DSO599: Data Analysis Project #1

Prepared by: Faris Alfaadhel and Saif Ur Rehman

IBM Watson Application

IBM Watson has been doing a great job of revolutionizing the way healthcare is performed through its healthcare analytics. One specific example which we were impressed with was the application of IBM Watson in Floyd's health care system[1]. Floyd is a "304-bed acute-care hospital and a regional referral center that covers more than 40 medical specialties, a behavioral health center, an outpatient surgery center, a physical therapy and rehabilitation center, and more" [1]. It serves more than 350,000 patients across six counties in Alabama and Northern Georgia. Floyd had a goal to obtain the National Committee for Quality Assurance (NCQA) Patient-Centered Medical Home (PCMH) recognition and wants to move toward population health management[1]. To obtain the PCMH recognition, the organization must demonstrate improvement in six process measures[1].

To tackle the issue and improve upon their population health management strategy, Floyd hospital started utilizing patient summaries in order to better pre-plan patient visits and better manage patients' health[1]. They measured their performance against three preventative clinical measures: breast cancer screening, cervical cancer screening, and colorectal screening, along with 16 diabetic measures[1]. After seeing the results of utilizing the patient summaries, more and more offices started utilizing them[1]. Nine months after utilizing IBM Watson Health, Floyd found that pre-visit planning had become a habit at all of Floyd's offices and the number of patients who underwent preventative screenings increased by as much as 41 percent in some offices[1]. The overall average improvement in preventative screenings across all offices was 22 percent. In addition to that, "diabetes process measures were up an average of nearly six percent, with one office seeing an increase of 18.6 percent" [1]. These improvements helped Floyd achieve the Level 3 PCMH recognition in September 2016[1].

What we really liked about this application is that the hospital utilized IBM Watson in order to move towards a more patient-centric approach. Through IBM Watson's technology, the hospital was able to better prepare for handling their patients, increase their preventative screenings, and although not mentioned in the article, it would ultimately lead to healthier patients and lower overall costs in the healthcare system.

LGBT Center OC

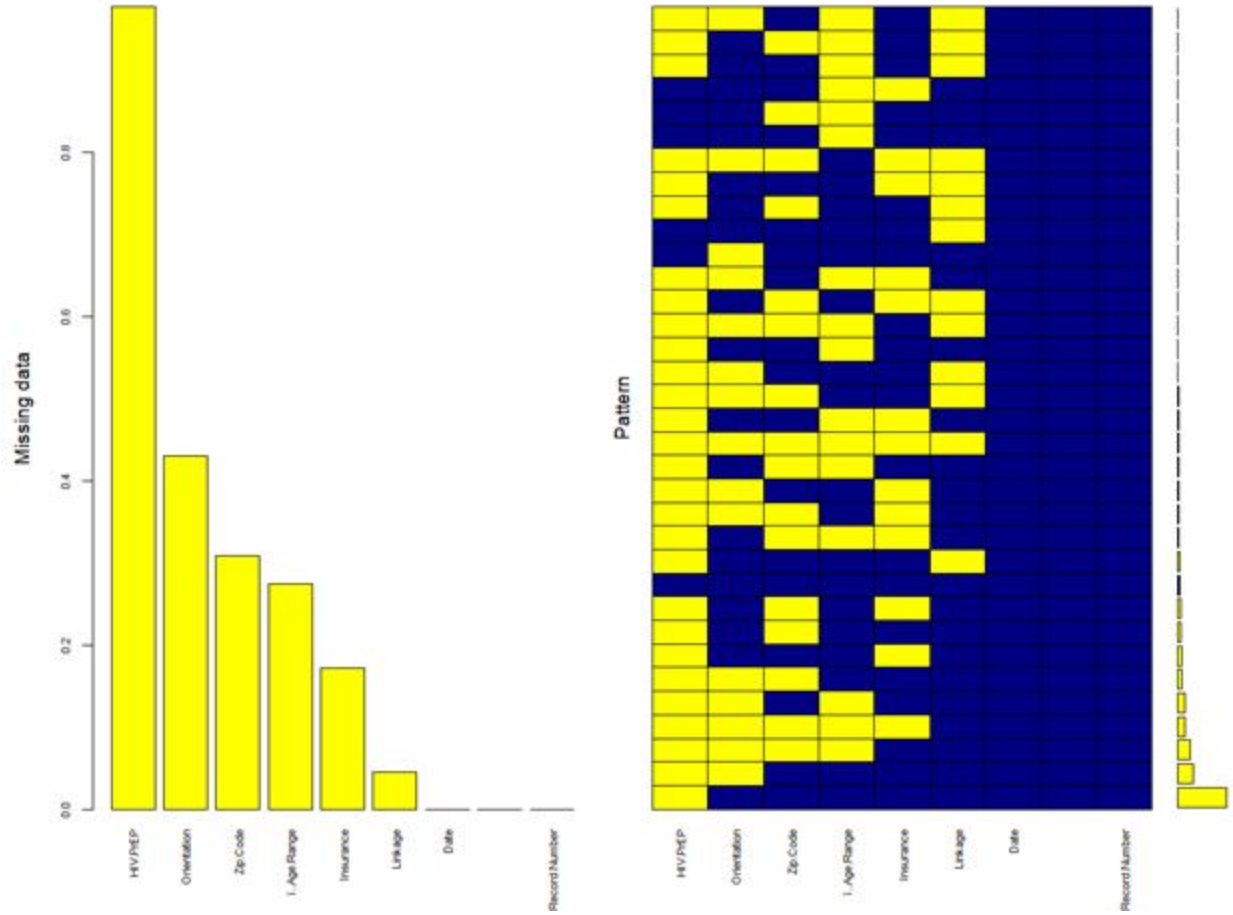
Insights

A-level Insights

Missing Data

Missing data is one of the hallmarks of this dataset. However, missing data itself can give us some valuable insights. The most obvious insight is the distribution of the missing data across the different questions. As shown by the figure, 97.73% of the HIV/PrEP column contains missing values, followed by orientation which has 43.05% missing values, then zip code with 30.84%, age with 27.39%, and insurance with 17.14%.

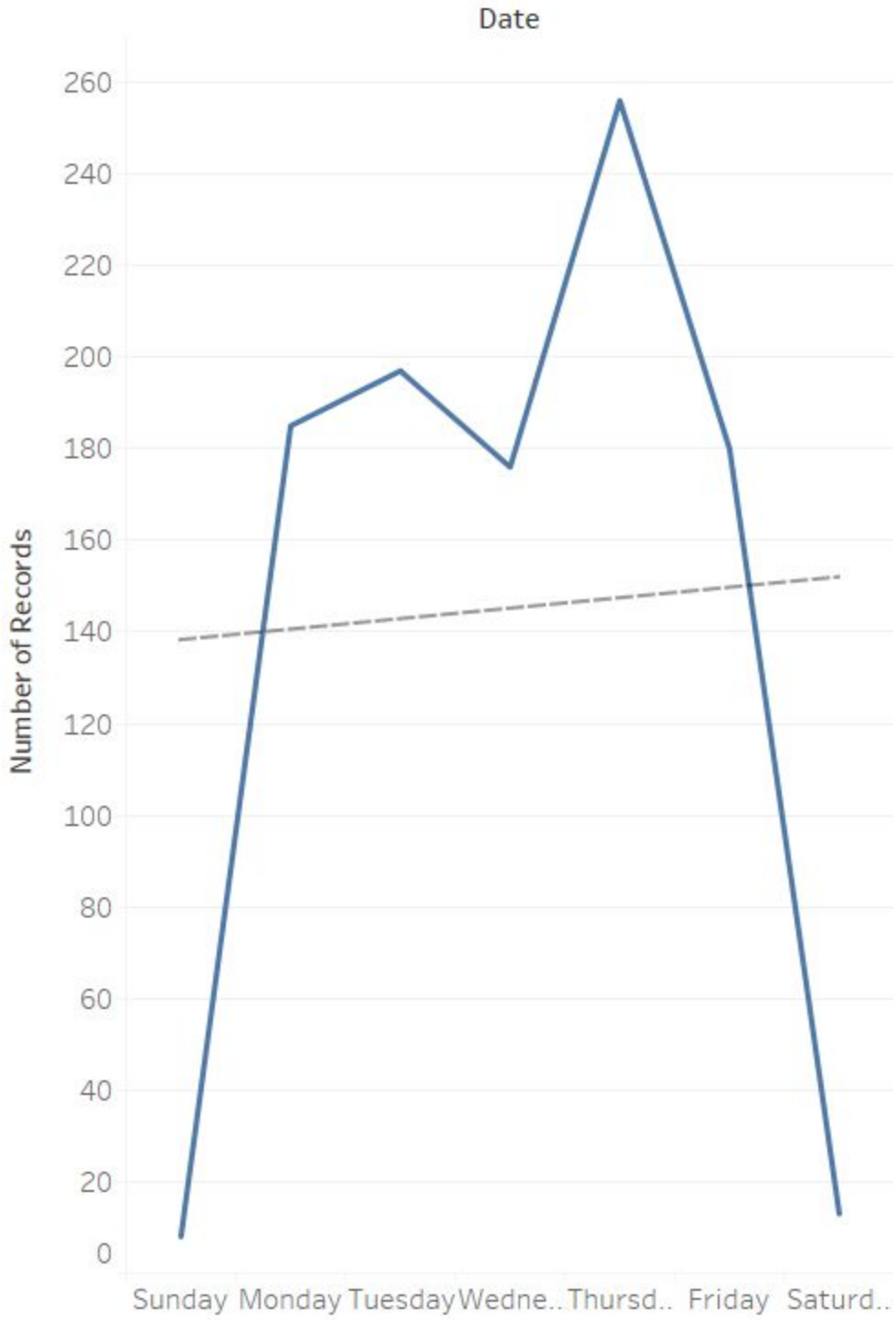
A more interesting and insightful observation from the dataset is the distribution of the missing values. The most frequent pattern in missing values is that the HIV/PrEP column is missing, while the rest of the data values are properly recorded. This is followed by both HIV/PrEP and Orientation columns being missing, with the rest of the data being present. The third most frequent pattern is four columns being missing: HIV/PrEP, Orientation, Zip Code, and Age.



Weekday Trends

Another interesting insight obtained from the data is the distribution of the visits throughout the week. As shown by the figure below, there is a clear spike in visits on Thursdays as compared with the rest of the weekdays. This could be due to external marketing that is performed by the clinic, or it can be a pattern in patient visits which will help the clinic better manage the forecasting of patient visits. Other than the spike, it makes sense that the number of patient visits during the weekend are lower during the weekday and that may be related to the clinic's operating schedule. It is an important distinction to note that there were indeed some patient visits on Saturday and Sunday so if the clinic is not operating at these times, then these might represent data entry errors or outlier patients which required immediate attention that took place on a weekend.

Weekday Trends

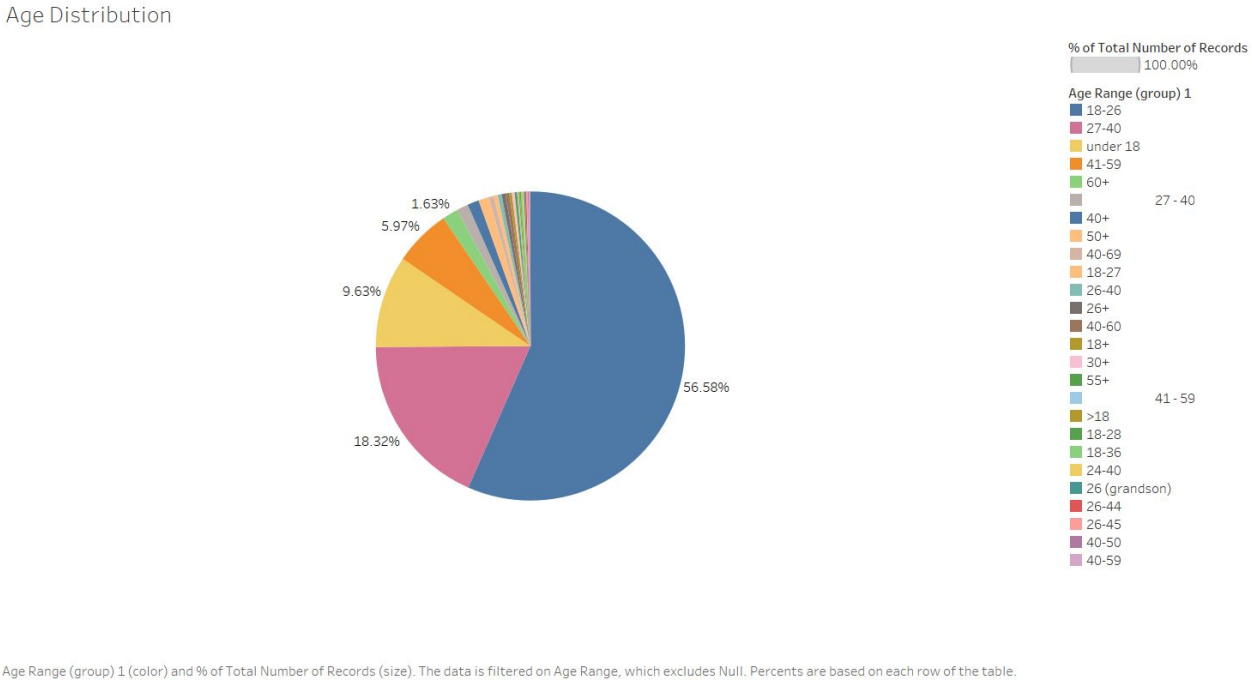


The trend of sum of Number of Records for Date Weekday.

B-level Insights

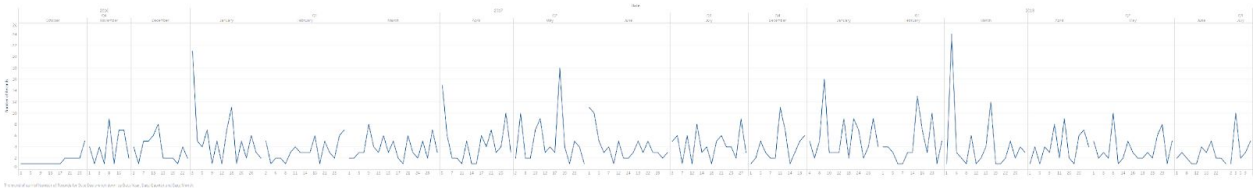
Age Demographics

An important aspect of any clinic is to have a greater understanding of the demographics of the patients that are visiting the clinic. This is very important because it helps the clinic tailor the care of their patients and helps in how to train staff to effectively administer treatments and remedies to their patients. If the clinic uses a one-size-fit-all approach, it generally leads to lower patient satisfaction and lower success rates and helping the clinic get a better understanding of the age demographics will help combat this approach. To get an accurate gauge of this, we determined that the most prominent age group that visits this clinic is between the ages of 18-26 consisting of 56.58% of the patients that visited and filled out the patient forms for age. The next four most popular age groups are 27-40, under 18, 41-59, and 60+ consisting of 18.32%, 9.63%, 5.97%, and 1.63% of the patient visits respectively.



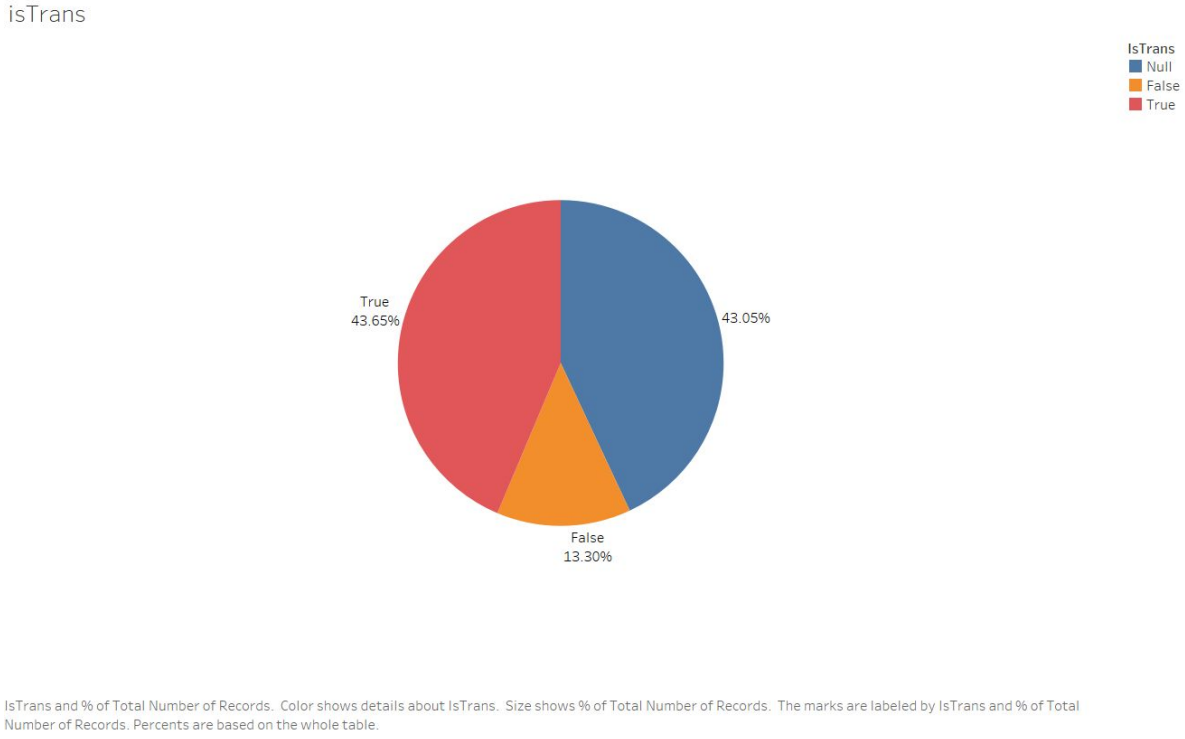
Missing data for certain periods

One insight that we have uncovered is that there is no data collected from 08/01/2017 until 11/30/2017. This could be due to storage issues, data collection issues, or a combination of multiple reasons. However, this does have a significant impact on the analysis that is performed since missing values for a big chunk of the periods prevents seasonality from being detected if performing a time-series analysis for patient visits.



Transsexual Demographics

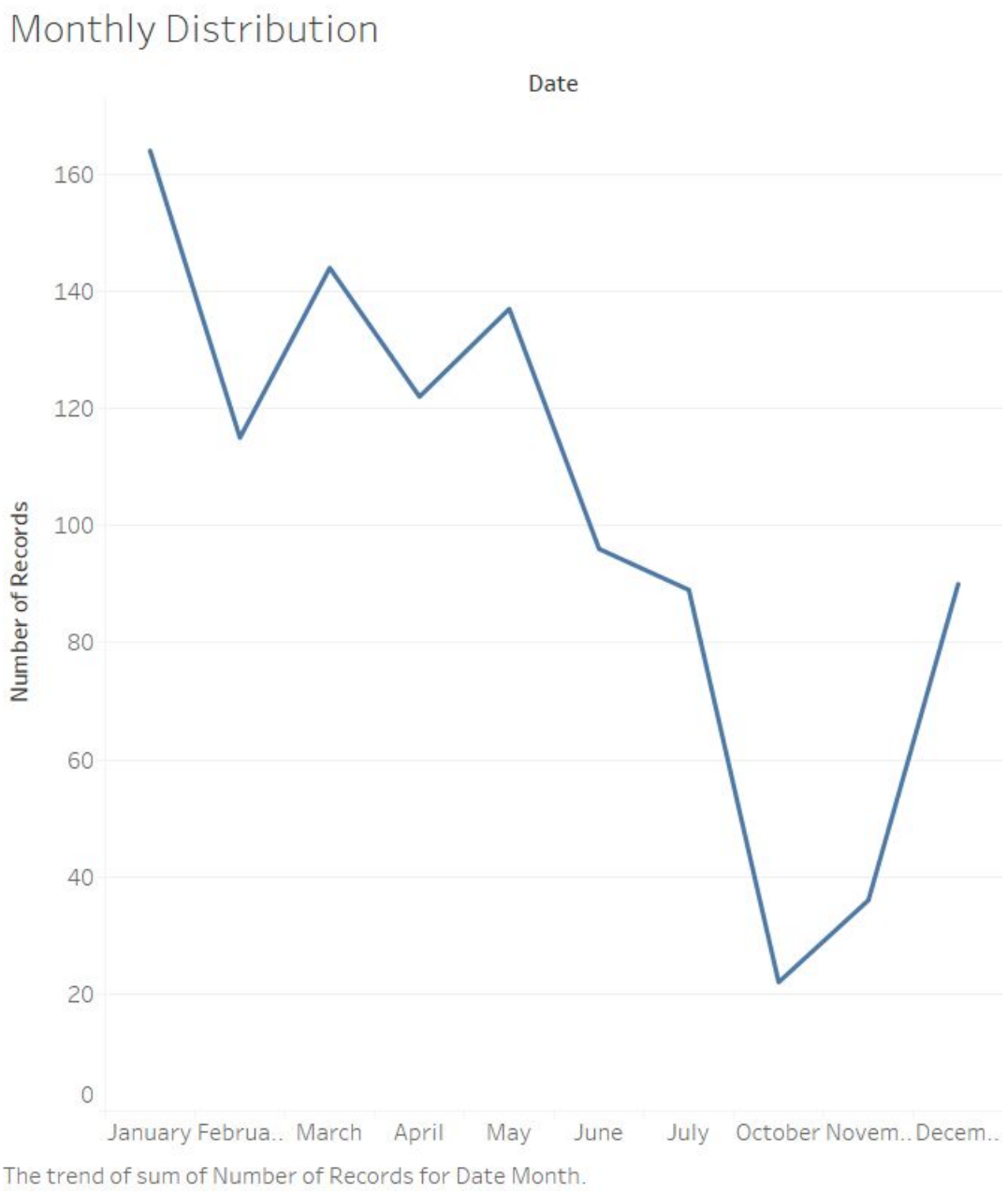
When evaluating the responses for orientation, the respondents show that the majority of the patients identify as transsexual (43.65% of the total number of patients). 43.05% of the patients did not indicate whether they are transsexual or not, while 13.30% did not indicate they were transsexual in their orientation response. This is an interesting insight as the orientation column seems to be an open comment column where patients could have initially listed anything as their orientation. The fact that most of the patients identified as transsexual could be very important because it could mean that the clinic receives the most visits from transsexual patients. However, this could also mean that other patients did not feel comfortable explaining their responses under the orientation column as the question was not framed correctly. More data is needed to extract more precise insights for the reasons of this distribution.



Monthly Distribution of Visits

One insight that was quite interesting was that there seems to be a downward trend as the year progresses in terms of the number of visits and increases in December. There is not enough conclusive

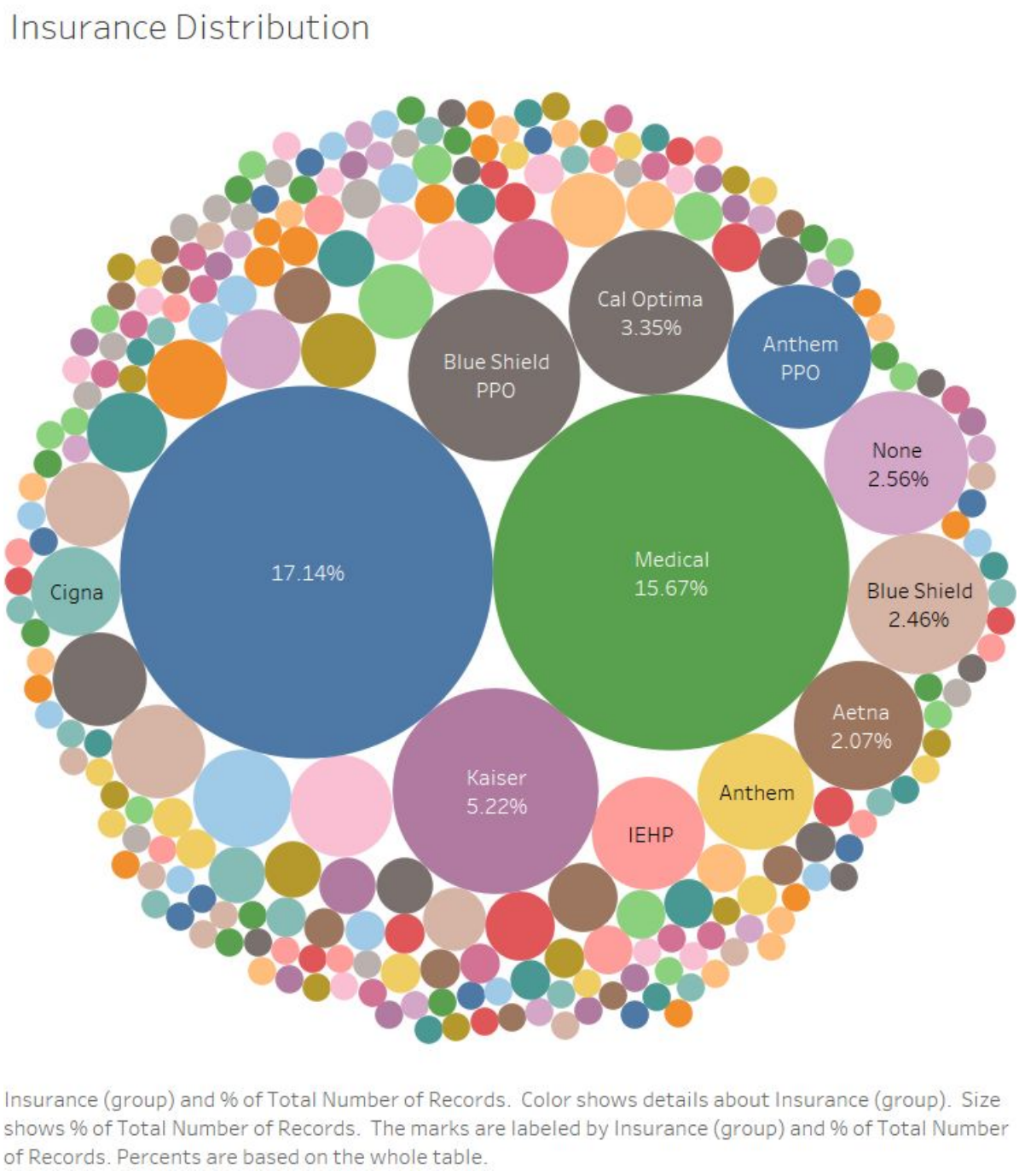
evidence to determine a single factoring cause for the trend, however, it is worthy to note that since data points between 08/01/2017 and 11/30/2017 are missing, this may be a contributing factor to the apparent downward trend during those months.



Insurance Distribution

Another interesting insight is the distribution of the insurance companies. There are 17.14% null values which means that there are some patients that did not report their insurance company. However, the

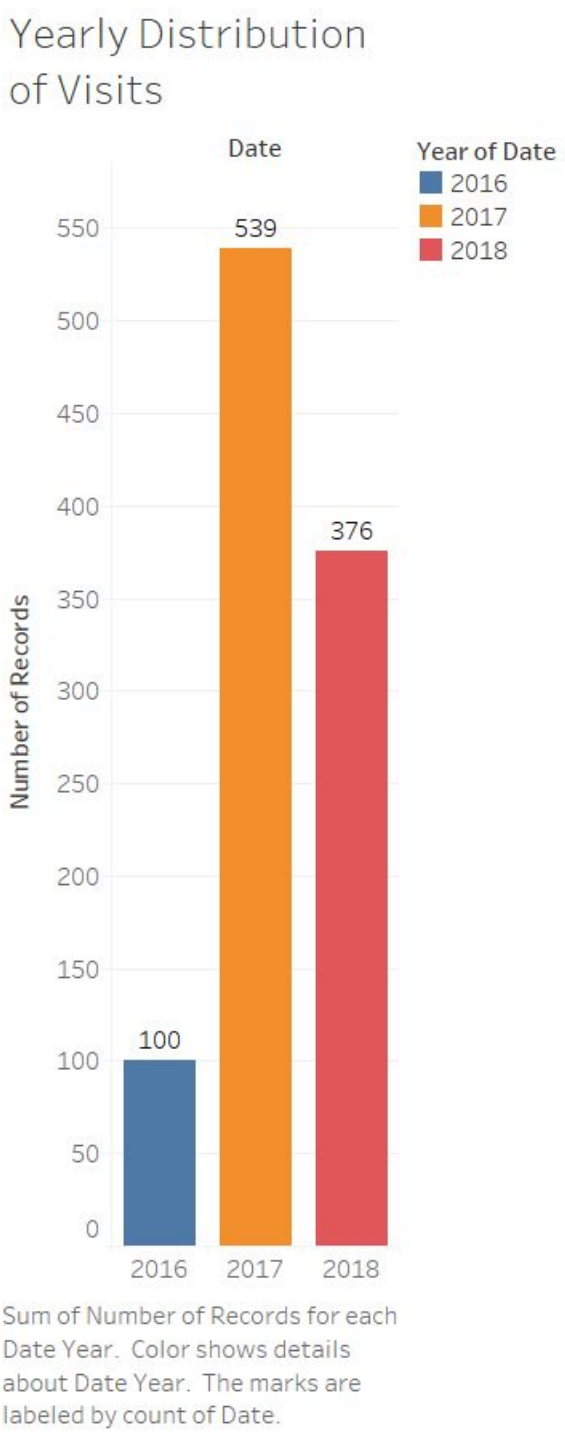
major insurance companies for patients that come into the clinic are Medical, Kaiser, and Blue Shield PPO which constitute 15.67%, 5.22%, and 3.65% of the patients' insurances.



C-level Insights

Yearly Distribution of Visits

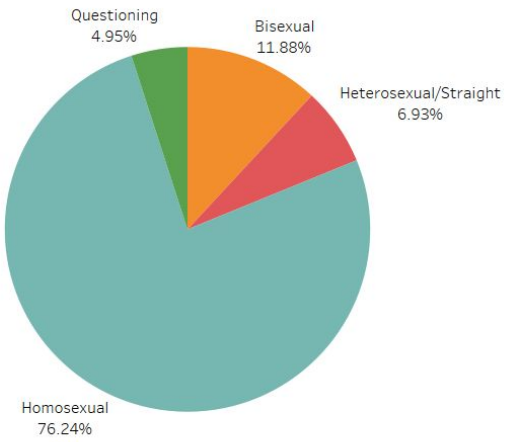
Another insight is that there is a higher number of visits in 2017 as compared with 2018 and 2016. In 2017, there were a total of 539 visits that were recorded, as compared with 2018 and 2016 which have 376 and 100 respectively. However, this is related to the fact that we only have data which starts in October of 2016, therefore contributing to a low 2016 number of visits. Similarly, 2018 has not finished yet, so we only have data until July of 2018 contributing to the low number of visits in 2018.



Sexual Orientation Demographic

Furthermore, when evaluating the sexual orientation of the patients, out of all of the patients that responded, 76.24% of the patients identified as homosexual, 11.88% identified as bisexual, 6.93% identified as heterosexual, and 4.95% identified as questioning.

Sexual Orientation

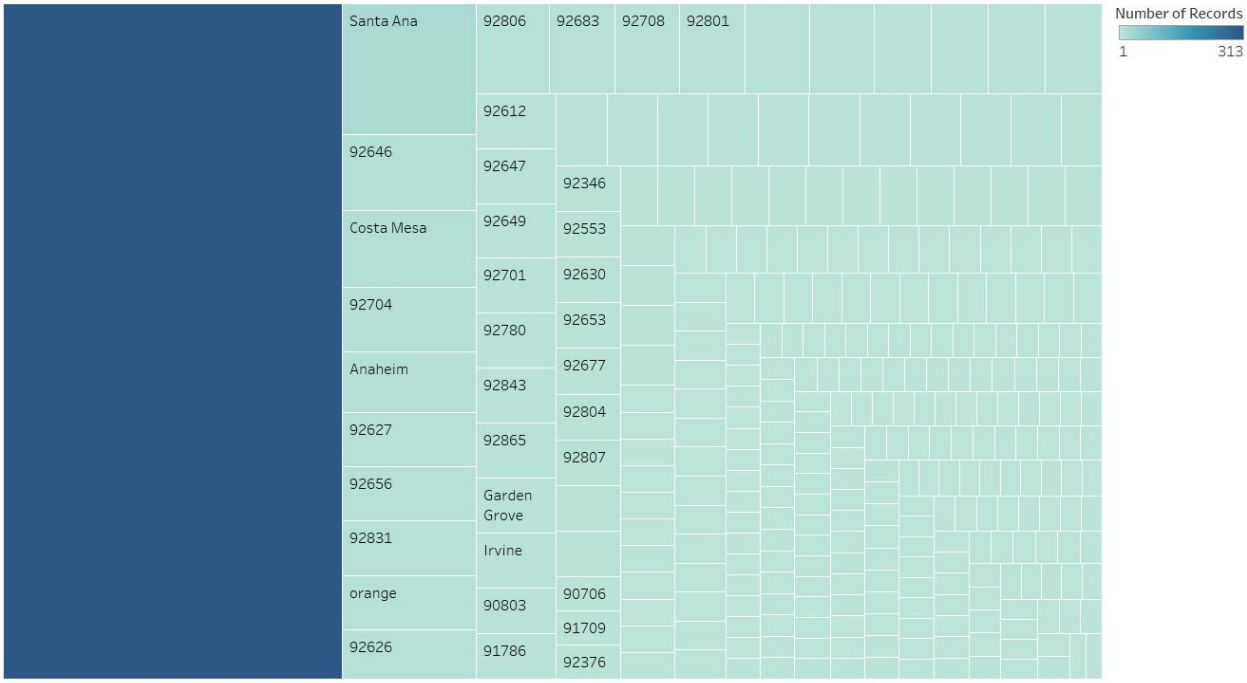


Categorized Orientation as an attribute and % of Total Number of Records. Color shows details about Categorized Orientation. Size shows % of Total Number of Records. The marks are labeled by Categorized Orientation as an attribute and % of Total Number of Records. Details are shown for Categorized Orientation. The view is filtered on Categorized Orientation, which keeps Bisexual, Heterosexual/Straight, Homosexual and Questioning. Percents are based on each column of the table.

Geographic Distribution

Knowing the geographic distribution of the patients is also an important insight for the clinic. The figure shows that the majority of the patients did not respond to this part, however, for the patients that did respond, they are mainly from areas such as: Santa Ana, zip code 92646, Costa Mesa, zip code 92704, Anaheim, and zip codes 92627, 92656, and 92831.

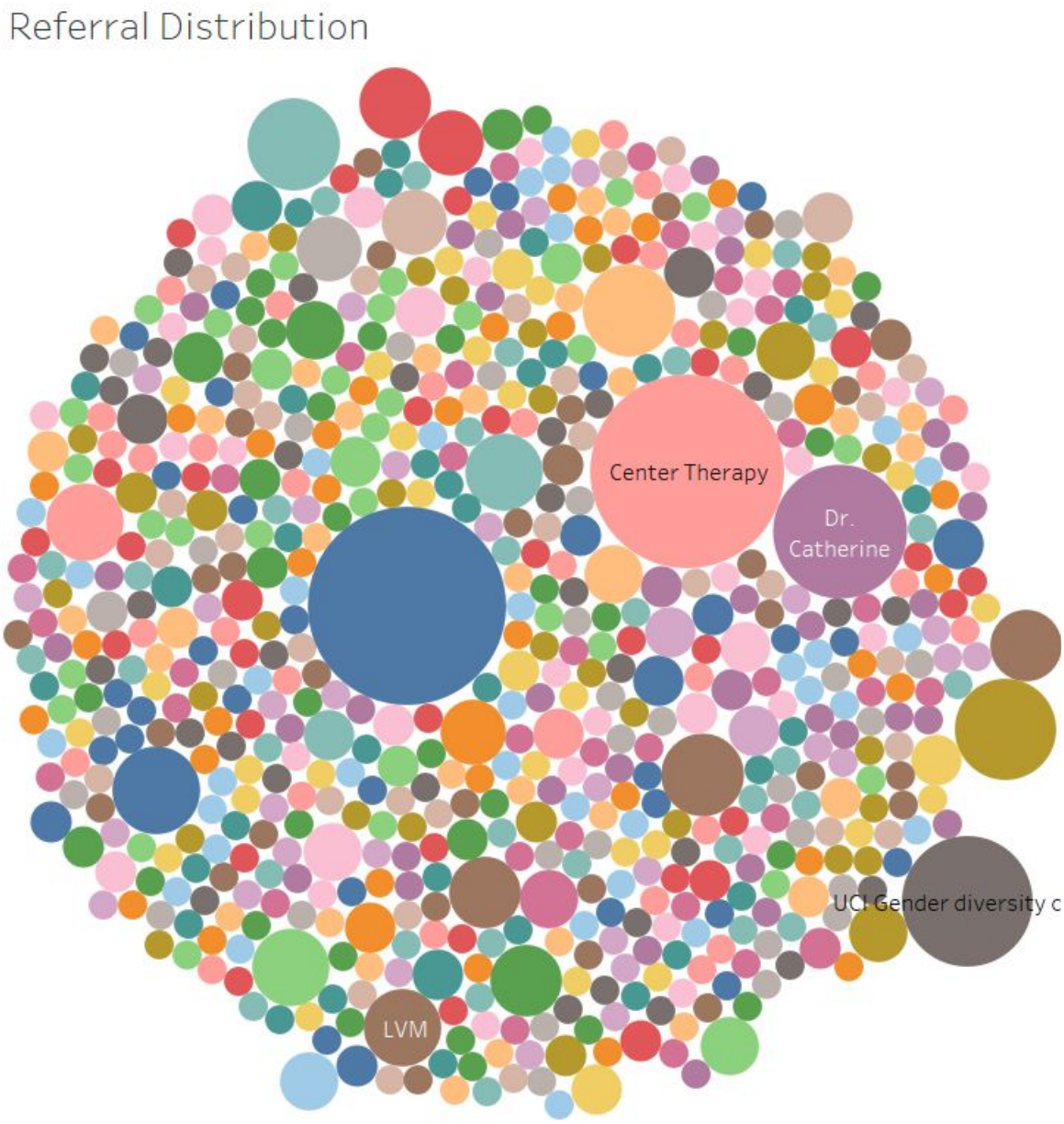
Zipcode Distribution



Zip Code. Color shows sum of Number of Records. Size shows sum of Number of Records. The marks are labeled by Zip Code.

Referral Distribution

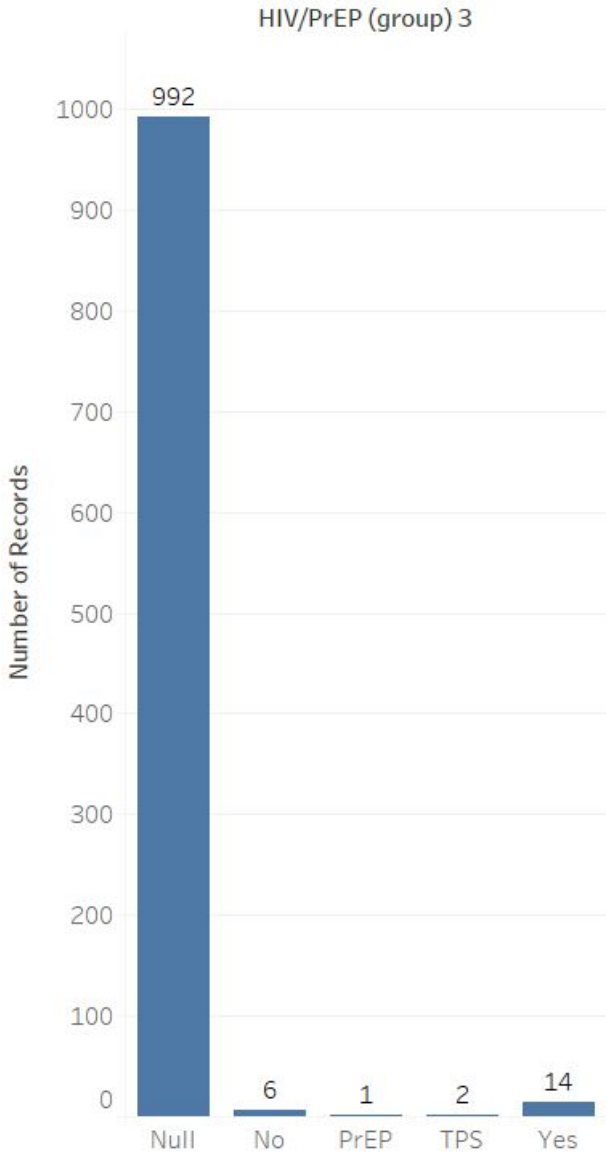
When evaluating the distribution of the referrals, it is clear that the biggest source of referrals come from: Center Therapy, Dr. Catherine Garcia, and the UCI Gender Diversity Clinic.



HIV and PrEP Responses

Out of the patients that responded, only 14 stated that they are HIV positive. Furthermore, 1 patient has reported that they have had PrEP, two patients have reported that they have had TPS, while six patients have reported that they do not have HIV. A majority of the patients, 992 of them, have not filled out this section of the questionnaire.

HIV/PREP



Sum of Number of Records for each HIV/PrEP (group) 3. The marks are labeled by sum of Number of Records.

Recommendations

A-level Recommendations

Data Collection and Integrity

One of the first and important recommendations that we have for the clinic is to improve their data collection processes in order to obtain better insights. To get a better idea of why that is so valuable, we present a couple of insights that we could have obtained from the data if there was more data collected and if the data integrity was preserved.

First, one of the most insightful analytics we could have provided are forecasts for future patient visits. This would drastically help the clinic better optimize their work schedules and staffing requirements for the day. However, this would only be possible with accurate data collected about patient visits. This forecast is contingent upon the accuracy of the data collected.

Currently, there is a significant gap in data collected from 08/01/2017 until 11/30/2017. Furthermore, there are sporadic schedules for the clinic in terms of visits. The clinic does not seem to be open on a set schedule during the week; however, this conclusion may be inaccurate since there are gaps in the data and the integrity of the data. It may just be that the data is not accurately collected for these periods. There are also periods where there are only one patient visit throughout the day, and some days where there are 24 patient visits. We tried performing a time-series forecast using ARIMA modeling and LSTM's, however, the results were not useful or accurate because of the data variability and not enough data points to accurately determine seasonality.

Some other insights that we could have provided from the data are regressions and correlations between the missing data and a deeper understanding of the correlations of the patients' characteristics. Furthermore, we could figure out what the exact demographics of the total patients are so that the clinic can better market and tailor their needs towards their patients and move towards a more patient-centric model.

Missing Data

As shown from the insights, there were also patterns in the missing data reflecting questions that the clinic can specifically focus on. The two main patterns from the missing data are that the question about the HIV/PrEP and question about orientation are the most commonly left blank. This is an indication that these two questions are the ones that the clinic should focus on rephrasing and improving the most as they represented the biggest patterns in terms of missing values.

Data Collection Software

To completely tackle the issue of missing data, we recommend the clinic to use a software such as Google Forms or SurveyMonkey so that missing data is not an issue. Using these software suites or any comparable data collection software, we recommend the clinic to require patients to fill out the specific question as opposed to allowing the patient to skip it. We understand the necessity of allowing the patients alternative responses and allowing anonymity, so to tackle this, the clinic should have some prespecified responses available for the patient to select, while also allowing a "N/A" and "Other"

responses to be inclusive to the patient diversity. These can be adapted as the clinic deems necessary so that more data insights can be extracted.

B-level Recommendations

Data Collection Specific Recommendations

To give more specific recommendations to the clinic in order to improve the data integrity for future data analysis projects, we recommend the center do the following steps:

1. Standardized age bins

We recommend the clinic to set up standardized bins for collecting age so that the patient can quickly fill in their age bracket. A recommended set of bins based on the data would be: Under 18, 18-24, 24-30, 30-40, 40-50, 50-65, over 65. These bins can be adapted as the center deems fitting. However, it is very important to set up standardized bins so that patients can easily just circle or bubble in their appropriate age. Another option would be to allow the patient to write the exact age; however, that will lead to longer data processing times and may result in another set of data integrity issues as typos and mistakes can occur frequently. Binning the ages will also decrease the occurrences of missing values. Another suitable alternative would be for the receptionist to add in the patient's birthday as part of their patient information so that age would be a calculated number. However, this is conditional upon the setup of the center and data collection processes of the clinic.

2. Reformat the questions about Gender and Sexuality

Currently, there is only one question about sexual orientation which is causing problems in the number of responses and confusion for the patients of the clinic. These should be appropriately broken up into separate questions regarding gender identity, sexual orientation, and transsexual.

One of the questions should be predetermined responses for gender identity: "What do you determine is your current gender identity", which includes the options: "Male", "Female", "Non-binary", "Questioning".

The next question should be: "Do you identify as being transsexual?" with the responses being: "Yes" or "No".

The follow up question should be "If yes, what do you identify as?" with the responses being: "MTF", "FTM", "Questioning", "Other", and "N/A". This question should only be asked if the patient has indicated "Yes" on the previous question about transsexuality, however, there is a "N/A" if mistakes occur.

Now it should lead on to questions about sexual orientation with the first question being: "What is your current sexual orientation" with options that include: "Straight/Heterosexual", "Gay/Lesbian", "Bisexual/Pansexual", "Asexual", "Questioning", "Other (includes genderfluid, queer, etc.)"

For all of these questions, the responses can be expanded to be more specific depending on the demographic composition of the patients visiting the center, however, they should be binned as absolute categorical responses to prevent confusion and loaded questions.

3. Separate the questions for HIV and PrEP

We recommend the clinic to separate the questions regarding if the person is HIV positive and whether they have taken PrEP. This can be separated by asking “What is your HIV status” with responses including: “HIV Positive”, “HIV Negative”, “Unknown” and another question asking “Have you taken PrEP” with responses including: “Yes”, “No”, “Unknown”.

4. Bin Insurance Information

Another aspect of the data that needs to be better collected is the insurance information. This can be done by separating out the questions and allowing only certain responses for the most popular insurance companies in the area and allocating the rest of them as “Other”.

For example, the first question should be “Are you on MediCal/MediCare”. With the available responses being “Yes”, “No”, and “Unknown”.

The next question should then ask, “If no, do you have PPO or HMO” with available responses of “PPO” or “HMO”, “I do not have insurance”, and “Unknown”.

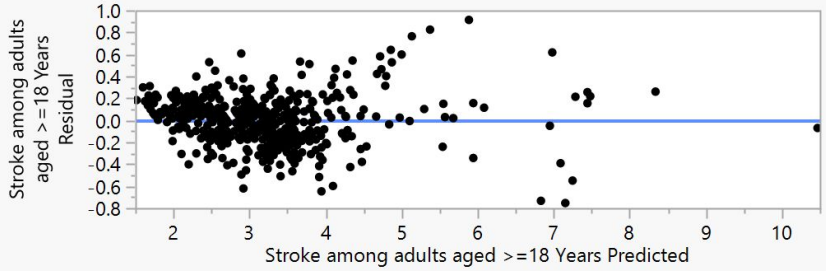
The final question should then give a list of the top 10 popular insurance companies and include a “Other” and “N/A” option which allows the person to write down the insurance company’s name. This can then be adapted when certain insurance companies’ are appearing in the data more frequently as they can be added to the list of popular insurance companies. This helps narrow down the question to certain lists because currently there are conflicting names and some people have the HMO version while others have the PPO version which is causing difficulty in aggregation and data analysis.

Analysis of a chosen dataset

Background

As part of the project requirements, we are asked to find a dataset and extract meaningful insights. This was no easy task, as multiple datasets contained databases larger than 200MB and requiring access to special software to analyze. We finally settled in on the 500 Cities: Local Data for Better Health dataset provided by the CDC[2]. It shows a variety of health outcomes by city for 500 cities in the US. Our primary motivation was to try to see if any of the health factors are predictive of others. Using JMP software, we chose the prevalence of Strokes in adults as our dependent variable, and tested against a variety of outcomes. The initial output is found below.

Residual by Predicted Plot



Summary of Fit

RSquare	0.957396
RSquare Adj	0.95657
Root Mean Square Error	0.228968
Mean of Response	3.213502
Observations (or Sum Wgts)	474

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	9	546.64781	60.7386	1158.554
Error	464	24.32578	0.0524	Prob > F
C. Total	473	570.97359		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.017864	0.047147	-0.38	0.7049
Binge drinking among adults aged >=18 Years	-0.007998	0.00523	-1.53	0.1269
Current lack of health insurance among adults aged 18–64 Years	0.0199555	0.002912	6.85	<.0001*
Current smoking among adults aged >=18 Years	0.024528	0.005573	4.40	<.0001*
High blood pressure among adults aged >=18 Years	0.0961387	0.006028	15.95	<.0001*
High cholesterol among adults aged >=18 Years who have been screened in the past 5 Years	-0.04703	0.006328	-7.43	<.0001*
Mental health not good for >=14 days among adults aged >=18 Years	0.1075102	0.009159	11.74	<.0001*
No leisure-time physical activity among adults aged >=18 Years	-0.016356	0.004645	-3.52	0.0005*
Obesity among adults aged >=18 Years	0.0118189	0.004795	2.47	0.0141*
Visits to doctor for routine checkup within the past Year among adults aged >=18 Years	-0.003995	0.002867	-1.39	0.1642

Insights

A-level Insights

Propensity to binge drink is not significant in predicting stroke prevalence

As we can see from the parameter estimates part of the JMP output above, whether a person drinks or gets a regular doctor check-up do not help us explain whether or not they will be at risk for stroke. At a p-vlaue of 0.1269 and 0.1642, respectively, the indicator is not statistically significant at the 5% significance level. However, once we remove binge drinking as a variable, we see that doctor check-ups are indeed significant. It is quite surprising that binge drinking is not a factor to increase stroke risk, as many public health agencies advertise it as such. More detailed analysis combined with critical trials is necessary to come to a scientific conclusion.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	-0.02644	0.04688	-0.56	0.5730	.
Current lack of health insurance among adults aged 18–64 Years	0.0199051	0.002916	6.83	<.0001*	4.337267
Current smoking among adults aged >=18 Years	0.0223003	0.005387	4.14	<.0001*	11.689347
High blood pressure among adults aged >=18 Years	0.100515	0.005313	18.92	<.0001*	20.927145
High cholesterol among adults aged >=18 Years who have been screened in the past 5 Years	-0.05029	0.005967	-8.43	<.0001*	20.542173
Mental health not good for >=14 days among adults aged >=18 Years	0.1081639	0.009162	11.81	<.0001*	11.059683
No leisure-time physical activity among adults aged >=18 Years	-0.013701	0.004315	-3.18	0.0016*	14.402357
Obesity among adults aged >=18 Years	0.0102642	0.004692	2.19	0.0292*	18.791881
Visits to doctor for routine checkup within the past Year among adults aged >=18 Years	-0.006141	0.002504	-2.45	0.0146*	17.701347

Poor mental health is the 2nd largest explainer of stroke prevalence

Effect Summary

Source	LogWorth	PValue
High blood pressure among adults aged >=18 Years	58.893	0.00000
Mental health not good for >=14 days among adults aged >=18 Years	27.588	0.00000
High cholesterol among adults aged >=18 Years who have been screened in the past 5 Years	15.355	0.00000
Current lack of health insurance among adults aged 18–64 Years	10.565	0.00000
Current smoking among adults aged >=18 Years	4.383	0.00004
No leisure-time physical activity among adults aged >=18 Years	2.797	0.00160
Visits to doctor for routine checkup within the past Year among adults aged >=18 Years	1.837	0.01456
Obesity among adults aged >=18 Years	1.535	0.02920

As we can see from the effect summary table above, high blood pressure explains a lot of the variance in stroke prevalence, which is expected based on contemporary studies[3] showing the correlation between the two. However, the fact that mental health leads to increased stroke prevalence is particularly interesting as it is not usually emphasized as a leading cause of strokes or increases stroke risk.

B-level Insights

Each 1% increase in the amount of adults who had been diagnosed with high cholesterol reduces prevalence of stroke by 0.05%

This is a great result and incredibly insightful. It seems that when adults are diagnosed with high cholesterol, they tend to stay off the path of getting a stroke. This makes sense as the patient now has a higher risk level of stroke and would take medication and preventative steps to avoid that adverse outcome.

1% increase in prevalence of high blood pressure leads to 0.1% increase in stroke risk

High blood pressure has the strongest magnitude and explanatory value of any of the indicators we analyzed in our regression. As we mentioned before, this makes sense, but the magnitude of the effect makes it incredibly important to manage high blood pressure and identify potential individuals with risk.

Recommendations

A-level Recommendations

Focus prevention messaging away from alcohol drinking and to mental health

Our first recommendation would be to try to rework stroke prevention marketing materials and messaging away from alcohol consumption and towards focusing on mental health and wellness. There has been a renewed focus on mental health and wellness, along with a proliferation of mobile applications to address these issues. If these apps can reduce the prevalence of stroke, it could generate huge cost savings for the cities' healthcare systems, as the average lifetime cost of a stroke patient is $60,000[4] compared to these apps being free.

B-level Recommendations

Provide free cholesterol examinations to high-risk populations

The negative relationship between being screened for cholesterol and the prevalence of stroke was a great insight. Based on this, the health departments of these cities need to provide free examinations which can be as cheap as $30 - exactly 0.05% of the previously mentioned lifetime costs of stroke. This makes it a break-even proposal at best, but the improvement to patient care within the constituencies of the city should make it well worth it.

Works Cited

[1]Floyd health care system. (2017). Retrieved from https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=ASC12460USEN&

[2]500 Cities Project: Local data for better health | Home page | CDC. (2017, November 28). Retrieved from https://www.cdc.gov/500cities/index.htm

[3]Hornsten, C., Weidung, B., & Et.al. (2016). High blood pressure as a risk factor for incident stroke among very old people: a population-based cohort study. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5398900/

[4]Stroke: Incidence and Cost in the United States. (n.d.). Retrieved from https://www.utsouthwestern.edu/edumedia/edufiles/education_training/programs/stars/unwin-stroke.pdf