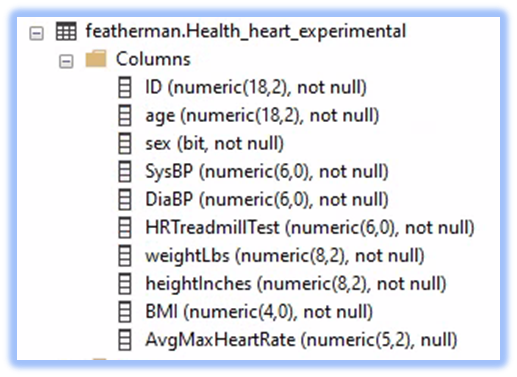
Practice With Arrays

Currently on DEVST05, but will be copied to DEVST06

This assignment has two parts as repetition is needed to practice with arrays and deriving insights and recommend-ations from a dataset. **Here arrays are required and are leveraged to produce more insightful analytics**. This is a muck around in the data assignment, so please do not expect to finish it in one sitting. Rather give yourself time to analyze the data in different ways until you find an insight worth reporting (in support of the questions provided). Columns from the table in the image to the right are pulled into an array, and new columns of metrics are created and used to analyze the dataset of > 70,000 patients.

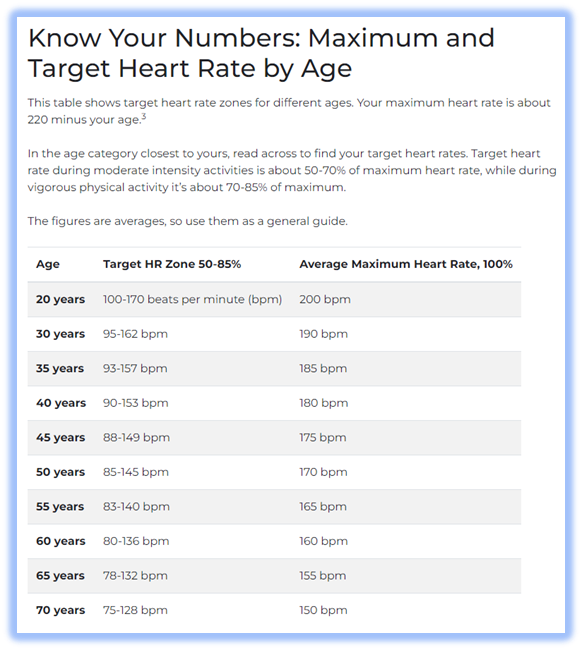
Part 1 – The health clinic was very happy with your report of the heart healthcare data. The health clinic is charged with guiding members towards healthier and happier lives. Start with your query from the prior related assignment and using the data in the table shown above load an array. As you work through this assignment you will have to create new columns in your array and do calculations using UPDATE SET commands to load data into the array. The goal is to build a great array that can then be analyzed using both SQL and PowerBI.  
  
*Inside your array, add new columns and then populate them by modifying your INSERT INTO or UPDATE SET statements. We are creating columns of analytics that can be used to build interesting charts to answer the following medical concerns:*

***Spring 2022 students and others that did not do the prior heart healthcare data analysis – here are the columns that you will need to add to your array.***

1. Write a SQL SELECT query that pulls all the columns from the table above (currently in the featherman\_analytics database on cb-ot-devst05.ad.wsu.edu…..also check devst06 as needed).

2. Notice the AvgMaxHeartRate is empty in the [Health\_heart\_experimental] table. When you query this table use a CASE statement to fill this column with values. The maximum heart rate for each of age group is shown in the image in the appendix. Th AvgMaxHeartRate column should display 85% of the published number for their age group.

3. USE SQL CASE statements to ***create three new columns*** each that categorizes the patients  
  
a) Make a new column that has age groups that match the groups defined in the appendix  
b) Make a new column that that has BMI groups that match the groups defined [here](https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html) or similar BMI chart.  
c) Make a new column that categorizes the results of each patients treadmill test. As discussed above a person whose treadmill test hit 199 these people are much less healthy than others whose treadmill test remained below the max for their age group (85% of the number in the image below). You create the categories for this column  
  
d) Add other columns to your SQL Select statement to perform some analysis (the analysis can either be performed in SQL or in PowerBI).

1. You should now have a column of the actual heartrate from a treadmill test, and a column of what the maximum heartrate should have been when exercising (.85 \* the # in the image to the right, for each age group).

Now let’s analyze the subjects of the study categorizing the subjects based on their treadmill test. Create a new column in your array that holds text. USE CASE() to write a new metric that puts subjects into groups based on their actual max heart rate on the treadmill test as compared to the max their age group. For example if the max heart rate for a 30 year old is 161, then you could put people into groups based on their treadmill test such as:  
  
a) reached or exceeded max heartrate (these are the subjects that should walk more often)  
b) reached 90% of max heart rate  
c) reached 70% of max heart rate  
d) reached only 50% of max heart rate on treadmill test (these are the really healthy subjects, walking on a treadmill did not spike their heart rate, because they are accustomed to the activity).

2. One of the medical directors read a study that reported that better than BMI, the simple metric of weight in pounds divided by height in inches yields a very accurate predictor of health. You are asked to create this Weight/Height metric. Do some analysis, comparing this metric to the groupings on the treadmill test

a) First create this new metric as a continuous variable, saving it to a new array column (numeric).  
  
b) To create categories based on the Weight/height metric add another new column to your array (text based), then use a case statement to break this variable into categories of your own design. You could then use these groups in a stacked column chart (added to the legend) or as the rows or columns in   
  
c) To further analyze subjects add another new column to your array (numeric). Alter your INSERT INTO statement to load this new column with the results of an NTILE(4) statement (that pulls the columns from the database) to create quartiles based on this metric. We need to check if the quartiles based on this health predictor are more insightful than the groups that we made with the CASE statement. Note: place the NTILE statement within the INSERT INTO @Array SELECT statement.

d) Analyze the data using these metrics and categories, and provide your analysis. Please do not complete this assignment in one sitting, rather give some thought to your analysis and insights. The difference between the grades B and an A is depth and insight. Try to make some inferences based on the available data, such as people in group 1 of the weight/height metric did the best on the treadmill test, and subjects in group 4 of the height/weight metric did the worst on the treadmill test).

3. The medical directors at the clinic are questioning the predictive validity of the treadmill stress test to identify subjects that are at risk for heart conditions. The treadmill stress test requires subjects get on a treadmill and walk until their heartrate peaks and they need to stop. According to conventional wisdom when the subject’s peak heart rate is more than 85% of the average maximum heart rate for their age group, then a blockage may be occurring in the subject’s heart.  
  
a) Create a column of your own design that categorizes the subjects into groups based on their treadmill results. Perform some research into the dataset using these categories.  
  
b) finally use an NTILE(4) statement to create quartiles based on this metric. Perform some research into the dataset using these categories. Note: place the NTILE statement within the INSERT INTO @Array SELECT statement.

e) Analyze the data using these metrics and categories, and provide your analysis.

4. One of the young doctors at the clinic discounted claims of an America’s obesity epidemic and claimed that the BMI scale is outdated and is not a valid predictor of heart health. Using the metrics created (and perhaps others of your own design) analyze the treadmill data (Treadmill % of Maximum heart rate), and whether there is a ‘correlation’ with BMI, WeightHeight ratio, or other metric or grouping that you can envision.

5. The same doctor that is questioning the BMI scale made the claim that many people now are large yet still aerobically very healthy (strong cardio-vascular systems) as evidenced by low treadmill results (much lower peak heart rate than then 85% cut-off). Another doctor claimed that they may believe this demographic change but not for subjects over 30 or 35 years old.   
  
A) provide your analysis of the treadmill results comparing that for different age groups. As this is a medical study based in facts, do not inject any of your opinions into the analysis or write-up.

**Part one Turn in**

The Director has asked for you to create a brief report to answer the questions posed above. Add additional analysis of your own design. Use a charts and tables adding a paragraph of text to answer each question posed above. Add your SQL code in the appendix of your turn-in using the heading Part One SQL code.

**Part Two**  
This portion of the assignment gives practice creating a big array and loading data into it, then doing some fun analysis, so try to have fun with this. Scenario: You and your friends have decided to take a post-graduation international travel adventure, but even after several conversations cannot decide on a location. Being known as an analyst that helps compile data to make better decisions, you have been asked to analyze some data to help with the decision of location choice. The following six tables have been uploaded to cb-ot-devst05.ad.wsu.edu (check 06 as needed) to start your analysis. (You may search out and upload different data).

|  |  |
| --- | --- |
| Text  Description automatically generated | Text  Description automatically generated |
| Text  Description automatically generated | Text  Description automatically generated |
| Graphical user interface, text  Description automatically generated | Text  Description automatically generated |

In this part of the assignment, you are asked to create an array that loads the data together as shown below. Use an INSERT INTO SELECT FROM query to choose the columns from one of the tables, then use UPDATE SET statements (use a WHERE statement that connects the array to the Country column of the database tables). (Alternatively, you can find your own data and upload it to your database using procedures shown [here](https://www.youtube.com/watch?v=K5_u6Xrbl_s) and in similar videos).

Graphical user interface, application

Description automatically generated

**Part 2 Turn-in Requirements**  
1. Add your SQL query into the appendix of your research document, using the heading Part Two SQL Statement.  
  
2. After you create an array, that holds all the columns shown above, add five new columns to your array. ***Either*** create new columns of metrics deriving new insights from a combination of the current columns ***or*** categorize the continuous data into new categories (i.e. grouping countries by GDP per capita because you want to go to a low GDP country where the hotel rooms are probably much less expensive).  
  
4. Add your interpretation of the insights that the data shown in the charts and tables presents. Merge together your insights to formulate the decision on vacation destination (i.e. we chose to include GDP per capita to guide our decision to go to a more affordable location, as Airbnb rates would probably be lower, than high GDP countries).  
  
5. At the end of your report, add your recommendation for top 3 vacation destinations, so that you give a variety of options to your group of friends.  
  
Higher grades given for completeness, professionalism, formatting and depth.  
  
**Final note**: While you may argue that the data could have been brought into Excel quite easily, with just a few hours of copy/paste; this practice merging datasets into a big array is an important skill. The person that can build datasets (typically a DBA) enjoys much higher job security and pay than the Excel copy/paste/format workers. The goal is eventually automated reporting using SQL processing. In practice, data can be transformed in a multitude of ways when it is integrated, allowing the creation of many new metrics, that would be difficult in Excel. Remember 99% of workers needing to produce insight are reliant on the 1% that can merge datasets and build columns of analytics.