ANALYSIS OF NO-SHOW APPOINTMENTS IN BRAZIL TASK 3

A. PROJECT OVERVIEW

A1. Research Question

Is there a correlation between patients receiving a reminder text for their appointment and if they show up to their appointment?

A2. Project Scope

The scope of this project will include analysis and visualizations from a Python application. This project will use patient appointment data as input. The output of this application will be visualizations and numbers that explore the correlation and causation between text reminders and no-shows. The scope of this project will not include other factors that may have a correlation or causation with patient no-shows. The scope will also be limited to data provided from healthcare appointments in Brazil, not data from other countries or other types of organizations.

A3. Solution Overview

This project will use Jupyter Notebook, a Python integrated development environment. Python is very commonly used to gather, clean, and analyze data since it has a variety of libraries perfectly suited for data analysis. Jupyter Notebook will be used specifically since the project can use well defined chunks of code with output directly underneath it, which helps to keep all outputs and visualizations organized. I will use descriptive data analysis techniques to find if there is a correlation between text reminders and no-show appointments. For example, I will employ bootstrapping and logistic regression to see if there is any statistical significance between receiving a text reminder or not receiving one. I will use Python and custom code to help in creating, cleaning, and analyzing the dataframe. Descriptive analytics is used to help determine trends in data, which is exactly what this project aims to do with determining the correlation between text reminders and patient no-shows.

B. Project Plan

B1. Project Plan

I executed the plan without change. All goals, objectives, and deliverables were completed as described below and in Task 2.

The goal of the project is to see if there is a correlation and a possible causation between receiving a text reminder and patient no-shows. The objectives are as follows:

- Determine if there is a correlation between text reminders and no-shows.
    - The deliverable for this objective is the findings of what type of correlation, if any, exists between the two. This will consist of a bar chart that compares the average rate of no-shows of when a patient receives a text reminder versus receiving no text reminder.
- Determine if text reminders have a causal relationship with no-shows.
    - The deliverable for this objective is to find the p-values from several different methods to see if text reminders cause a decrease in patient no-shows. The several different methods of calculating the p-value are as follows. One, an A/B test that provides a sampling distribution of 10000 simulations under the null hypothesis (receiving or not receiving a text reminder has no effect on no-shows) of that will be visualized as a bell curve with a red line to represent the actual observed difference between the average no-show rates. Second, a Z-test will be used to calculate the z-score and p-value based on the data. Third, logistic regression will be used and a model will be fit using the Noshow data and a summary of the model will be output.

B2. Project Planning Methodology

This project will use the Agile project methodology. This methodology consists of four major principles with those being: adaptive, goal-oriented, integrated, and learnable. This project will be adaptive and change as needed throughout the analysis process. This can include many things that may need to be flexible throughout the project, such as what libraries and techniques used in analysis, or changing how the findings are visualized. This project is incredibly goal-oriented, with the two clear goals set of exploring the correlation and causation between text reminders and no-shows. These goals will be met within their specific timeframes as well. While this project does not apply as much to the integrated principle since there is only one person working on it, the results and applications from this project can be applied on different projects, such as healthcare appointment data from other countries, or data from non-healthcare related appointments. The final principle, learnable, is very important to this project. This project is focused on learning about the relationship between reminders and no-shows, and how to further learn and improve this analysis throughout. This project is the

first step to many of diving in and exploring the different relationships between various factors and patient no-shows.

B3. Project Timeline and Milestones

The project timeline was very similar to the projected timeline, and all milestones were completed as described. The actual start and end dates for each milestone differed from the estimated timeline.

Milestone	Projected Start Date	Projected End Date	Duration
Establish requirements for analytics process	6/30/2023	6/30/2023	1 day
Acquire and clean data	6/30/2023	6/30/2023	1 day
Code analysis and visualizations	7/01/2023	7/02/2023	2 days
Test application	7/01/2023	7/02/2023	1 day
Code any revisions/final touches	7/03/2023	7/03/2023	1 day

C. Methodology

C1. Data Selection and Collection Process

This data was collected by downloading the csv file of the dataset off of Kaggle.com and then reading it into a dataframe in Python. This plan did not change when the data was collected. This dataset was selected since it provides data on appointment details from healthcare organizations in Brazil. No other data needed to be collected since this dataset had all the necessary information.

C2. Obstacles to Data Collection

There were no obstacles to collecting the data. Downloading the csv file from Kaggle.com went smoothly and was read into a dataframe without issue.

C3. Data Governance Issues

Since the data was public on Kaggle.com, this project is freely able to use it under the license listed in D1 as long as it is not for commercial use. This data does have the name and age of each patient, as well as some chronic health issues, but these columns are not used or included in this project. There are no major privacy, security, ethical, or legal compliances that this project needs to follow since the data is publicly available and is using columns that do not include any personal information.

C4. Advantages and Limitations of Data Set

The advantages of this data set are that all the necessary information for my analysis was already available, so no other data was needed. Another advantage is that the data was easily downloaded from a public source with no major changes needed. Very minor changes were made to the data to ensure for easier analysis. A major disadvantage would be the size of the data set and where the data is collected. The data set has around 111,000 rows, which is relatively small for the entire population of Brazil. More rows would lead to more accurate analysis that better reflects the entire country. Since the data only pertains to Brazil, it cannot be used to accurately show healthcare appointment trends around the world. Having more data from nations across the world would help to show if text reminders could be a useful tool for healthcare organizations everywhere.

D. Data Extraction and Preparation Process

The data extraction process was downloading the csv file from Kaggle.com and then using the Pandas library in Python to read the file into a dataframe. With the dataframe in Python, I first renamed the ‘No-show’ column to ‘Noshow’ for easier access, then dropped the few rows that had outliers in the ‘Age’ column. The next thing I did to prepare the data was replace the values in the ‘Noshow’ column. Initially, ‘no’ and ‘yes’ were used to indicate whether a patient arrived to the appointment or not, but I changed the values so it would be 0 instead of ‘no’ and 1 instead of ‘yes’. 0 represents a patient who did arrive to their appointment, while 1 represents a patient who did not show. The final thing done was drop the unnecessary columns, since they were not used or considered at all in the analysis. These columns all focused on preexisting conditions and if the patient was on a scholarship, which is Brazil’s version of assisted healthcare.

E. Data Analysis Process

E1. Data Analysis Methods

I will use descriptive data analysis techniques to find if there is a correlation between text reminders and no-show appointments. For example, I will employ bootstrapping and logistic regression to see if there is any statistical significance between receiving a text reminder or not receiving one. I will use Python and custom code to help in creating, cleaning, and analyzing the dataframe. Descriptive analytics is used to help determine trends in data, which is exactly what this project aims to do with determining the correlation between text reminders and patient no-shows.

E2. Advantages and Limitations of Tools and Techniques

The main tool used in this analysis is Python, which provides a great environment for data cleaning and analysis. Python has many different libraries which aide in this process, such as Pandas, Numpy, Matplotlib, and Statsmodels in order to create, manipulate, visualize, and model the data as needed. A disadvantage of Python is the runtime, as Python can run slower than other languages and applications. The two main techniques used were bootstrapping and logistic regression. Benefits of bootstrapping includes that it makes no assumptions about the underlying data and it can be widely applied to many different scenarios and datasets. Disadvantages include that it is incredibly computationally intensive, as it creates resampled datasets for each iteration, and in my analysis there are 10,000 iterations. The quality of the estimates also depends on the original data, so any imbalances or quality issues can be reflected in the bootstrapped estimates. With logistic regression, it has the advantages of being easily interpreted and is computationally efficient. Disadvantages include that it assumes a linear relationship between the variables and can have overfitting issues.

E3. Analytical Method Applications

This project applied the analytical methods mentioned above as follows:

- After cleaning the data, I found the average no-show rates for the whole dataset and the subsets of patients who received a text and those who did not
- I calculated the observed difference between the rates of those who received a text (treatment group) and those who did not (control group)
- Find the total number of rows of both groups, the treatment group and control group
- Coded a for loop that found the sampling distribution of no-shows under the null hypothesis for 10,000 different resampled datasets
- Plotted the sampling distribution bell curve and the actual observed difference between the treatment and control groups
- Created two new columns, ‘intercept’ and ‘ab_page’ for logistic regression model

- Fit logistic regression model to ‘Noshow’ column and two new ones
- Output the summary of model to see the outcomes of model

Several assumptions were made when employing both techniques. The first assumption was that the null hypothesis was the average rate of no-shows for the whole dataset, which was used when creating the resampled data in the bootstrapping technique. Another assumption made was that the data was balanced and fair, so no imbalances or quality issues would arise in the distribution. When using the logistic regression technique, it’s most appropriate on binary outcomes, so I ensured that the outcomes were either a patient arriving to their appointment or not showing up. The final assumption I made with the regression model is that no overfitting occurred.

F. Project Success

F1. Statistical Significance of Analysis

To determine the statistical significance of this project, I calculated the p-values from the bootstrapping and logistic regression techniques. The null hypothesis states that receiving a text reminder made no significant difference on no-show rates. If the p-values were less than 0.05, I could reject the null hypothesis and say that text reminders do influence patient no-shows. With both the bootstrapping and logistic regression, I calculated a p-value of 0.00. Since these values are less than 0.05, I can reject the null hypothesis and conclude that text reminders do affect patient no-shows.

F2. Practical Significance of Analysis

The practical significance of this solution is that text reminders have a direct cause on a patient showing up to their appointment. If sending a reminder to every patient before their appointment will decrease the number of no-shows, this can help healthcare organizations to combat the increase in patient no-shows. This will help providers know whether implementing a text reminder system can result in saved time and money for their organizations. In this case, since there is a positive correlation between receiving a text reminder and patient no-shows, it helps healthcare organizations know that implementing text reminders will not help combat no-show rates.

F3. Overall Success and Effectiveness

Overall, I would conclude that this analysis was a moderate success. All of the criteria were met, since a correlation was found and p-values were calculated. However, the correlation between text reminders and patient no-shows was positive instead of

negative. This can help organizations make more informed decisions on what policies to implement in their practices.

G. Key Takeaways

G1. Conclusions

This project was aimed at finding the possible correlation and causation between text reminders and patient no-shows in Brazil. The project was able to determine a positive correlation between the two variables, indicating that patients who receive a text reminder are more likely to not show up to their appointment. This project was also able to calculate multiple p-values that were less than 0.05 to show that text reminders are a cause of patient no-shows.

G2. Effective Storytelling

The visual representations used in this project were the best methods since they clearly demonstrate the correlation and causation between the two variables. The bar chart clearly shows the average no-show rates of the group that received a text and the group who did not receive a text. The bell curve visualization of the bootstrapping distribution with a line to show the actual observed difference helps to show how the observed difference is much different than the difference under the null hypothesis. Both visualizations effectively demonstrate what they are intended to.

G3. Courses of Action

The first course of action would be to further explore what patients are more likely to receive text reminders. If there is a clear correlation between who receives a text reminder and another variable, such as age, that can help to see if there are any other factors that contribute to the positive correlation between text reminders and no-shows. The second course of action would be to collect data on healthcare appointments from other countries as well. This dataset only used information from Brazil, so integrating data from other places could help to see any common no-show trends worldwide. That way, healthcare organizations across the world could benefit from the results of the analysis.

H. Panopto Recording

Appendices and Sources

Dataset: https://www.kaggle.com/datasets/joniarroba/noshowappointments