**To Show or Not To Show**

**William N. Wagar III**

**Western Governors University**

**Capstone Project**

**Task 3**

**Project Overview**

**A.**

* The research question that this investigation set out to shed light on and derive information about is the rollout of a text-message reminder program for medical appointments. Through a descriptive and diagnostic analysis the initial results of the first month of collected data have been processed and insights into the population and sub-populations have been uncovered so that stakeholders and decision makers will be able to make decisions regarding changes to the program and which demographics to target for improvement and what additional questions can be asked for further investigation?
* The scope of this project was 40 days data regarding medical appointments. The dates used were from the 29th of April 2016 thru the 8th of June 2016. The geographic scope is to Rio de Janeiro, Brazil. The working scope of the project included the collection of the data set, cleaning and wrangling, splitting data set into smaller, logical components to investigate/drill-down further. The scope of the intention of the project is descriptive and diagnostic.
* The investigation uncovered the general descriptive statistics for the newly rolled out SMS-Reminder program that can be used as a baseline for improvements going forward. The overall method was diagnostic, to observe how the SMS-reminder program was functioning in an operational capacity. The general descriptive statistics will be used to ‘diagnose’ where the program is underperforming and what other questions should stakeholders and decision makers be asking to improve the overall impact of the SMS-reminder program. The tools used were as planned: Python3, Jupyter Notebook, PyCharm IDE, third party libraries (Pandas, Numpy, and Matplotlib most of all), a laptop with lots of RAM.

**Project Plan**

**B.**

* The project plan was executed almost as was anticipated in Task 2. In Task 2 the initial plan called for an analysis of the top and bottom 10 neighborhoods WRT to No-Show and Presented appointments. As the investigation progressed it became apparent data that granular was outside the scope of this investigation. General descriptive statistics for age, age bracket, gender, and days between the day the appointment was scheduled and the actual appointment day are needed first as well as an idea of what could be done with the granular neighborhood data in place is needed before expending the man-hours. Also considered in Task 2 was finding the day of the week of the appointment to be used in some descriptive statistics for the baseline. That attribute was created and kept in all csvs and DFs for later potential use but this analyst decided that investigating the appointment day with respect to the other variables in this investigation was too granular for the same logic above with respect to the ‘Neighborhoods’.
* The Waterfall methodology was used in this investigation as planned but with some iterative nature of Agile that was not anticipated in Task 2. The main element of a waterfall methodology used was the concept that one Jupyter Notebook investigating a specific part of the dataset had to be finished before going forward as the structure of the notebook, code used, and insights gleaned were necessary for the next step. In section C below this analyst will delve further into the process. The iterative nature presented itself as this analyst went through the dataset, as well as different sub-sections and as that was happening, different ideas and questions were considered, and some kept. Those new statistics to calculate or process change decided on would have to be retroactively inserted into a Jupyter Notebook and then the investigation could continue until a new concept or idea presented itself and then the process would repeat (thought/idea, consider value of idea/is it worth keeping, moving on or going back in the notebook/code to integrate new element).

**Project Timeline and Milestones**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Milestone** | **Expected Start Date** | **Actual Start Date** | **Expected End Date** | **Actual End Data** | **Duration (Man-Hours)** |
| **Obtain Task One Approval** | **3/2/2023** | **3/2/2023** | **3/3/2023** | **3/3/2023** | **4 Hours** |
| **Obtain Task Two Approval** | **5/15/2023** | **5/15/2023** | **5/16/2023** | **5/16/2023** | **8 Hours (revisions, instructor appointments and feedback)** |
| **‘Birds-Eye’ Stat Summary** | **5/16/2023** | **5/17/2023** | **5/17/2023** | **5/19/2023** | **9 Hours** |
| **Gender and Age Analysis** | **5/18/2023** | **5/20/2023** | **5/20/2023** | **5/22/2023** | **6 Hours** |
| **Conditions and Time Factor Analysis** | **5/21/2023** | **5/23/2023** | **5/23/2023** | **5/24/2023** | **6 Hours** |
| **Complete Post-Investigation Report** | **5/24/2023** | **5/26/2023** | **5/25/2023** | **5/27/2023** | **8 Hours** |
| **Final Submission** | **5/26/2023** | **5/27/2023** | **5/27/2023** | **5/27/2023** | **0 – Just Submitting** |

**Methodology**

**C.**

* There was no difference from the Task 2 plan for data selection and collection. The data set was vetted and found to contain what was needed and collected with no issue from Kaggle.
* There were no obstacles encountered while collecting the data.
* There were no data governance issues as the data set was obtained from Kaggle. As stated in Task 2, ‘There is no governance, privacy, security, ethical, legal, or regulatory compliance considerations.’ The data was anonymized, and any communicating of findings will not violate any privacy guidelines or reveal any information that would allow a person whose data is represented to be identifiable.
  + The advantages of this data set are a large sample size with descriptive attributes that can be used in a descriptive and diagnostic analysis, the data is already in an anonymized state regarding PII or Health Data, some of the columns can be used to create other columns that contain valuable information.
  + The limitations of this dataset are that is only for a 40-day period, between April 29, 2016, and June 8, 2016. This is just a snapshot and limits the insights that can be gained. Additional attributes would have been helpful in painting a more detailed picture, such as income, income brackets, marital status, does patient have children, is patient a child, etc.

**D.**

The extraction/wrangling process was done primarily using Juptyer Notebooks and PyCharm IDE. PyCharm was appropriate and useful to use when creating the code used in the Jupyter Notebook as PyCharm helps users keep track of variables across a project and visually assists the user by providing snippets of what the code coming next likely will be given what has already been typed on that line. This is appropriate for a student as there were many times when this analyst had used a function or method before but was not entirely sure about the syntax. PyCharm was very useful in this respect as it would prompt on the screen what the generic syntax structure is supposed to be and the mandatory arguments, etc.

Python 3 is the industry standard programming language and was used for that reason in this analysis. Specifically, this analyst made use of the ‘**df.loc[()]**’ syntax quite frequently when extracting and filtering data; **.loc[()]** made filtering the dataset based on specific conditions extremely easy.

Jupyter Notebooks were the end of the journey for the code created in PyCharm. Jupyter Notebooks running a Python3 kernel were the optimal choice to visualize, annotate, and present code in clean, professional manner. Jupyter Notebooks make the presentation of the extracted data visualizations/graphs much more professional and easier to view than the visualizations and manner of viewing the visualizations than PyCharm offers.

The data was extracted from the original data set from Kaggle. From this point it was necessary to wrangle and extract information viewed through specific lenses. From the original data set all the rows that were eligible to receive a reminder were separated from those that were not eligible to receive a SMS reminder. From that point, rows that DID receive a SMS reminder were extracted and separated from those that DID NOT receive a SMS reminder. From that point each of those respective data sets (SMS-Eligible-SMS-Received and SMS-Eligible-No-SMS) were filtered further so that it was possible to view in isolation the populations that PRESENTED AT THEIR APPOINTMENT or NO-SHOW with respect to whether or not they received a SMS reminder.

**E.**

1. The methods used in this investigation were segmentation analysis and descriptive statistics combined with data visualization. As mentioned earlier in this report, the data set was segmented into smaller and smaller segments involved filtering on more and more conditions the farther down that was drilled.
2. Segmenting the dataset into specific populations allows the analyst to view each segmented set in isolation. It is necessary to segment the data to isolate descriptive statistics about specific populations and specific conditions. The visualizations allow the viewer/stakeholder/decision maker to instantly connect with the findings of the segmentation.
3. The initial step was to choose the lens this analyst was going to view the data set; on whether an eligible appointment row receive a reminder or not. As mentioned earlier the eligible appointment rows had to be segmented from non-eligible (same-day, less than 24 hours between ScheduledDay and Appointment Day). The next logical segmentation was to filter the eligible reminder rows into 2 segments, those that received a reminder and those that did not.
   1. From this point descriptive statistics were obtained using Python3 code for the segments as aggregate (ALL ROWS ELIGIBLE AND RECEIVED SMS and ALL ROWS ELIGIBLE AND NO SMS REMINDER).
      1. Each of the two segments listed in point ‘a.’ above were further segmented into populations that PRESENTED or NO-SHOW for each segment.
         1. These segments were the most granular and descriptive statistics were gleaned from them ( 4 total, 2 for RECEIVED SMS, and 2 for NO-SMS RECEIVED)
   2. The statistics obtained from the actions in point ‘i’ and its sub-point, ‘1’ were used to paint an overall picture of how the program is performing.
   3. As numerous rows were filtered into different data sets, there are numerous checks throughout the notebooks and code of ensuring that the numbers always added up to what they should have been. (ex. Finding the sum of 2 newly filtered data frames to make sure they equaled the total of the data frame they were created from so no rows were lost).
   4. As general statistics were the name of the game for this investigation, this analyst created a function to calculate the percentage of a condition in a column with respect to the total rows in that data frame. This function ensured that user of the function only needed to know the name of the data frame, the name of the column they were trying to calculate a condition percentage on, and the condition itself, See function definition code at the end of the section:

1. Once the structure of the notebooks were outlined this analyst calculated descriptive statistics on the different populations of this dataset viewed through the lens of whether or not a received a SMS reminder and if they presented for their appointment or not. At the end of every notebook is a bold markdown cell with all the descriptive statistics obtained from that population. Each notebook contains its own set of statistics at the end and make this very easy for anyone to compare different populations.

def column\_percentage\_calculator(df, column\_name, condition):

total\_rows = df.shape[0]

num\_matched\_rows = len(df.loc[(df[column\_name] == condition)])

percentage\_matches\_condition = (num\_matched\_rows / total\_rows) \* 100

return percentage\_matches\_condition

This function was defined at the beginning of any Jupyter Notebook that was heavily involved with finding descriptive statistics.

**RESULTS**

**F.**

1. Significant Statistics gleaned below:
   1. The main question this investigation is trying to answer is what the initial effectiveness of the SMS-Reminder program is. This investigation showed that of appointments eligible to receive an SMS-Reminder **THOSE THAT RECEIVED AN SMS REMINDER HAD A 5.12% HIGHER PRESENT AT APPOINTMENT RATE.**
   2. The make up of the 2 populations investigated (SMS-Eligible-Received-SMS and SMS-Eligible-No-SMS) are very similar to each other. Both of these populations show that overwhelmingly the appointment rows in this investigation are for appointments scheduled between 2 and 30 days. This trend persists regardless of receiving a SMS- Reminder or Presenting/No-Show.
   3. Women make up the majority of appointment rows by a 2-1 ratio (or close to) in every population.
   4. Of those appointment ROWS THAT PRESENTED REGARDLESS OF SMS-Reminders, those mean and median ages were noticeably older (by 5 years give or take). As one would expect given the higher mean and median ages, the age brackets of ‘Adult’ and ‘Older Adult & Seniors’ make up the majority of the population.
   5. Of those appointment ROWS THAT NO-SHOW REGARDLES OF SMS-Reminders, those mean and median ages were noticeably lower than those that PRESENTED (by 5 years give or take). As one would expect given the lower mean and median ages, the age brackets of ‘Youth’ and ‘Young Adult’ make up the majority of the population.
   6. Of those rows that Presented (regardless of SMS-Reminder) the average days between ScheduledDay and AppointmentDay were slightly lower than the rows that were No-Show.
   7. Appointment rows that suffered from Hypertension had the biggest difference between Present and No-Show. Populations that Presented contained appointment rows with much higher concentrations of hypertension sufferers.
2. Practical Significance:
   1. The practical significance of this investigation is that we can see that the SMS-reminder program is having an impact on No-Show appointments. As we saw at the beginning of this section, those appointment rows that received a SMS-Reminder had a 5.12% higher present rate than those that did not receive a SMS-Reminder, this suggests that the program is effective in improving appointment attendance.
   2. The observation that older individuals (with higher mean and median ages) were more likely to present at appointment, regardless of SMS reminders, suggests that age is a significant factor in appointment attendance. This finding highlights the importance of considering age-related strategies when implementing changes or improvements to the SMS-Reminder program.
   3. The lower mean and median ages of No-Show appointments, regardless of SMS-reminders, suggests that younger individuals have a higher likelihood of missing their appointments. This shows the importance of understanding the unique challenges and preferences of younger age groups to address ever-changing specific needs.
   4. The lower average days between the day the appointment was scheduled and the actual appointment day in the population of presented appointments compared to higher average days for No-Show appointments suggests a practical significance that optimization of the scheduling process could be implemented to reduce the number of days between appointments for all patients.
3. This analyst feels that this project was a resounding success. From a mass of data information has been extracted that has practical and statistical significance. From these baseline analytics changes to the program can be made and now there is data to compare the results of those changes against.

**G.**

1. This analyst can conclude that the SMS-Reminder program is having an impact on appointment attendance as predicted. At present it can be determined with the data analyzed that that impact is a 5.12% greater appointment attendance when a SMS-Reminder is sent. We can conclude that women are almost twice as likely to be the patients attending the appointment. Patients who are older comprise the majority of the segments that had the best present-rate while the two younger age groups comprise the majority of the segments with the higher no-show rate.
2. The visualizations in the notebooks attached with this project all follow a color theme. Any consumer of this information would be able to put these visualizations next to each other and see how they change or stay the same very easily. A stakeholder would be able to see visualizations for the entire data set and compare them to smaller and smaller drilled down populations. That tells a story of change between viewing a trend for a particular population in isolation compared to how that trend behaves outside isolation with that particular population.
3. Based on this investigation this analyst would recommend the following:
   1. Conduct a separate investigation for the appointments scheduled over 30 days out. As shown in the investigation, the percentage of appointments between 2-30 days is so much greater than the other two time periods that to make more informed decisions on how to increase the impact the SMS-Reminder program separate analysis is likely needed for longer timeframes between appointments.
   2. To paint a better picture, in the next iteration of data collection for appointment rows, collect more demographic information about the patient such as marital status, do they have children, is the patient under the age of majority, ethnicity, education level, etc.
   3. Continue to conduct this analysis month-by-month so data can be gathered for each month of appointments and trends regarding seasonality can be investigated.
   4. Design a predictive model to attempt to predict whether the patient will no-show and design steps for what to do if a patient is predicted as a no-show.

**SOURCES**:

* No third-party sources were used in the creation of this report or in this investigation. There are no in-text quotes used that need to be referenced. All work produced in relation to this project, apart from articles quoted in Task 2, that were properly quoted and cited, are the creation of this analyst. Any and all code, summarization, paraphrasing or referencing in this report is this analyst’s intellectual creation or references this analyst’s intellectual creation.