Infusion Center ABC Dataset Analysis
Internal Report

Mac Mai

March 2023

1 Potential Issues

1.1 Dataset Issues

- Many appointments have checked in, but have never been marked as completed.
  - This could lead to iQueue thinking these appointments are still ongoing or that the chairs are still in use. This could cause increased wait times as staff would have to manually check that a patient has left and that the chair is free again.
- None of the appointments have a chair out time.
  - This can lead to difficulty in calculating appointment lengths and wait times, as well as calculating what chairs are free.
- Many of the appointments don't have a check out time.
  - Similar to the previous issues, this can cause problems with knowing when a patient has actually left.
- Many of the dates and times reported used differing formats.
  - For example: DD/MM/YY, DD/MM/YYYY, YYYY-MM-DD, HH:MM, HH:MM AM/PM
  - This causes huge issues, as having mismatched data means that times cannot be compared as easily. It also can lead to miscommunication about when an event is actually meant to occur. For example: "03/04/22 8:30". This could be either March 4$^{\text{th}}$, 2022 or April 3$^{\text{rd}}$, 2022, and it could be occurring in either the morning or the evening.
- On November 4$^{\text{th}}$, there was an extra column of data.
  - By inserting the current date in a column where it should not have been, data was pushed over. This caused multiple columns to have incorrect data, not just the one error.
- Some appointments don't have descriptions of what infusion they received.
  - This could potentially lead to a patient's death. If it is not accurately recorded what medications they have received, they could be receiving an incorrect medication.

- Some appointments' infusion descriptions have extra notes added in.
  - While this is not a huge issue, it could lead to problems later on if you try to look at data for the rates of medications received. There should ideally be one value for the infusion description, and another column for additional notes.
  - This could even be expanded on, with splitting it into even more columns, one for each variable. For example: Medication Name, Dosage, Type. This would allow you to group by medication and see how much you use of each.

1.2 Database Issues

There are some other issues that pertain more to the design of the database itself.[1]

- There are many repeated appointment IDs due to patients receiving multiple infusions. Ideally, this should be structured using multiple tables, using INPATIENT_DATA_ID_x as a primary key in an appointment table and a foreign key linking it in an infusion table. This would reduce redundancy and make the database able to be more easily used in an SQL environment. I would also recommend ensuring that the database is in a normal form (Boyce-Codd Normal Form is the typical standard and would remove the vast majority of redundancies.[2])
- CHAIR_START should be renamed to CHAIR_IN. As it is currently, it can lead to slight confusion as chair start doesn't quite match with chair out. Chair in would be a more descriptive name, as nothing is being started yet, they have just entered the chair.

2 Solutions

2.1 Dataset Issues

The issue I'd like to talk about the most was that of formatting all the dates and times. While this in of itself was relatively easy[3], the issues that came with this were interesting. My first idea was to put dates in YYYY-MM-DD, times in HH:MM, and datetimes in YYYY-MM-DD HH:MM. This worked great, up until I wanted to compare times and calculate differences.

Because I had formatted them back into strings[4], I would need to convert them back into datetimes to compare. So, I instead changed the times into datetimes as well and left them all as datetimes instead of converting back into strings. This did have the drawback that seconds were included as well, but I decided that having a more accurate time that was easier to work with was preferable to displaying slightly fewer characters.

I'd love to talk about the rest of the things I tried for other problems, but that was the one I thought was the most interesting.

[1] This is assuming that I have received the entire database, not a concatenated version of all the data.
[2] Don't ask me to explain it without Google though; I just learned about it the other day but I thought it was cool.
[3] using to_datetime(infer_datetime_format = True)
[4] using .dt.strftime()

2.2 Database Issues

The biggest issue that I did not try to solve in my program is that of the database design. If I had more time, I would try to normalize the data more. I would start by splitting it into separate databases, each corresponding to an entity in this system: Appointments and Infusions. In Appointments, I would only keep the columns that pertain to the appointment itself: checkin, checkout, scheduled, cancelled, etc. In Infusions, I would put the rest of the data: chairin, chairout, infusionstart, infusionend, orderdescription, etc. Both of these would also have the apptID in order to link them.

You can do this in pandas by slicing the dataframe and assigning it to a new dataframe. This would lead to having to restructure code, as you would need to do actions more similar to how you would use SQL, like doing inner joins and grouping. I would also put restrictions in place that would check data as they're inputted, and only allow it if the times lined up. I'm not quite as familiar with doing this in pandas as I am with SQL, but as far as I'm aware it has all the same functionality.

3 Conclusions

After removing all the erroneous and redundant rows, I was left with 157 individual appointments. This is about .4 of the original count (769 infusions across 423 appointments). Here is what I found:

- After checking in, the average patient waits 00:49:15 before entering the chair.[5]
- The average patient is 00:05:47 late to their appointment.[6]
- The average appointment is 00:30:40 longer than scheduled.[7]

[5] Calculated as the difference in check in time and chair in time.

[6] Calculated as the difference in check in time and scheduled appointment time.

[7] Appointment length was calculated as the difference between entering the chair and ending the last infusion. Taking the difference of that and the appointment length column gives us this statistic.