D211: Advanced Data Acquisition

Performance Assessment

Part 1: Data Dashboard

1. See attached for both datasets. The medical dataset was obtained by WGU and the heart dataset was obtained on Kaggle.
2. For this analysis I decided to try using PowerBI instead of Tableau so the spreadsheet using PowerBI and pivot tables is attached. I also decided to try using Google Colab for my SQL coding since it offers a lot of tools to create databases and push data into a SQL server.
3. To navigate this dashboard, the user simply load up the xlsx file, and then look over sheets. The “combo sheet” should not be touched since that is the sheet where all the data was combined to be analyzed. The first tab will be “Heart Disease by Age and Gender” which counts the number of patients based on gender and high blood pressure with an age axis. This has 3 selectors that will allow you to filter by gender, high blood, or age to dynamically see the results. The next tab “Heart Disease by Gender” counts the number of cases of heart disease by gender. The button at the bottom allows the user to select which sex they want to home in on. The final sheet is “Gender & Income on Blood” which compares incomes to gender and sees if they have high blood pressure. The buttons near the bottom allow the user to select if the patient has high blood pressure or not along with gender and can compare incomes that they have.
4. See attached for SQL code.

Part 2: Demonstration

1. Video Link: <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=316cc26a-d026-4f9a-a306-ae24015a1a52>

Part 3: Report

1. The purpose of this analysis was to see if there were patterns that could be found in the medical dataset that compares income, gender, and having high blood pressure, to another dataset that has age, sex, and heart disease diagnosis. By comparing these 2 datasets, I would be able to compare incomes, to heart disease, to gender, and to high blood pressure, which would really give me a deeper understanding of the data.
2. I decided to try using Microsoft PowerBI/Pivot tables for this analysis since I wanted to try learning how to use other analytics software that wasn’t Tableau in order to diversify my skillset. Pivot tables are useful since it will dynamically display and represent my data tables that I created in SQL as well as allowing me to connect databases and other SQL servers to help with this analysis.
3. In order to clean this dataset, I first had to isolate the columns that I wanted to analyze. I used python to convert the columns I wanted into their own database, I then converted the columns and datafiles into SQL databases in order to do the manipulation necessary for this analysis. My first task was to check out that my operations were done correctly, I so checked the databases to make sure they were populated. The next task was to check for null values and to address them. Thankfully there were no null values so I then began just doing some general data exploration. I ended up joining the age groups together to help expand the dataset for analysis. I then updated the databases to have similar variables for the “Sex” variables. I updated all sex/gender to be Male/Female using SQL to help combine the datasets. Finally, with gender, I combined the datasets. I then finally did a few advanced operations in order to explore the data and help support the dashboards that I developed. I used SQL to check the number of people with certain ranges of income, heart disease, gender, and high blood pressure in order to help figure out how these factors affect each other.
4. In order to create my dashboards, I first had to export my database. I used a little SQL and python to convert the database into a csv file and then imported it into excel in order to use PowerBI and Pivot tables. I then double checked the data to make sure it transferred over correctly, and then started playing around with the Pivot Tables in order to see what kind of insights I could find.
5. After playing around with my Pivot tables, I was able to find that Males are more likely to have heart diseases than women. It also seemed to show that people with higher incomes across both genders have a lower risk of high blood pressure which has a relationship to heart disease. This analysis would help an executive decision maker decide on which treatments would be most cost effective and would help prepare options for those who are most likely to have heart problems.
6. One of the main issues with my datasets is that the sizes between them are quite different. The medical dataset is a lot larger than the heart dataset. The medical one contains around 10,000 records while the heart dataset contains only around 1,000. Since the heart dataset contains only 10% of the medical dataset, it can cause our results to be skewed quite heavily towards the medical dataset. Another issue that has come up with the dataset is the basic SQL joining that had to be done. Since the datasets are interpreting data differently (i.e. using Yes/No vs 1/0) it can cause a few problems in translation because how Tableau, or PowerBI, or Pivot Tables interpret the datatypes can cause weird glitches in joining the databases, as well as checking for certain datatypes.

Part 4: Sources

<https://colab.research.google.com/?utm_source=scs-index> Google Colab, Google

<https://www.w3schools.com/sql/sql_join.asp> SQL join commands, W3 schools

<https://stackoverflow.com/questions/37649/swapping-column-values-in-mysql> Swapping column values in MySQL, Paul Dixon & Artem Russakovskii

<https://stackoverflow.com/questions/1632792/how-do-i-compare-two-columns-for-equality-in-sql-server/1632831> How do I compare two columns for equality in SQL Server?, Ken White