Capstone Project - Predicting Patient No-Shows Using Appointment Data

```
library(tidyverse)
library(lubridate)
library(caret)
library(randomForest)
library(GGally)
```

Read data and assign to *appointments*

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

```
## Parsed with column specification:
## cols(
##   kept_status = col_character(),
##   appt_date = col_character(),
##   appt_time = col_time(format = ""),
##   appt_length = col_integer(),
##   date_scheduled = col_character(),
##   patient_age = col_integer(),
##   patient_gender = col_character(),
##   billing_type = col_character(),
##   prior_missed = col_integer(),
##   prior_kept = col_integer(),
##   patient_distance = col_integer(),
##   office_zip = col_character(),
##   provider_specialty = col_character(),
##   remind_call_result = col_character()
## )
```

```
zipcodes <- read_csv("zipcodes.csv")
```

```
## Parsed with column specification:
## cols(
##   office_zip = col_character(),
##   county_code = col_character(),
##   city_size = col_integer()
## )
```

Data Summary and Structure

```
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
##
##
##
##
```

```
str(appointments)
```

```
## 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 2880
##   .. ..- attr(*, "units")= chr "secs"
##  $ 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
##  - attr(*, "spec")=List of 2
##   ..$ cols   :List of 14
##   .. ..$ kept_status       : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ appt_date         : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ appt_time         :List of 1
##   .. .. ..$ format: chr ""
##   .. .. ..- attr(*, "class")= chr  "collector_time" "collector"
##   .. ..$ appt_length       : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ date_scheduled    : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ patient_age       : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ patient_gender    : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ billing_type      : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ prior_missed      : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ prior_kept        : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ patient_distance  : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ office_zip        : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ provider_specialty: list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ remind_call_result: list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   ..$ default: list()
##   .. ..- attr(*, "class")= chr  "collector_guess" "collector"
##   ..- attr(*, "class")= chr "col_spec"
```

```
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
```

```
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
```

```
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
```

```
# Check for NAs
purrr::map(appointments, ~sum(is.na(.)))
```

```
## $kept_status
## [1] 0
##
## $appt_date
## [1] 0
##
## $appt_time
## [1] 0
##
## $appt_length
## [1] 0
##
## $date_scheduled
## [1] 0
##
## $patient_age
## [1] 0
##
## $patient_gender
## [1] 0
##
## $billing_type
## [1] 0
##
## $prior_missed
## [1] 0
##
## $prior_kept
## [1] 0
##
## $patient_distance
## [1] 974
##
## $office_zip
## [1] 0
##
## $provider_specialty
## [1] 0
##
## $remind_call_result
## [1] 0
```

patient_distance variable has 972 NA values.

Data Dictionary

```
variable_description <- 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)

as_data_frame(cbind(c(1:length(variable)),variable, variable_description))
```

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

Will combine the appointment time and date into one variable, appt_datetime.

```
appointments_2 <- appointments %>%
    mutate(appt_datetime = lubridate::mdy_hms(paste(appt_date, appt_time)))

appointments_2$date_scheduled <- as.POSIXct(appointments_2$date_scheduled,
                                     format = "%m/%d/%y")
```

Calculating percent of missed appointments overall. Will first create a logical variable *missed*, where 1 represents a missed appointment and 0 represents a kept appointment.

```
appointments_2 <- appointments_2 %>%
    mutate(missed = ifelse(appointments_2$kept_status == "Missed", 1, 0))
missed_rate <- mean(appointments_2$missed)
missed_rate
```

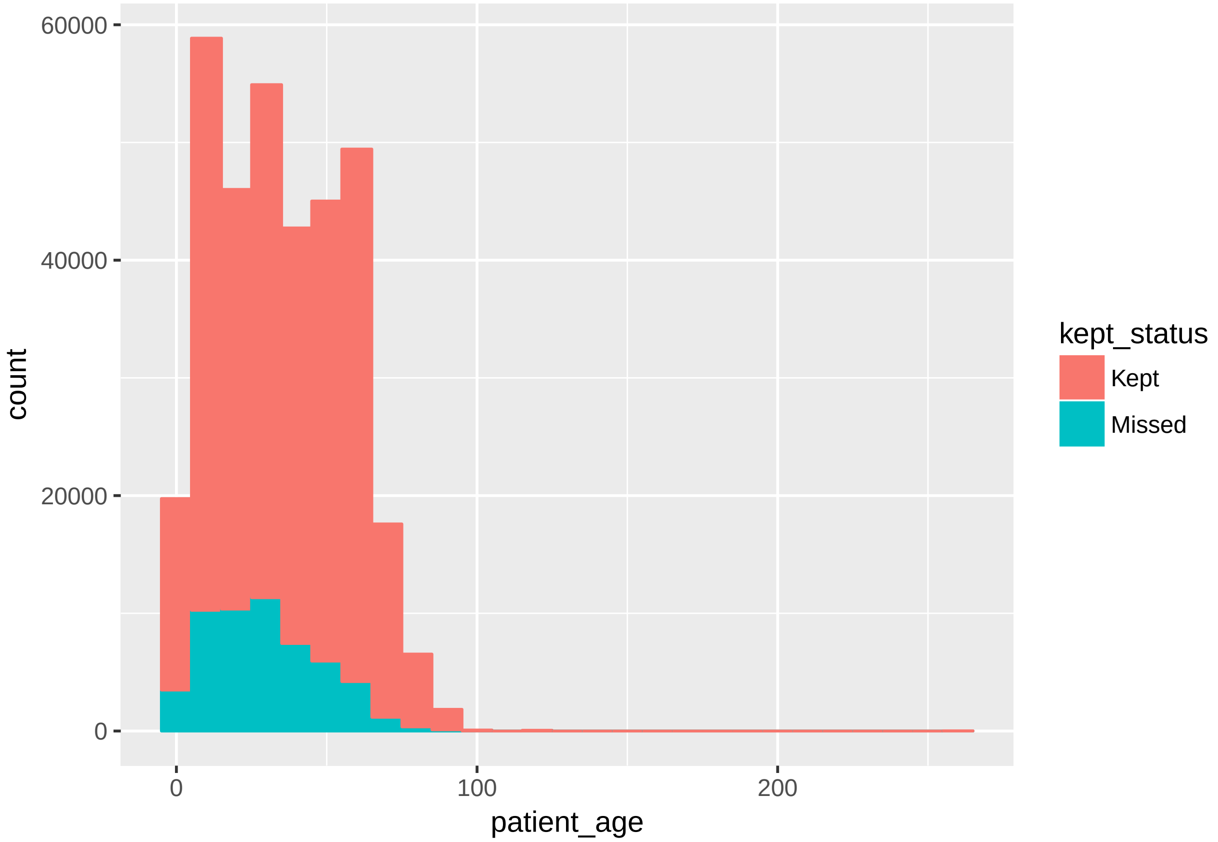
```
## [1] 0.1592944
```

About 16% of the total appointments are missed.

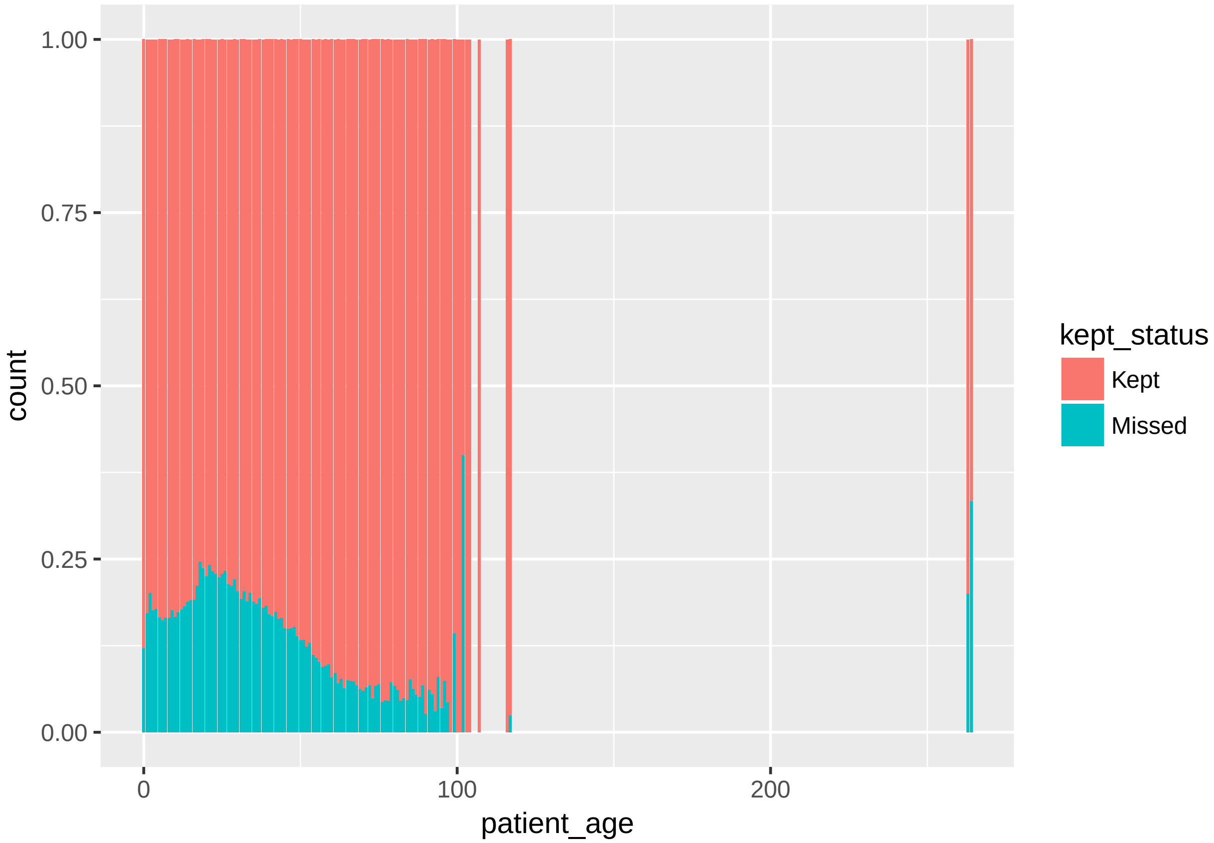
Data Exploration

patient_age

```
ggplot(
    data = appointments_2,
    mapping = aes(x = patient_age, col = kept_status, fill = kept_status)
) +
    geom_histogram(binwidth = 10)
```



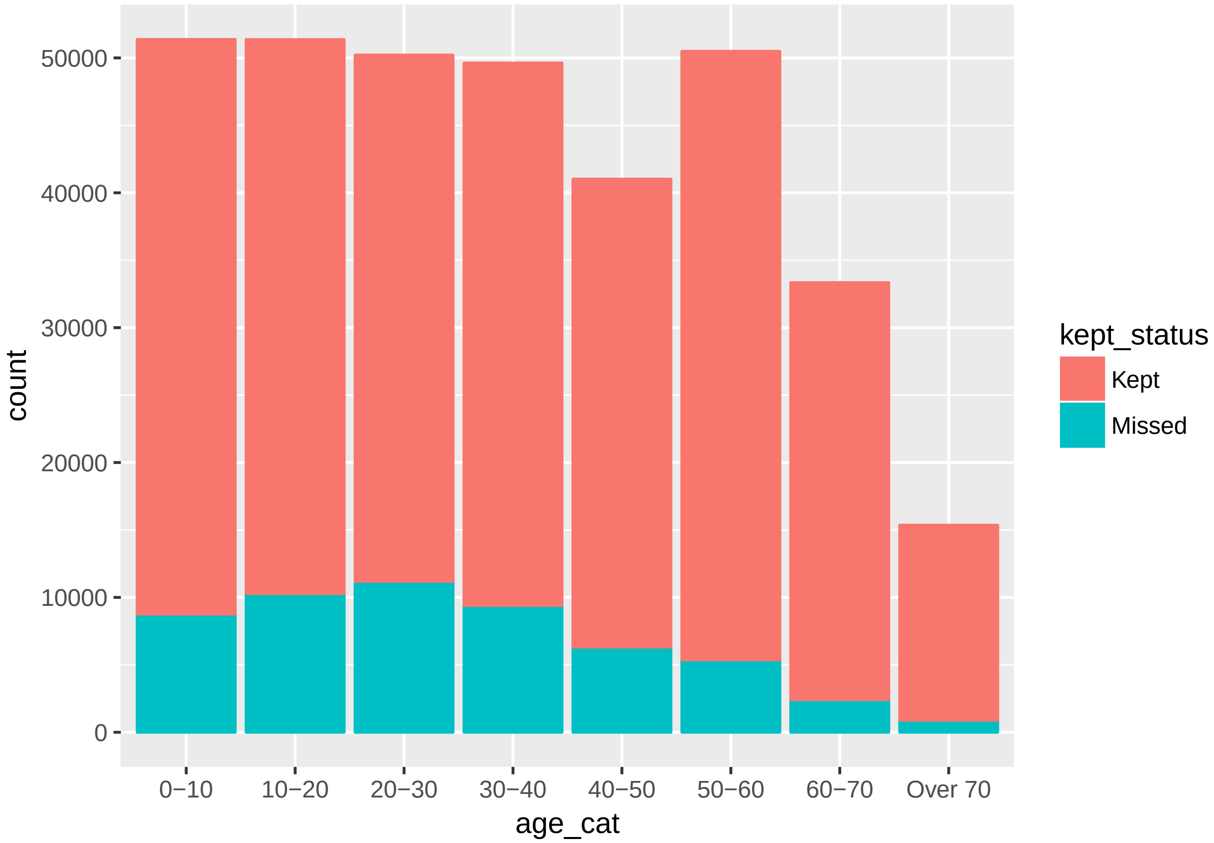
```
ggplot(
    data = appointments_2,
    mapping = aes(x = patient_age, fill = kept_status)
) +
    geom_bar(position = "fill")
```



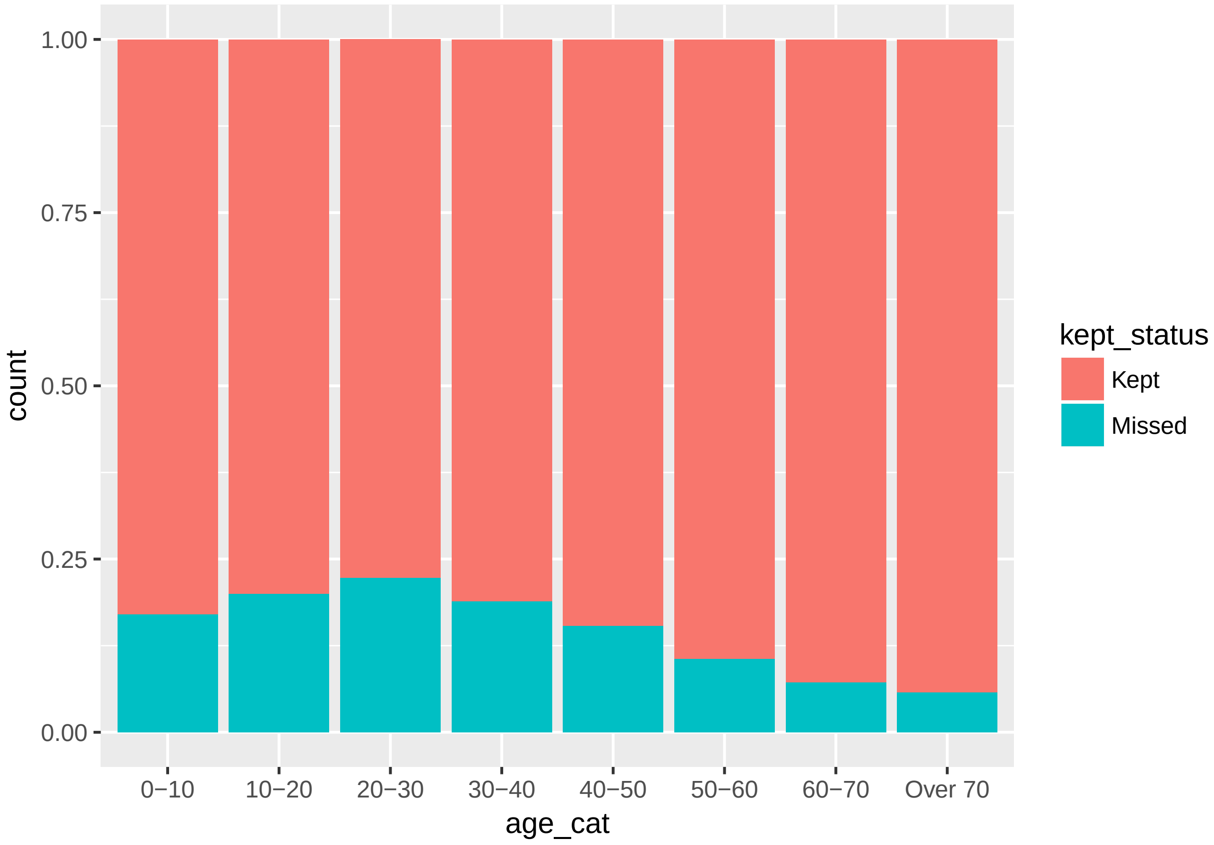
Ranges from 0-264, so there are obviously a few impossible values. Ratio of missed appointments decreases with age in general.

Removing obervations of ages greater than 100, creating categorical age groups and replotting.

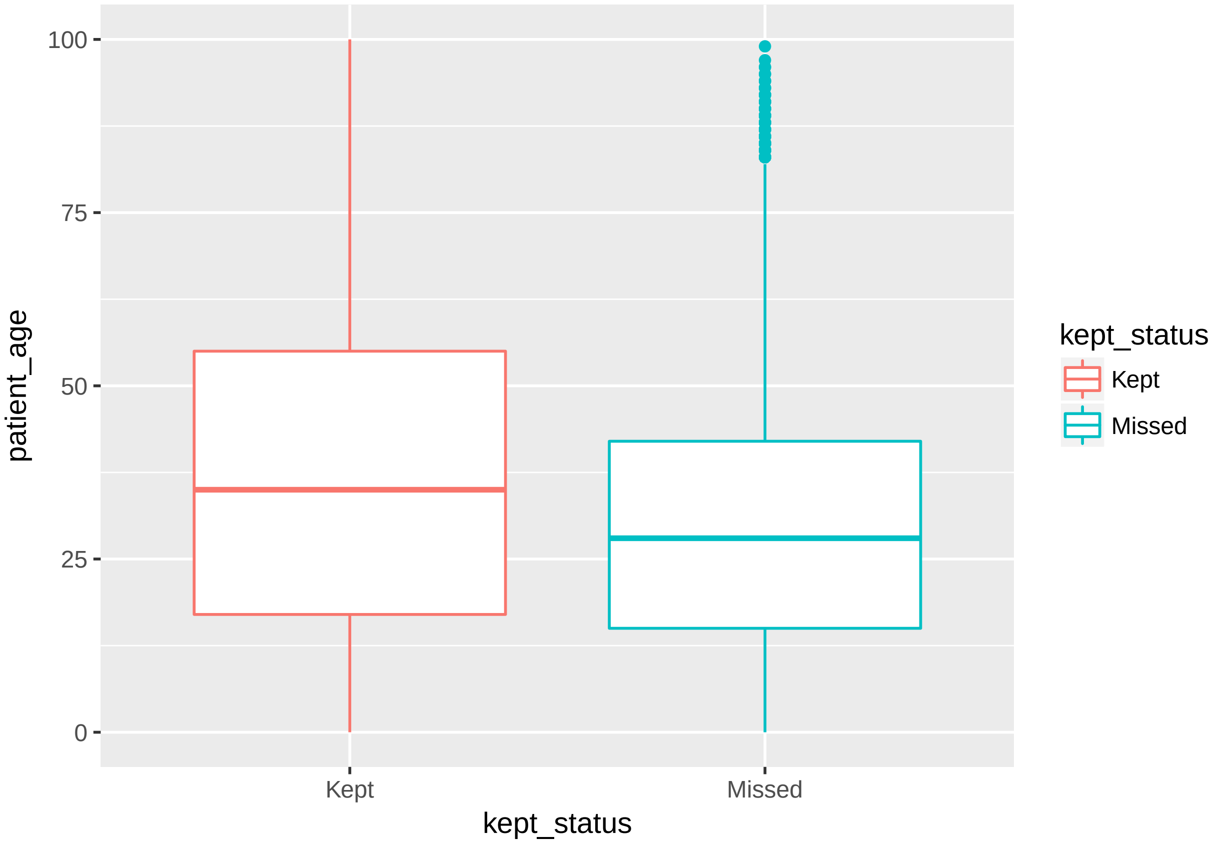
```
age_labels <- c("0-10", "10-20","20-30", "30-40", "40-50", "50-60", "60-70",
                "Over 70")
age_breaks <- c(-1, 10, 20, 30, 40, 50, 60, 70, 101)

appointments_2 <- appointments_2 %>%
    filter(patient_age <= 100) %>%
    mutate(
        age_cat = cut(patient_age, breaks = age_breaks, labels = age_labels))


ggplot(
    data = appointments_2,
    mapping = aes(x = age_cat, col = kept_status, fill = kept_status)
) +
    stat_count()
```



```
ggplot(
    data = appointments_2,
    mapping = aes(x = age_cat, fill = kept_status)) +
    geom_bar(position = "fill")
```



```
ggplot(
    data = appointments_2,
    aes(x = kept_status, y = patient_age, col = kept_status)
) +
    geom_boxplot()
```

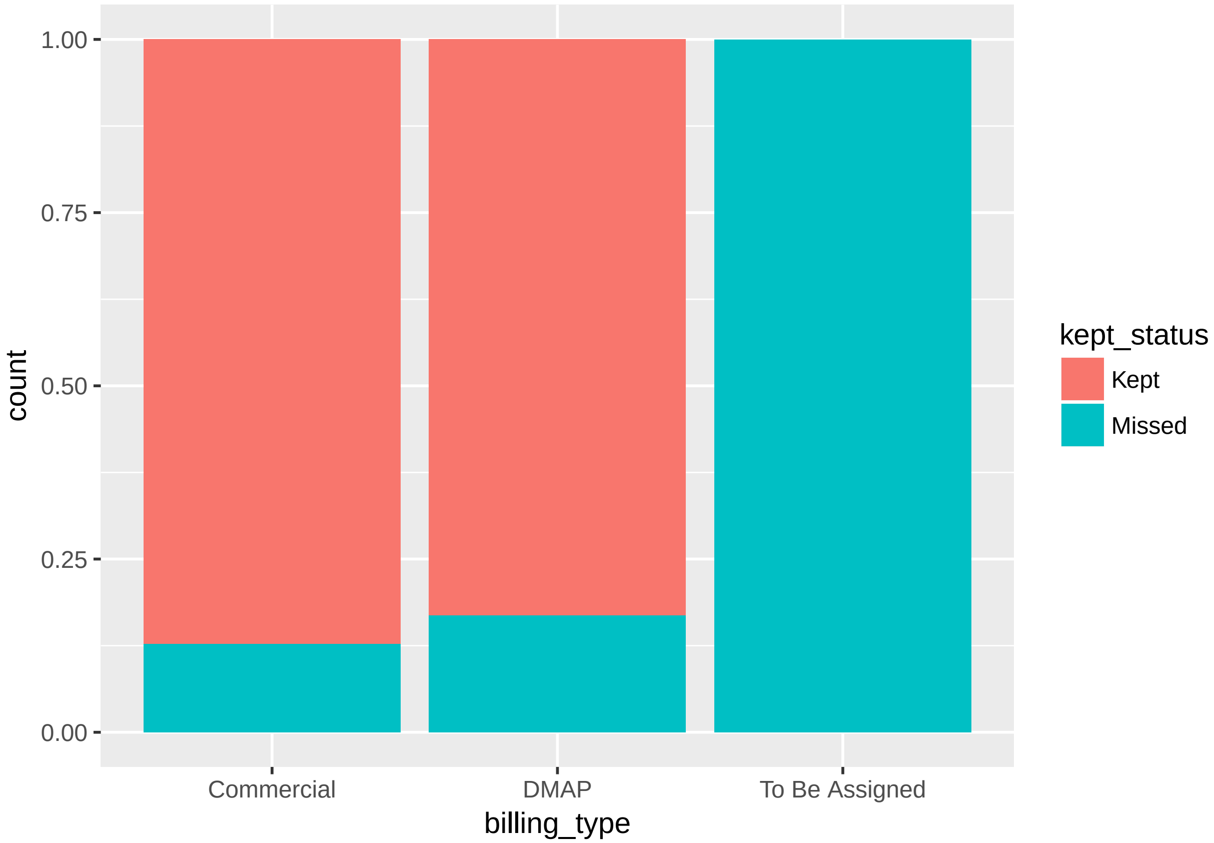


billing_type

```
table(appointments_2$billing_type)
```

```
##
##     Commercial           DMAP To Be Assigned
##          78278         264486              1
```

```
ggplot(
    data = appointments_2,
    mapping = aes(x = billing_type, fill = kept_status)
) +
    geom_bar(position = "fill")
```



Only one row has *To Be Assigned* value and will just be removed There is a minor difference between billing types. DMAP has a higher proportion of missed appointments

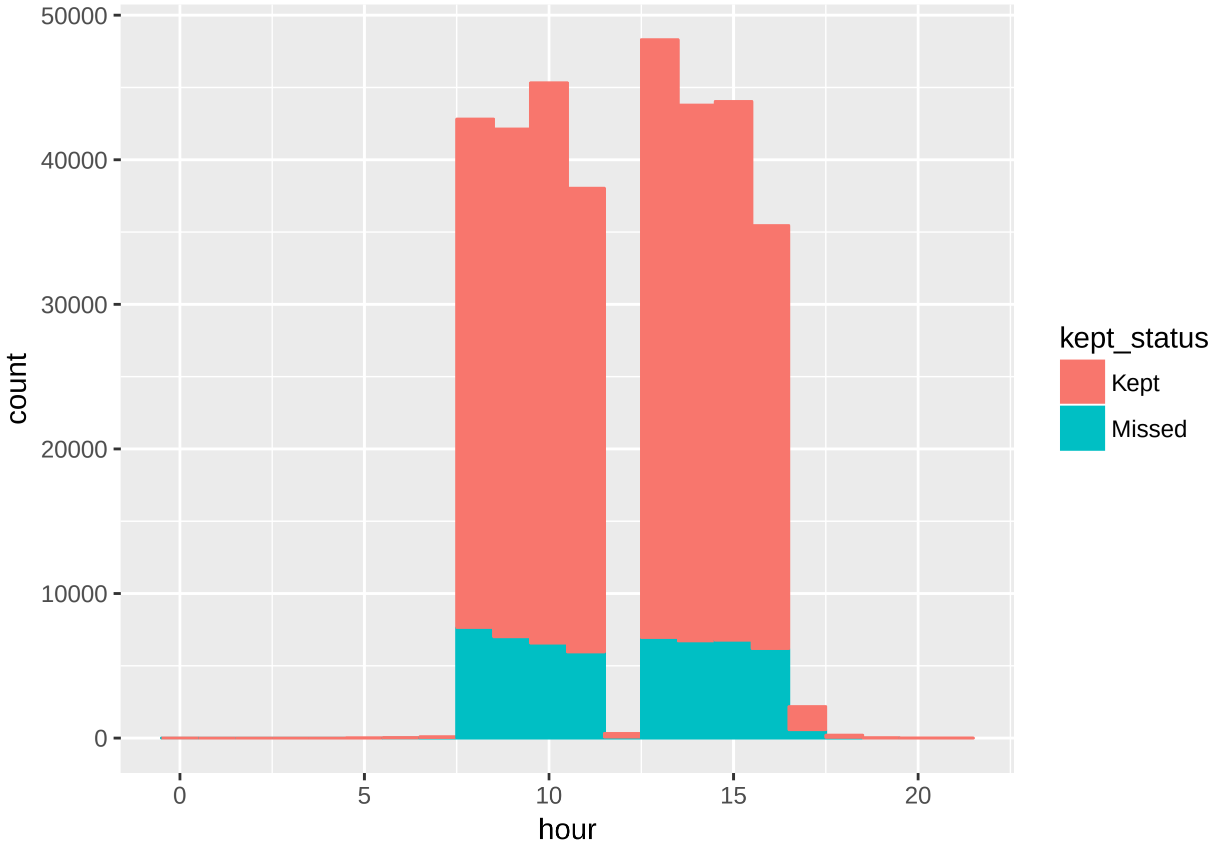
```
appointments_2 <- subset(appointments_2,
                         appointments_2$billing_type != "To Be Assigned")
```

Check coding style above

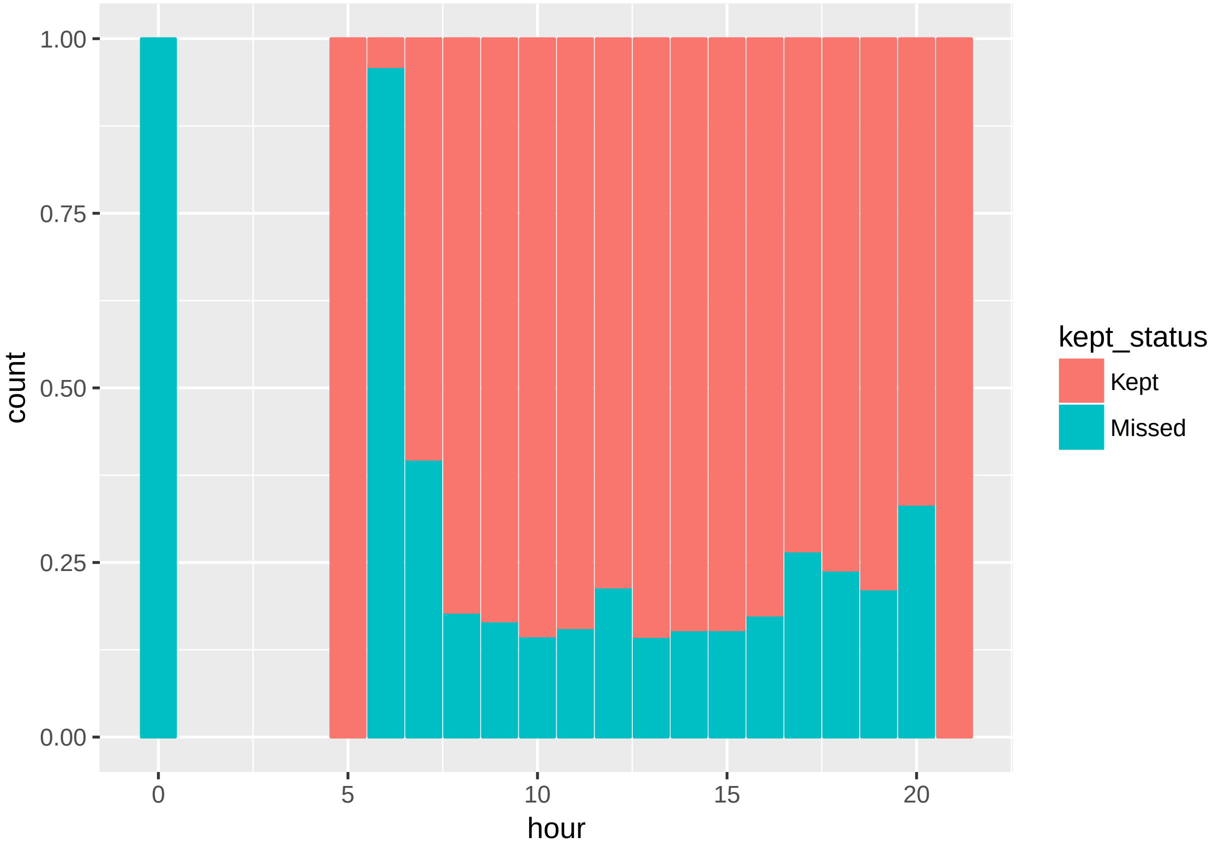
appt_datetime

Creating new *hour* variable and plot by hour

```
appointments_2 <- appointments_2 %>%
    mutate(hour = lubridate::hour(appointments_2$appt_datetime))

ggplot(data = appointments_2,
    mapping = aes(x = hour, col = kept_status, fill = kept_status)
) +
    geom_histogram(binwidth = 1)
```



```
ggplot(data = appointments_2,
    mapping = aes(x = hour, col = kept_status, fill = kept_status)
) +
    geom_bar(position = "fill")
```



Ranges from 00:00:00 to 21:00:00. More appointments are missed in the early morning, late afternoon and early evening, and around lunchtime, however, there are very few appointments at these times. During main scheduling periods, the variation is less significant.

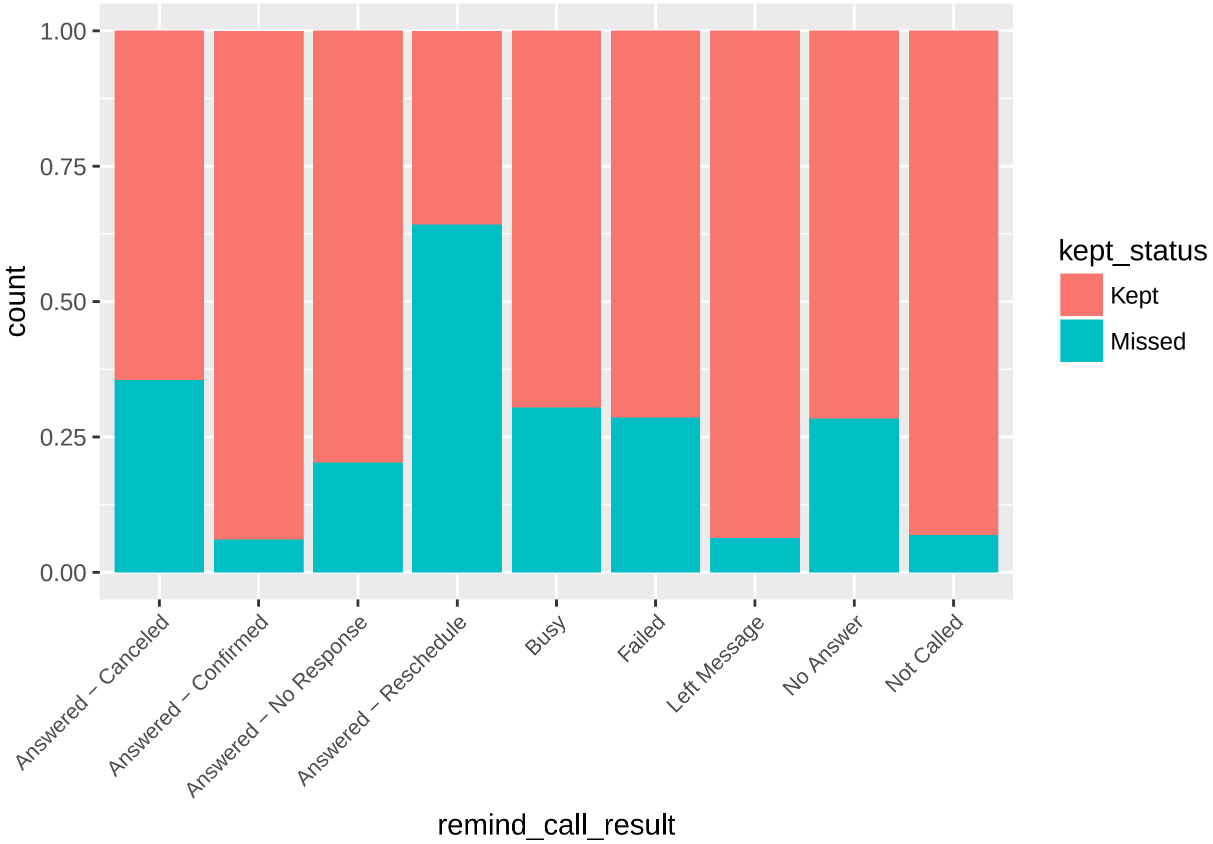
remind_call_result

```
table(appointments_2$remind_call_result)
```

```
## 
##    Answered - Canceled   Answered - Confirmed Answered - No Response 
##                    152                  49108                 180860 
##  Answered - Reschedule                   Busy                 Failed 
##                   1369                   1104                  27943 
##           Left Message              No Answer             Not Called 
##                  18429                    377                  63422
```

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

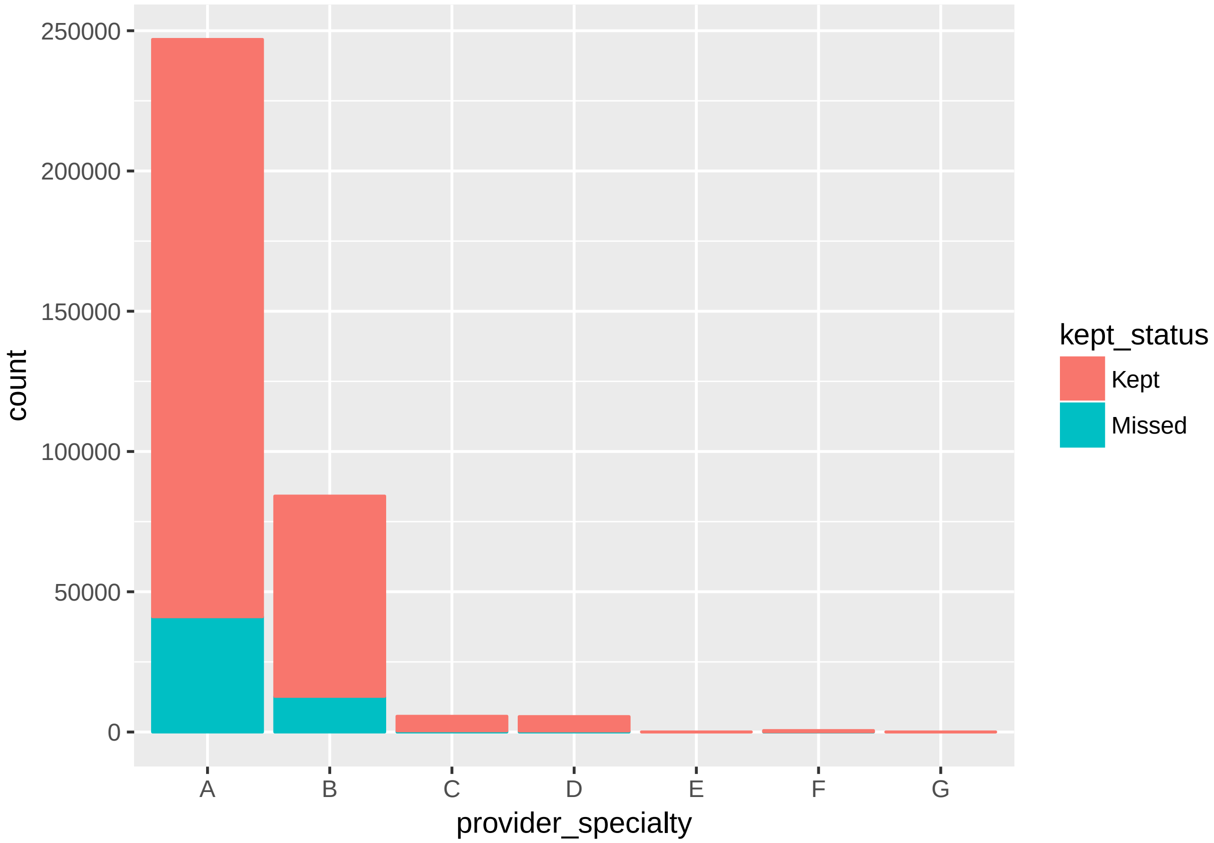
```
ggplot(data = appointments_2,
    mapping = aes(x = remind_call_result, fill = kept_status)
) +
    geom_bar(position = "fill") +
    theme(axis.text.x = element_text(size = 8, angle = 45,hjust = 1, vjust = 1))
```



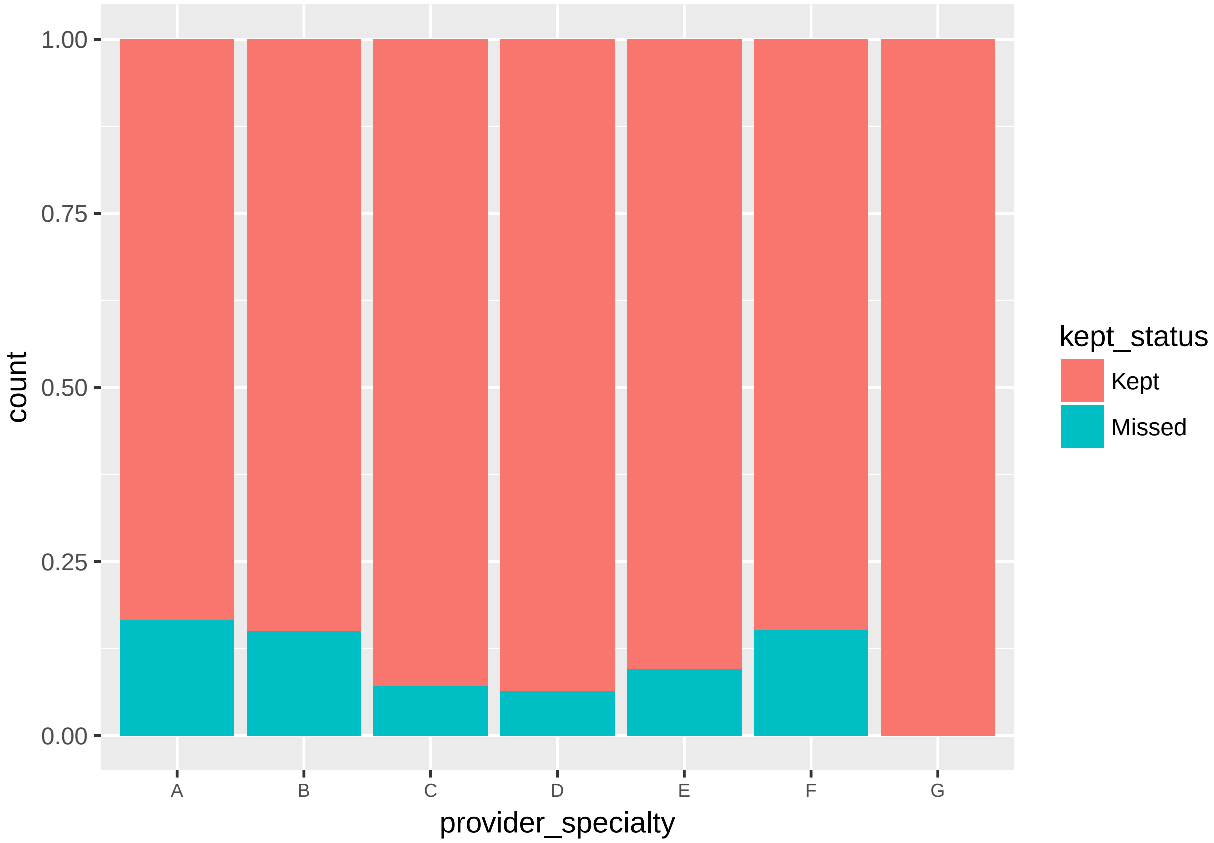
~65% of appointments with “Answered - Cancelled” and ~35% with “Answered-Reschedule” still kept their appointments, however, very few observations in these categories.

provider_specialty

```
ggplot(
    data = appointments_2,
    mapping = aes(x = provider_specialty, col = kept_status, fill = kept_status)
) +
    stat_count()
```



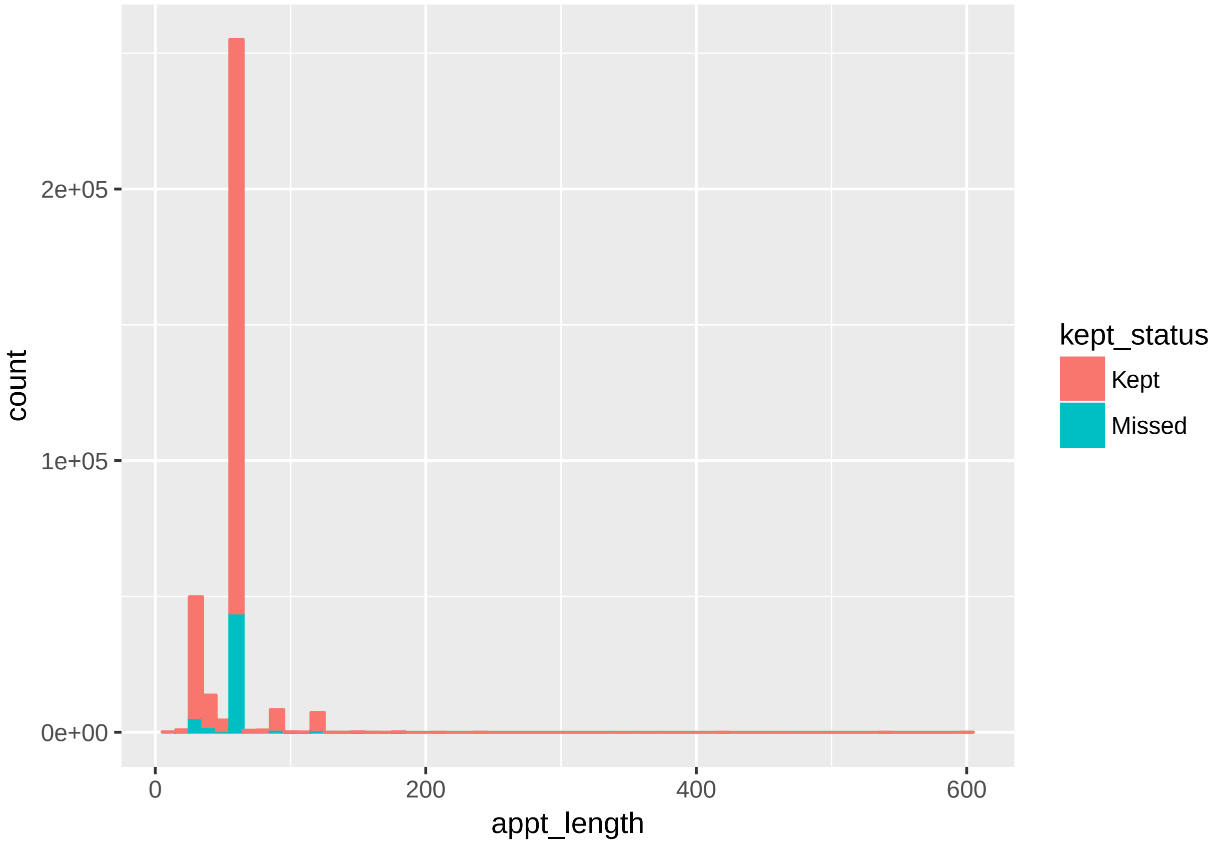
```r
ggplot(data = appointments_2) +
    geom_bar(
        mapping = aes(x = provider_specialty, fill = kept_status),
        position = "fill"
) +
    theme(axis.text.x = element_text(size = 7))
```



C, D, and E provider specialties have lower proportion of missed appointments,

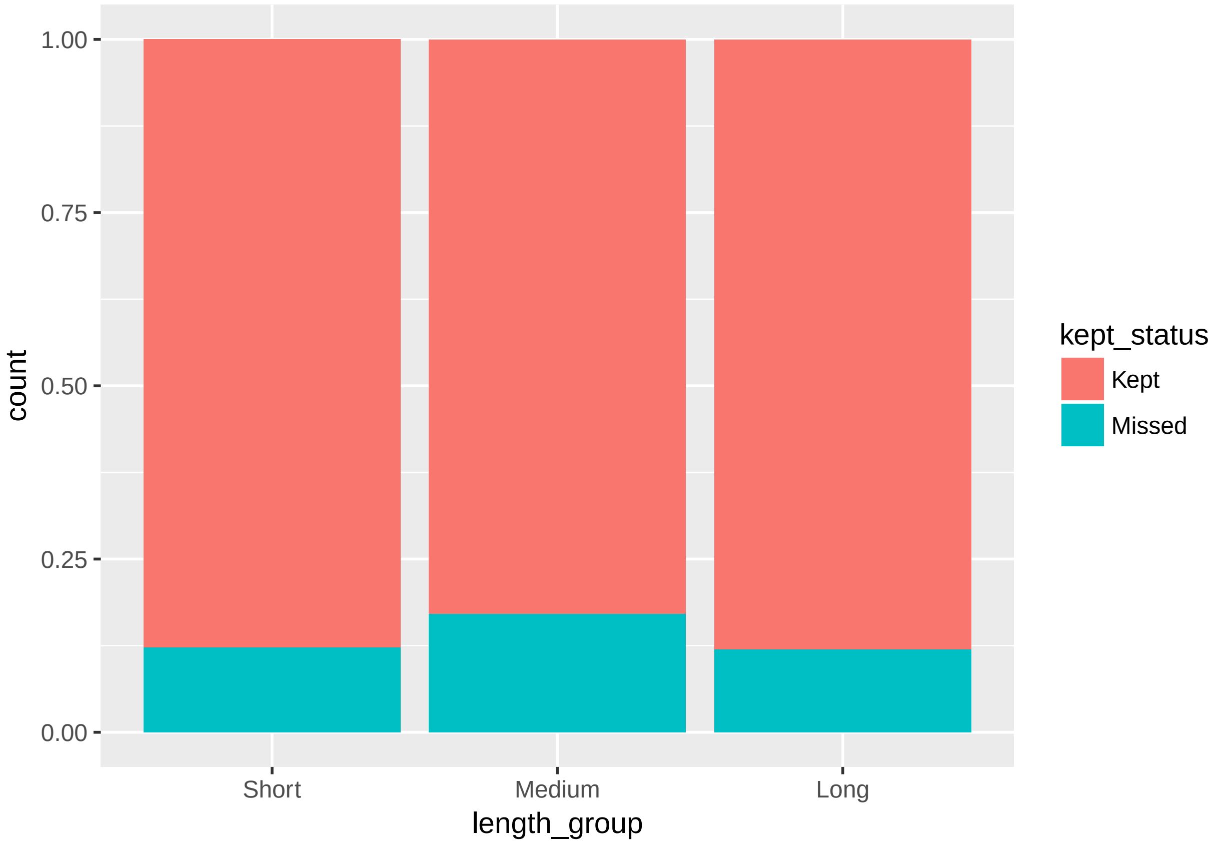
appt_length

```
ggplot(
    data = appointments_2,
    mapping = aes(x = appt_length, col = kept_status, fill = kept_status)
) +
    geom_histogram(binwidth = 10)
```



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

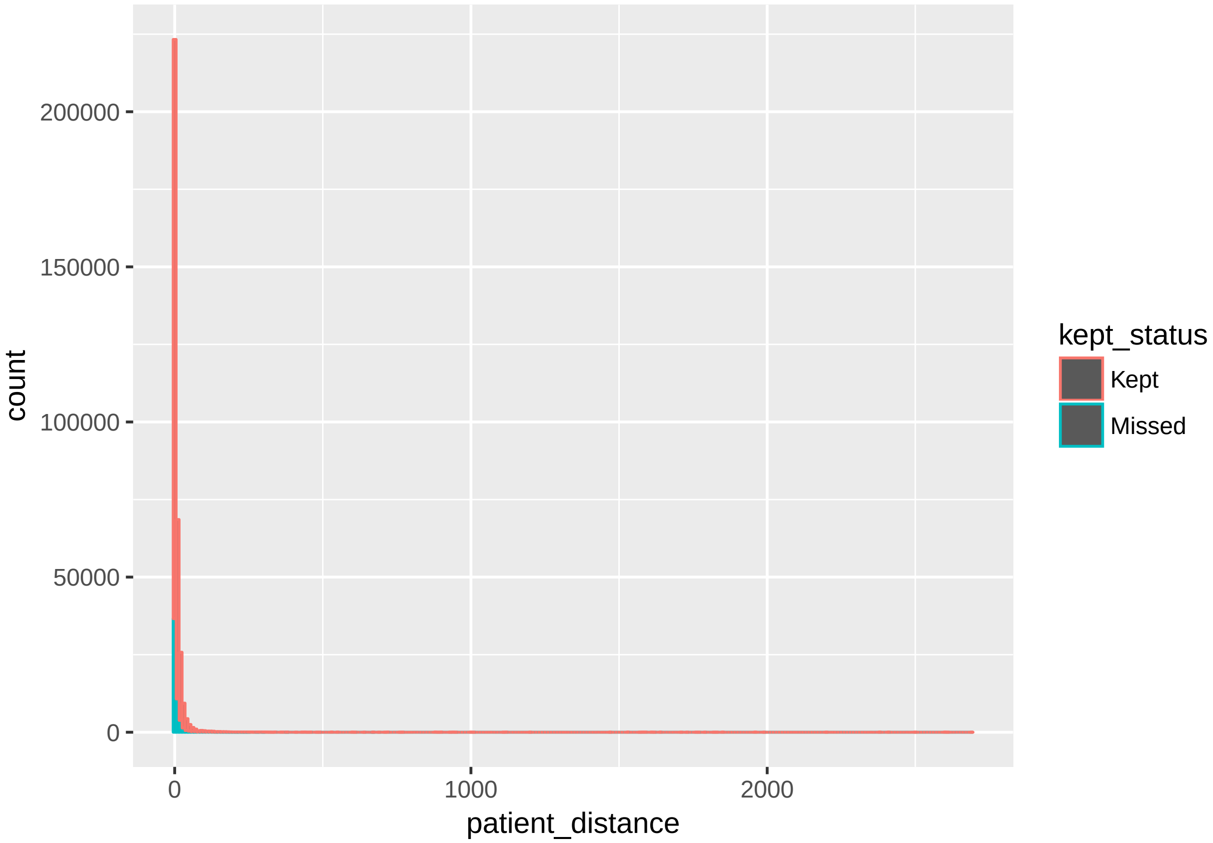
```
length_breaks <- c(-1, 45, 75, 1000)

length_labels <- c("Short", "Medium", "Long")

appointments_2 <- appointments_2 %>%
    mutate(
        length_group = cut(
            appt_length, breaks = length_breaks, labels = length_labels)
        )

ggplot(
    data = appointments_2,
    mapping = aes(x = length_group, fill = kept_status)
) +
    geom_bar(position = "fill")
```



patient_distance

```r
ggplot(
    data = appointments_2,
    aes(x = patient_distance, group = kept_status, col = kept_status)
) +
    geom_histogram(binwidth = 10)
```

```
## Warning: Removed 972 rows containing non-finite values (stat_bin).
```



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

```
appointments_2$patient_distance <- appointments_2$patient_distance %>%
    tidyr::replace_na(median(appointments_2$patient_distance, na.rm = TRUE))
```

Create new variables

percent_missed = percent of prior appointments missed. new represents represents first time appointments appt_lead_time is the difference between the day the appointment was scheduled and the day of the appointment.

```
appointments_3 <- appointments_2 %>%
    mutate(percent_missed = prior_missed / (prior_missed + prior_kept)) %>%
    mutate(new = ifelse(prior_missed == 0 & prior_kept == 0, 1, 0)) %>%
    mutate(appt_lead_time = date(appt_datetime) - date(date_scheduled)) %>%
    mutate(weekday = strftime(appt_datetime, "%A"))
```

Add city_size and county_code from zipcode data.

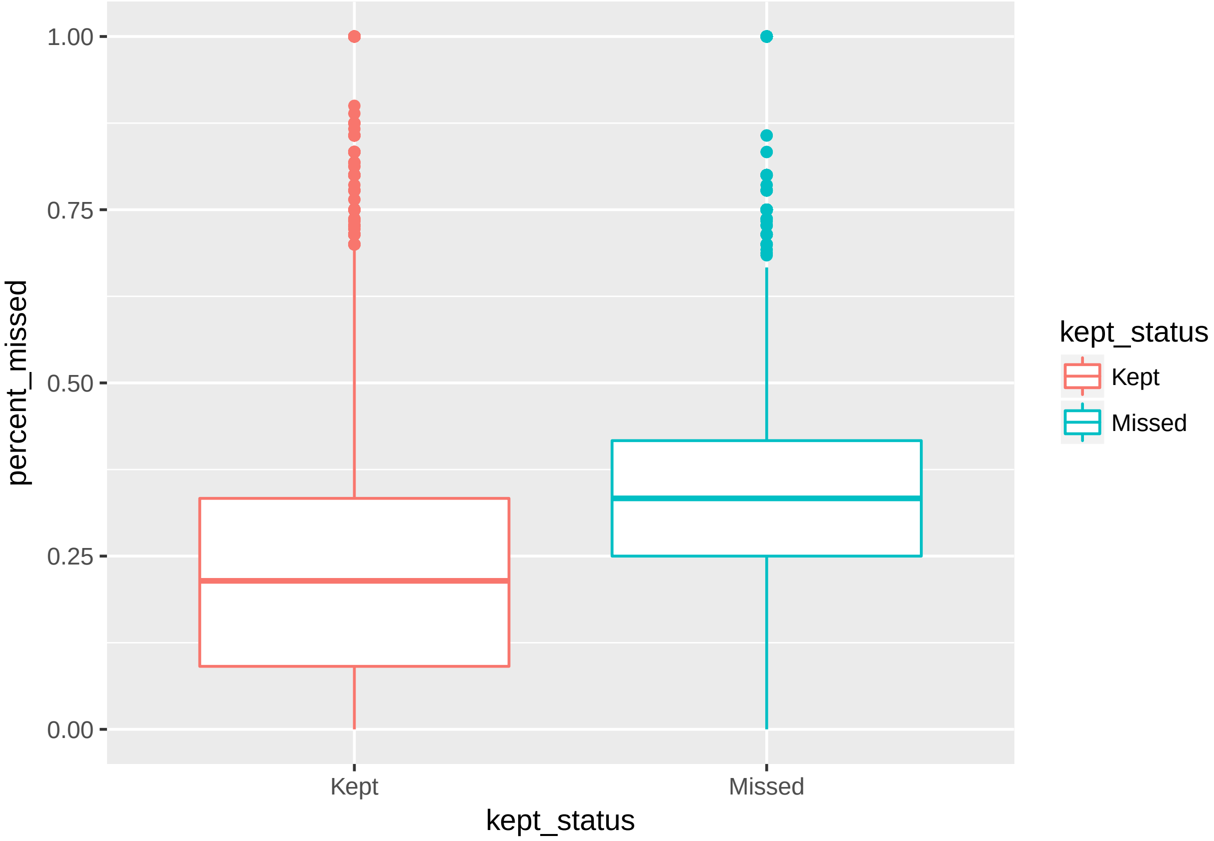
```
appointments_3 <- dplyr::left_join(appointments_3, zipcodes, by = "office_zip")
```

percent_missed

Create random subset and plot

```
ggplot(
    data = appointments_3,
    aes(x = kept_status, y = percent_missed, col = kept_status)
) +
    geom_boxplot()
```

```
## Warning: Removed 22338 rows containing non-finite values (stat_boxplot).
```

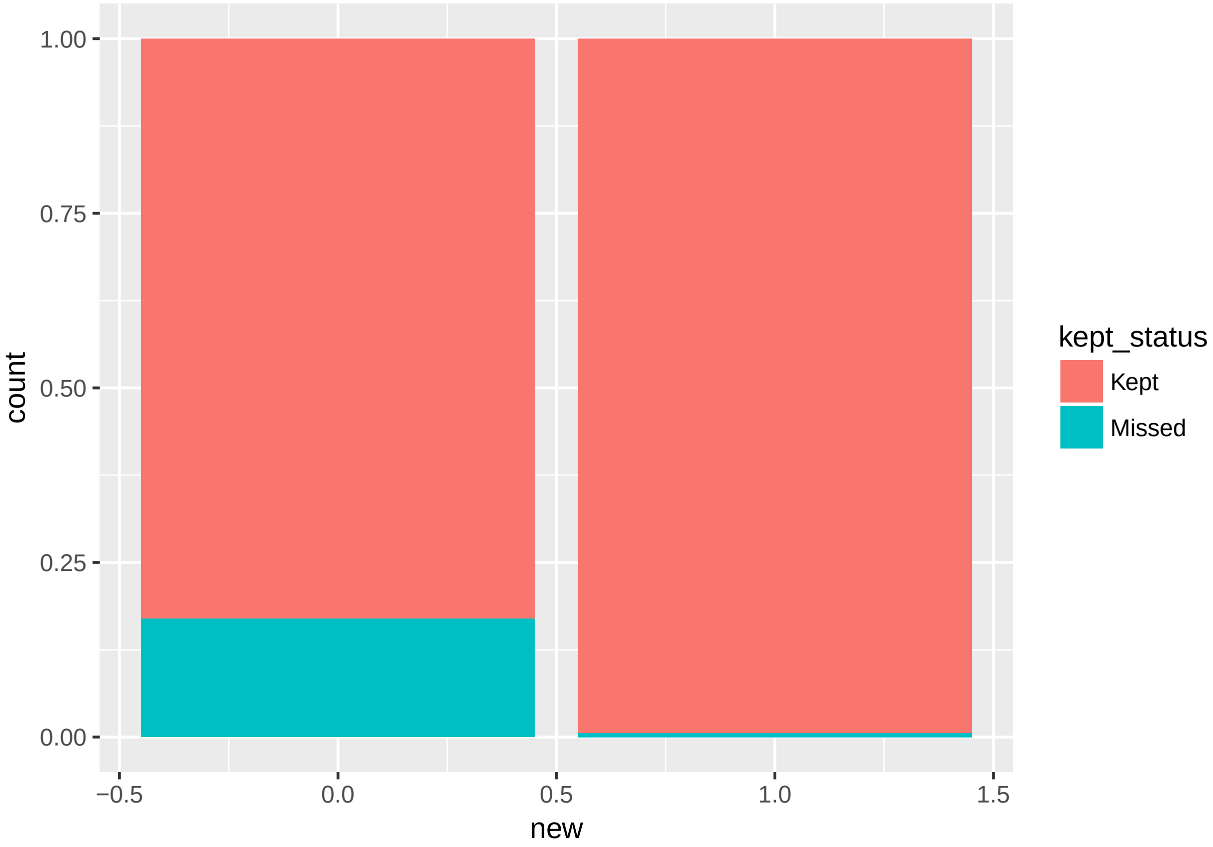


new

```
table(appointments_3$new)
```

```
##
##      0      1
## 320426  22338
```

```
ggplot(
    data = appointments_3,
    mapping = aes(x = new, fill = kept_status)
) +
    geom_bar(position = "fill")
```

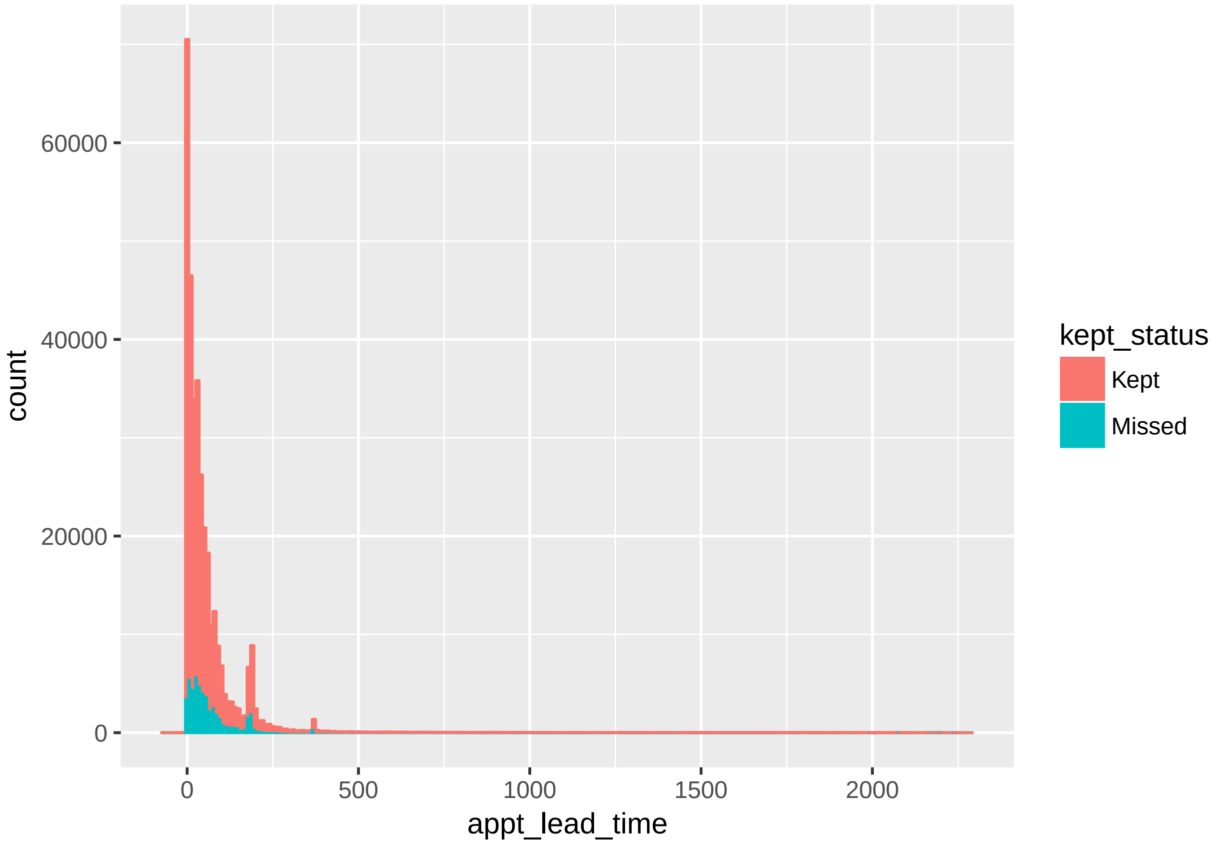


New patients have a very high percentage of kept appointments. 22k of 342k appointments are first-time, or about 6.4%

appt_lead_time

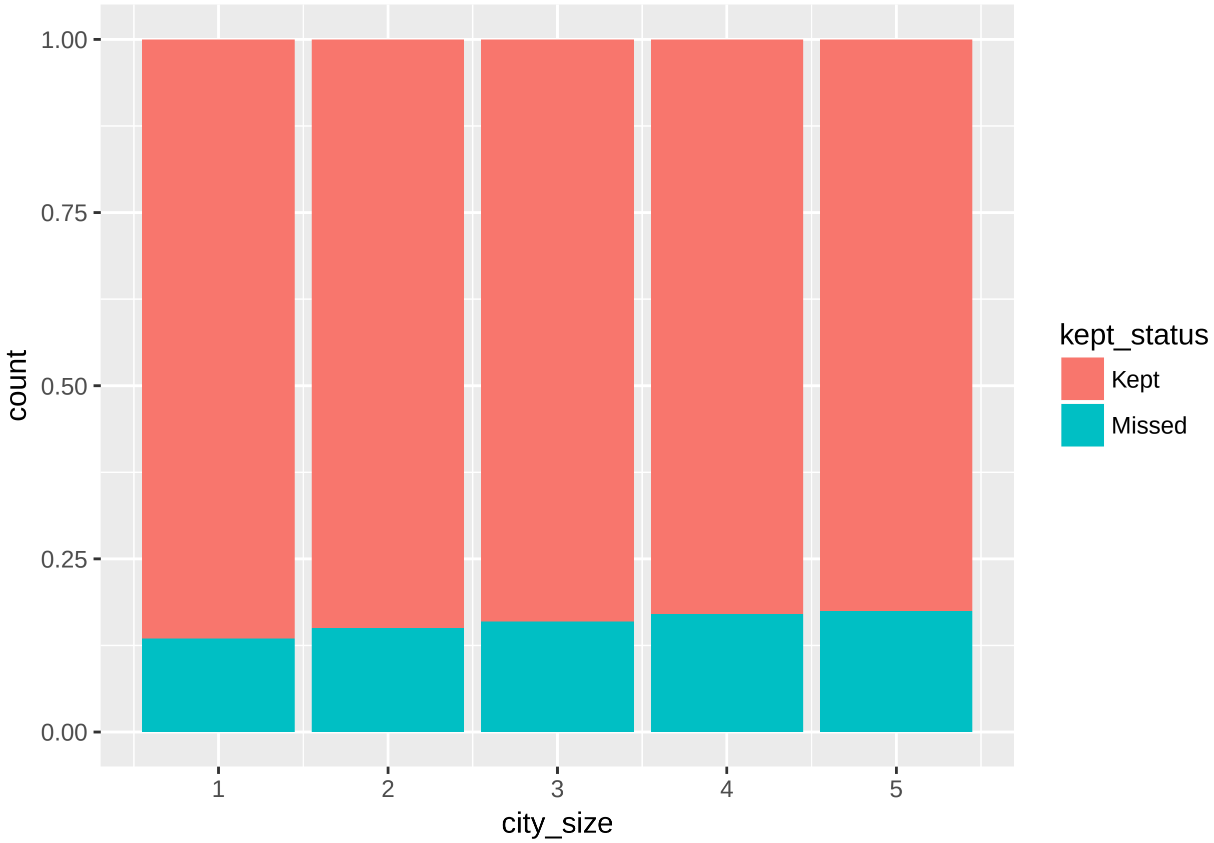
```
ggplot(
    appointments_3,
    aes(x = appt_lead_time, col = kept_status, fill = kept_status)
) +
    geom_histogram(binwidth = 10)
```

```
## Don't know how to automatically pick scale for object of type difftime. Defaulting to continuous.
```



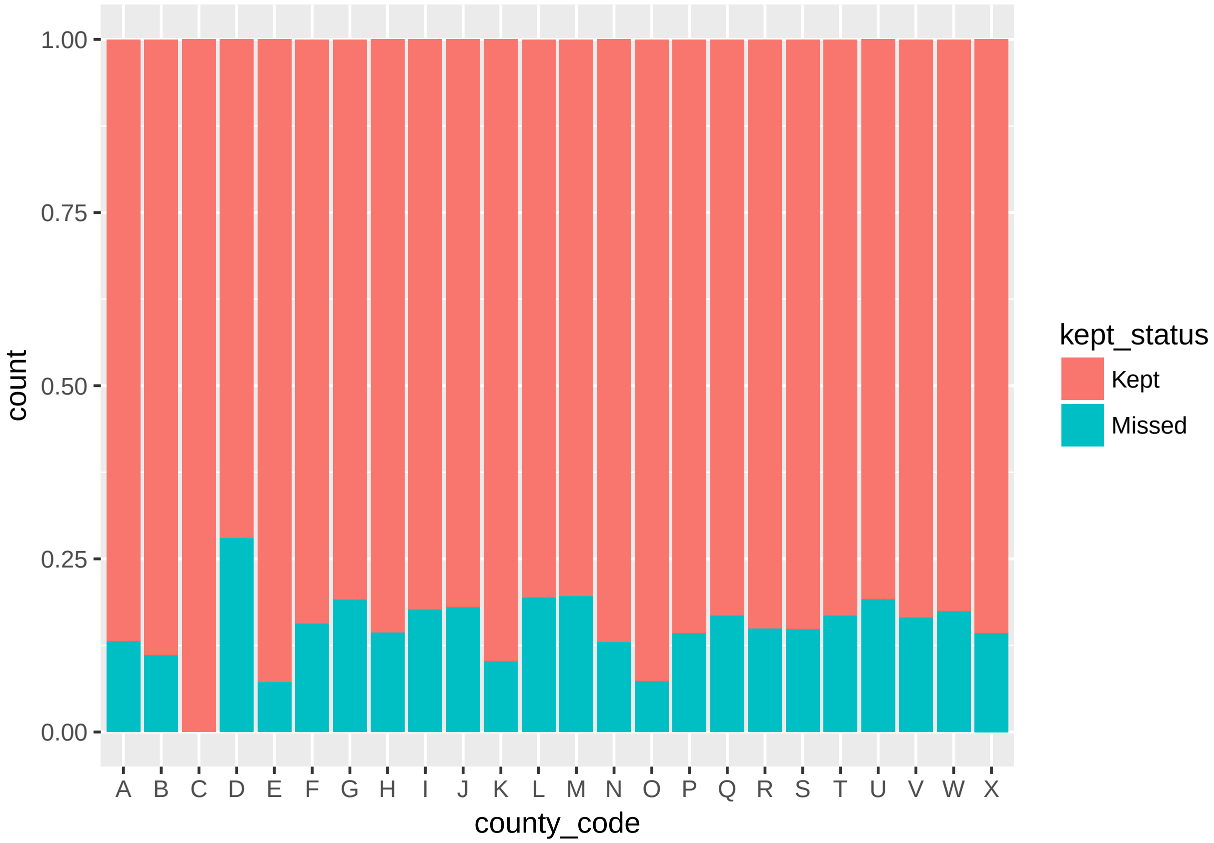
city__size

```
ggplot(
    data = appointments_3,
    mapping = aes(x = city_size, fill = kept_status)
) +
    geom_bar(position = "fill")
```



county_code

```
ggplot(
    data = appointments_3,
    mapping = aes(x = county_code, fill = kept_status)
) +
    geom_bar(position = "fill")
```



```
#appointments_sample_025 <- appointments_3 %>%
#   sample_frac(size = 0.025, replace = FALSE)
#ggpairs(data = appointments_sample_05[,20:24], cardinality_threshold = 50)
```

Modeling

Create Modeling Data

```
model_data <- appointments_3 #%>%
model_data$new <- as.factor(model_data$new)
model_data$percent_missed <- as.integer(model_data$percent_missed * 100)
#Replace NAs with mean
model_data$percent_missed <- model_data$percent_missed %>%
    tidyr::replace_na(mean(model_data$percent_missed, na.rm = TRUE))
factor_columns <- c("kept_status", "patient_gender", "billing_type",
"office_zip", "provider_specialty", "remind_call_result", "hour", "weekday",
"county_code", "length_group")
model_data[factor_columns] <- lapply(model_data[factor_columns], factor)
#Check for NAs
purrr::map(model_data, ~sum(is.na(.)))
```

```
## $kept_status
## [1] 0
##
## $appt_date
## [1] 0
##
```

```
## $appt_time
## [1] 0
##
## $appt_length
## [1] 0
##
## $date_scheduled
## [1] 0
##
## $patient_age
## [1] 0
##
## $patient_gender
## [1] 0
##
## $billing_type
## [1] 0
##
## $prior_missed
## [1] 0
##
## $prior_kept
## [1] 0
##
## $patient_distance
## [1] 0
##
## $office_zip
## [1] 0
##
## $provider_specialty
## [1] 0
##
## $remind_call_result
## [1] 0
##
## $appt_datetime
## [1] 0
##
## $missed
## [1] 0
##
## $age_cat
## [1] 0
##
## $hour
## [1] 0
##
## $length_group
## [1] 0
##
## $percent_missed
## [1] 0
##
```

```
## $new
## [1] 0
##
## $appt_lead_time
## [1] 0
##
## $weekday
## [1] 0
##
## $county_code
## [1] 0
##
## $city_size
## [1] 0
```

```
model_data <- model_data %>%
    select(kept_status, patient_age, remind_call_result, provider_specialty, billing_type,
           hour, percent_missed, appt_length, patient_gender, patient_distance,
           new, appt_lead_time, weekday, county_code, city_size)
```

Divide model_data into train, validate, and test sets

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

```
##
##   Kept Missed
## 174601  31059
```

```
train2 <- train[168738:205660,]
table(train2$kept_status)
```

```
##
##   Kept Missed
##  31059   5864
```

```
train_kept <- train2[train2$kept_status == "Kept",]
train_missed <- train[train$kept_status == "Missed",]

train_balanced <- rbind(train_kept, train_missed)
table(train_balanced$kept_status)
```

```
##
##   Kept Missed
##  31059  31059
```

Logistic Regression Model

```
glm_train <- caret::train(kept_status ~ ., data = train_balanced, method = "glm")
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading
```

```
glm_train$finalModel
```

```
##
## Call:  NULL
##
## Coefficients:
##                                (Intercept)
##                                  1.288e+01
##                                patient_age
##                                 -1.244e-02
##   `remind_call_resultAnswered - Confirmed`
##                                 -1.254e+00
## `remind_call_resultAnswered - No Response`
##                                 -1.667e-01
##  `remind_call_resultAnswered - Reschedule`
##                                  2.338e+00
##                     remind_call_resultBusy
##                                  5.352e-01
##                   remind_call_resultFailed
##                                  3.890e-01
##           `remind_call_resultLeft Message`
##                                 -1.200e+00
##              `remind_call_resultNo Answer`
##                                 -1.074e+00
##             `remind_call_resultNot Called`
##                                 -1.442e+00
##                        provider_specialtyB
##                                 -3.163e-01
##                        provider_specialtyC
##                                 -7.081e-01
##                        provider_specialtyD
##                                 -3.095e+00
##                        provider_specialtyE
##                                  1.281e+01
##                        provider_specialtyF
##                                         NA
##                        provider_specialtyG
##                                         NA
##                            billing_typeDMAP
##                                   3.257e-02
##                                       hour5
##                                  -2.877e+01
##                                       hour6
##                                  -1.276e-01
##                                       hour7
##                                  -1.100e+01
##                                       hour8
##                                  -1.326e+01
##                                       hour9
##                                  -1.333e+01
##                                      hour10
##                                  -1.343e+01
##                                      hour11
##                                  -1.329e+01
##                                      hour12
##                                  -1.220e+01
##                                      hour13
##                                  -1.339e+01
##                                      hour14
##                                  -1.341e+01
##                                      hour15
##                                  -1.346e+01
##                                      hour16
##                                  -1.339e+01
##                                      hour17
##                                  -1.270e+01
##                                      hour18
##                                  -1.207e+01
##                                      hour19
##                                   5.317e-01
##                                      hour20
##                                  -9.188e-01
##                                      hour21
##                                          NA
##                              percent_missed
##                                   4.913e-02
##                                 appt_length
##                                   2.461e-03
##                          patient_genderMale
##                                   7.012e-02
##                         patient_genderOther
##                                  -2.362e-01
##                       patient_genderUnknown
##                                   8.413e-01
##                            patient_distance
##                                  -6.363e-05
##                                        new1
##                                  -3.246e+00
##                              appt_lead_time
##                                   3.580e-03
##                               weekdayMonday
##                                   1.985e-03
##                               weekdaySunday
##                                   8.878e+00
##                             weekdayThursday
##                                  -1.129e-01
##                              weekdayTuesday
##                                  -1.691e-01
##                            weekdayWednesday
##                                  -2.877e-01
##                                county_codeB
##                                  -2.087e-01
##                                county_codeC
##                                  -1.260e+01
##                                county_codeD
##                                          NA
##                                county_codeE
##                                  -1.123e+00
##                                county_codeF
##                                   1.944e-01
##                                county_codeG
##                                   1.506e+00
##                                county_codeH
##                                   5.828e-01
##                                county_codeI
##                                   1.373e+00
##                                county_codeJ
##                                   6.609e-01
##                                county_codeK
##                                   1.226e+00
##                                county_codeL
##                                   8.800e-01
##                                county_codeM
##                                   1.368e+00
##                                county_codeN
##                                   1.059e+00
##                                county_codeO
##                                   9.599e-01
##                                county_codeP
##                                   5.798e-01
##                                county_codeQ
##                                   4.113e-01
##                                county_codeR
##                                   4.244e-01
##                                county_codeS
##                                   1.350e-01
##                                county_codeT
##                                   1.351e+00
##                                county_codeU
##                                   1.763e+00
##                                county_codeV
##                                   2.832e-01
##                                county_codeW
##                                   5.056e-01
##                                county_codeX
##                                   5.174e-01
##                                   city_size
##                                 -3.431e-01
##
## Degrees of Freedom: 62117 Total (i.e. Null);  62051 Residual
## Null Deviance:       86110
## Residual Deviance: 66800     AIC: 66930
```

```r
confusionMatrix(glm_train)
```

```
## Bootstrapped (25 reps) Confusion Matrix
##
## (entries are percentual average cell counts across resamples)
##
##           Reference
## Prediction Kept Missed
##     Kept   35.4   11.3
##     Missed 14.5   38.7
##
##  Accuracy (average) : 0.7417
```

```r
##p_glm <- predict(glm, train)
#caret::confusionMatrix(p_glm, train$kept_status)
```

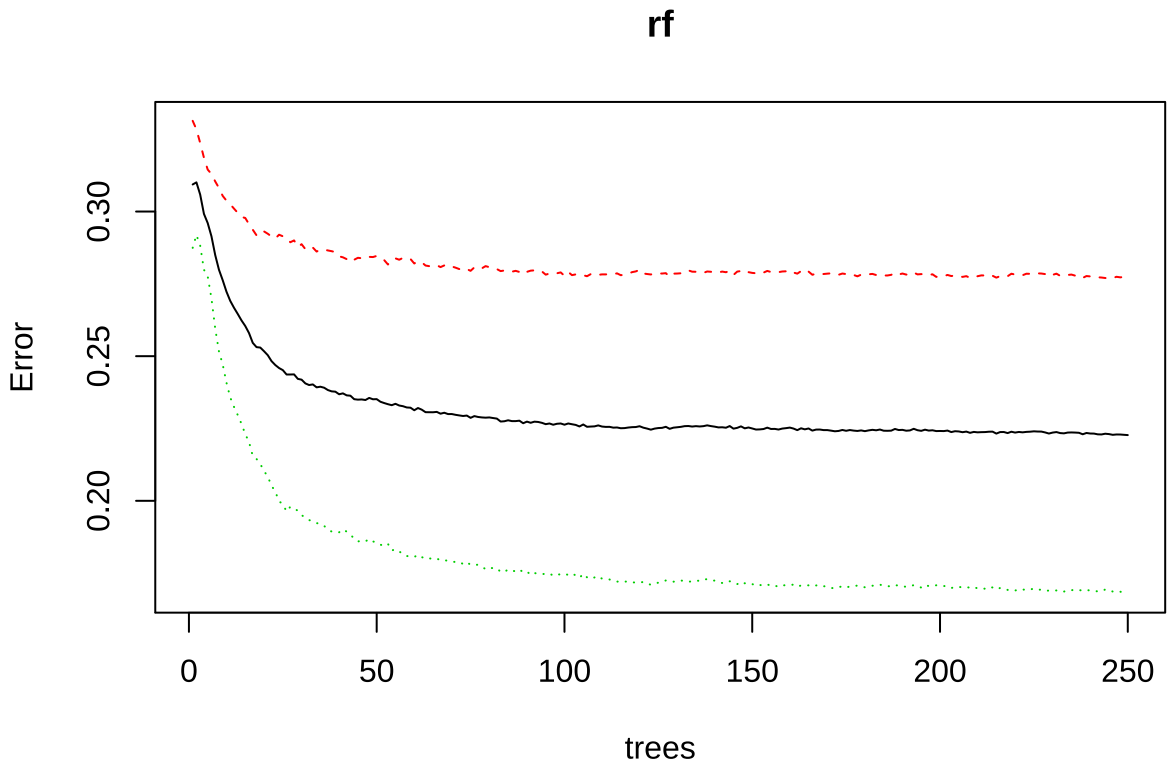
Random Forest Model

Using randomForest Package

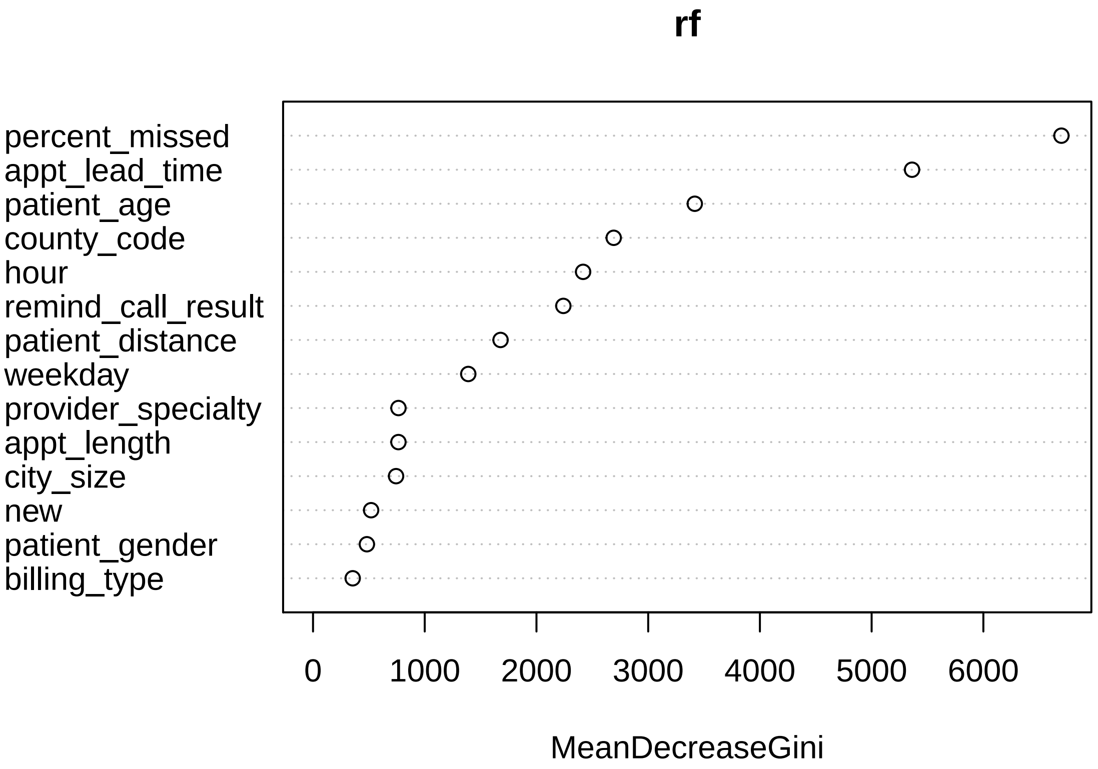
```r
rf <- randomForest(kept_status ~ ., data = train_balanced, ntree = 250)

print(rf)
```

```
##
## Call:
##  randomForest(formula = kept_status ~ ., data = train_balanced,      ntree = 250)
##                Type of random forest: classification
##                      Number of trees: 250
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 22.27%
## Confusion matrix:
##         Kept Missed class.error
## Kept   22438   8621   0.2775685
## Missed  5214  25845   0.1678740
```

```r
plot(rf)
```



```
varImpPlot(rf)
```



Using caret Package

```
# Look at number of cvs and repeats for faster run-time
control <- caret::trainControl(method = "cv", number = 2)
```

```
seed <- 7
metric <- "Accuracy"
set.seed(seed)
mtry <- 3
tunegrid <- expand.grid(.mtry = mtry)
```

```
rftrain <- caret::train(kept_status ~ ., data = train_balanced, method = "rf", metric = metric, tuneGri

caret::confusionMatrix(rftrain)
```

```
## Cross-Validated (2 fold) Confusion Matrix
##
## (entries are percentual average cell counts across resamples)
##
##            Reference
## Prediction Kept Missed
##     Kept   33.5    8.3
##     Missed 16.5   41.7
##
##  Accuracy (average) : 0.7513
```