**Problem Statement: Ohio Clinic: Meeting Supply and Demand:**

Let’s say you are the clinic supervisor at an Ohio clinic, which was facing a big problem; it  had been facing losses from the past three years in spite of having the best doctors and no lack of scheduled appointments. .To be reassured about the financial side of things, you even hired a third-party firm to audit the finance department. However, the firm found no foul play.

You are given a dataset and its feature description is given below:

|  |  |
| --- | --- |
| **Feature** | **Description** |
| Age | Age of patient |
| Gender | Gender of patient |
| AppointmentRegistration | Date on which appointment was issued to the patient |
| ApointmentData | Date for which appointment was issued to the patient |
| DayOfTheWeek | Day of the week for which appointment was issued |
| Status | Day of the week for which appointment was issued (i.e., response variable) |
| Diabetes | Whether the patient has diabetes or not |
| Alcholism | Whether the patient is affected by Alcoholism or not |
| HyperTension | Whether the patient has Hypertension or not |
| Handicap | Whether the patient is handicapped or not |
| Smokes | Whether the patient smokes or not |
| Tuberculosis | Whether the patient has tuberculosis or not |
| Scholarship | Whether or not a patient has been granted scholarship from a social welfare organization or not. Poor families may benefit by receiving financial aid. |
| Sms\_Reminder | Whether SMS reminder for appointment has been issued to the patient or not |
| Awaiting Time Awaiting Time | Appointment Registration – ApointmentData |

Dataset link- <https://www.kaggle.com/joniarroba/noshowappointments>

Can you figure out the main reason behind such losses being incurred so that resources may be allocated accordingly to minimize them? Going one step ahead can you build some kind of model to predict if an appointee will come or not?

Solution:

Summary:

The aim of this problem statement is to analyse and build a prediction model to predict if an appointee will come or not.

Machine learning Pipeline: Steps to solve above mentioned two problems are as given below –

1. Understanding the dataset

The first and foremost step to solve any machine learning problem should be to thoroughly understand the data. Looking into the dataset given in the link, its feature description are given below:

|  |  |
| --- | --- |
| **Feature** | **Description** |
| Age | Age of patient |
| Gender | Gender of patient |
| AppointmentRegistration | Date on which appointment was issued to the patient |
| ApointmentData | Date for which appointment was issued to the patient |
| DayOfTheWeek | Day of the week for which appointment was issued |
| Status | Day of the week for which appointment was issued (i.e., response variable) |
| Diabetes | Whether the patient has diabetes or not |
| Alcholism | Whether the patient is affected by Alcoholism or not |
| HyperTension | Whether the patient has Hypertension or not |
| Handicap | Whether the patient is handicapped or not |
| Smokes | Whether the patient smokes or not |
| Tuberculosis | Whether the patient has tuberculosis or not |
| Scholarship | Whether or not a patient has been granted scholarship from a social welfare organization or not. Poor families may benefit by receiving financial aid. |
| Sms\_Reminder | Whether SMS reminder for appointment has been issued to the patient or not |
| Awaiting Time Awaiting Time | Appointment Registration – ApointmentData |

There are total 110528 records in the dataset with no missing values.

Data columns (total 14 columns):

PatientId 110527 non-null float64

AppointmentID 110527 non-null int64

Gender 110527 non-null object

ScheduledDay 110527 non-null object

AppointmentDay 110527 non-null object

Age 110527 non-null int64

Neighbourhood 110527 non-null object

Scholarship 110527 non-null int64

Hipertension 110527 non-null int64

Diabetes 110527 non-null int64

Alcoholism 110527 non-null int64

Handcap 110527 non-null int64

SMS\_received 110527 non-null int64

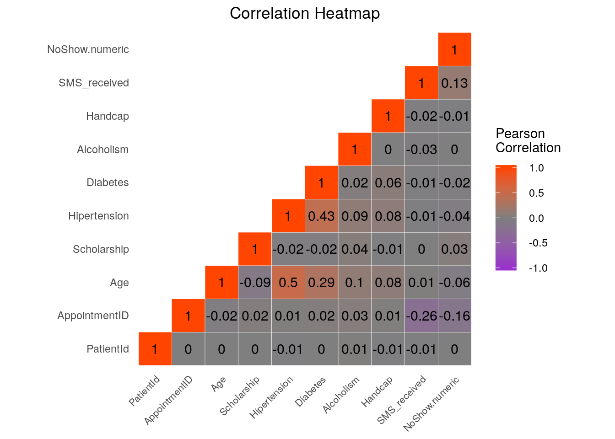
No-show 110527 non-null object

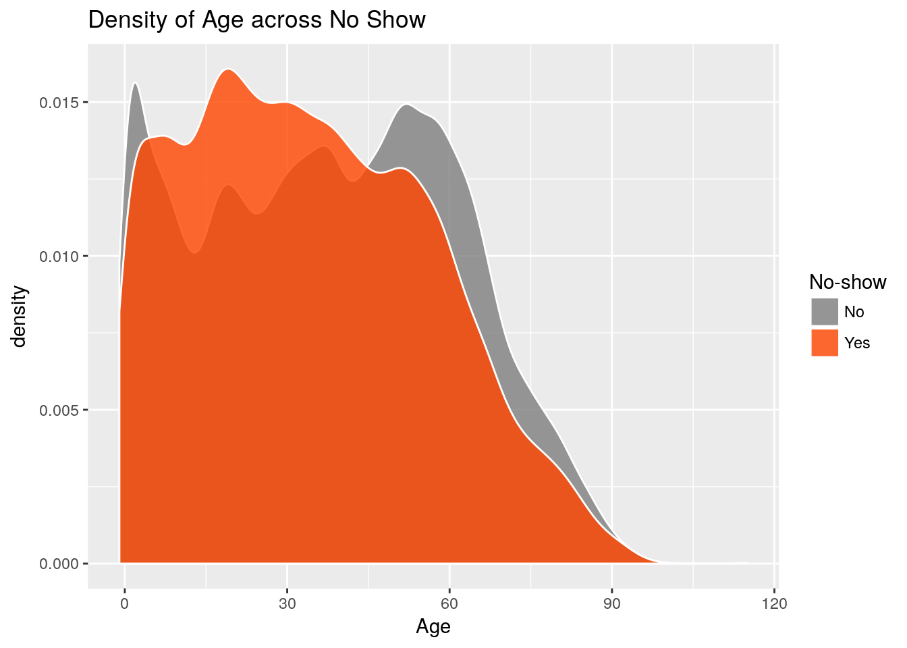
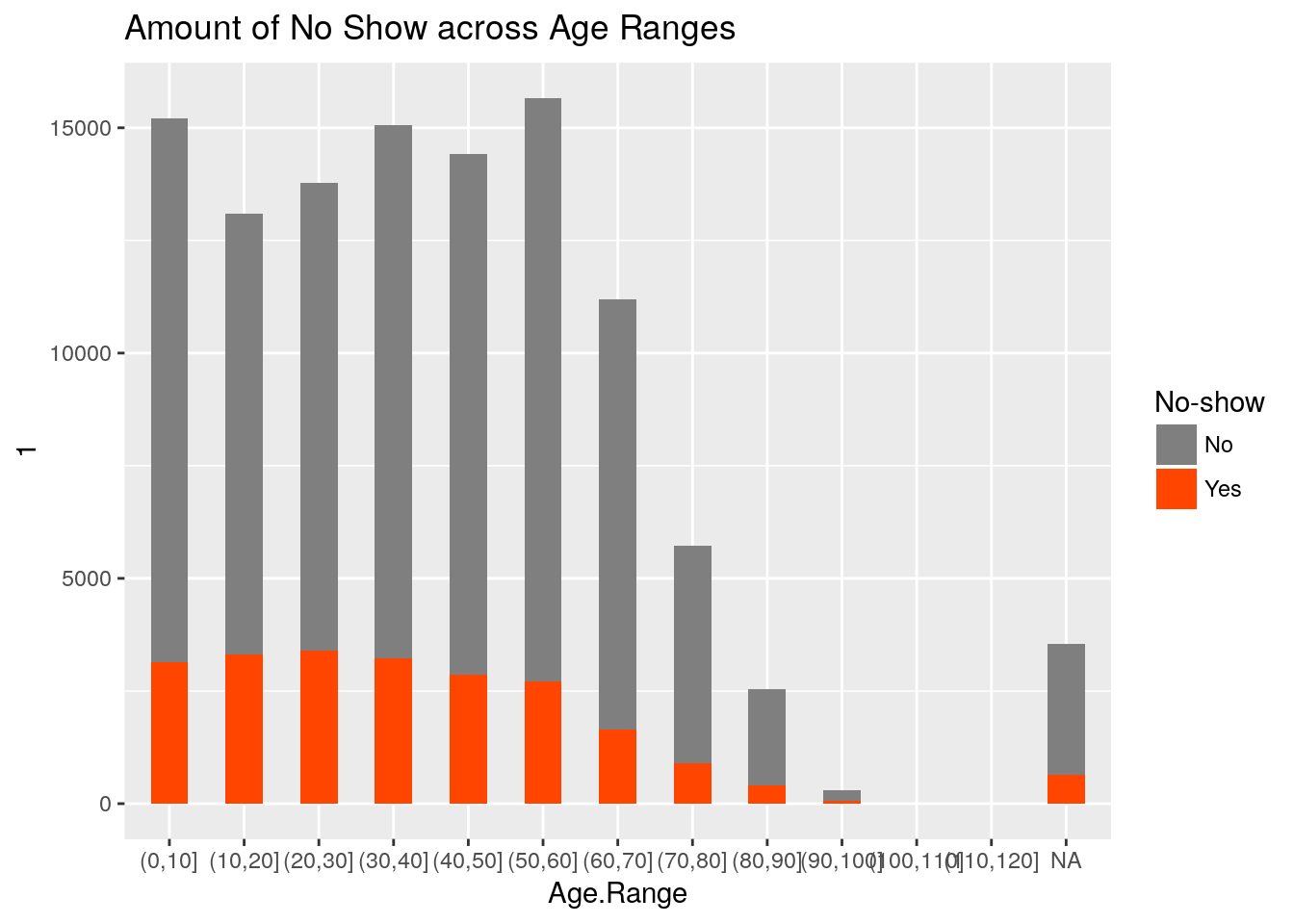
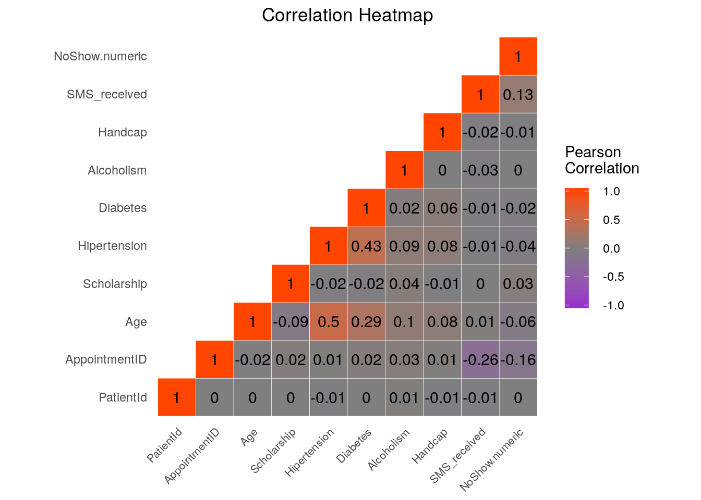
1. Feature Exploration

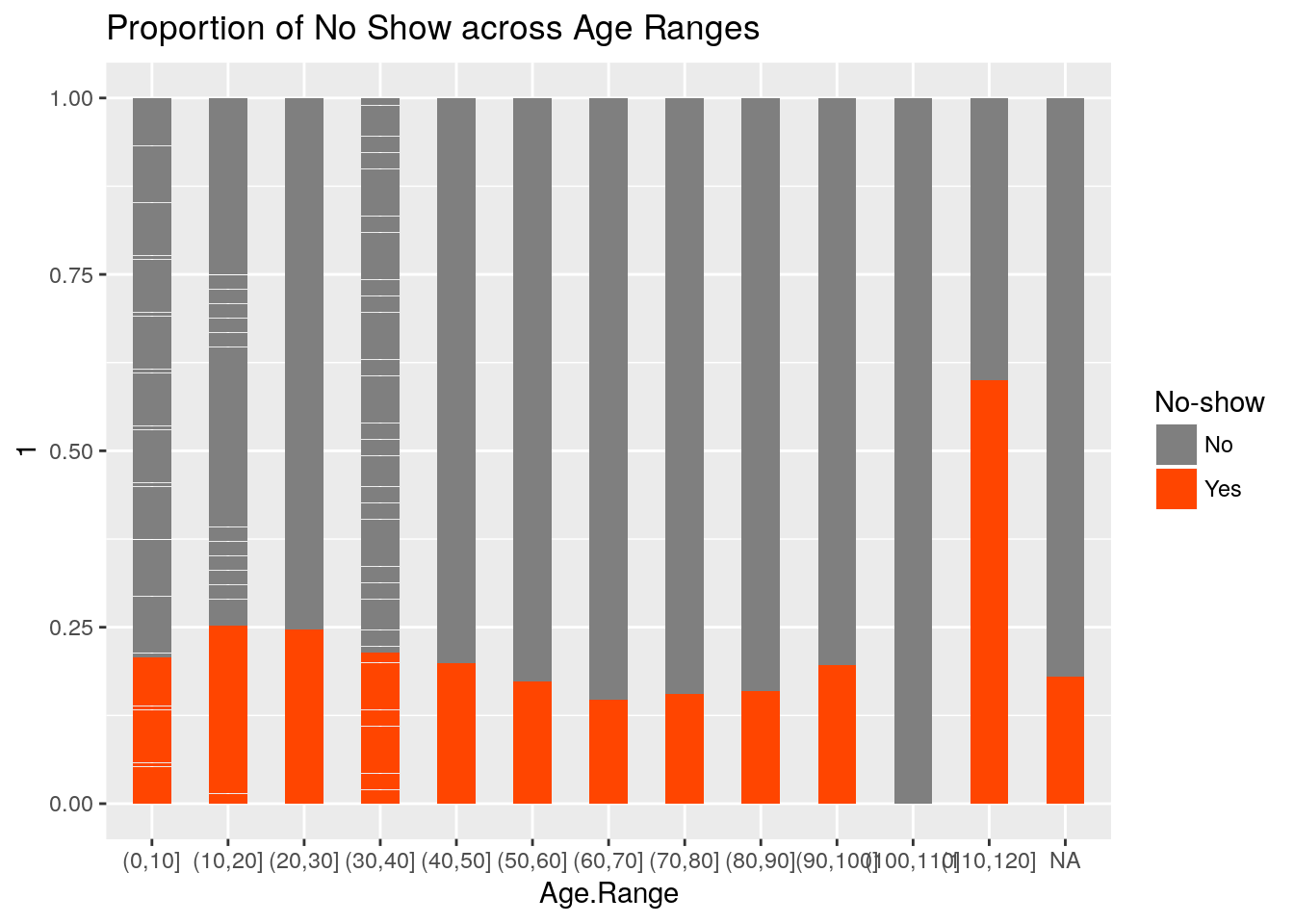
On exploring the dataset available, it can be figured out

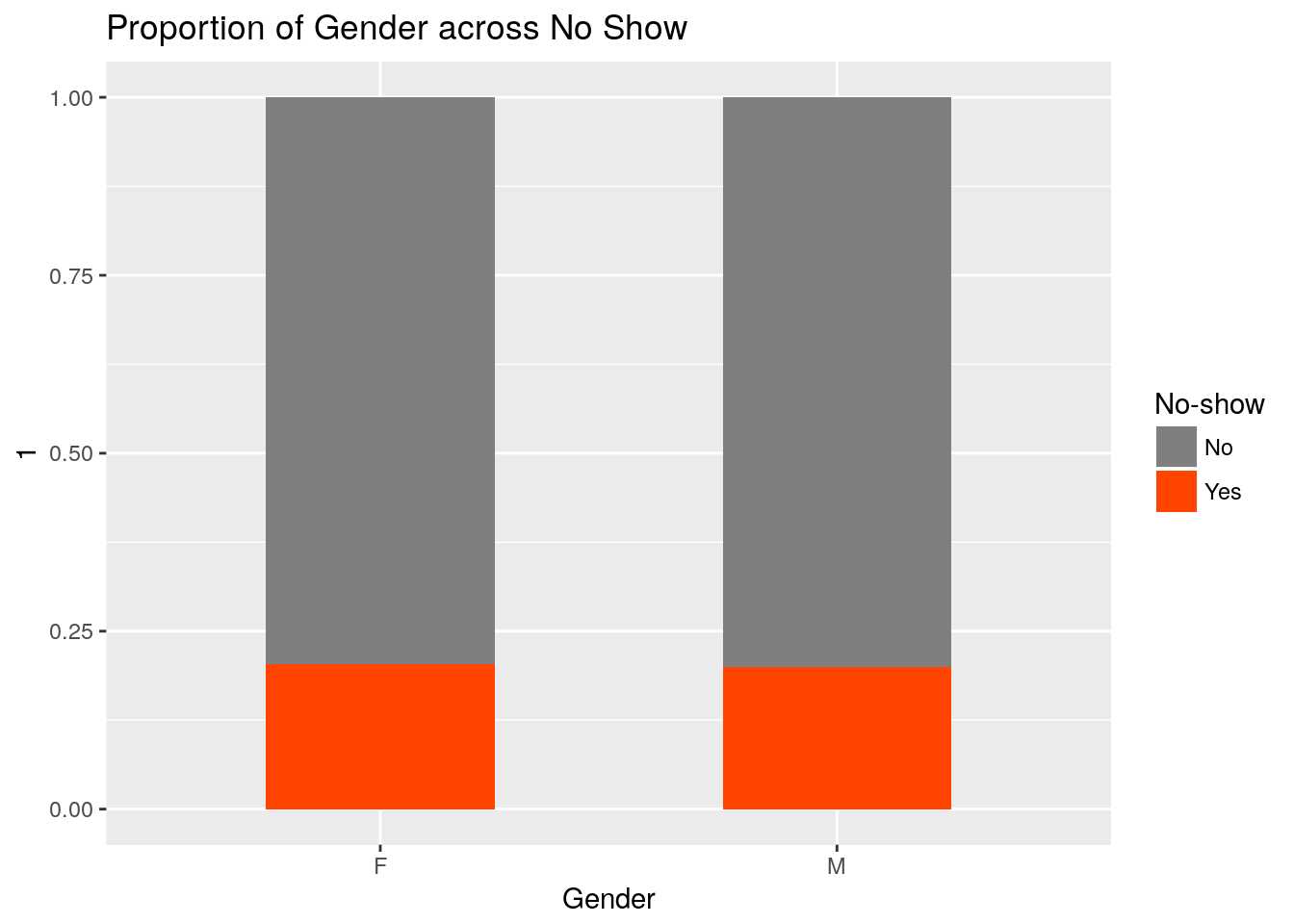
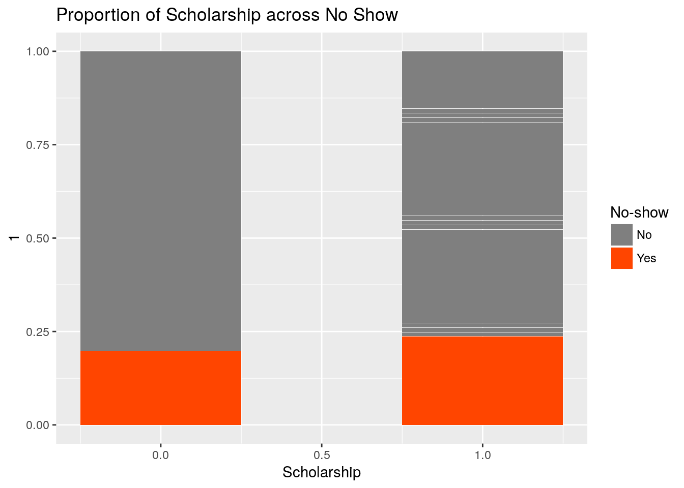
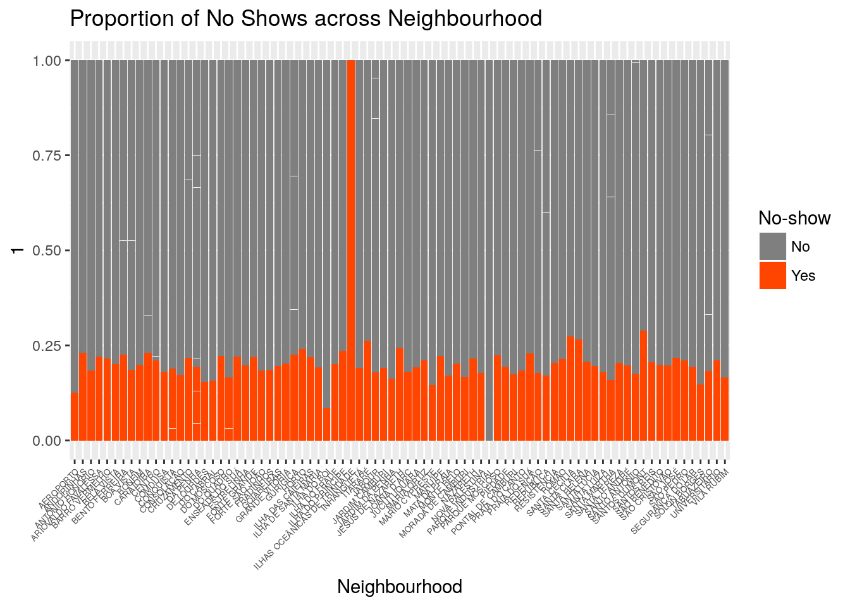
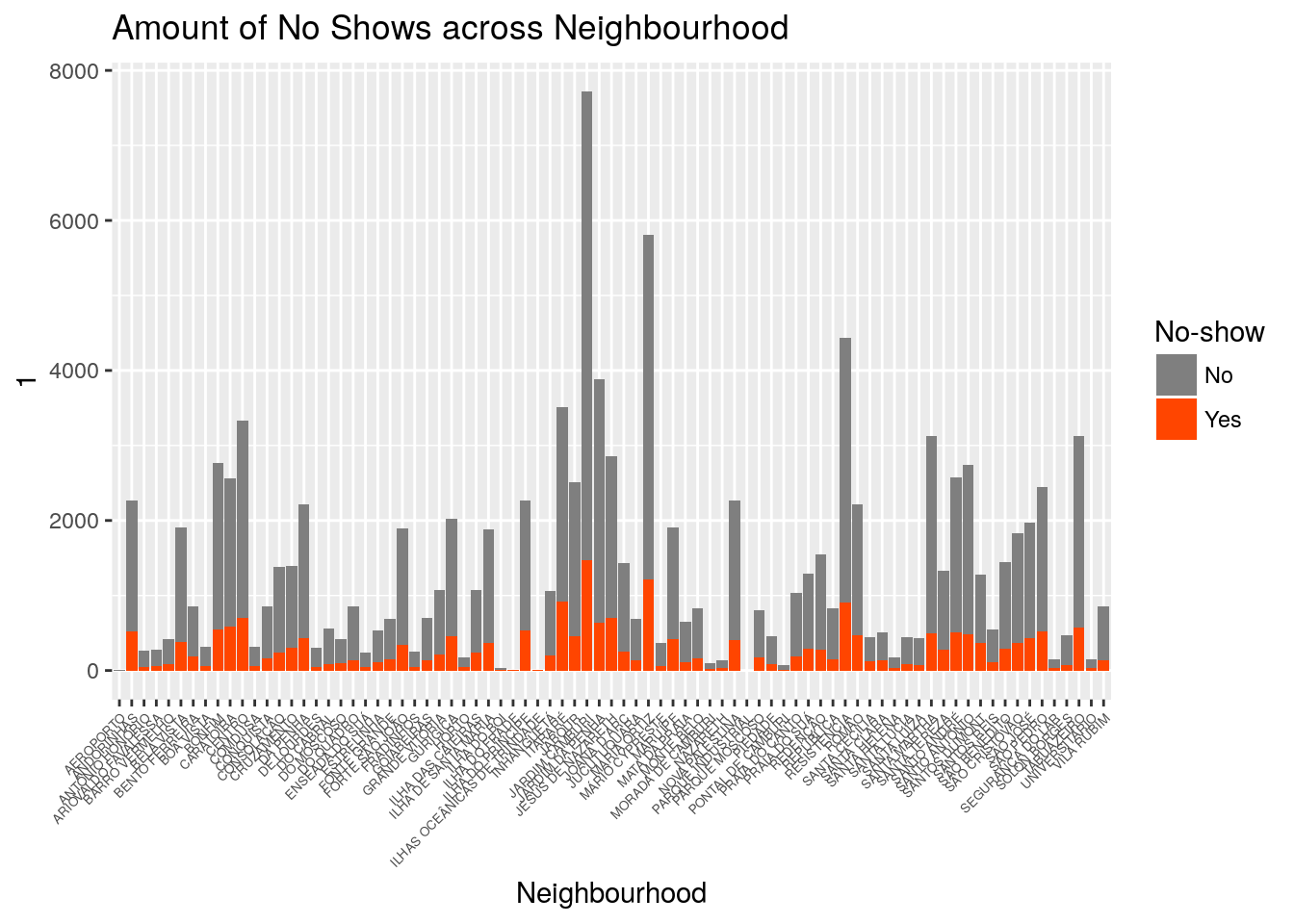
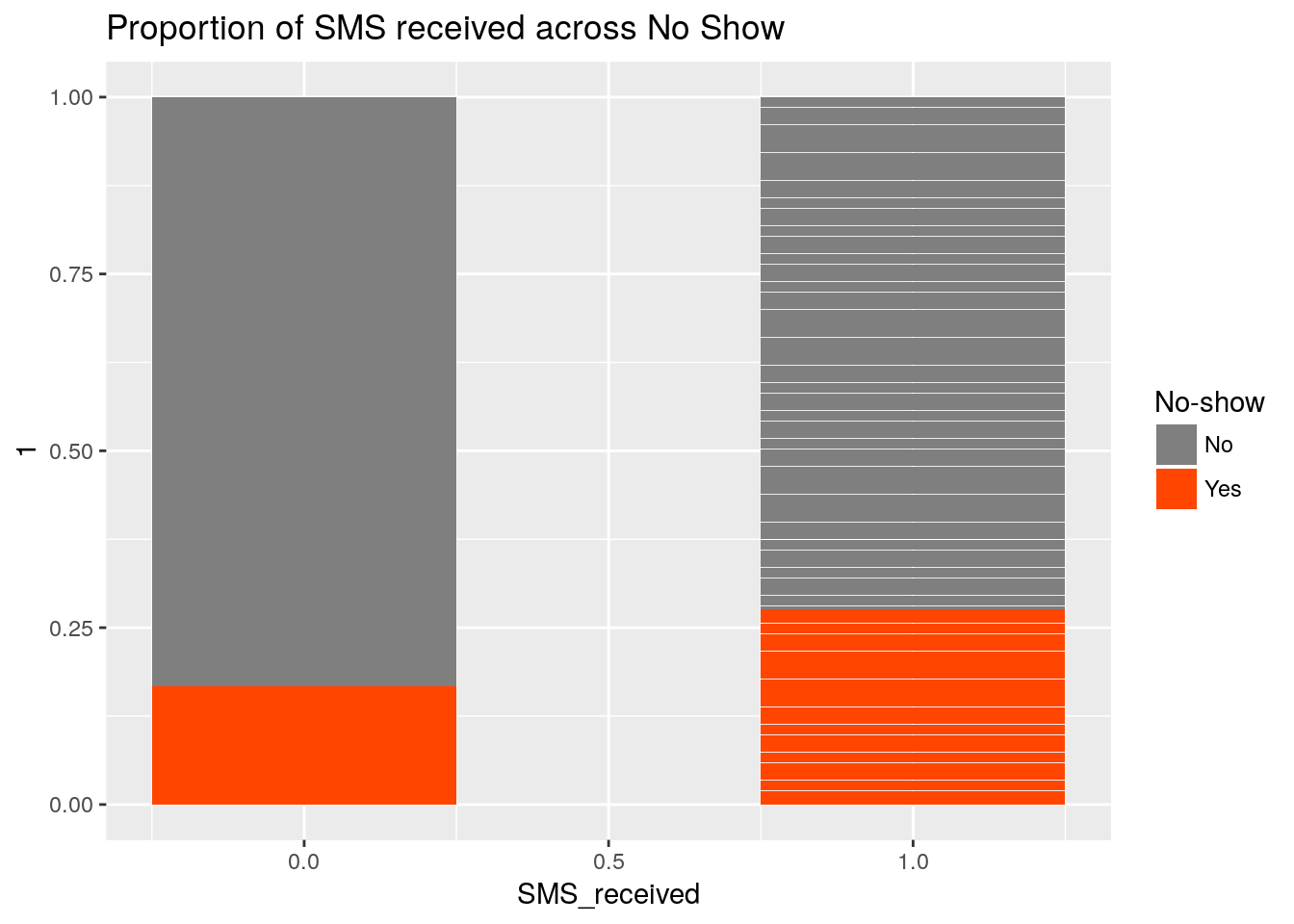
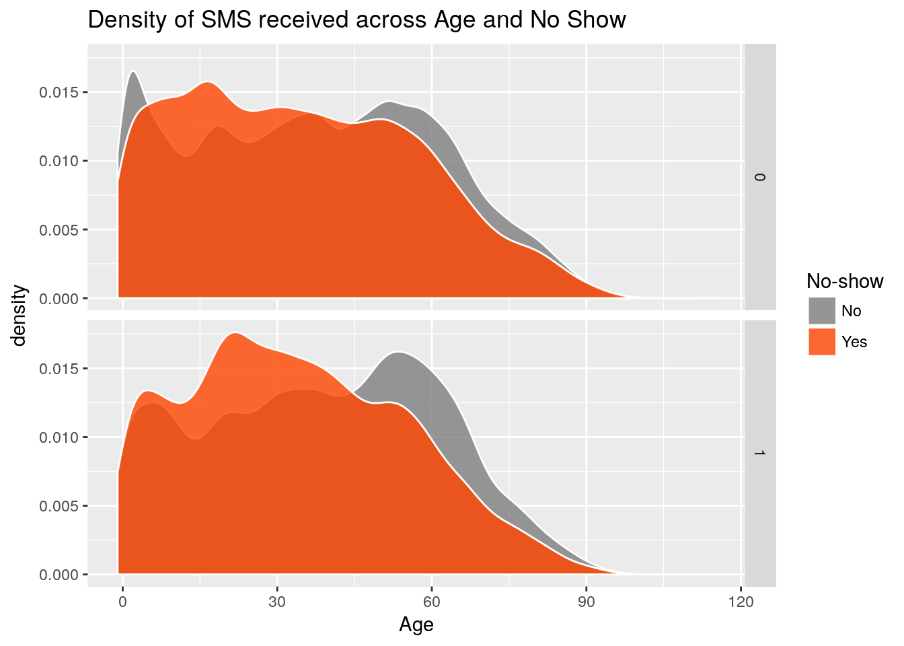
1. Patient ID and Appointment ID are irrelevant columns and does not give any insight about the overall data and hence can be ignored
2. Feature Engineering
3. Day of the week and day of the month can be derived from ScheduleDay and AppointmentDay to understand the correlation of day of the week for both scheduled day the appointment day with no-show.
4. A new feature ‘Days left’ which calculates the number of days left for the appointment. This can be calculated by subtracting AppointmentDay by ScheduleDay
5. Exploratory Data Analysis

Various graphs can be plotted to understand the relationship of different parameters against the ‘no show’ variable









1. Modelling

After preparing and analysing the data, a predictive model is prepared. Since we need to predict if a patient will show up for the appointment or not, this is a classification model which gives binary output -yes and no.Classification models such as logistic regression and Decision Trees can be used

1. Evaluation of the model

Confusion matrix is created to evaluate the model performance. A model is good enough if it has a higher recall and precision.