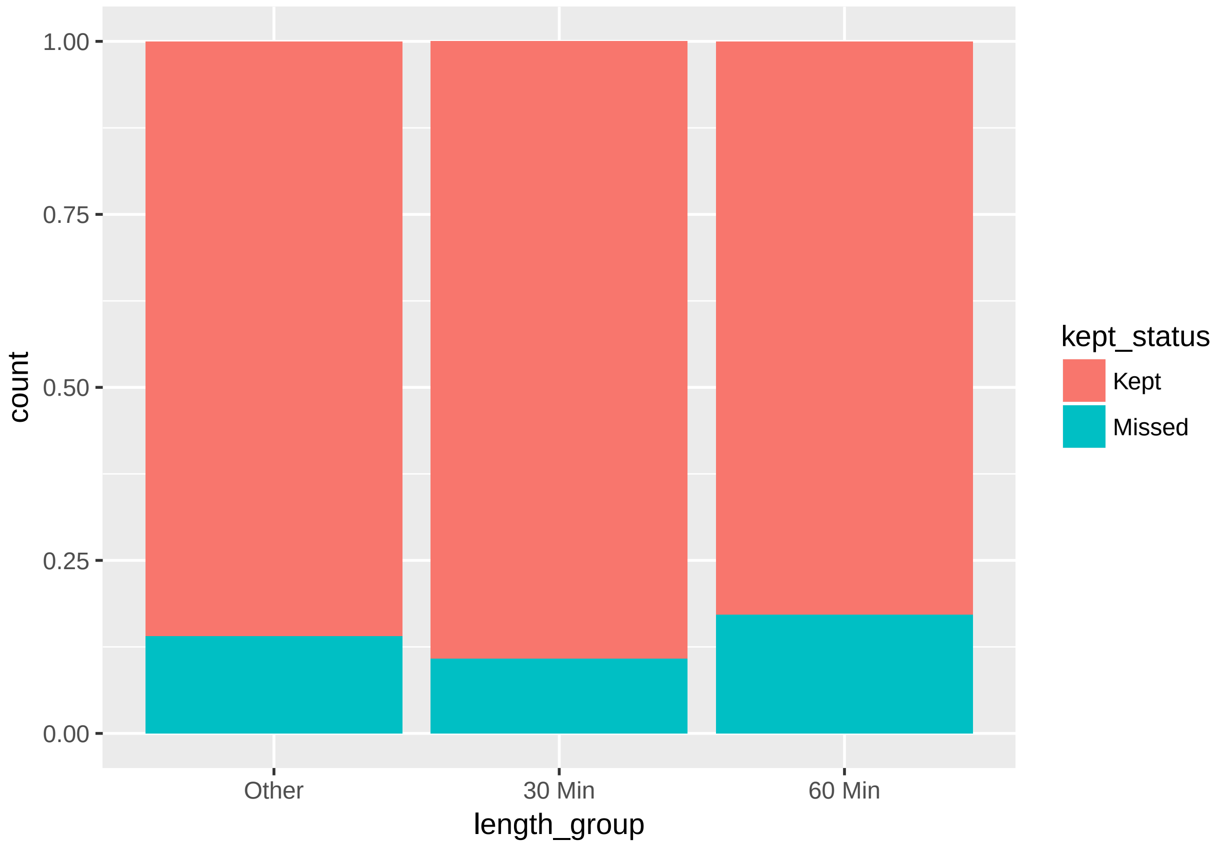
appt_length

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



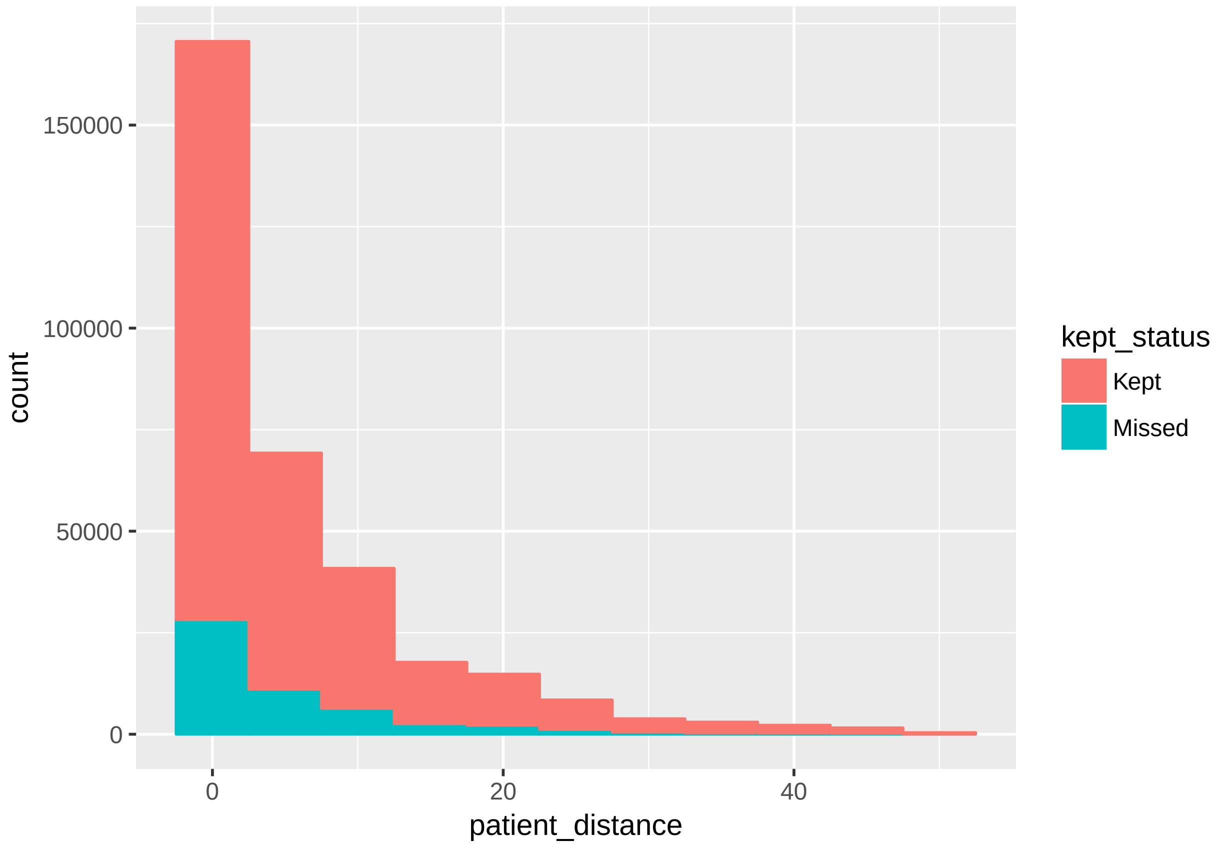
Most Appointments are 60 minutes long. 30-minute appointments are next most common.

```
length_breaks <- c(-1, 29, 30, 59, 60, 1000)

length_labels <- c("Other1", "30 Min", "Other2", "60 Min", "Other3")

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"))

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



patient_distance

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



patient_distance is right-skewed, therefore NA values will be replaced with median rather than mean.

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

New Variables

In addition to the original variables, there are several additional variables that can be calculated based on the originals.

The `percent_missed` variable is the percentage of prior appointments missed, calculated by dividing the prior missed appointments by the total number of prior appointments. For new patients, this calculation will result in an error because it will be attempting to divide by zero. The errors will be replaced with zero.

The variable `is_new_patient` will specify whether a patient is new, represented by a 1, or existing, represented by 0. My hypothesis is that new patients are more likely to keep their appointments, since I think it is human nature to try to give a good first impression. This is calculated by searching for appointments where `prior_missed` and `prior_kept` are both 0.

The variable `appt_lead_time` will calculate how far in advance an appointment was booked. This is calculated by taking the difference between `date_scheduled` and `appt_date`. If people are more likely to forget appointments booked farther in advance, or they are more likely to be for less urgent preventative care than last minute appointnemtns, this will pick that up.

The variable `appt_weekday` is the day of the week the appointments occurs, and `weekday_scheduled` is the date the appointment was booked.

```
appointments <- appointments %>%
    mutate(percent_missed = prior_missed / (prior_missed + prior_kept)) %>%
```

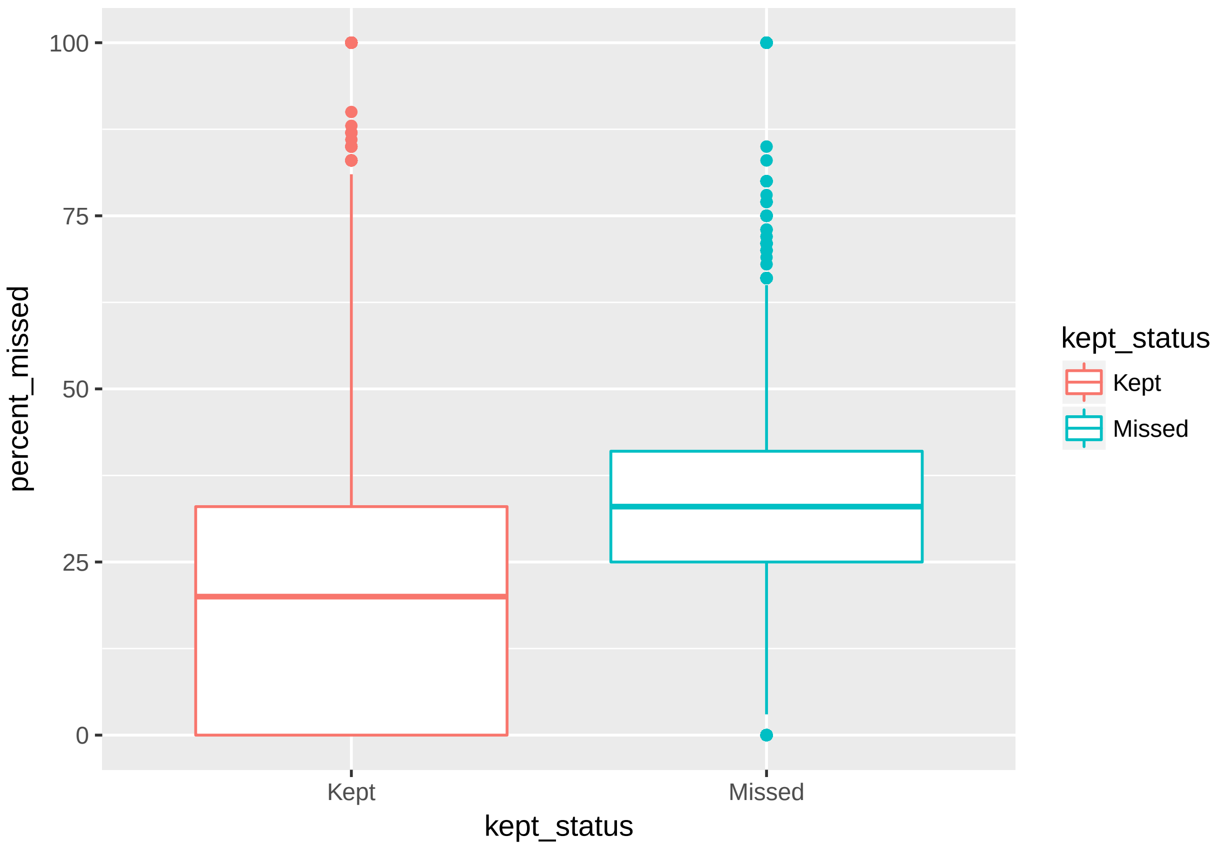
```
    mutate(
        is_new_patient = ifelse(prior_missed == 0 & prior_kept == 0, 1, 0)) %>%
    mutate(appt_lead_time = date(appt_datetime) - date(date_scheduled)) %>%
    mutate(appt_weekday = strftime(appt_datetime, "%A")) %>%
    mutate(weekday_scheduled = strftime(date_scheduled, "%A"))

appointments$percent_missed <- as.integer(appointments$percent_missed * 100)
appointments$percent_missed <- appointments$percent_missed %>%
    tidyr::replace_na(0)
```

Add `county_code` from zipcode data.

```
appointments <- dplyr::left_join(appointments, zipcodes, by = "office_zip")
```

percent_missed

```
ggplot(
    data = appointments,
    aes(x = kept_status, y = percent_missed, col = kept_status)
) +
    geom_boxplot()
```

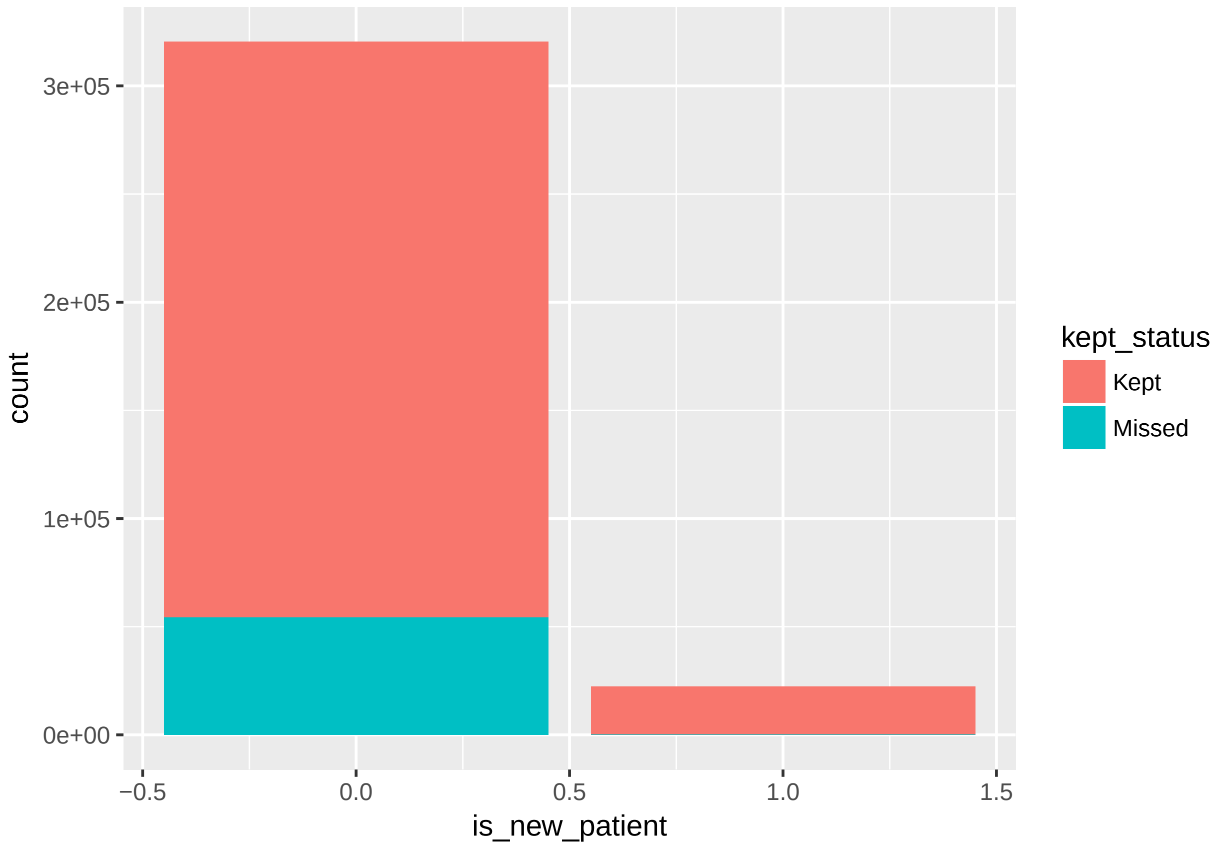


is_new_patient

```r
table(appointments$is_new_patient)
```

```
##
##      0      1
## 320442  22340
```

```r
ggplot(
    appointments,
    aes(x = is_new_patient, fill = kept_status)
) +
    geom_bar()
```



New patients have a very high percentage of kept appointments, but make up a small percentage of the total appointments.

appt_lead_time

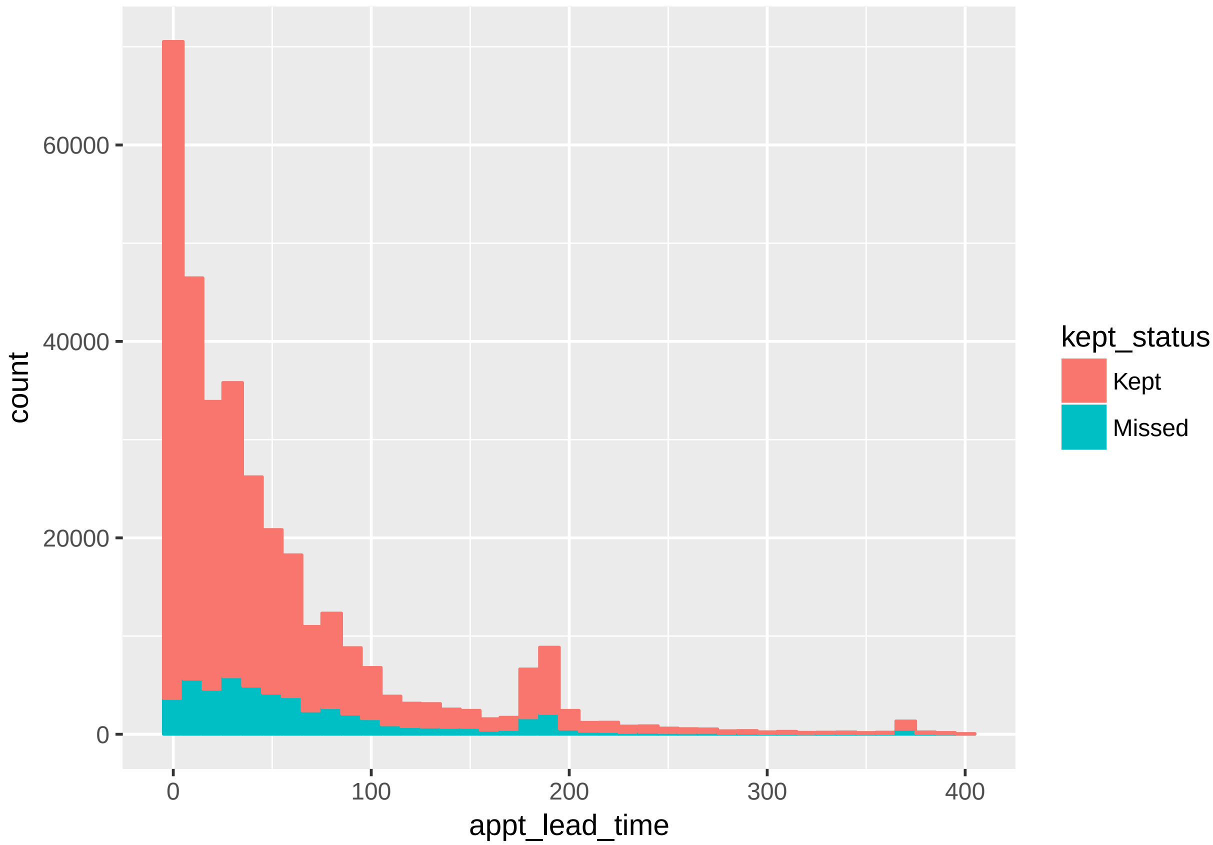
First I will check for negative values. Negative values suggest the appointment occured before it was scheduled, which wouldn't be possible and must represent an entry error.

```r
length(which(appointments$appt_lead_time < 0))
```

```
## [1] 82
```

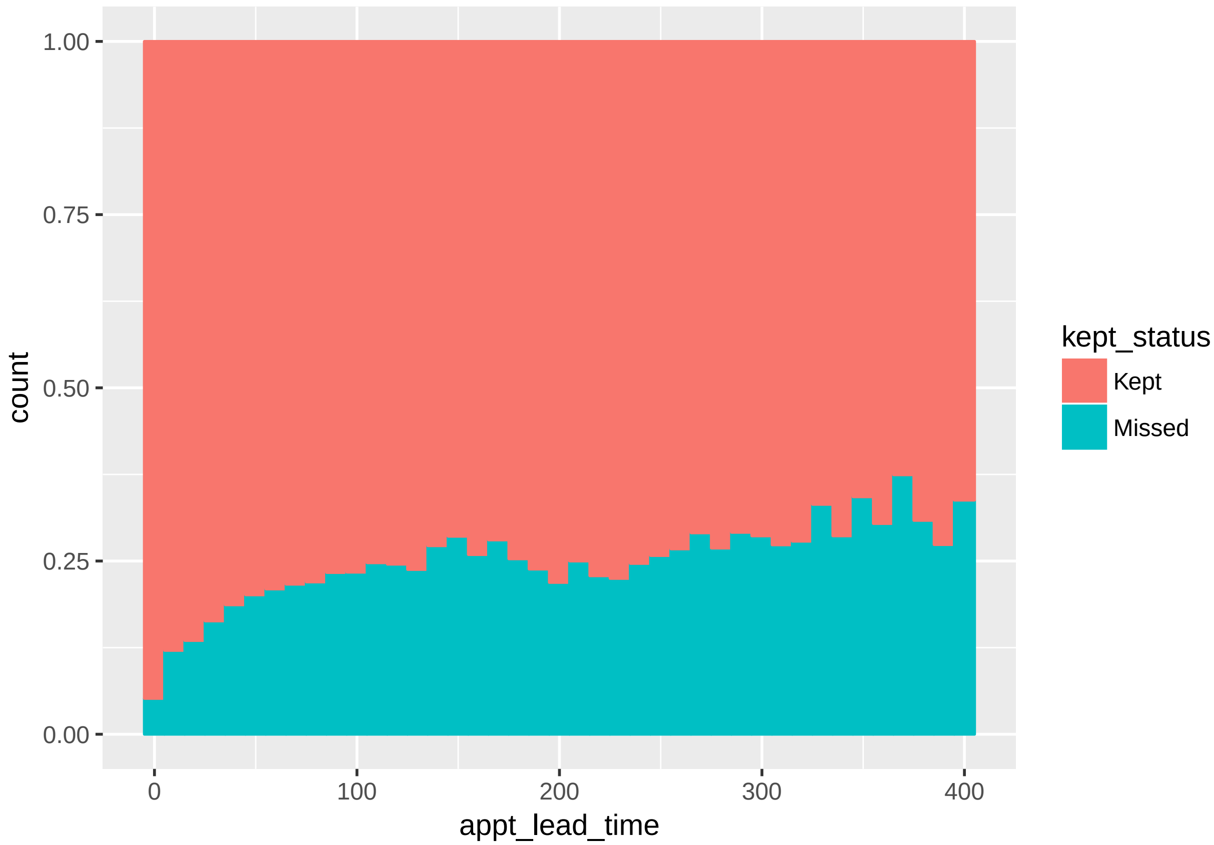
There are 82 negative values, which will be replaced with zero.

```
appointments$appt_lead_time <- ifelse(appointments$appt_lead_time < 0, 0, appointments$appt_lead_time)

appointments %>%
    filter(appt_lead_time >= 0 & appt_lead_time < 400) %>%
    ggplot(
        aes(x = appt_lead_time, color = kept_status, fill = kept_status)
    ) +
            geom_histogram(binwidth = 10)
```




```
appointments %>%
    filter(appt_lead_time >= 0 & appt_lead_time < 400) %>%
    ggplot(
        aes(x = appt_lead_time, color = kept_status, fill = kept_status)
    ) +
            geom_bar(binwidth = 10, position = "fill")
```

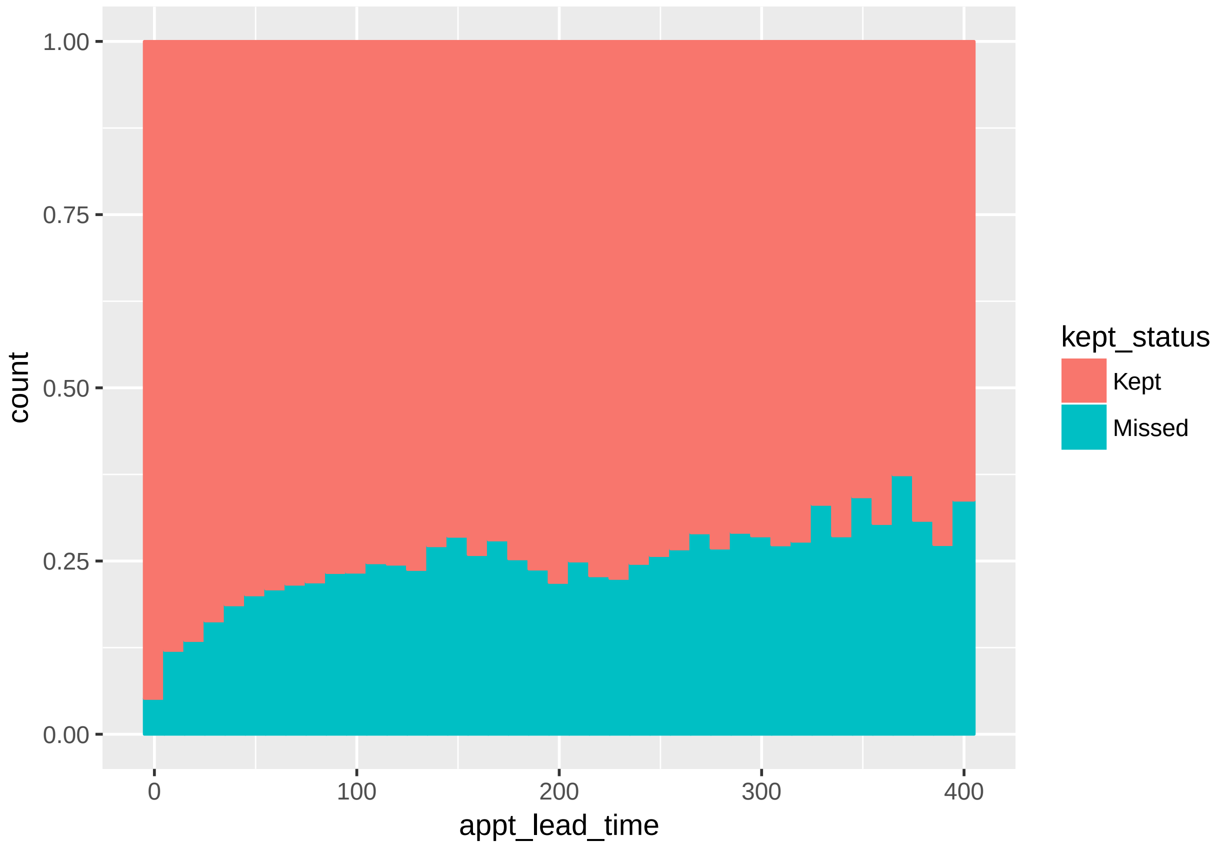
```
## Warning: `geom_bar()` no longer has a `binwidth` parameter. Please use
## `geom_histogram()` instead.
```



There is a lower proportion of missed appointments among those with the shortest lead times. This backs up my theory that shorter lead times could indicate more urgent health matters which I wouldn't expect a patient to be as likely to skip. Also, there are small bumps in the histogram around 180 and 360 days, which is probably indicitive of regular 6-month and 12-month checkups.

```
appointments %>%
    filter(appt_lead_time >= 0 & appt_lead_time < 400) %>%
    ggplot(
        aes(x = appt_lead_time, color = kept_status, fill = kept_status)
    ) +
            geom_bar(binwidth = 10, position = "fill")
```

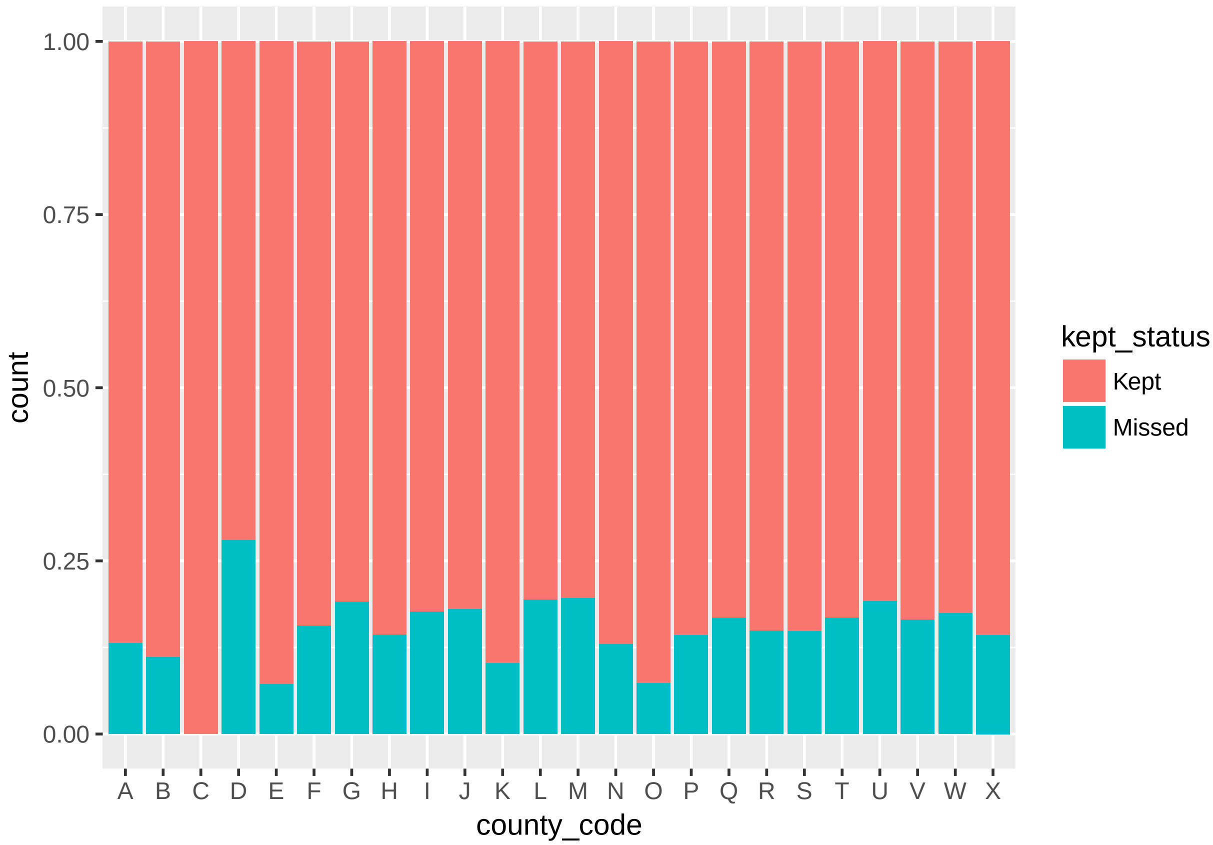
```
## Warning: `geom_bar()` no longer has a `binwidth` parameter. Please use
## `geom_histogram()` instead.
```



Looking at `appt_lead_time` proportionally, the drop in missed appointments among those with the shortest lead tims is easier to see.

county_code

```
ggplot(
    appointments,
    aes(x = county_code, fill = kept_status)
) +
    geom_bar(position = "fill")
```



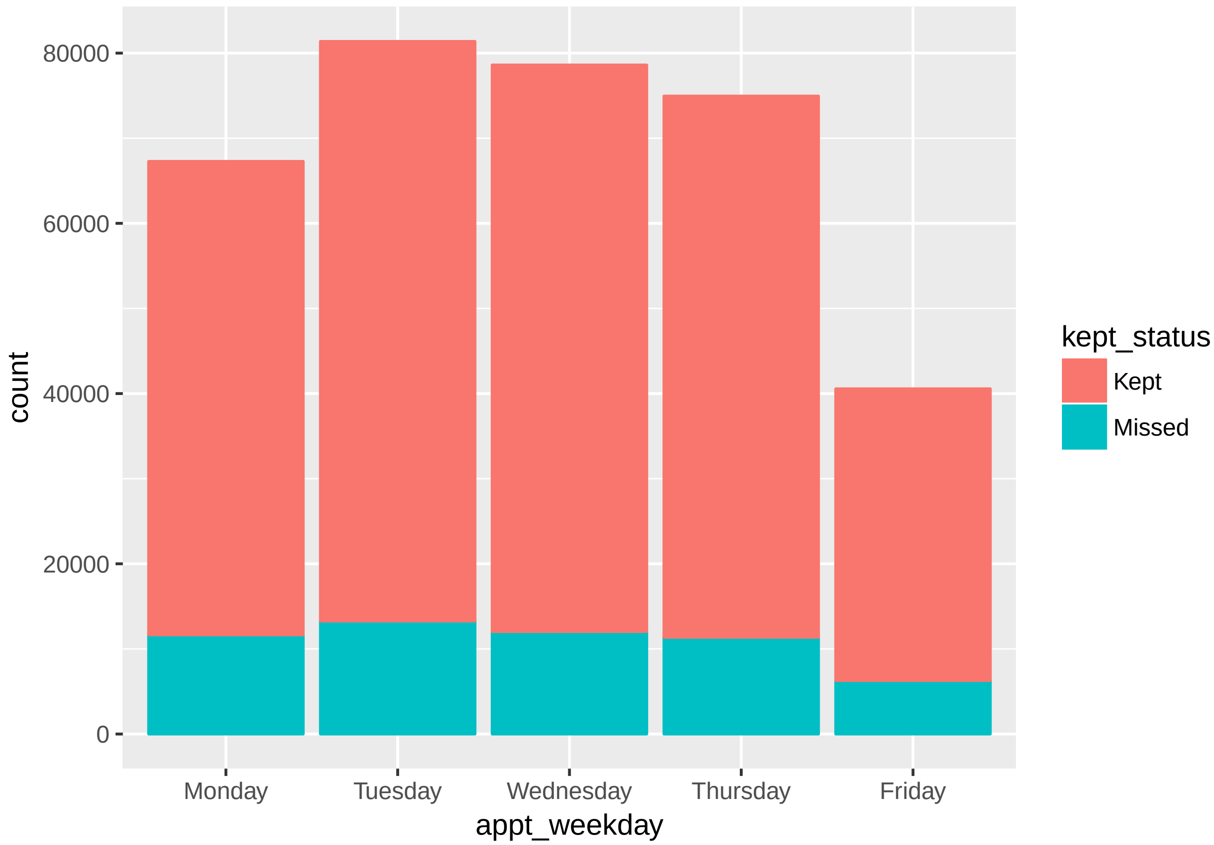
appt_weekday

```
table(appointments$appt_weekday)
```

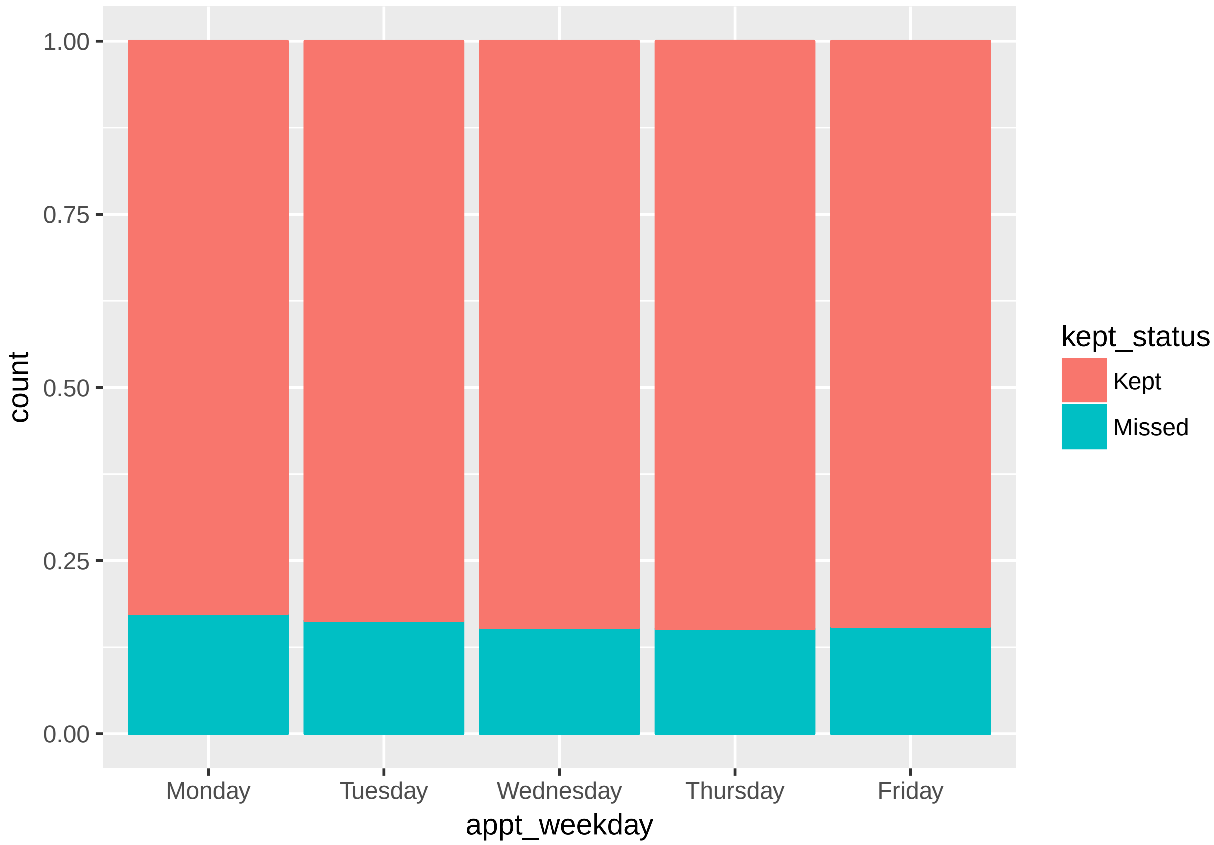
```
##
##    Friday    Monday    Sunday  Thursday   Tuesday Wednesday
##     40558     67280        12     74952     81376     78604
```

There are only 12 Sunday appointments, so I will remove the observations. I will also convert the character variable to an ordered factor to see the days of the week in the correct order.

```
appointments <- appointments %>%
    filter(appt_weekday != "Sunday")


appointments$appt_weekday <- factor(
    appointments$appt_weekday,
    levels = c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday")
)

ggplot(
    appointments,
    aes(x = appt_weekday, color = kept_status, fill = kept_status)
) +
    geom_bar()
```

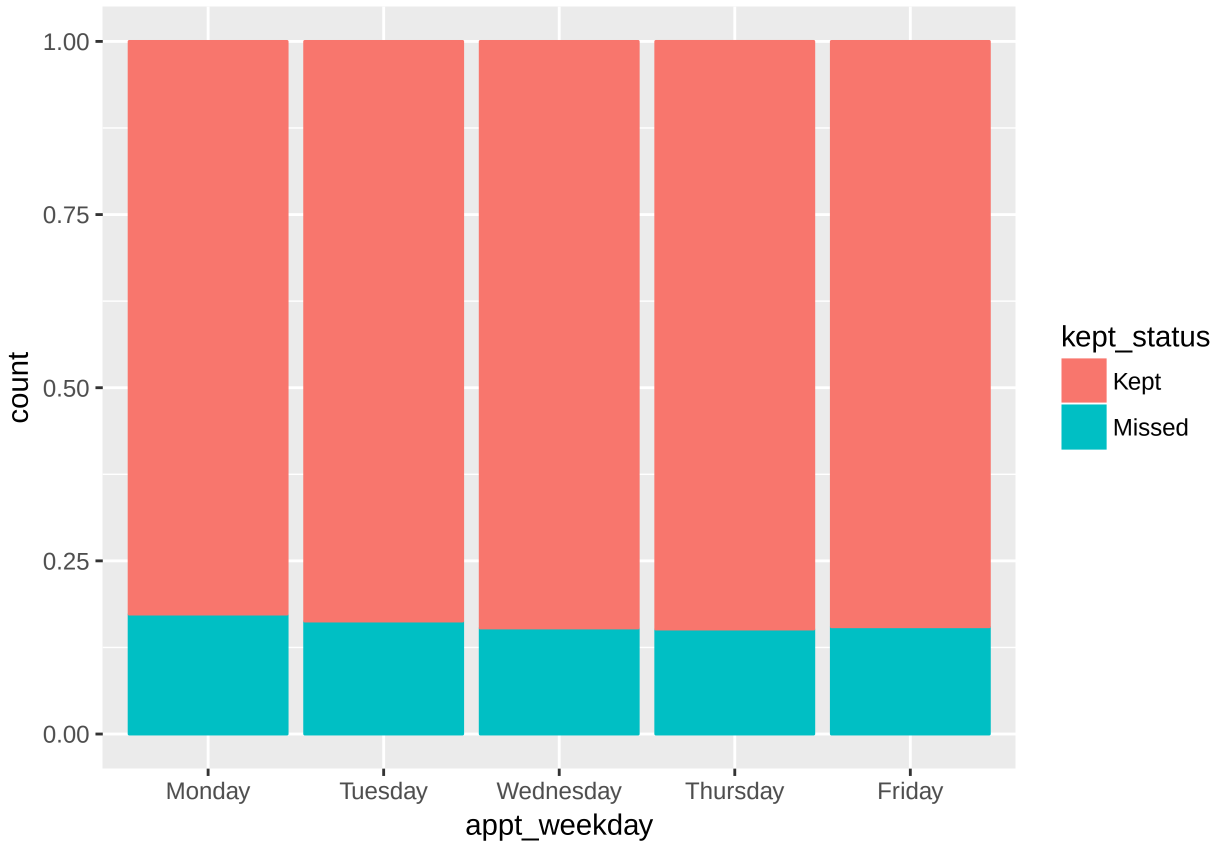



```
ggplot(
    appointments,
    aes(x = appt_weekday, color = kept_status, fill = kept_status)
) +
    geom_bar(position = "fill")
```



Most appointments occur on Tuesdays, and trail off later in the week with Friday having the fewest. Due to the difference between the number of appointments each day, it is difficult to see the ratio of missed appoiinments each day, so I will create a proportional plot.

```
ggplot(
    appointments,
    aes(x = appt_weekday, color = kept_status, fill = kept_status)
) +
    geom_bar(position = "fill")
```



There is a small variance in the ratio of missed appointments across the days of the week. The highest ratio of missed appointments occurs on Monday, and drops off slightly through Wednesday before leveling off.

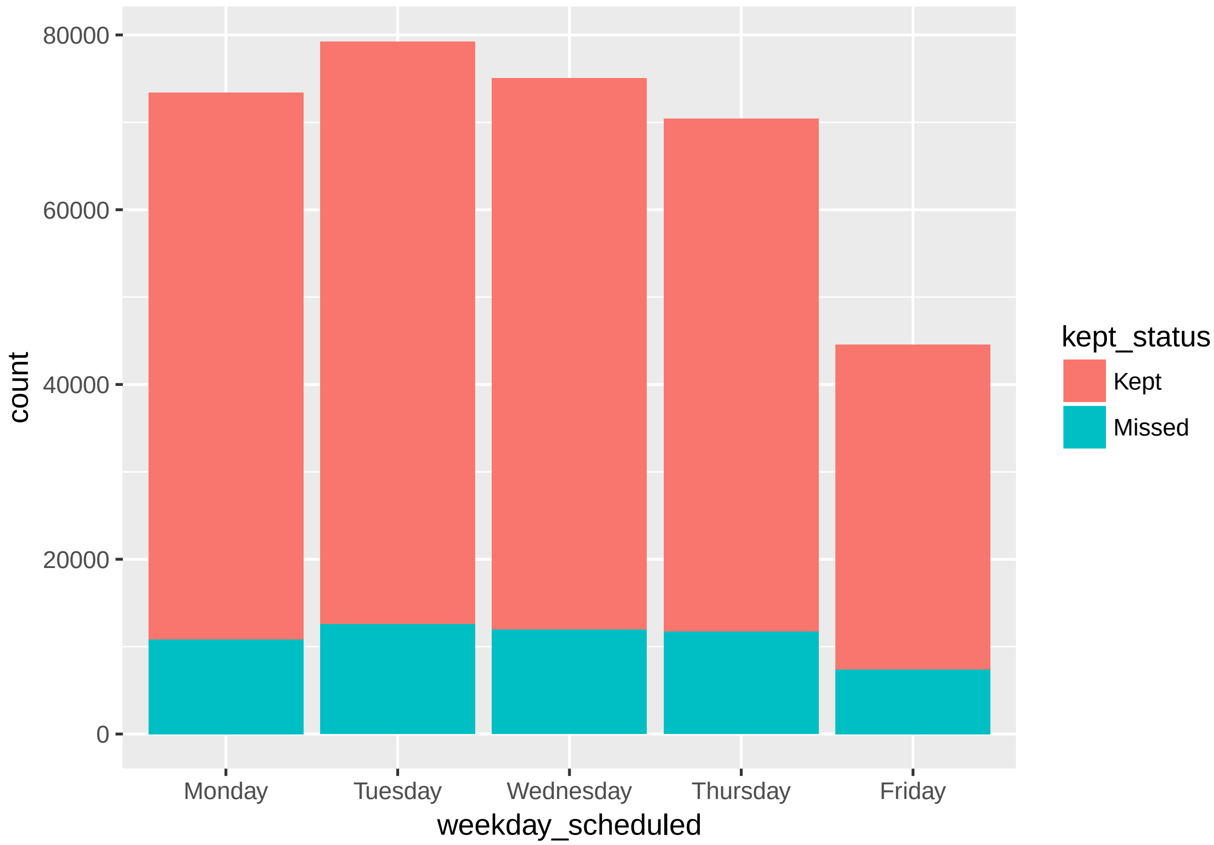
weekday_scheduled

```
table(appointments$weekday_scheduled)
```

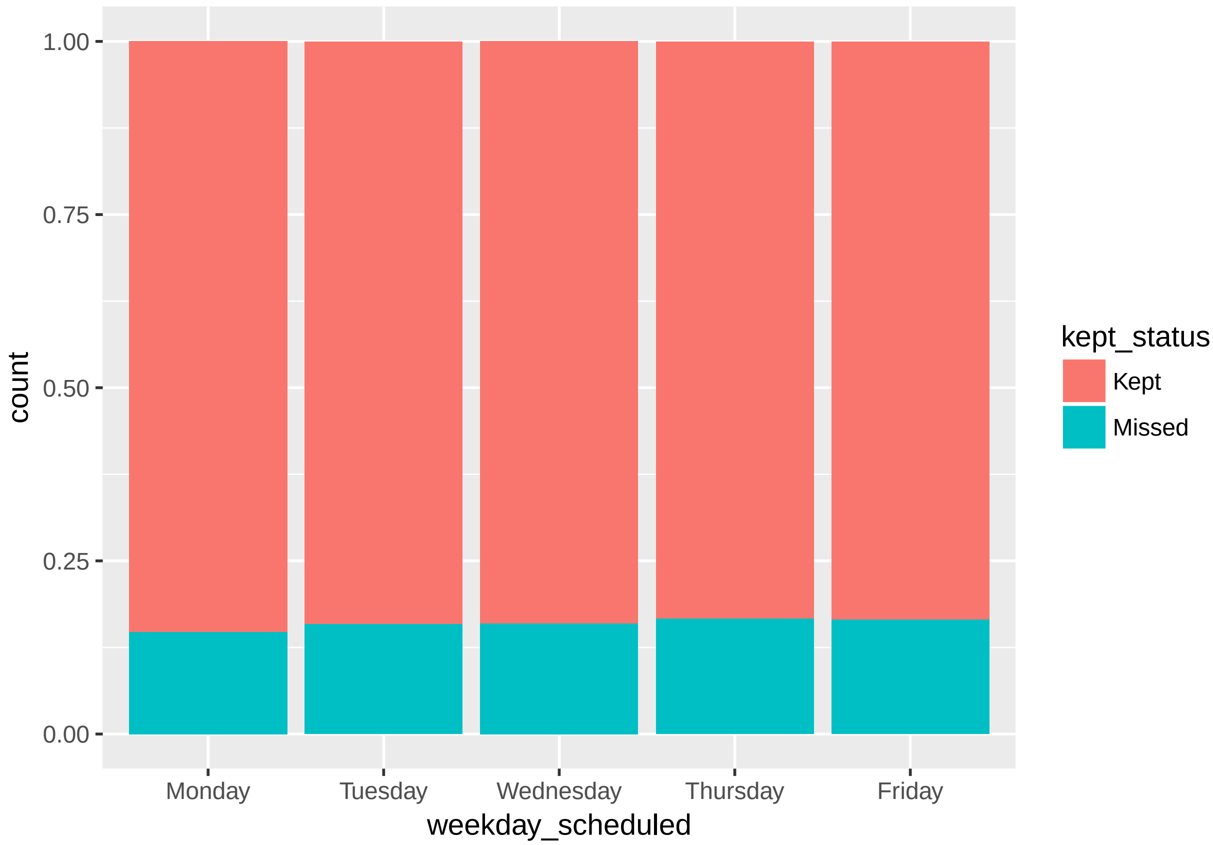
```
##
##     Friday     Monday   Saturday     Sunday   Thursday    Tuesday Wednesday
##      44553      73413         43          7      70441      79261     75052
```

A very small percentage of the observations occur on Saturday or Sunday, so I will remove them.

```
appointments <- appointments %>%
    filter(weekday_scheduled != "Sunday") %>%
    filter(weekday_scheduled != "Saturday")

appointments$weekday_scheduled <- factor(
    appointments$weekday_scheduled,
    levels = c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday")
)

ggplot(
    appointments,
    aes(x = weekday_scheduled, fill = kept_status)
) +
    geom_bar()
```




```
ggplot(
    appointments,
    aes(x = weekday_scheduled, fill = kept_status)
) +
    geom_bar(position = "fill")
```



The weekday the appointments are scheduled follows a similar pattern as the weekday the appointments occur, peaking on Tuesday and trailing off as the week progresses. However, the weekday the appointment is made has the opposite effect as the weekday the appointment occurst, with the ratio of missed appointments gradually rising throughout the week.

Modeling

Create Modeling Data

I will select the data to be used in modeling and assign to `model_data`, then convert the categorical information to dummy variables to help with the modelling.

```
model_data <- appointments %>%
    select(
        kept_status, appt_length, patient_age, patient_gender, billing_type,
        patient_distance, provider_specialty, remind_call_result, hour,
        percent_missed, is_new_patient, appt_lead_time, appt_weekday,
        weekday_scheduled, county_code)


factor_columns <- c(
    "kept_status", "patient_gender", "billing_type",
    "provider_specialty", "remind_call_result", "hour", "is_new_patient",
    "appt_weekday", "weekday_scheduled",
    "county_code")
model_data[factor_columns] <- map(model_data[factor_columns], factor)
```

```r
dummy_vars <- caret::dummyVars(~ ., data = model_data)

model_data_dummy <- data.frame(predict(dummy_vars, newdata = model_data))
```

The next step is to look for linear combinations, highly correlated variables, and variables with near zero variance. These variables add complexity to the model without providing any significant information, so I will remove them to help the model work better and more efficiently.

First I will look for linear combinations.

```r
linear_combos <- caret::findLinearCombos(model_data_dummy)

colnames(model_data_dummy[, linear_combos$remove])
```

```
## [1] "patient_gender.Unknown"        "billing_type.DMAP"
## [3] "provider_specialty.Other"      "remind_call_result.Not.Called"
## [5] "hour.21"                       "is_new_patient.1"
## [7] "appt_weekday.Friday"           "weekday_scheduled.Friday"
## [9] "county_code.X"
```

```r
model_data_dummy <- model_data_dummy[, -linear_combos$remove]
```

Next I will check for highly correlated variables.

```r
cor_matrix <- cor(model_data_dummy)
high_cor <- as.data.frame(which(abs(cor_matrix) > 0.90, arr.ind = TRUE))
cm_index <- high_cor %>% filter(row != col)

cor_matrix[cm_index[, 1], cm_index[, 2]]
```

```
##                         kept_status.Kept kept_status.Missed
## kept_status.Missed          -1.000000000        1.000000000
## kept_status.Kept             1.000000000       -1.000000000
## patient_gender.Male         -0.003350511        0.003350511
## patient_gender.Female        0.003832990       -0.003832990
## provider_specialty.B         0.013063134       -0.013063134
## provider_specialty.A        -0.030816952        0.030816952
##                         patient_gender.Female patient_gender.Male
## kept_status.Missed               -0.003832990         0.003350511
## kept_status.Kept                  0.003832990        -0.003350511
## patient_gender.Male              -0.997632962         1.000000000
## patient_gender.Female             1.000000000        -0.997632962
## provider_specialty.B             -0.004807503         0.004841524
## provider_specialty.A              0.008441594        -0.008524185
##                         provider_specialty.A provider_specialty.B
## kept_status.Missed               0.030816952         -0.013063134
## kept_status.Kept                -0.030816952          0.013063134
## patient_gender.Male             -0.008524185          0.004841524
## patient_gender.Female            0.008441594         -0.004807503
## provider_specialty.B            -0.915215490          1.000000000
## provider_specialty.A             1.000000000         -0.915215490
```

```r
cbind(
    cm_index[, 1],
    colnames(model_data_dummy[cm_index[,1]]),
    cm_index[,2],
    colnames(model_data_dummy[cm_index[,2]]))
```

```
##      [,1] [,2]                    [,3] [,4]
## [1,] "2"  "kept_status.Missed"    "1"  "kept_status.Kept"
## [2,] "1"  "kept_status.Kept"      "2"  "kept_status.Missed"
## [3,] "6"  "patient_gender.Male"   "5"  "patient_gender.Female"
## [4,] "5"  "patient_gender.Female" "6"  "patient_gender.Male"
## [5,] "11" "provider_specialty.B"  "10" "provider_specialty.A"
## [6,] "10" "provider_specialty.A"  "11" "provider_specialty.B"
```

Creating dummy variables for `kept_status` duplicated the information as there are only two possible values, therefore I will remove the dummy variable `kept_status.Missed`.

Most cases of `patient_gender` are `male` or `female`, with few values of `other` and `unknown`, leading to a high negative correlation between male and female. Similary, most cases of `provider_specialty` are `A` or `B`. Due to the dominance of two possible values, there is a high correlation between them and one will be eliminated. I will eliminate the dummy variables `patient_gender.Male` and `provider_specialty.B`.

```
model_data_dummy <- model_data_dummy[,-c(2, 6)]
```

Finally, I will check for variables with a variance near zero, and remove them.

```
near_zero_var <- caret::nearZeroVar(model_data_dummy)

colnames(model_data_dummy[, near_zero_var])
```

```
##  [1] "patient_gender.Other"
##  [2] "provider_specialty.C"
##  [3] "provider_specialty.D"
##  [4] "remind_call_result.Answered...Canceled"
##  [5] "remind_call_result.Answered...Reschedule"
##  [6] "remind_call_result.Busy"
##  [7] "remind_call_result.No.Answer"
##  [8] "hour.0"
##  [9] "hour.5"
## [10] "hour.6"
## [11] "hour.7"
## [12] "hour.12"
## [13] "hour.17"
## [14] "hour.18"
## [15] "hour.19"
## [16] "hour.20"
## [17] "county_code.A"
## [18] "county_code.B"
## [19] "county_code.C"
## [20] "county_code.D"
## [21] "county_code.E"
## [22] "county_code.G"
## [23] "county_code.H"
## [24] "county_code.K"
## [25] "county_code.L"
## [26] "county_code.M"
## [27] "county_code.N"
## [28] "county_code.O"
## [29] "county_code.Q"
## [30] "county_code.R"
## [31] "county_code.S"
## [32] "county_code.T"
```

```
## [33] "county_code.V"
## [34] "county_code.W"
```

```r
model_data_dummy <- model_data_dummy[, -near_zero_var]
```

Divide model_data into train, validate, and test sets

The data will be divided into three sets: train, validate, and test. I will use 60% of the tata for training, and 20% each for validation and test.

Because the data is sorted by date, I want to use the most recent data for the test data, the next oldest for validation, and the oldest for training. Since I am trying to predict future appointments, testing on the most recent data will result in the best measure of the model's performance on future appointments.

```r
model_data_dummy$kept_status.Kept <- as.factor(model_data_dummy$kept_status.Kept)
model_data_dummy$kept_status.Kept <- fct_recode(
    model_data_dummy$kept_status.Kept, "Kept" = "1", "Missed" = "0")

train <- model_data_dummy[1:205660,]
validate <- model_data_dummy[205661:274200,]
test <- model_data_dummy[274201:nrow(model_data_dummy),]
table(train$kept_status.Kept)
```

```
##
## Missed   Kept
##  31050 174610
```

```r
train_balance_subset <- train[168750:205660,]
table(train_balance_subset$kept_status.Kept)
```

```
##
## Missed   Kept
##   5861  31050
```

```r
train_kept <- train_balance_subset[train_balance_subset$kept_status.Kept == "Kept",]
train_missed <- train[train$kept_status.Kept == "Missed",]
train_balanced <- rbind(train_kept, train_missed)
table(train_balanced$kept_status.Kept)
```

```
##
## Missed   Kept
##  31050  31050
```

Set Up Model Parameters

```r
rf_control <- caret::trainControl(method = "cv", number = 2, classProbs = TRUE)
seed <- 7
metric <- "Accuracy"
set.seed(seed)
mtry <- 3
tunegrid <- expand.grid(.mtry = mtry)
```

Logistic Regression Model

```r
glm_control <- caret::trainControl(method = "none")

glm_model <- caret::train(
    kept_status.Kept ~ .,
    data = train_balanced,
    method = "glm",
    trControl = glm_control
)

summary(glm_model)
```

```
##
## Call:
## NULL
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1351  -0.9204   0.0226   0.9112   4.4960
##
## Coefficients:
##                                          Estimate Std. Error z value
## (Intercept)                             4.9675554  0.1759886  28.227
## appt_length                            -0.0021217  0.0005820  -3.646
## patient_age                             0.0110427  0.0005104  21.634
## patient_gender.Female                   0.0748451  0.0187680   3.988
## billing_type.Commercial                 0.0355343  0.0238987   1.487
## patient_distance                        0.0001074  0.0001374   0.782
## provider_specialty.A                   -1.8598458  0.0778798 -23.881
## provider_specialty.B                   -1.5509413  0.0794919 -19.511
## remind_call_result.Answered...Confirmed    0.0851678  0.0408951   2.083
## remind_call_result.Answered...No.Response -0.9946251  0.0298761 -33.292
## remind_call_result.Failed              -1.5622067  0.0395231 -39.526
## remind_call_result.Left.Message         0.0358584  0.0576422   0.622
## hour.8                                  0.7843766  0.0881687   8.896
## hour.9                                  0.8419953  0.0882685   9.539
## hour.10                                 0.9489091  0.0881305  10.767
## hour.11                                 0.7969438  0.0887809   8.977
## hour.13                                 0.9100060  0.0880188  10.339
## hour.14                                 0.9199228  0.0883180  10.416
## hour.15                                 0.9779254  0.0882228  11.085
## hour.16                                 0.9042273  0.0888692  10.175
## percent_missed                         -0.0471094  0.0006684 -70.486
## is_new_patient.0                       -2.1175291  0.1271382 -16.655
## appt_lead_time                         -0.0037272  0.0001148 -32.466
## appt_weekday.Monday                    -0.0495366  0.0350647  -1.413
## appt_weekday.Tuesday                    0.1734662  0.0334379   5.188
## appt_weekday.Wednesday                  0.2811529  0.0339214   8.288
## appt_weekday.Thursday                   0.1283543  0.0345337   3.717
## weekday_scheduled.Monday                0.0348621  0.0329362   1.058
## weekday_scheduled.Tuesday              -0.0937567  0.0322156  -2.910
## weekday_scheduled.Wednesday            -0.0691246  0.0326217  -2.119
## weekday_scheduled.Thursday             -0.1165339  0.0327970  -3.553
```

```
## county_code.F                             0.3775954  0.0364214  10.367
## county_code.I                            -0.2772360  0.0302845  -9.154
## county_code.J                             0.0077711  0.0348045   0.223
## county_code.P                            -0.0884900  0.0259861  -3.405
## county_code.U                            -0.3226888  0.0364713  -8.848
##                                          Pr(>|z|)
## (Intercept)                               < 2e-16 ***
## appt_length                              0.000267 ***
## patient_age                               < 2e-16 ***
## patient_gender.Female                    6.67e-05 ***
## billing_type.Commercial                  0.137048
## patient_distance                         0.434435
## provider_specialty.A                      < 2e-16 ***
## provider_specialty.B                      < 2e-16 ***
## remind_call_result.Answered...Confirmed   0.037289 *
## remind_call_result.Answered...No.Response  < 2e-16 ***
## remind_call_result.Failed                 < 2e-16 ***
## remind_call_result.Left.Message          0.533885
## hour.8                                    < 2e-16 ***
## hour.9                                    < 2e-16 ***
## hour.10                                   < 2e-16 ***
## hour.11                                   < 2e-16 ***
## hour.13                                   < 2e-16 ***
## hour.14                                   < 2e-16 ***
## hour.15                                   < 2e-16 ***
## hour.16                                   < 2e-16 ***
## percent_missed                            < 2e-16 ***
## is_new_patient.0                          < 2e-16 ***
## appt_lead_time                            < 2e-16 ***
## appt_weekday.Monday                      0.157738
## appt_weekday.Tuesday                     2.13e-07 ***
## appt_weekday.Wednesday                    < 2e-16 ***
## appt_weekday.Thursday                    0.000202 ***
## weekday_scheduled.Monday                 0.289840
## weekday_scheduled.Tuesday                0.003611 **
## weekday_scheduled.Wednesday              0.034092 *
## weekday_scheduled.Thursday               0.000381 ***
## county_code.F                             < 2e-16 ***
## county_code.I                             < 2e-16 ***
## county_code.J                            0.823318
## county_code.P                            0.000661 ***
## county_code.U                             < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 86089  on 62099  degrees of freedom
## Residual deviance: 68325  on 62064  degrees of freedom
## AIC: 68397
##
## Number of Fisher Scoring iterations: 6
```

Random Forest Model

```r
rf_model <- caret::train(
    kept_status.Kept ~ ., data = train_balanced, method = "rf", metric = metric,
    tuneGrid = tunegrid, trControl = rf_control)
```

Model Comparison

```r
pred_glm <- predict(glm_model, validate)

conf_mat_glm <- caret::confusionMatrix(
    pred_glm, validate$kept_status, positive = "Missed")

conf_mat_glm
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Missed  Kept
##     Missed   8158 15448
##     Kept     3298 41636
##
##                Accuracy : 0.7265
##                  95% CI : (0.7231, 0.7298)
##     No Information Rate : 0.8329
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.3101
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.7121
##             Specificity : 0.7294
##          Pos Pred Value : 0.3456
##          Neg Pred Value : 0.9266
##              Prevalence : 0.1671
##          Detection Rate : 0.1190
##    Detection Prevalence : 0.3444
##       Balanced Accuracy : 0.7207
##
##        'Positive' Class : Missed
##
```

```r
conf_mat_glm$byClass["F1"]
```

```
##        F1
## 0.4653471
```

```r
pred_rf <- predict(rf_model, validate)

caret::confusionMatrix(rf_model)
```

```
## Cross-Validated (2 fold) Confusion Matrix
##
## (entries are percentual average cell counts across resamples)
```

```
##
##            Reference
## Prediction Missed Kept
##     Missed   41.5 15.1
##     Kept      8.5 34.9
##
##  Accuracy (average) : 0.764
```

```r
conf_mat_rf <- caret::confusionMatrix(
    pred_rf, validate$kept_status, positive = "Missed")

conf_mat_rf
```

```
## Confusion Matrix and Statistics
##
##            Reference
## Prediction Missed  Kept
##     Missed   9062 16999
##     Kept     2394 40085
##
##                Accuracy : 0.7171
##                  95% CI : (0.7137, 0.7204)
##     No Information Rate : 0.8329
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.3268
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.7910
##             Specificity : 0.7022
##          Pos Pred Value : 0.3477
##          Neg Pred Value : 0.9436
##              Prevalence : 0.1671
##          Detection Rate : 0.1322
##    Detection Prevalence : 0.3802
##       Balanced Accuracy : 0.7466
##
##        'Positive' Class : Missed
##
```

```r
conf_mat_rf$byClass["F1"]
```

```
##        F1
## 0.4830877
```