Data Wrangling Steps: No-Show Data Set

Dovid Burns

1/1/18

The data set that we are working with is relatively clean. We ran the pandas command .info() which showed no NaN's in the dataframe. Looking at each row we did find some problems with individual data points. In the age column we found one age of -1. This is definitely an error, thus we have removed it from the data set.

We used the appointment date column to find out the day of the week that the appointments took place using .dt.weekday_name in order to assess whether there exists a correlation between the day of the week and appointment no-shows. We found that most of the appointments were scheduled to occur Monday through Friday, with only a few exceptions. There were 39 appointments scheduled to take place on Saturday, and, since the next least-common scheduling day had over 17,000 appointments in our dataset, we have decided to exclude the Saturday appointments as outliers.

We examined the time that passed between when the client scheduled the appointment and when the appointment took place. Strangely, there were some time spans with negative values, meaning the appointment was scheduled after the appointment was due to have taken place. Since this is impossible, and there are only five appointments where this occurred, we have removed these from the data set. It is interesting to note that all five of these appointments were no-shows.

Looking at value counts of neighborhoods we discovered that some neighborhoods had very few appointments. These neighborhoods were treated as outliers from the dataset. We will

only keep neighborhoods with greater than 50 appointments so this excludes appointments from ILHA DO BOI, ILHA DO FRADE, AEROPORTO, ILHAS OCEÂNICAS DE TRINDADE, and PARQUE INDUSTRIAL. These neighborhoods together comprise less than sixty appointments, so we simply removed them from our dataset without a significant loss of data.