**Data Science Study: Medical Appointment No-Shows**

**Authors: Wenbo Liu, Heather Wang, Shiying Zheng, Shanshan Qiao, Wei Tao**

Business Understanding

At a hospital in Vitória, Brazil, a patient makes a doctor's appointment and receives all the instructions but fails to show up to the doctor's appointment. This phenomenon causes great inconvenience for doctors who plan their workdays and patients who need medical attention. The appointment slots are wasted due to the no-shows. Since the spaces are occupied and missed by patients deliberately or accidentally, patients who actually need medical help cannot utilize the opportunities, which turns out to be a waste of medical resources. On the other hand, doctors might have prepared all the equipment needed for the appointment but are only greeted with no-shows, which is also a waste of the doctor’s time. Therefore, hospitals need to figure out how to identify individual cases where doctors’ appointment no-shows are likely to happen and find a solution that can improve the efficiency of using medical resources. For example, hospitals can adopt backup plans, allow patients to register multiple slots, and alert patients who do not show up to the doctor’s appointment without notifying the hospital in advance.

This hospital collects a variety of data characteristics for its doctor’s appointments, including no-shows. Many factors of patients’ conditions can contribute to whether the patients will show up to the appointments or not. Setting our target variable as no-shows and conducting our data analytics through data mining, we will have a better understanding of how patients’ physical and mental health conditions may have an impact on whether they will show up to a doctor’s appointment or not. Although we know that we should not judge anyone by their health problems, we need to admit that some conditions will influence their performance on normal daily tasks. After learning which kinds of patients have more tendencies to dodge doctors’ appointments, we would be able to make plans that can minimize the costs and losses.

Data Understanding and Preparation

Aquarela Advanced Analytics provided the data we have used to address our business problem on Kaggle: Medical Appointments No Shows:

<https://www.kaggle.com/datasets/joniarroba/noshowappointments> .

The data characteristics included in the dataset are demonstrated in Exhibit 1.

Based on the nature of the dataset, we will be analyzing the data as a classification problem because our goal is to predict which doctor’s appointment has more possibilities of being dodged. The dataset has 14 variables and 110526 doctor’s appointments documented. After looking into our data more closely, it becomes clearer that our target variable should be “No-show” – whether the patients show up to the doctors’ appointments or not.

Through our EDA, we found no null values or duplicated items in the original dataset, which worked in our favor. However, some variables like “patient\_id” and “appointment\_id” are irrelevant, so we dropped them early in our analysis.

We also created new variables from the existing ones in the dataset to help with our prediction of how likely is one patient to fail to show up to a doctor’s appointment. The variable “day” was created to represent the doctor’s appointment's weekday. The original dates are presented in a date-time format, so we added a new weekday variable. Another new variable we created is “timespan,” representing the time elapsed between the patients’ call to schedule the appointment and the actual appointment date. The variable was calculated by subtracting the “Scheduledday” variable from the “Appointmentday” variable. We made this new variable because we thought that the elapsed time between the two days could impact the no-show rates. For example, a longer timespan might result in a higher no-show rate because the patients might have forgotten about the appointments.

Visualizations

We produced some bar charts and box plots through R to observe some general trends and rough relationships between ‘no shows’ with other variables, and we noticed that:

Most appointments are scheduled on weekdays (Exhibit 2);

The absence rates on weekends are slightly higher than that on weekdays (Exhibit 3);

The number of female patients who made appointments is almost two times the number of male patients (Exhibit 4);

Whether patients are alcoholics or not nearly has no impact on their attendance results (Exhibit 5);

Patients with more severe handicaps or disabilities are less likely to show up (Exhibit 6);

Patients who received SMS are less likely to show up (Exhibit 7);

The average age of patients who attended the appointment is higher than that of those who did not show up (Exhibit 8);

Timespan (The number of days between the appointment being made and the appointment day) shows many outliers in the box plot, and there is an evident trend that patients with a lower timespan are more likely to show up on appointments (Exhibit 9);

Patients who are recorded to have certain diseases, such as diabetes or hypertension, are more likely to show up (Exhibit 10);

Modeling

After observing the target variable, we knew that it should be a classification problem with the result of either yes or no. First, we created a holdout sample to split the dataset into training data (88422 instances) and testing data (22104 instances), which took up 80% and 20% of the whole dataset.

Based on what we learned in this course, we decided to apply a logistic model to fit the dataset. First, all the variables were put into the function to fit the model. Then we used the summary function to make feature selections, where variables with a P-value less than 0.05 would be chosen from the original dataset and refit the model. They are ‘timespan,’ ‘age,’ ‘scholarship,’ ‘hypertension,’ ‘diabetes,’ ‘alcoholism,’ and ‘sms\_received’ respectively. According to the coefficients of different variables, if there is one unit increased in timespan, the log odds of the probability of “no show” would increase by around 0.0230. If there is one unit increased in age, the log odds on the probability of “no show” would decrease by around 0.0077.

Our second model is a classification tree. Similarly, we used variables chosen from the feature selection to build a tree model. We set the cp to 0, the maxdepth to 6, and the minsplit to 100. Then we applied the rpart.plot function to visualize the tree (Exhibit 11). According to the plot, there is only one path going to the “Yes” result: First, the patient should have a time span longer than one day; then, they must be under the age of 49; the time span should be shorter than ten days; next, the patient should have accepted social welfare (the “scholarship” variable in the dataset); and lastly, the patient should be an alcoholic. Other paths all head to the result of “No.”

The third classification model is the random forest, the ensemble of multiple decision tree models. We put all significant variables into the random forest model and preliminarily set “ntree” equal 500 and cutoff equals 0.5 and 0.5. Then we executed a tuning process using the function “tuneRF” with the value of “mtryStart” equals 2. After that, we noticed that when “mtry” equals 2, the model would have the lowest OBB error, which is less than 0.202 (Exhibit 12). We eventually built the best random forest model with a “mtry” equals 2, and Exhibit 13 shows the features’ importance for the random forest model in ascending order.

We chose SVM as the last model with all significant variables used in the function SVM. We set the kernel as “linear,” and the cost equals 0.1. Then we produced a plot of distribution for the SVM model (Exhibit 14).

Evaluation

For the first logistic model, we initially set the threshold as 0.5 after calculating the propensities. According to the confusion matrix, even if the accuracy is 0.7946, the false-positive cases (patients who are predicted to attend but actually were not) are too many, with a number of 4391. Only 72 patients who did not show up have been predicted correctly by the model, and 149 patients who showed up at their appointments are predicted as “no show.” Next, we set the threshold value to 0.3, and more “no show” patients were correctly predicted (868 cases). However, in the sacrifice of the accuracy, which decreased to 0.7567. Then we produced the ROC plot (Exhibit 15) with the AUC equaling 0.660.

Then we tested the performance of the decision tree model. Unfortunately, there were only nine instances in the testing dataset that met the condition of “Yes”, and the model got all patients as “No” despite the accuracy of 0.7980. 4467 patients were mispredicted. As we changed the threshold of the classification, the accuracy decreased, despite the false positive would increase. The AUC of the tree model is 0.710, which is larger than that of the logistic model. The ROC plot of the model is given (Exhibit 16). Generally, the model struggled in filtering patients who did not attend.

Next, we evaluated the performance of the random forest model. After optimizing the OBB error, the model presented almost the same accuracy as the decision tree, with a value of 0.7982, which had slight progress. According to the output, 4461 patients got the wrong predictions, which were 6 fewer people than the decision tree. Also, based on its AUC, which equals 0.672, the random forest model performed even worse than the decision tree even if it is an ensemble of three models. The ROC plot is given in Exhibit 17.

The SVM model was similar to the previous models. It created 35718 support vectors to classify different target variables, but it predicted every instance as “No.” According to Exhibit 14, we observed that all instances were distributed around 1, and some even went beyond. Then we tried different algorithms and set the kernel as “polynomial” and “sigmoid”; however, the model's performance did not improve either.

Conclusion

After running four models, we can retrieve some of the most critical factors impacting the no-show rate. Variables like “Timespan,” “age,” and “sms\_received” have more significance to the prediction than others. It means that the time elapsed between the day when patients scheduled their appointments and the actual appointment dates, the age of the patients, and whether they have received notification can affect whether the patients decide to show up to their doctors’ appointments or not. Other factors related to patients’ health conditions are not too important in the predictions.

In terms of accuracy, the models’ performances were similar to each other and basically all showed good predictions. Particularly, the random forest model showed the highest accuracy rate of 79.82%. Other models also got good values above 75%.

In terms of AUC, the logistic model had the worst performance with a value of 0.660, followed by the random forest, with a value of 0.672. The decision tree model showed the greatest AUC of 0.710. It is not surprising to see the random forest worked less effectively than the decision tree because the random forest is likely to overfit the data. Thus, a possible explanation could be overfitting.

The common drawback that the models have shared is that the true negative instances were barely filtered out. In most cases, the models tended to classify all patients as “No,” which means everyone would attend the appointment but actually, they did not. To optimize the model, we tried different threshold values on these models, and it did help reduce the false positive error with more “no show” patients being recognized. But the total accuracy rate decreased as more true “No” cases were predicted as “Yes”. Therefore, a trade-off should be made regarding different business situations and goals. In this case, finding patients who did not attend the appointment is our primary goal, so we recommend tuning down the threshold of the classifiers appropriately from 0.5 to around 0.3 to distinguish more real “no show” patients from the dataset with little accuracy loss.

To reduce the potential waste of medical resources, improve the model, and acquire better insights into the possible reason why patients decide not to show up for their appointments, we suggest that the hospital could gather more information about patients. For example, the hospitals could collect patients’ income levels, the weather conditions of the appointment day, and the distance between the patient’s neighborhood and the hospital. According to the “AppointmentDay” variable, the dataset provides ranges only over one month, which is insufficient for conducting convincing analysis and predictions. Time details are also missing from “AppointmentDay.” However, it is quite essential to know what time of the day those scheduled appointments were at because early or late times may have a negative impact on the probability of patients showing up. We assume that appointments from noon to afternoon may have a higher show-up rate than other times in the day. Besides, knowing which exact departments of no-show appointments are can also help us and the hospital do better analysis and prediction on this dataset. For example, patients who scheduled appointments for their toothache would be more likely to dodge than patients who scheduled their pregnancy checkups.

Appendix:

Exhibit 1: Data Characteristics from Kaggle.com for Medical Appointment No-Shows

PatientId - Identification of a patient

AppointmentID - Identification of each appointment

Gender - Male or Female

Schedule\_day - The day someone called or registered the appointment

Appointment\_day - The day of the actual appointment

Age - How old is the patient.

Neighborhood - Where the appointment takes place.

Scholarship - True or False. Observation, whether the patient takes social welfare or not. Bolsa Família

Hypertension - True or False

Diabetes - True or False

Alcoholism - True or False

Handicap - how many disabilities a patient has

SMS\_received - 1 or more messages sent to the patient.

No-show - True or False.

Exhibit 2:

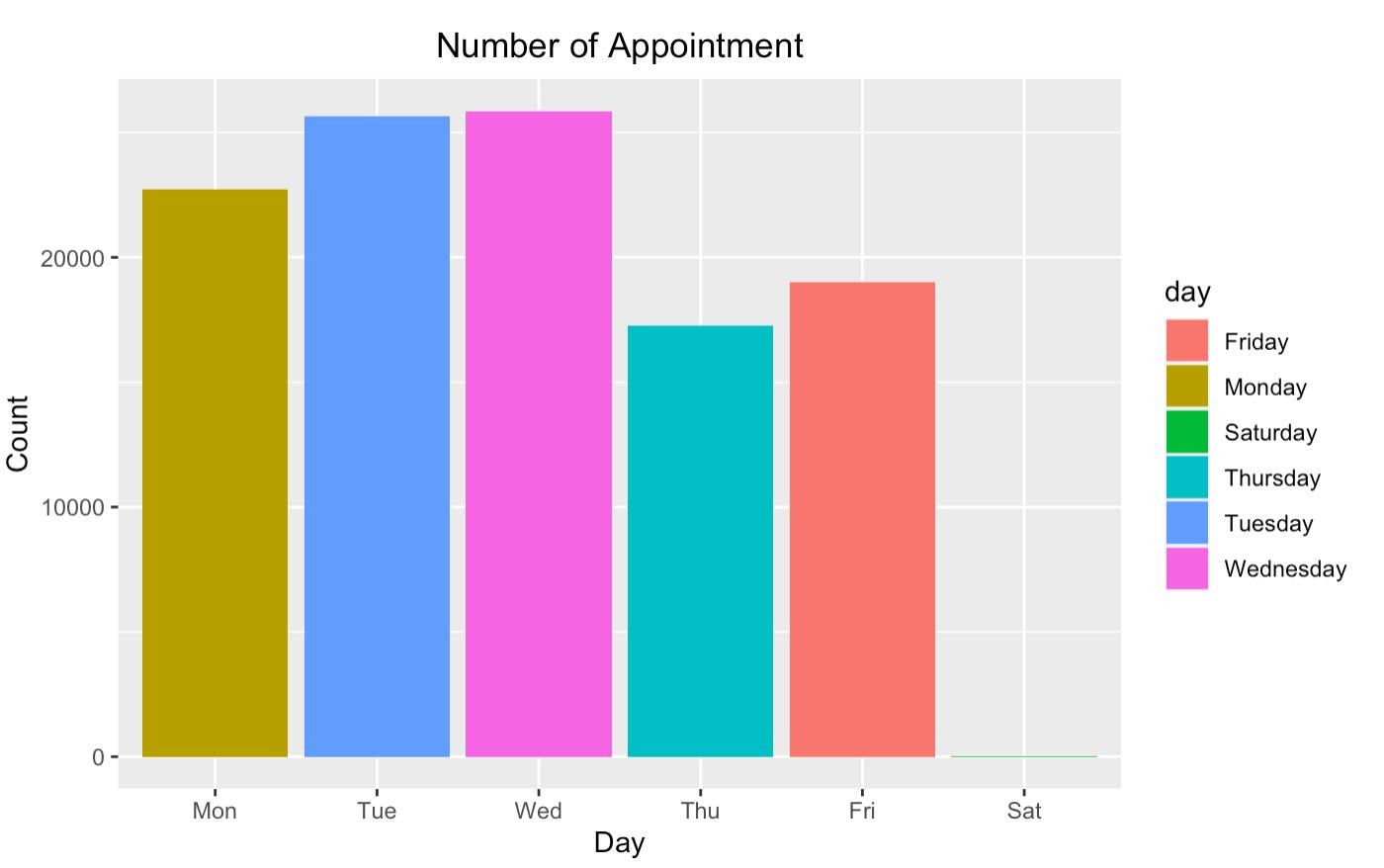


Exhibit 3:

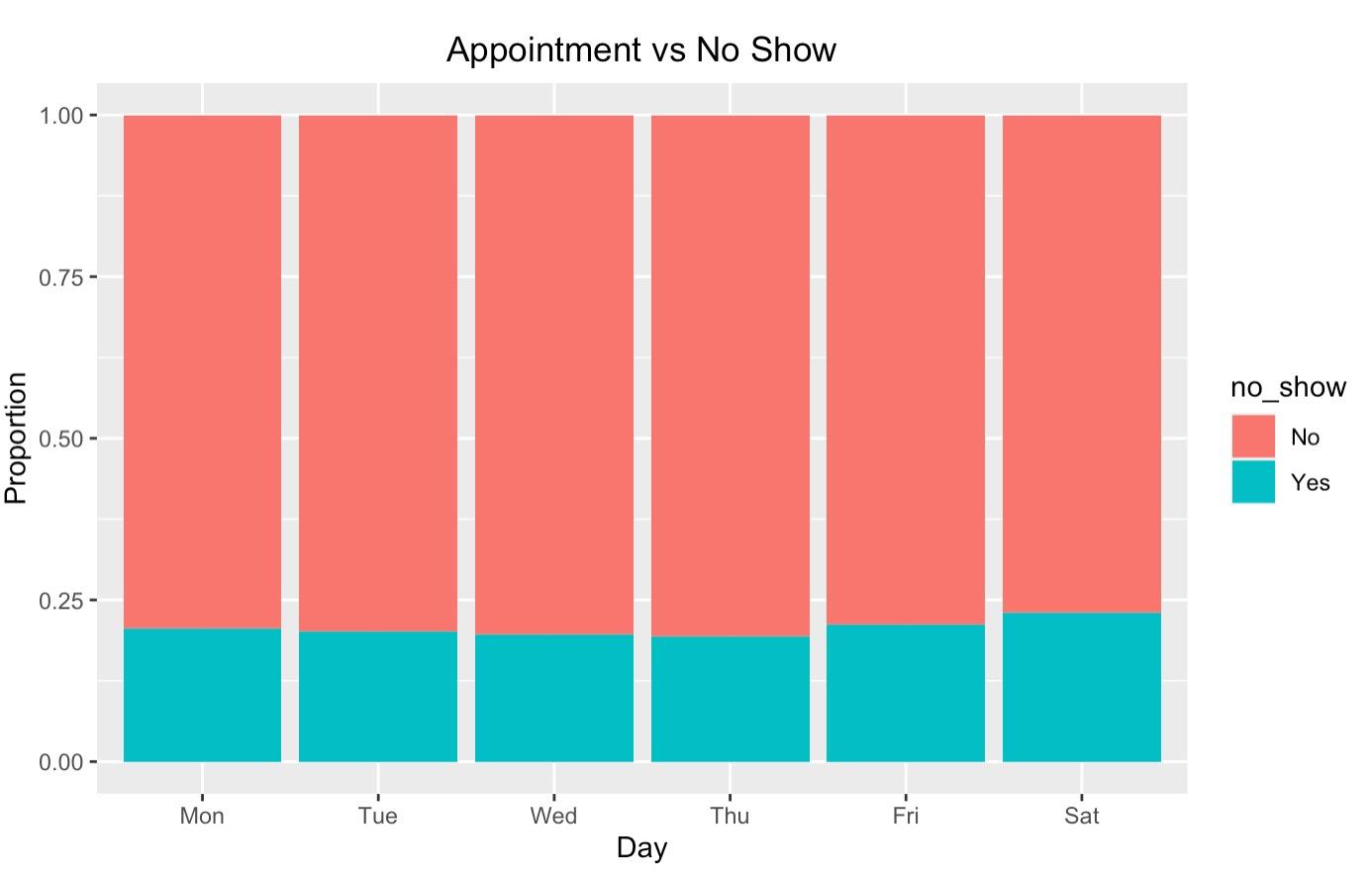


Exhibit 4

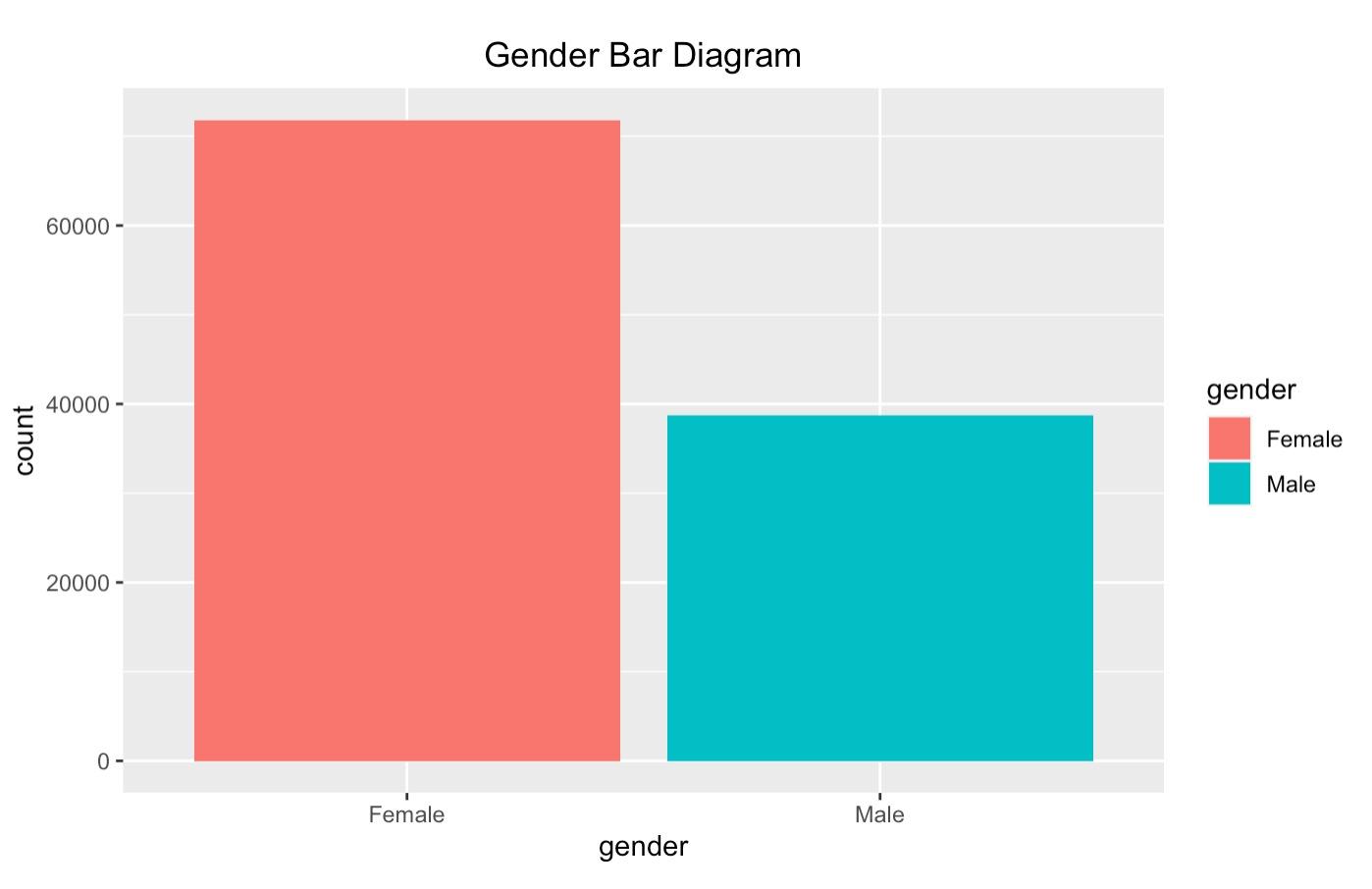
:

Exhibit 5:

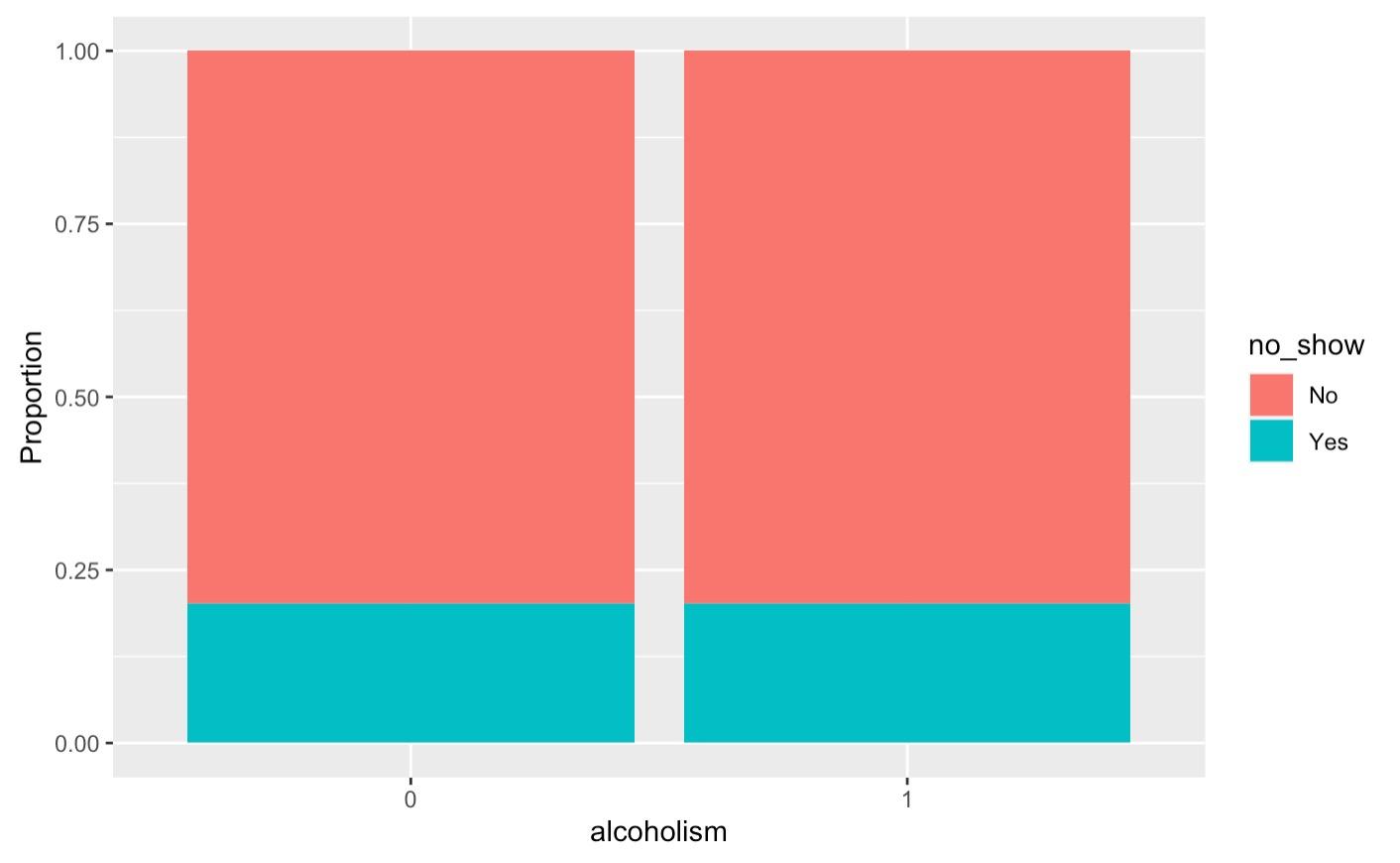


Exhibit 6:

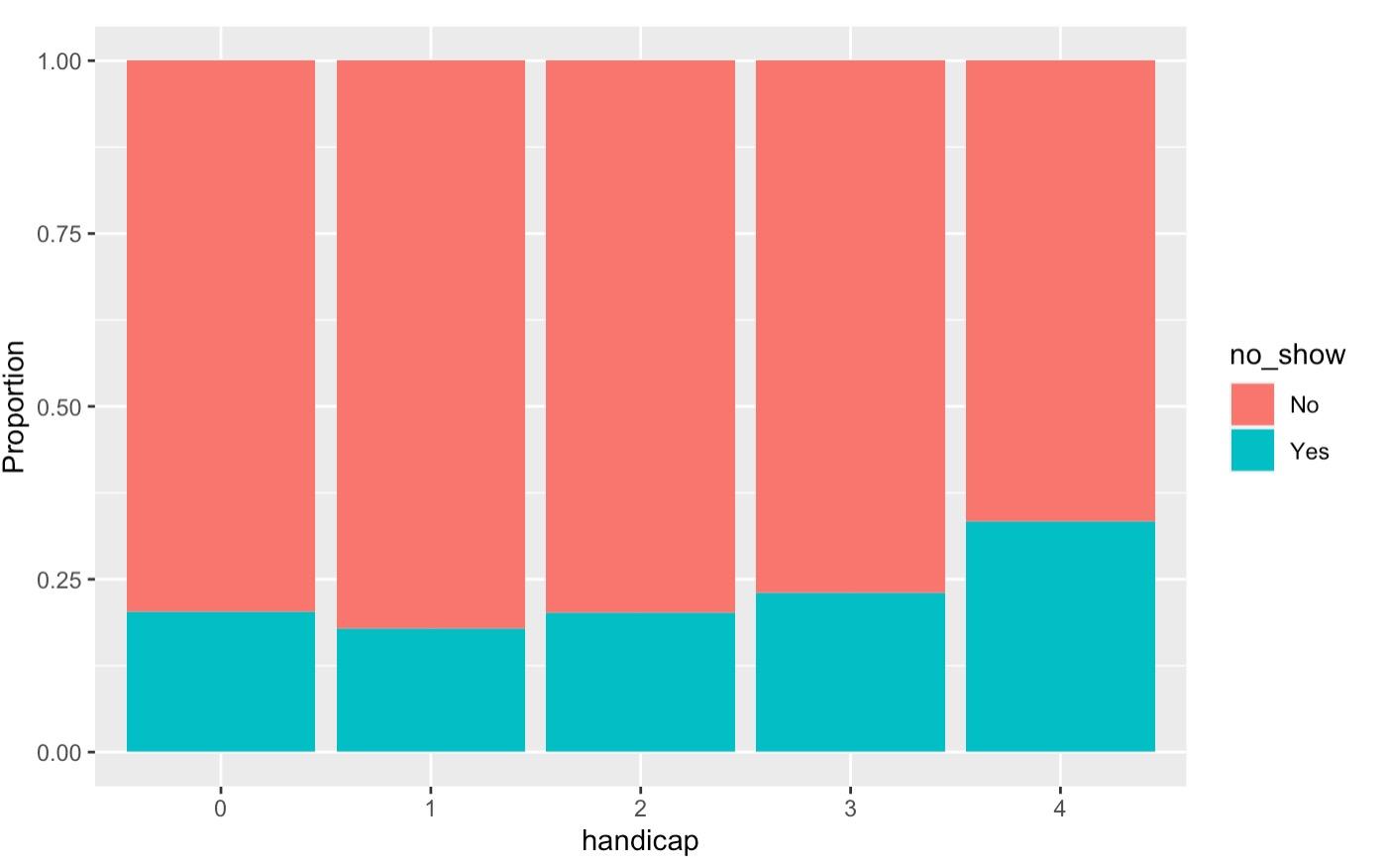


Exhibit 7:

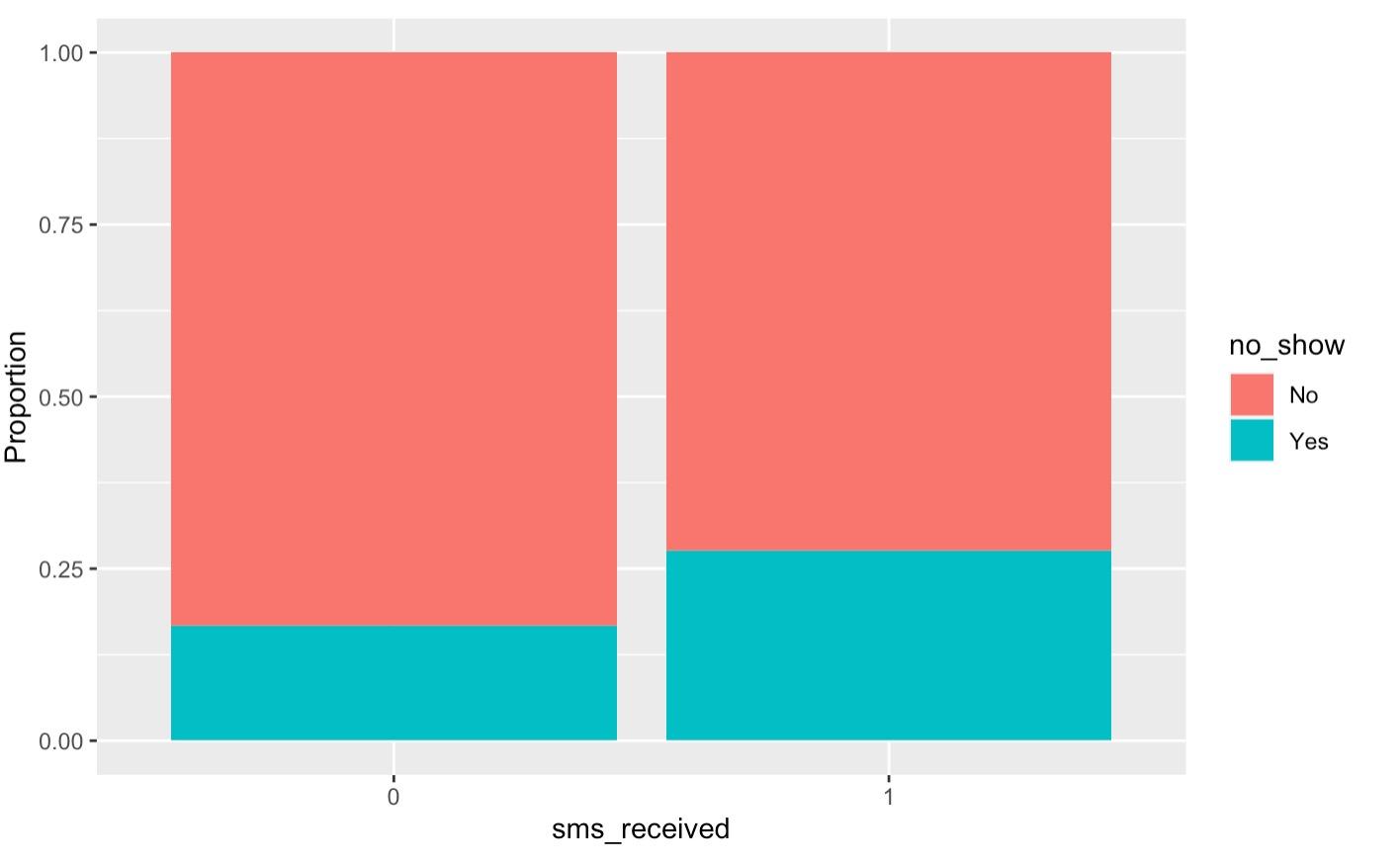


Exhibit 8:

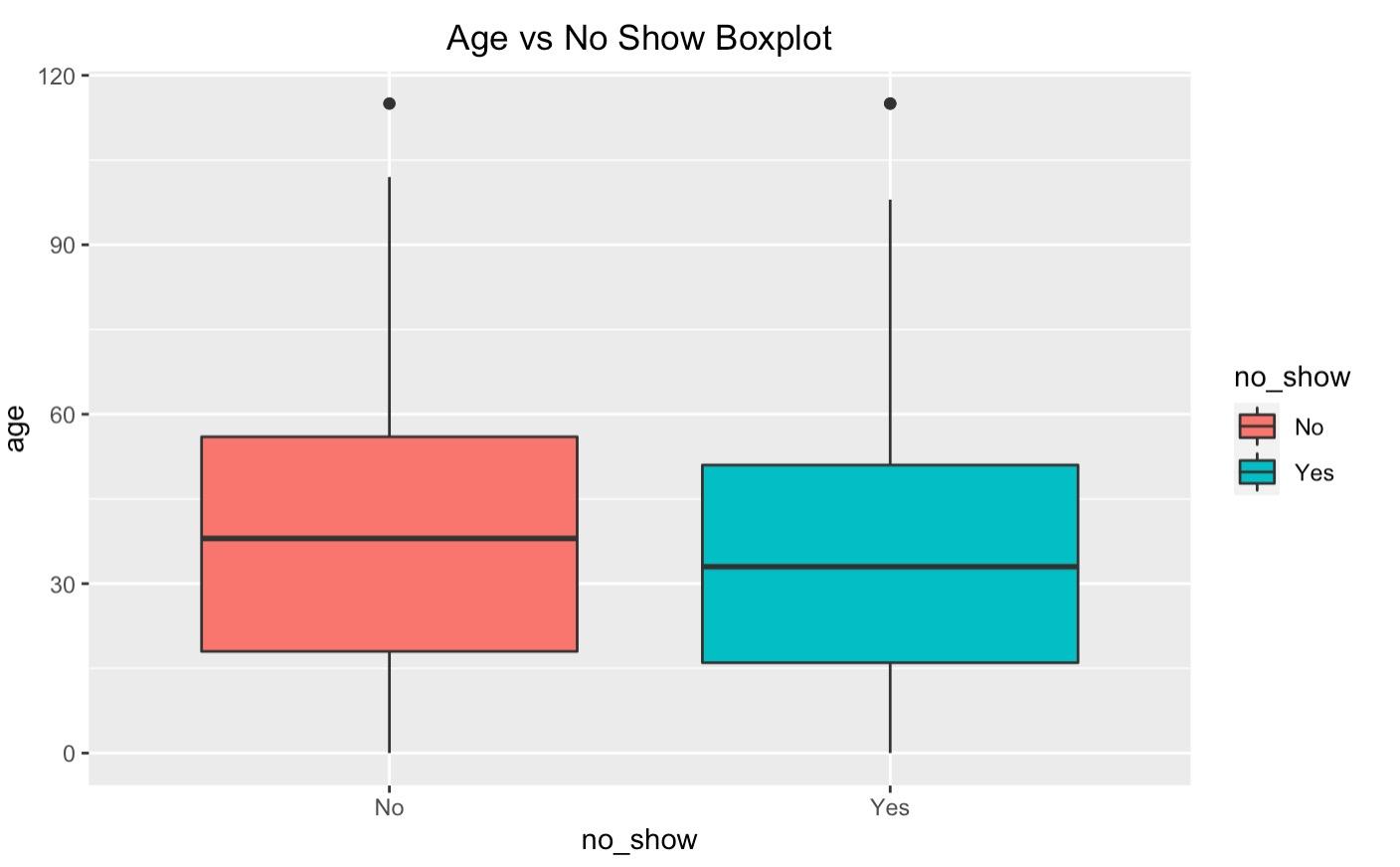


Exhibit 9:

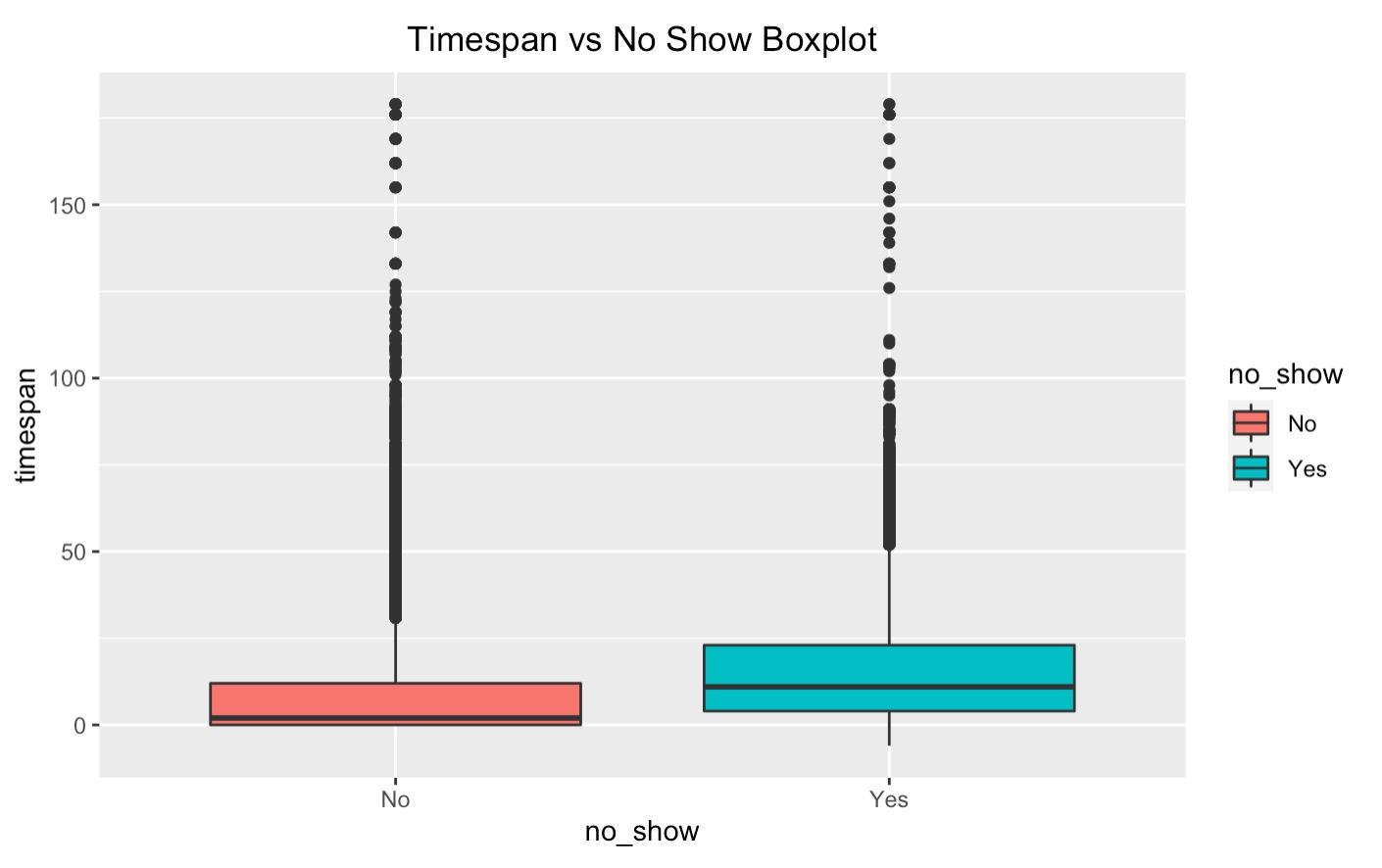
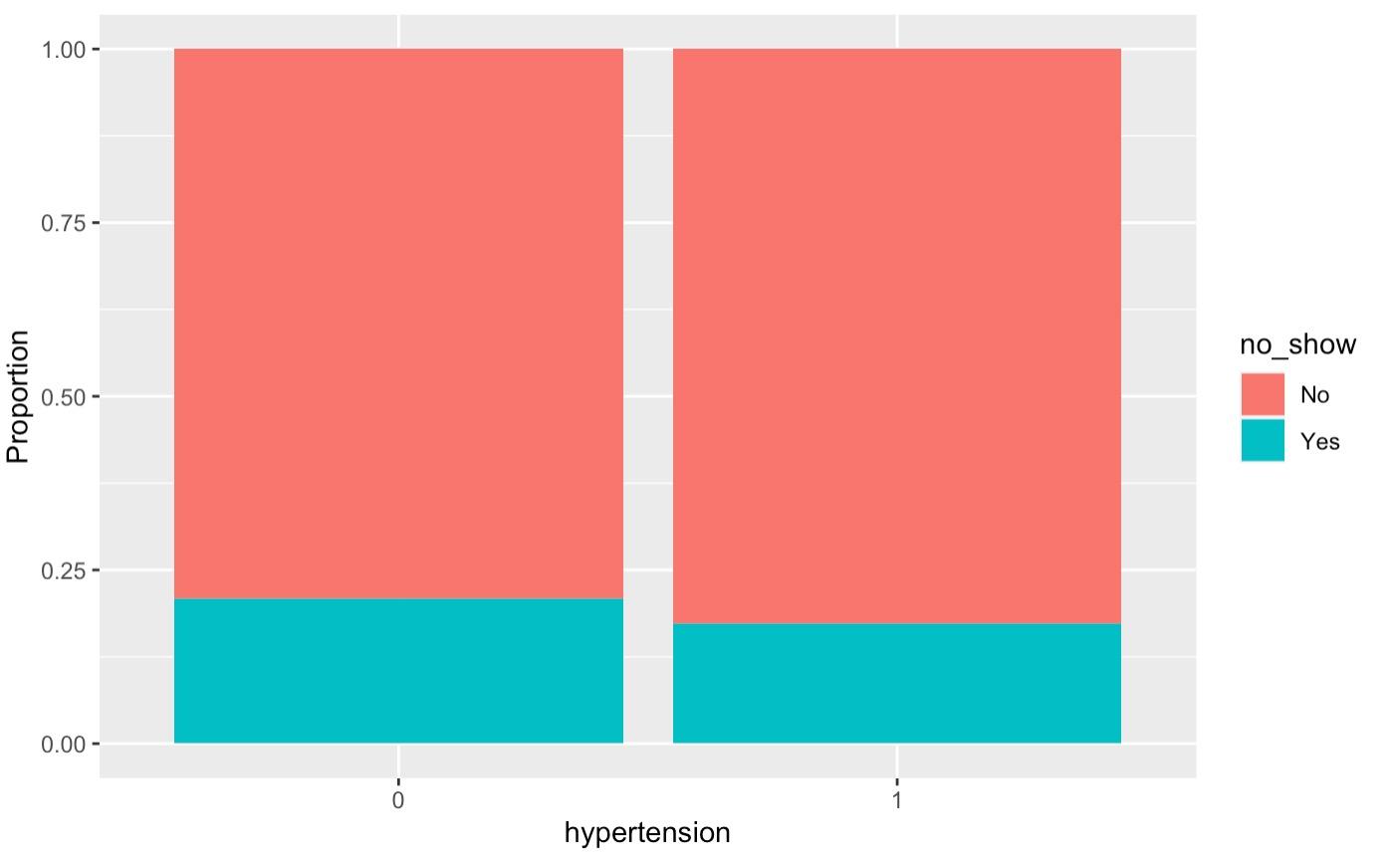


Exhibit 10:



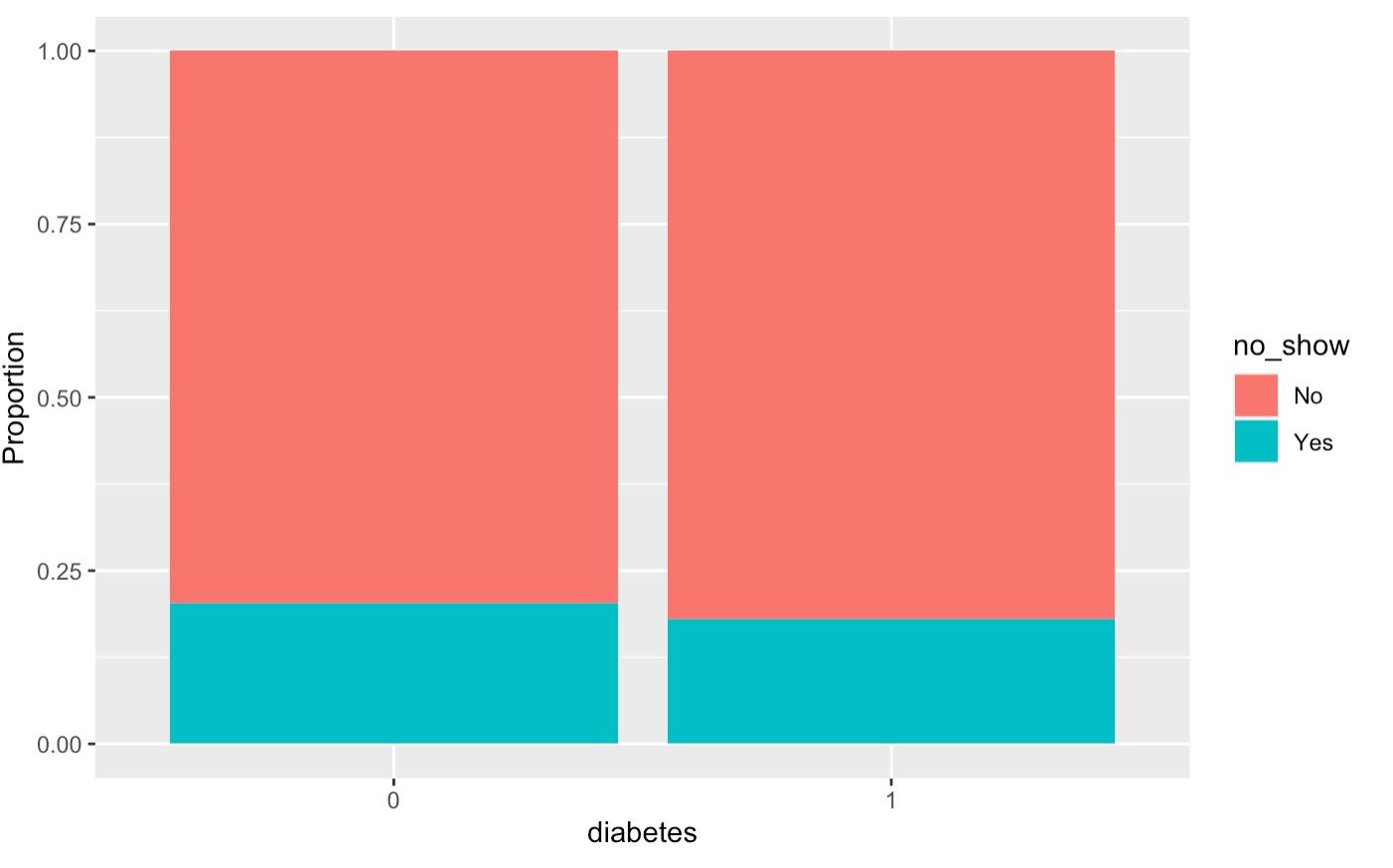


Exhibit 11:

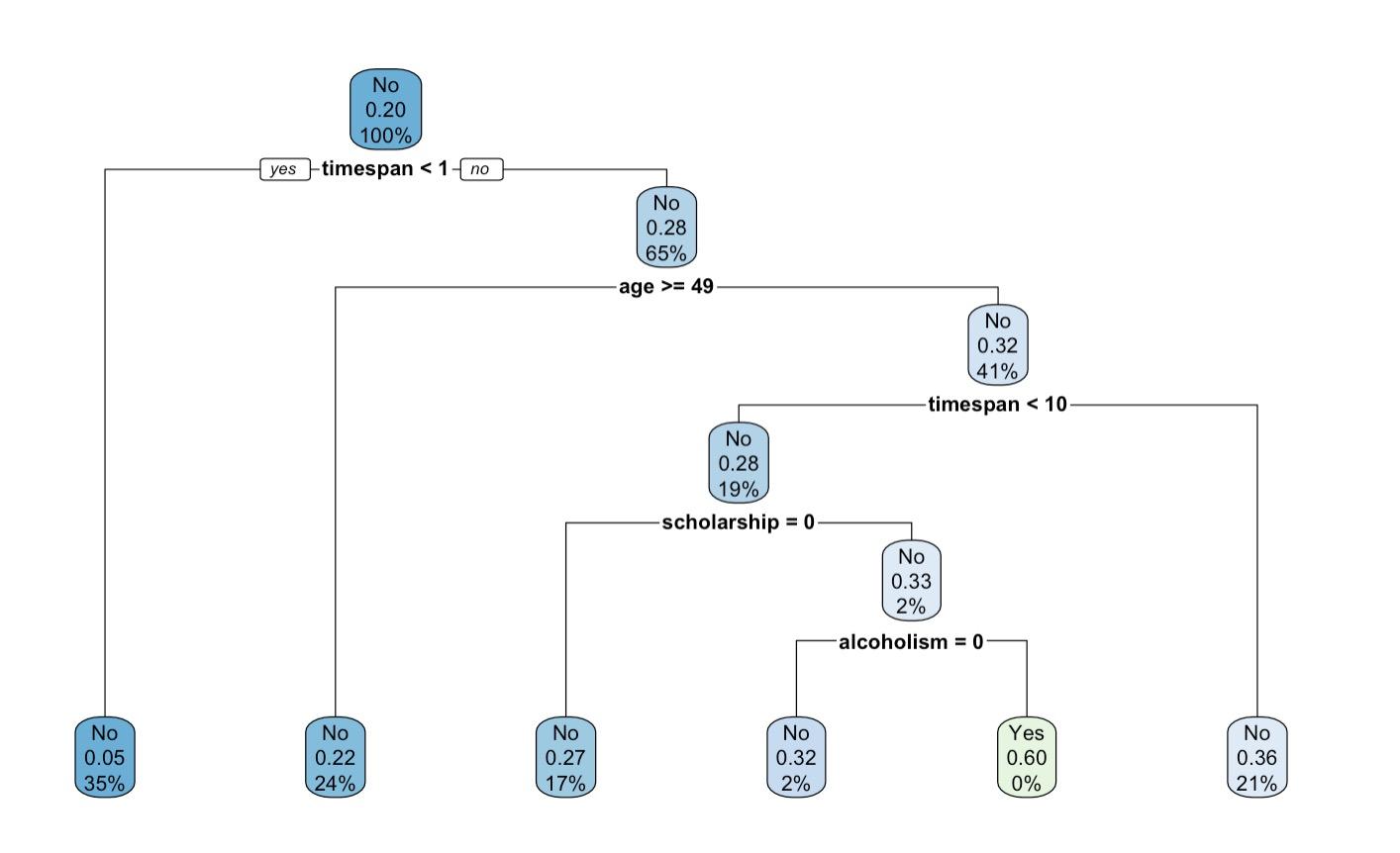


Exhibit 12

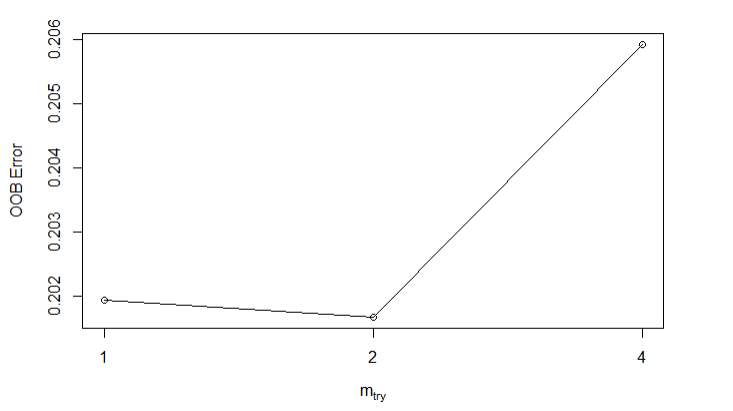


Exhibit 13

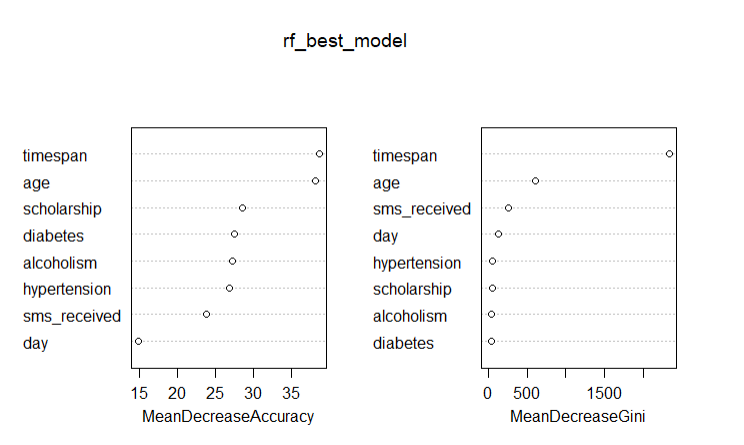


Exhibit 14

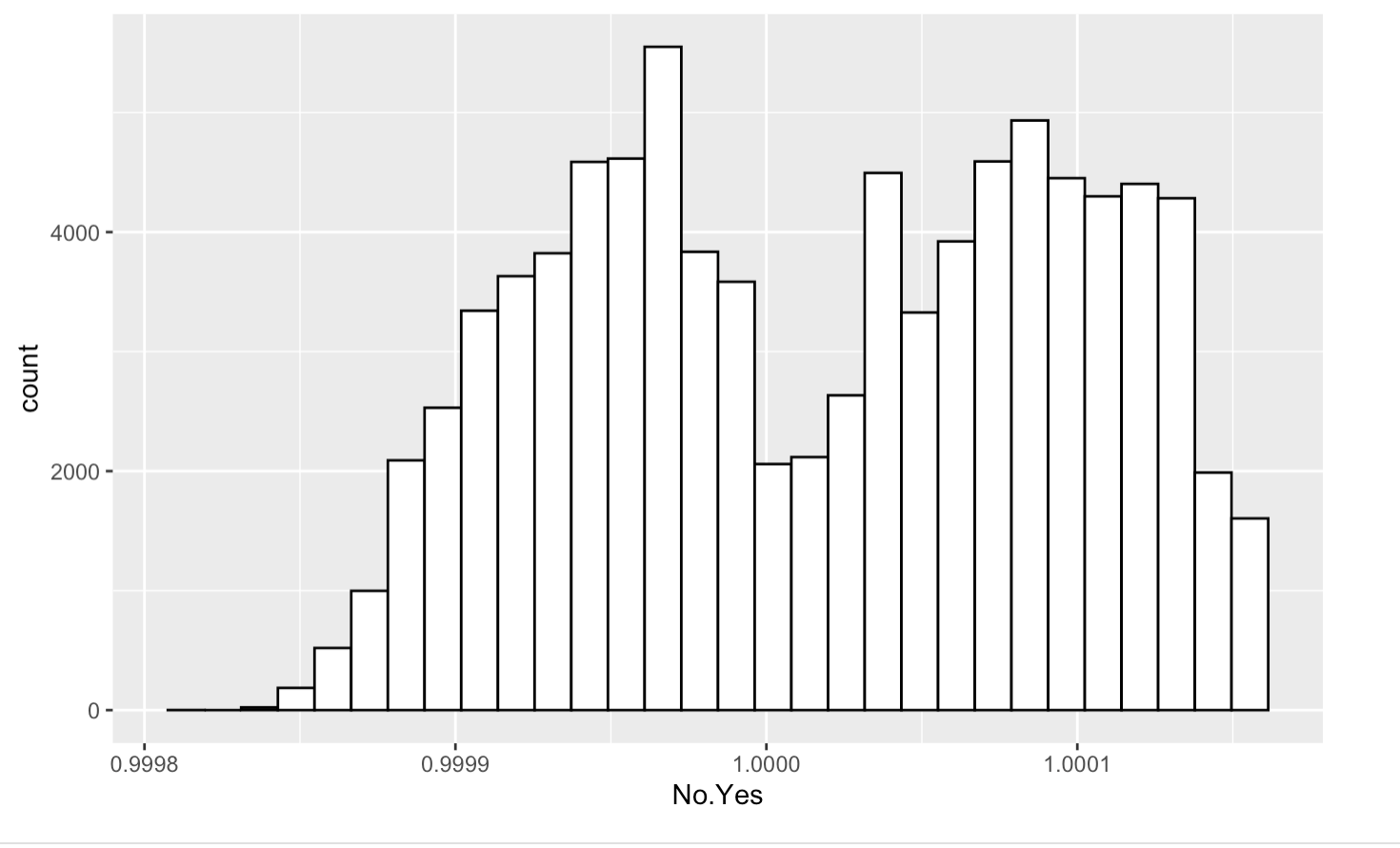


Exhibit 15

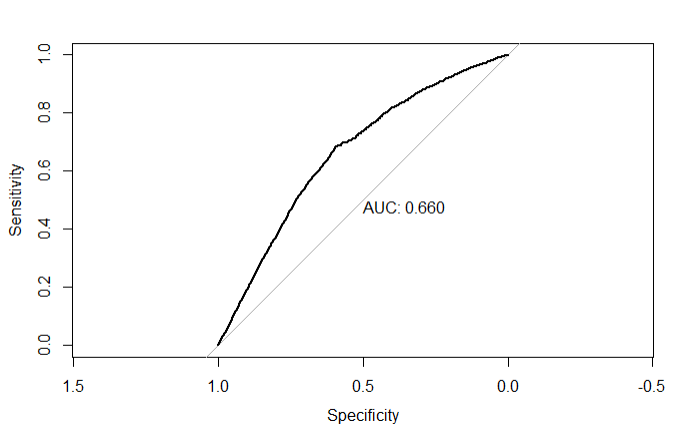


Exhibit 16

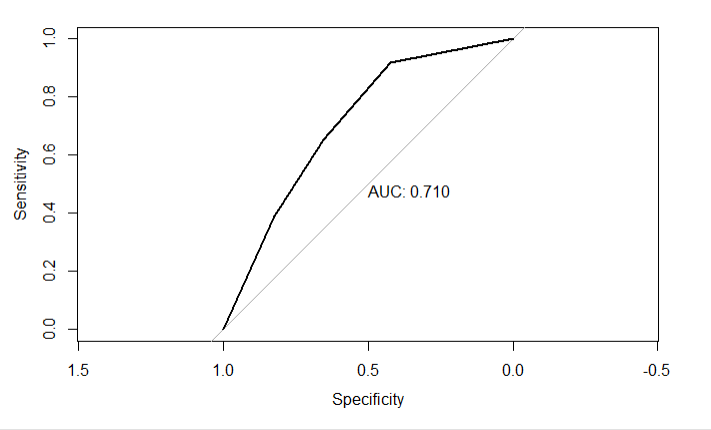


Exhibit 17

