Data Wrangling for Appointments Capstone

Derek Samsom

4/1/2018

The data was fairly clean to begin with. I checked for any missing values, and only the `patient_distance` variable had any. There were 974 missing out of 342,000 observations. Plotting this variable revealed a right-skewed distribution, so I replaced the missing values with the median.

A table of the `billing_type` variable values showed one observation of "To Be Assigned", or basically unknown. Since this was only one observation, I removed it from the data set.

The appointment date and time were separate variables, that I used to create a combined POSIXct variable.

The `patient_age` variable showed some impossibly high values when plotting. The maximum observation was 264. I removed the observations with an age greater than 110, as I expect anything greater to be likely an error. This removed approximately 80 observations.

The final data wrangling step included adding the following calculated variables, as well as a variable obtained by joining another table:

- `percent_missed` - Calculated by dividing prior number of missed appointments by total number of prior appointments. Calculation resulted in "NA" value for first time appointments, which were converted to zero.
- `new_patient` - Specifies which patiens are new by determing pateints where both prior number of kept and prior number of missed appointments is zero.
- `appt_lead_time` - Difference between the day the appointment was scheduled and the date of the appointment, in days.
- `appt_weekday` - Weekday the appointment occured.
- `weekday_scheduled` - Weekday the appointment was scheduled.
- `county_code` - County (encoded for confidentiality) where the appointment took place, joined from a separate data set.