INST 327 Section 0202 Team 0202-021

Team: Michelle Doan, Mo Konduru, Brittany Kupcho, John Adjani-Aldrin, Perry Gonzalez

Introduction

This semester we covered a vast array of techniques to be displayed through the analysis of a Montgomery County crime dataset. Our group was tasked to gain insight into the dataset by transforming the information in our file into a usable database with detailed tables. We sought to answer numerous questions such as: the most crime ridden cities, the least safe cities in the county, the districts with the highest victim count, most common offense that occurs, and the most frequent time for crime to occur. These questions enabled us to look into the most secure places in the county and view the cities and districts in the county with the most criminal activity.

Our group chose this dataset due to the many questions that could be answered and the implications that this data could give in terms of criminal activity throughout the county. One thing about our data is that it doesn't give insight into the types of people that are committing the crimes which would then be able to give more implications into the demographics of certain areas. Yet on the other hand this is beneficial in terms of adhering to the necessary ethics regarding privacy of the victims and the individuals that committed such crimes. Nonetheless, our database analysis provides a starting point for further research in the future in order to gain a broader understanding of the crime patterns in Montgomery County.

Views / Queries

crime_name_safety answers the following questions:

1.) Which cities in Montgomery County are least safe in terms of property?

2.) Which cities in Montgomery County are least safe in terms of people?

3.) Which cities in Montgomery County are least safe in terms of society?

Safe_City answers the following questions:

1.) Where would a family looking for a safe city want to move to?

Avoid_City answers the following questions:

1) Where would a family looking for a safe city want to avoid?

Dangerous_Cities answers the following questions:

1.) What is the most dangerous city to reside in terms of overall crime?

Districts_Victims answers the following questions:

1.) How many victims are there in the most crime ridden districts?

Popular_Offense answers the following questions:

1.) What offense shows up the most in the city of Silver Spring?

2.) What offense shows up the least in Silver Spring?

High_Crime_Times answers the following questions:

1.) When should we make more safety resources available (police on duty)?
2.) What time of day on July 1, 2016 what times are the most crimes happening?
3.) What city did the crimes occur the most on that day?

Null_EndTime answers the following questions:

1.) What crimes don't have an end time recorded?

Requirements Table

View Name	Req A- At least four of your queries (saved as views) should involve multiple (two or more) tables, and thus involve JOIN clauses	Req B- At least three of your queries should involve some form of filtering (WHERE, HAVING, etc.) (Requirement B)	Req C- At least two of your queries should involve some form of aggregation over records (SUM, COUNT, AVERAGE, GROUP BY, etc.) These cannot be queries that simply count the number of rows in a given table, such as SELECT COUNT(invoice_id) FROM invoices.	Req D- At least one of your queries should involve a join (linking) table and both of its source tables.	Req E - At least one of your queries should use a subquery
crime_name_safety	X	X	X	X	
Dangerous_Cities	X			X	
Safe_City, Avoid City	X		X	X	
popular_offense	X	X	X	X	
High_Crime_Times	X	X		X	
null_endtime		X	X		X

Changes from original design

We changed a significant portion of our database from previous iterations. When we initially created our ERD model we just had a general idea of what we wanted. We created foreign keys and did not have the correct data types. We had many issues with our initial structure which we had to fix over time. We made more iterations to our ERD once we realized that the data types were wrong. When we started, a lot of data types were in VARCHAR and did not work when we tried to import them. We realized that we needed to change some to INT and change some of the VARCHAR values to allow for import. Most of the changes we made happened when we were trying to import. Attempting to import the data into our tables made us realize a lot of the issues that still persisted in our ERD. Many of the foreign key constraints weren't correct so we had to redo them. We also initially had too few rows in our csv files, so we expanded from about 30 to 200. This gave us a lot more variety in our data and made it easier to create the queries for the questions.

Once we expanded our rows, we had to change more data types since some were not importing correctly. We also noticed that we would benefit from adding a victim column to the offense table, since having the victim count for each crime can prove useful. We also moved around many of the columns in a way that would make more sense for the flow of the table. Now with this current iteration we are able to perform joins easily and are able to move around the table in an order that makes sense. We no longer have any issues with importing and our queries have been working as expected. After going over with the instructor over our ERD, we decided to make further improvements to our design. We made the final decision to remove the crime_time table since it was unnecessary, and we added the values to crime. We created a new resident_type table that includes information about the location of the resident. This gives us a lot more information to work with and makes the ERD make more sense.

It took a significant amount of time of trial and error and researching through class material to come to this point. There were many errors that did not seem fixable but we were able to eventually solve them. Our new design takes advantage of everything we have learned so far in class and considers the correct implementation of that knowledge.

Database Ethics Consideration

In terms of the database ethics we have evaluated how the data we used was collected, stored and exchanged. The collection method was through the Montgomery crime dataset which was a public record database that displayed crime within various cities within Montgomery County. We extracted certain pieces of data to draw conclusions, without manipulating or falsifying any of the information in order to prove a theory. The data was stored in SQL Workbench which we used to create the various tables and queries which suggests that there were no unethical concerns with the project as a whole. The data was exchanged between group members with no data manipulation and the conclusions that were drawn were not used to defame or harm a city or entity. It was solely used to gather information and provide support or help to a community.

Lessons learned

As a team of five, it was difficult collaborating with one another because we scheduled meeting times and made sure that everyone was contributing meaningfully to our project. Working in a team of five challenged us more because even though there were more people to contribute to the project we had to work together which is harder with more people.

The lessons we have learned on a technical side, the normalization and ERD model were our biggest challenges of the project because that laid the foundation of our database.

Normalization of the Montgomery county dataset was particularly difficult because a lot of the codes and names we did not understand what they meant so it took extra time to understand which codes would actually be relevant to connecting to each other. MySQL workbench overall was not difficult to use, but the user truly has to understand a lot of the functions that it offers and how to be as efficient as possible.

Once we figured out the normalization and created our ERD model, it was very tedious for us to make sure that the VARCHAR had the right number of characters because it messed up some of the imports and only allowed for partial data to be imported. Normalization overall is a difficult topic that a lot of my team members and I were struggling with, but once we understood it, it made the rest of the project easier. Importing was easy when we figured out which tables needed to be imported first which were tables without foreign keys restraints which we learned about later. When it came to us creating the queries for our dataset, it was difficult to write a view that would answer multiple questions that needed to be answered. Although it was not difficult to meet the requirements for each query written, it was difficult for us to create a subquery that would be complex enough to answer a question.

If our group were to have more time on the project, we could have created more queries that could have answered complex questions that would have helped families that want to move to Montgomery county. With more time we also could have added more data than only two hundred lines to also make sure that we could get more accurate information. The more data we had to provide the more accurate counting and averages would be.c

Potential future work

For the scope of this project, we have made a database with 200 records. In a real world setting, this is not completely representative of the data. Since crimes frequently occur, our 200 records only consist of the crimes that have taken place on July 1 and July 2 of 2016. In order for this to fully answer the questions we have set out to, there would need to be a wider variety of data.

Since safety in particular areas is ever changing, keeping the record of the past 10 years would be an ideal way to develop this database in the future. Data from too long ago would no longer be relevant, and it gives users a timeframe. Upkeep of the database by a database administrator would be necessary to consistently add new instances and change things as needed. An example of these changes may include If a police district changed to include different areas the crimes committed would have to be updated to include this change.

In order to make the filtering process easier in the future, we could implement a way to show comparative statistics. Since the way this database is intended to be used includes families picking a place to live or not live, tables that compare the statistics of multiple places would be a helpful tool. The database is also intended for police districts to know how and when to use their resources. While it may be helpful to view all of the data for the cities in Montgomery County, a way to filter by certain streets/areas within a city would be a helpful tool for police districts.

Overall, more ways to filter the data would be useful if we were to continue to improve and expand our database, as well as adding more diverse and up-to-date information. Keeping in mind that past records may change, and that the ways people would want to filter this information is very user-specific, there are a variety of extensions that can be added to the Montgomery County crime database.