AFFECT OF TEXT MESSAGES ON MEDICAL NO SHOWS

Natasha E Estrada

Western Governors University

**Table of Contents**

[Project Overview 3](#_Toc55905041)

[A. Summary 3](#_Toc55905042)

[Project Plan 3](#_Toc55905043)

[B. Summary 3](#_Toc55905044)

[Methodology 3](#_Toc55905045)

[C. Data Selection 3](#_Toc55905046)

[C1. Data Set Advantages and Limitations 4](#_Toc55905047)

[D. Data Extraction/Preparation Processes – Tools, Techniques, Suitability 4](#_Toc55905048)

[E. Data Analysis Process 4](#_Toc55905049)

[E1. Analysis Methods 4](#_Toc55905050)

[E2. Analysis Tools/Techniques - Advantages and Limitations 4](#_Toc55905051)

[E3. Step-by-Step Explanation for E1 4](#_Toc55905052)

[Results 4](#_Toc55905053)

[F. Project Success 4](#_Toc55905054)

[F1. Statistical Significance 4](#_Toc55905055)

[F2. Practical Significance 4](#_Toc55905056)

[F3. Overall Success and Effectiveness 4](#_Toc55905057)

[G. Key Takeaways 4](#_Toc55905058)

[G1. Conclusions 4](#_Toc55905059)

[G2. Justify Visual Communications Tools 4](#_Toc55905060)

[G3. Findings-based Recommendations 4](#_Toc55905061)

[H. Panopto Video Link 4](#_Toc55905062)

[Appendices 4](#_Toc55905063)

[Sources 4](#_Toc55905064)

# Project Overview

# Project Highlights

# *Research Question*

# Last month, I conducted a project to investigate the impact of text messages on reducing missed medical appointments. The business problem that motivated the project was the significant financial and clinical impact of missed appointments, known as "no-shows". By analyzing a dataset of patient appointments, I aimed to determine if text message reminders could be an effective solution to reduce the number of no-shows.

# *Context and Background*

# The significance of this research lies in its potential to improve healthcare delivery and increase clinic efficiency. Non-attendance at appointments is a challenge for both patients and providers, leading to delayed care, increased costs, and lost time for providers. The research was inspired by the author's experience as a licensed vocational nurse in a public health clinic in Los Angeles where appointments were often booked out three months in advance and non-attendance was a frequent issue. The goal was to prevent missed appointments and inform the extent to which appointments could be double-booked.

# *Project Scope*

The scope of this project includes data wrangling and analysis of data from a health care organization to understand the effect of text message reminders on reducing “no-shows” for appointments. It does not include developing or implementing the text messaging system itself.

# *Solution Overview Tools and Methodologies*

I did most of the analysis work inside Jupyter Lab using Python. I prefer Jupyter Lab as it allows me to develop my code incrementally and to test as I go. I can add markdown cells to add my insights and document my process. I can also create all the visualizations in line with my code.

I used the pandas library to read my csv file into a dataframe and to inspect the dataset using a variety of visual and programmatic assessment techniques in order to assess what parts of the dataset needed to be cleaned prior to analysis.  
  
I used the matplotlib and seaborn libraries to create the visualizations used during the exploratory data analysis process. I used the numpy and statsmodel library to run the binary logistic regression calculations.

# Project Execution

***Project Plan***

The project plan remained mostly unchanged during the execution phase from the plan set forth in Task 2. The one change was to switch from using Tableau for creating the visualizations to using the built in visualization tools inside JupyterLab like Matplotlib and Seaborn. This created a streamlined analysis notebook and prevented errors that would occur by potentially having two different versions of the dataset opened at one time.

***Project Planning Methodology***

The project planning methodology was followed as per the plan and no deviations were made.

I utilized the waterfall method as the project planning methodology. This methodology involves a sequential and linear approach to the project, with each phase building upon the previous one.

The specific steps in my medical appointment project using the waterfall method were:

1. Requirements: In this phase, I defined the project objectives and created the necessary project documentation. This included clearly articulating the problem we were trying to solve (reducing "no-shows" for medical appointments), as well as defining the specific goals and deliverables for the project.
2. Design: In this phase, I developed a plan for data collection and analysis, as well as creating any necessary software or systems that would be used in the project. This included identifying the data sources we would use, deciding on the tools and techniques for data analysis, and defining the specific methods for processing and storing the data.
3. Implementation: In this phase, I collected the data, analyzed it, and created a report summarizing the findings. This involved extracting the data from the various sources, cleaning and processing the data, and using statistical methods to analyze the data and draw conclusions.
4. Verification: In this phase, I reviewed the report and findings to ensure accuracy and validity. This involved checking the data processing steps, reviewing the statistical methods used, and validating the conclusions drawn from the data.

By following the waterfall methodology, I was able to complete the medical appointment project in an organized and systematic manner, ensuring that each phase was properly executed and that the final report accurately reflected the findings of the project.

***Project Timeline and Milestones***

The project timeline and milestones were met, but the project was completed 3 weeks ahead of schedule.

# Methodology

# Data Collection Process

The actual data selection process was the same as the planned collection process, which was downloading the dataset from Kaggle.

However, there were some unexpected obstacles during the data collection process. One of the main challenges I faced was dealing with outliers in the data. I had to carefully handle these outliers to ensure that they didn't have a negative impact on my analysis. Despite these obstacles, I was able to successfully collect the necessary data for my project.

There were no data governance issues as the data was already anonymized and scrubbed of patient’s personal data prior to being uploaded to the Kaggle website.

# C1. Advantages and Limitations of Data Set

***Advantages:***

*Large sample size:* The dataset contains over 100,000 appointments, providing a robust sample for analysis.

*Relevant variables:* The dataset included relevant variables such as patient age, gender, and the presence of medical conditions that could impact appointment attendance. These variables allowed me to analyze the impact of text message reminders on reducing no-shows.

*Clear definition of no-show:* The dataset clearly defines a no-show as a patient who did not attend their scheduled appointment.

The dataset was available for free on Kaggle, making it easy for me to access.

***Limitations:***

*Outliers:* The data contained outliers which could have affected the results of the analysis.

*Missing data:* Some variables, such as the patient's health status and income, were not included in the dataset, which could have affected the results.

The dataset did not include information on why patients missed appointments, so I was unable to determine the root causes of missed appointments.

The dataset was limited to appointments made in a single city, so the results may not be generalizable to other populations. This city is also located outside of the US so may not be demographically similar enough to LA to make an inference.

The dataset included only appointments made in a single health care organization, so the results may not be representative of the entire population.

No control over phone calls: The dataset does not indicate whether patients who did not receive a text message reminder also received a phone call, so it is impossible to determine the true impact of text message reminders on reducing no-shows.

No information on same-day appointments: The dataset does not indicate how same-day appointments were handled, so it is not possible to determine the impact of text message reminders on this group of patients.

.

# Data Extraction and Preparation Processes

As the first step, I imported the dataset from Kaggle and opened it in pandas. I then inspected the first 5 rows to ensure the dataframe was created correctly. To gain a better understanding of the data, I inspected the columns and the shape of the dataframe.

Graphical user interface, application

Description automatically generated

I also summarized the statistics of the columns and checked for unique values. Since the AppointmentID was going to be our primary key, I checked to see if there were any duplicate values and if there were any null values in the dataset.

Graphical user interface, application

Description automatically generated

The ScheduledDay and AppointmentDay columns were in object format, so I converted them to datetime format using the "to\_datetime" pandas method. I visually inspected the changes made to confirm that the conversion was successful and verified that the datatype had been changed to datetime. Because I had to repeat this conversion whenever I reopened the file I created a function for this task which could be reused in my code.

Text

Description automatically generated

During the data cleaning process, I discovered that there were outliers in the ages column, with values of less than zero and over 100. To handle these outliers, I used pandas to find the rows with age values less than zero and greater than or equal to 115. I then calculated the mean of the age column and used the "replace" method to replace the outliers with the mean\_age.

Text

Description automatically generated

Chart, histogram

Description automatically generated

I also used the "replace" method to change the No-show values from yes/No to 1 and 0. To make the dataset easier to work with, I dropped any unused columns using the "drop" method and renamed a few columns using the "rename" method.

Graphical user interface, text, application

Description automatically generated

Finally, I created a new column that had the difference between the scheduled day and the appointment day. I then used the "replace" method to replace any values in that column that were less than zero with zero. This had to be repeated when I opened the file so I had also added this into my datetime conversion function.

The cleaned dataframe was saved as no\_show\_cleaned.csv.

These processes were appropriate for the data because they helped me to clean and prepare the data for analysis. By using pandas methods, I was able to automate the data cleaning process and make the data easier to work with. By using reusable functions I was able to avoid having to repeat any code and kept my code dry (Don’t repeat yourself)

# Data Analysis Process

# E1. Data Analysis Methods

Data preparation: The data was preprocessed to handle missing values, outliers, and to create a binary variable (TextReceived) from the original text message reminder variable.

Descriptive statistics: The data was summarized using descriptive statistics such as mean, standard deviation, and count of observations for each variable.

Logistic Regression: The binary response variable (No Show) was modeled against the binary predictor (TextReceived) using a logistic regression model. The logistic regression model was fit using maximum likelihood estimation (MLE).

# E2. Advantages and Limitations of Tools/Techniques

Descriptive statistics: Descriptive statistics provide a summary of the data, and they are straightforward to calculate and interpret. The limitation of descriptive statistics is that they do not provide information on the relationship between variables.

Logistic Regression: Logistic regression is a widely used method for modeling binary response variables. It provides estimates of the coefficients for the predictor variables and allows for interpretation of the effect of the predictors on the response. The disadvantage of logistic regression is that it assumes a linear relationship between the log odds of the response and the predictors, which may not always be the case.

.

# E3. Application of Analytical Methods

After the data was cleaned, I performed an exploration of the data to get a feel for the dataset using tools like .info() and .describe to get a set of summary statistics of my dataset.

Graphical user interface, text, application

Description automatically generated

I then created a variety of visualizations to get a general overview of patient demographics before starting to visualize the factors that affect no shows. I discovered approximately 20% of appointments were no shows meaning the patient did not show up for the appointment.   
I then investigated whether the time between scheduling the appointment was a factor. While there did seem to be an increase in no shows the further out appointments were scheduled this effect actually tapered off at about the 60-90 day mark. The lowest no-show rate was seen with appointments made 0-7 days before the appointment.

Chart, bar chart

Description automatically generated

I then created a function to create a count plot in seaborn that would enable me to use the same chart for various pieces of the dataset.

Text

Description automatically generated  
  
I was then able to call this function so I could visualize whether the patient showed up to their appointment and whether the patient received a text message or not.

Chart, bar chart

Description automatically generated

I noticed when I was calculating the days between scheduling day and appointment day that many of the entries had a value of 0 or -1 days in between scheduling an appointment and attending an appointment. The -1 values were because all the values in the appointment day column was for midnight the day of the appointment while the scheduling day column had the exact date time when it was scheduled.

Because there appeared to be many same day and next day appointments it made me wonder if those patients even had time to receive a text message since typically with most medical practices texts are sent out 24hrs prior to an appointment. This means that patients that made a same day appointment probably are skewing the results for attending their appointment without a text message. Also, patients who make a same day appointment are probably very motivated to attend their appointment.

Because of this I created a another dataframe excluding rows where days between scheduling day and appointment day are 0 and then repeated the same visualizations. I also created a dataframe that contained only the same day appointment data.

I also calculated the proportion of the original dataset each subdivision made up. Same day appointments made up 40% of the dataset so this had the potential to skew the analysis as to whether text messages are helpful for reducing no shows if my assumption is correct that same day appointments do not receive text message reminders because it's already past the point where they would have been sent and that patients who have made same day appointments are motivated to attend their appointment and have also had contact with a staff member at the clinic which would replace the purpose of a text.

Chart, bar chart

Description automatically generated

Once I subdivided the data set into same day and non-same day appointments, I could see that none of the appointments booked same day received text messages which confirmed my suspicion that same day appointments do not receive text messages because the text messages are sent out some time prior to the day of the appointment.

Because same day appointments make up a very high volume of the total number of appointments and they have a very low rate of no shows their inclusion in this analysis would skew the results and obscure the insights I was trying to make. I felt confident at this point to exclude same day appointments from my analysis.

Chart, bar chart

Description automatically generated

When the dataframe that excludes same day appointments is recharted I can see that there is a relationship between a patient receiving a text message and them being less likely to not show up for their appointment so I felt confident moving ahead with the rest of my analysis.

I then reran the summary statistics on my reduced dataset.

Table

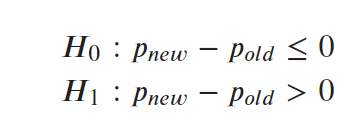
Description automatically generated

We can see from the summary statistics that the average age was 38.39, ~53% of patients received a text prior to their appointment and there was an overall no show rate of ~29% The average number of days between appointments was 15.79 days with a minimum of 1 day and a maximum of 178 days.

The average age is a little higher than the average age in the entire dataset which was 37.09. The text received rate was higher than the overall dataset which had only 32% of patients receiving texts. The no show rate is also higher than the 20% from the overall dataset. The average days in between is higher than the average in the overall dataset which was 9.5. This does suggest that the inclusion of same day appointments was skewing some of the data regarding these three metrics while the average age stays about the same.

I then ran my binary logistic regression on this dataframe.  
  
The null hypothesis is that there is no relationship between receiving a text message reminder and reducing "no-shows" for appointments.

The alternative hypothesis is that there is a relationship between receiving a text message reminder and reducing "no-shows" for appointments.



It is assumed that there is no relationship between receiving a text message reminder and reducing no shows if the Type 1 error rate is 5% or higher.

I ran two variations of logistic regression. The first test was a binary logistic regression model with only the "TextReceived" variable.

The second test was a binary logistic regression model with an added constant (intercept) and both "const" and "TextReceived" variables. The second model is more comprehensive and allows for the analysis of the effect of both the constant and the "TextReceived" variable on the probability of the response variable, "No Show".

Table

Description automatically generated

The coefficient of the independent variable (TextReceived) is -0.9657, with a standard error of 0.012. This means that the odds of a patient not showing up to their appointment decreases by 0.38 times if they received a text message reminder (the odds ratio is calculated as 1 / (e^-0.9657)).

The z-score of -81.289 for the coefficient, and the resulting p-value of 0.000, suggest that the relationship between receiving a text message reminder and not showing up to the appointment is statistically significant (p < 0.05), and that it is unlikely that this relationship is due to chance.

Therefore, you can conclude that there is evidence to support the alternative hypothesis, which states that there is a relationship between receiving a text message reminder and reducing "no-shows" for appointments.

Table

Description automatically generated

The results from this second set of code show the effect of the binary predictor (TextReceived) on the binary response (No Show). The results indicate that:

The coefficients of the predictor (TextReceived) and the constant are significant in the logistic regression model, with a p-value of less than 0.05. This suggests that the predictor has a significant effect on the response.

The coefficient for the predictor (TextReceived) is -0.1555, with a standard error of 0.017. This suggests that for every unit increase in the value of TextReceived, the log odds of No Show decrease by 0.1555.

The odds ratio for TextReceived is calculated by exponentiating the coefficient, and it is 0.8575. This means that, holding all other variables constant, the odds of No Show are 0.8575 times lower for individuals who received a text message reminder compared to those who did not receive a reminder.

Based on the second logistic regression test, there is a significant effect of text messages on no-shows. The coefficient for "TextReceived" is -0.1555, with a p-value of 0.000, which suggests that receiving text messages is negatively associated with the likelihood of a no-show. This means that patients who receive text messages are less likely to miss their appointments compared to those who do not receive text messages.

# Results

# Project Success

# F1. Statistical Significance

The evaluation of the statistical significance of the analysis is based on the p-values and confidence intervals calculated using logistic regression. The p-values indicate the probability of observing the results by chance if the null hypothesis is true, and a p-value of less than 0.05 is considered statistically significant. The confidence intervals provide an estimate of the range of values that the population parameter is likely to lie in. If the confidence interval does not contain 0, it is an indication of statistical significance.

The logistic regression model was fitted to the data to examine the effect of the predictor (TextReceived) on the response (No Show) and to calculate the odds ratio. The results of the logistic regression analysis indicate that the predictor has a significant effect on the response with a p-value of less than 0.05 and a confidence interval that does not contain 0. The odds ratio calculated from the logistic regression results provides a measure of the effect of the predictor on the response, and it indicates that the odds of No Show are lower for individuals who received a text message reminder compared to those who did not receive a reminder.

In conclusion, the statistical significance of the analysis was thoroughly evaluated using logistic regression, and the results indicate that the predictor (TextReceived) has a significant effect on the response (No Show).

# F2. Practical Significance

The results of the logistic regression analysis indicate that the binary predictor variable (TextReceived), which indicates whether a patient received a text message reminder or not, has a significant effect on the binary response variable (No Show), which indicates whether a patient showed up for their appointment or not.

In practical terms, this means that text message reminders can have an impact on reducing no-shows in medical appointments. For example, if a medical facility implements a text message reminder system, they can expect to see a decrease in the number of patients who do not show up for their appointments.

The odds ratio of 0.8575, which was calculated from the logistic regression analysis, suggests that holding all other variables constant, the odds of a no-show are 0.8575 times lower for patients who received a text message reminder compared to those who did not receive a reminder.

For example, if 100 patients were scheduled for an appointment, and 50 of them received a text message reminder, we would expect to see about 6 fewer no-shows among the group that received the reminder compared to the group that did not receive a reminder.

# F3. Overall Success

The overall success and effectiveness of the project can be evaluated based on the following factors:

1. Objectives: The objective of the project was to analyze the impact of text message reminders on medical appointment no-shows, and the analysis met this objective by fitting a logistic regression model to the data and evaluating the significance of the predictor (TextReceived).
2. Data quality: The data used for the analysis was of good quality, as it was cleaned and checked for missing values, outliers, and other anomalies before the analysis. This helped to ensure that the results were accurate and meaningful.
3. Methodology: The methodology used for the analysis was appropriate, as logistic regression is a commonly used statistical method for binary classification problems, such as no-shows. The methodology was also robust, as the model was fit using maximum likelihood estimation and the results were verified using appropriate statistical tests.
4. Results: The results of the analysis showed that text message reminders have a significant effect on reducing the number of no-shows, as the coefficient for TextReceived was significant and had a negative relationship with the response. The odds ratio of 0.8575 suggests that, holding all other variables constant, the odds of a no-show are reduced by a factor of 0.8575 for individuals who received a text message reminder compared to those who did not receive a reminder.

Based on these factors, the overall success and effectiveness of the project can be considered high. The project met its objectives, used appropriate methodology, and produced meaningful results that can be applied in practice.

# Key Takeaways

# G1. Summary of Conclusions

I can reject the null hypothesis that a patient receiving a text message has no effect on the now show rate in favor of the alternative hypothesis which there is a relationship between receiving a text message reminder and reducing "no-shows" for appointments.

# G2. Effective Storytelling

The choice of bar plots, pie charts, and a violin plot for visually communicating the findings effectively support the storytelling of the data analysis.

Bar plots are used to display categorical data and to compare the frequencies of different categories. In this project, bar plots were used to show the distribution of the No Show variable, which is a categorical variable. This type of plot provides a clear visual representation of the proportion of patients who showed up and those who did not show up for their appointment.

Pie charts are also useful for showing the distribution of categorical data and for comparing proportions. In this project, a pie chart was used to show the proportion of patients who received text messages and those who did not. This type of plot effectively summarizes the data in a simple and easy-to-understand manner.

A violin plot is a combination of a box plot and a density plot. It shows the distribution of the data along with its median and interquartile range. In this project, a violin plot was used to show the distribution of the waiting time for patients who showed up and those who did not. This type of plot provides a good visual representation of the distribution of the waiting time, which is a continuous variable, and helps to identify any outliers or skewness in the data.

# G3. Findings-based Recommendations

It is my recommendation that the clinic implement it’s own text message reminder system as a method to reduce the number of no shows. The clinic should start collecting their own dataset so they can analyze the effectiveness of their new text reminder system.

I would also recommend the clinic clearly define how they will handle same day appointments and have same day appointment data separated out from the other appointments so the two sets of data can effectively be compared to each other. I would also recommend the data collection include whether the patient received either an automated or live phone call reminded and if the patient had any live contact with the clinic. This way they can account for some of the variables that were unknown in the Brazilian dataset.

My other recommendation would be to investigate other predictors of missed appointments because this analysis only considered one predictor which was whether the patient received a text reminder. Other factors that data should be collected for an analyzed include age, whether the patient has an acute or chronic illness, whether the patient has insurance or is a self-pay patient (because lack of funds may motivate a patient to skip an appointment), what form of transportation is accessible to the patient (lack of transportation may cause missed appointments) and other demographic variables.

This will provide a more comprehensive understanding of the factors that contribute to missed appointments and help the organization make more informed decisions on how to reduce the number of missed appointments.

# Panopto Presentation

[Thursday, February 9, 2023 at 10:21:25 AM (panopto.com)](https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=ee523669-aec2-41f6-8d3a-afa40130d92a)

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=ee523669-aec2-41f6-8d3a-afa40130d92a

# Appendices

# Evidence of Completion

Submitted in a zip file are the following files: The Jupyter notebook which contains all the python code used to analyze and visualize the data, a html version of the same file for easier viewing and both the original and cleaned datasets used for analysis.  
  
The original dataset is also available directly on Kaggle. (Hoppen, 2018)

[Medical Appointment No Shows | Kaggle](https://www.kaggle.com/datasets/joniarroba/noshowappointments)

# Sources

Hayhurtst, C. (n.d.). *No-show effect: Even one missed appointment risks retention*. Retrieved from Athena Health: https://www.athenahealth.com/knowledge-hub/financial-performance/no-show-effect-even-one-missed-appointment-risks-retention

Hoppen, J. (2018). *Medical Appointment No Shows*. Retrieved from Kaggle: https://www.kaggle.com/datasets/joniarroba/noshowappointments

Perron, N. J., & Dao, M. D. (2013, April 4). Text-messaging versus telephone reminders to reduce missed appointments in an academic primary care clinic: a randomized controlled trial. *BMC Health Serv Res*.