Project 8.1

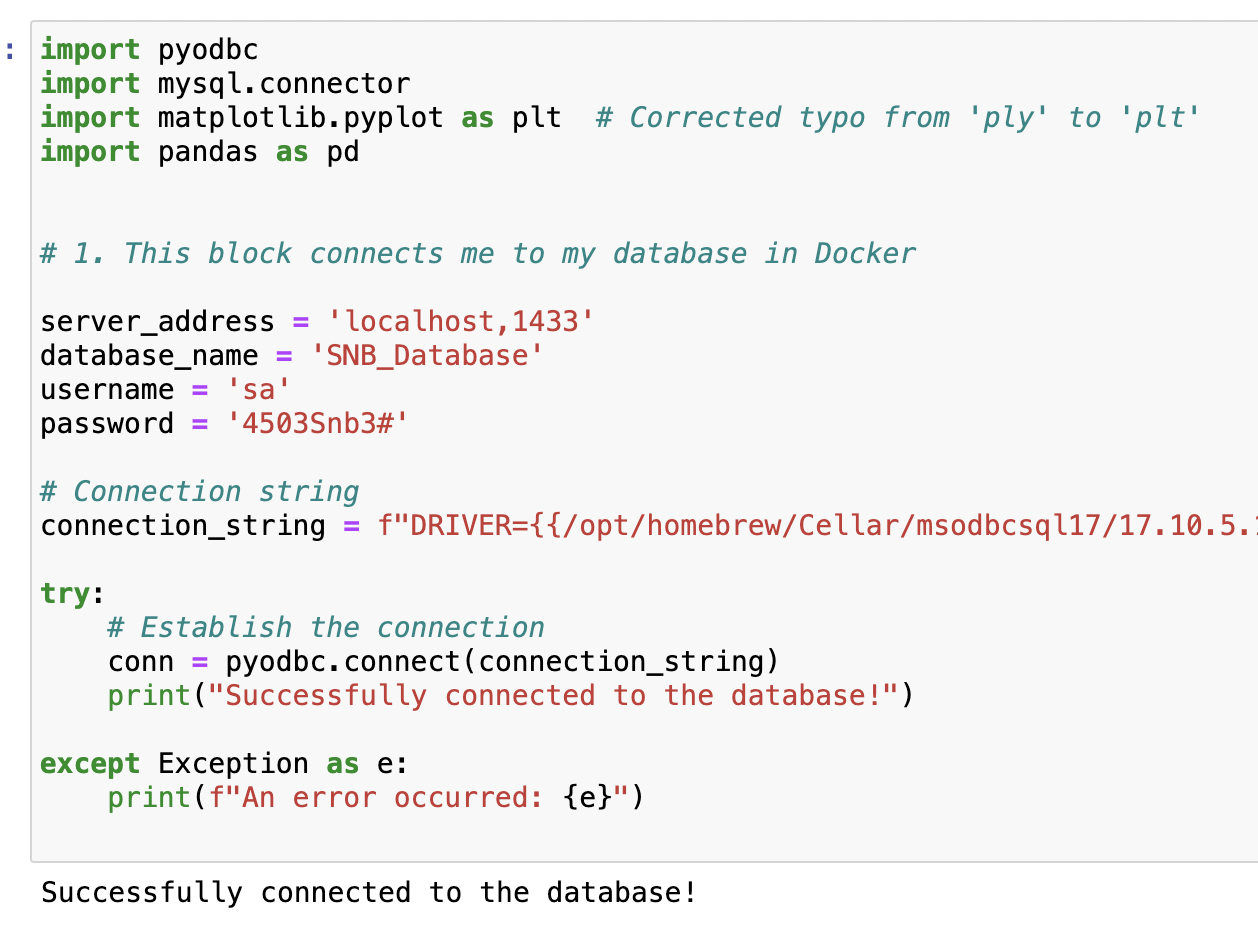
Here are the steps I took to execute the requirements of this project.

First, I ensured that I was able to connect Jupyter Notebook to my database in Docker. As mentioned in the Try it Activity 8.1, Docker wouldn’t work with my MacBook Air since my computer uses an Apple M1 chip. Apparently, there's a compatibility issue that's well know. I spent 20+ hours troubleshooting with Docker Hub, GitHub, Reddit, GPT4, etc. updating my Mac's operating system, downloading new (supposedly "M1-friendly") versions of Docker, and trying to create a container with server images compatible with ARM64/v8 architecture. Eventually, I found workarounds on MS’s website notes and Docker Hub to get an SQL server running in Docker to use with VS Code.

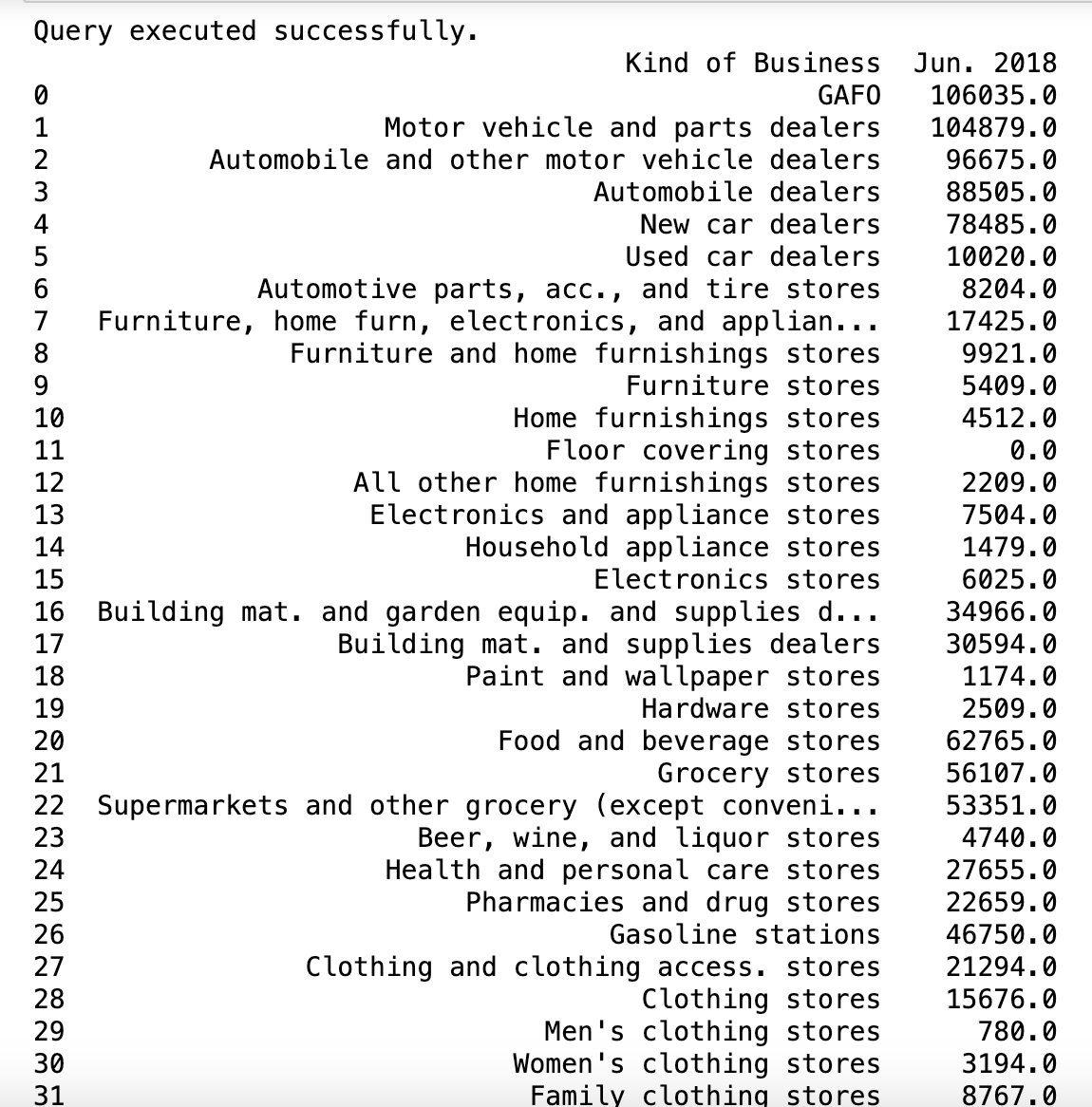
Once the server was running, however, it wouldn’t connect to VS Code or Jupyter Notebook due to an error with the ODBC Driver. Specifically, the driver was being stored on my computer (using homeBrew) in a different path than where VS Code and Jupyter Notebook expected it. Among other things, I attempted to do the work with Sqlite and Panda DataFrames, instead of SQL, but that had its own unique challenges.

Eventually, I got the ODBC Driver working after several hours.

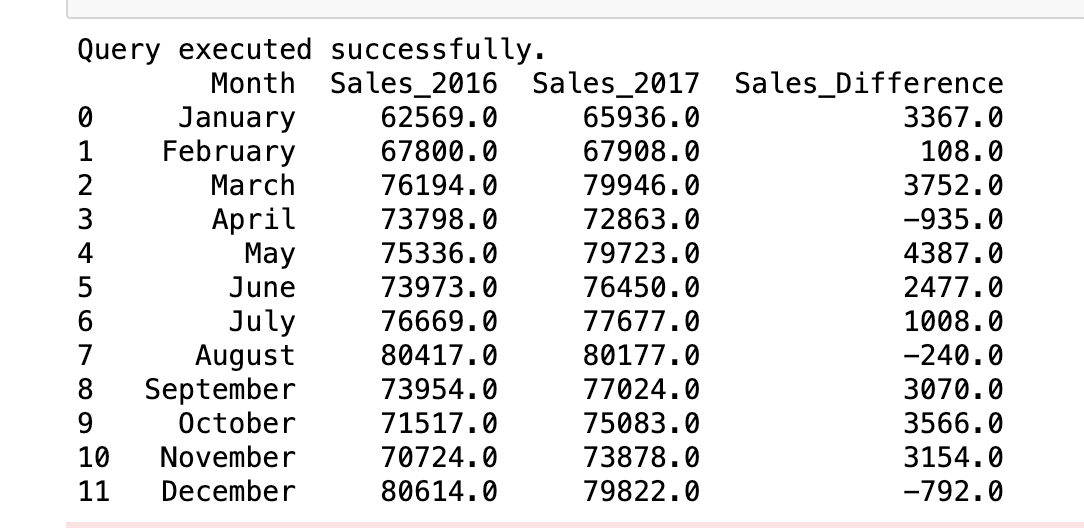
As you can see below, I was able to get Jupyter Notebook working with Docker for Project 8.1.



My next step was to run a simple query using python within Jupyter to get data for a particular month/ year for all the MRTS categories, which I did (please see code in my Jupyter Notebook submission).

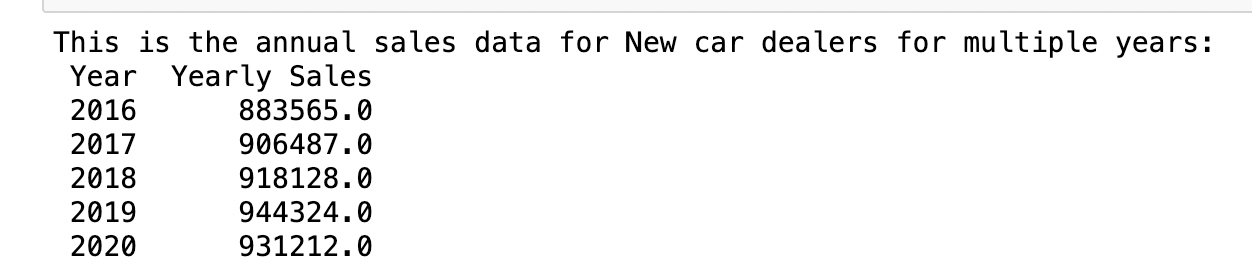


Once that was done, I expanded my queries to include a year-over-year analysis, by month, of total sales for all categories. Here is the output:

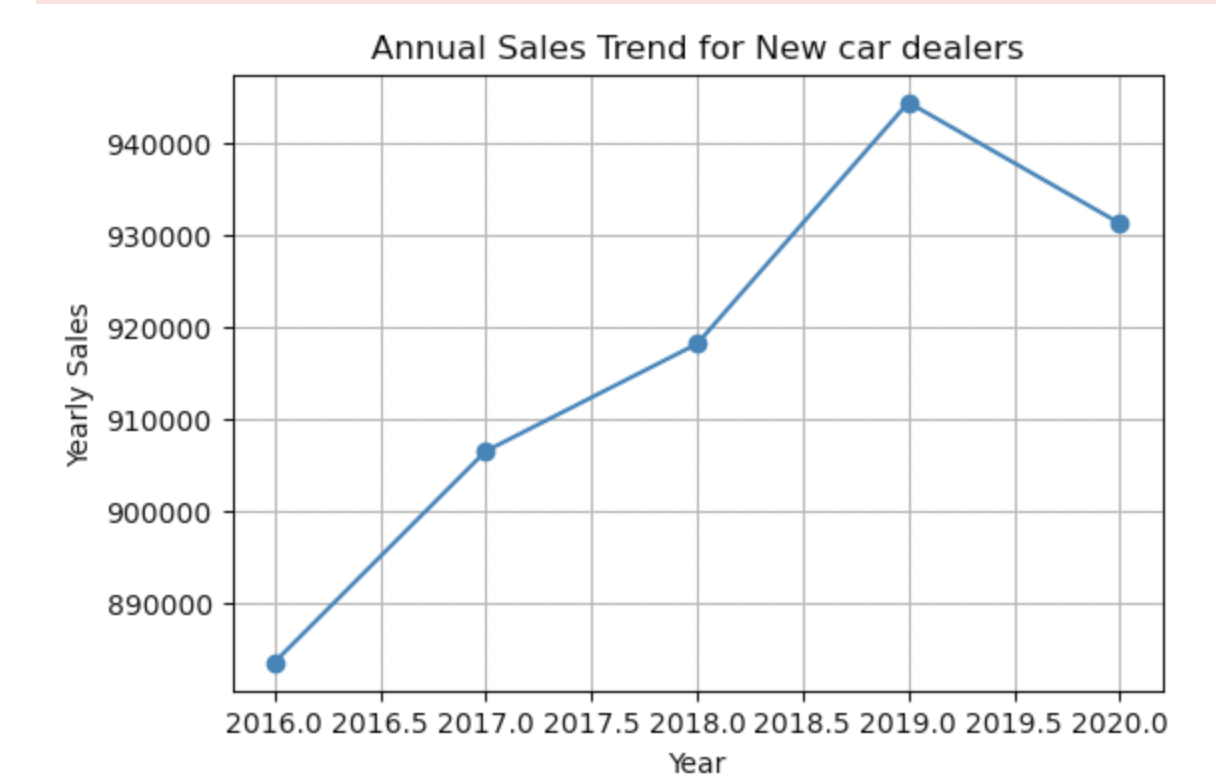


Please see Jupyter Notebook for the python code.

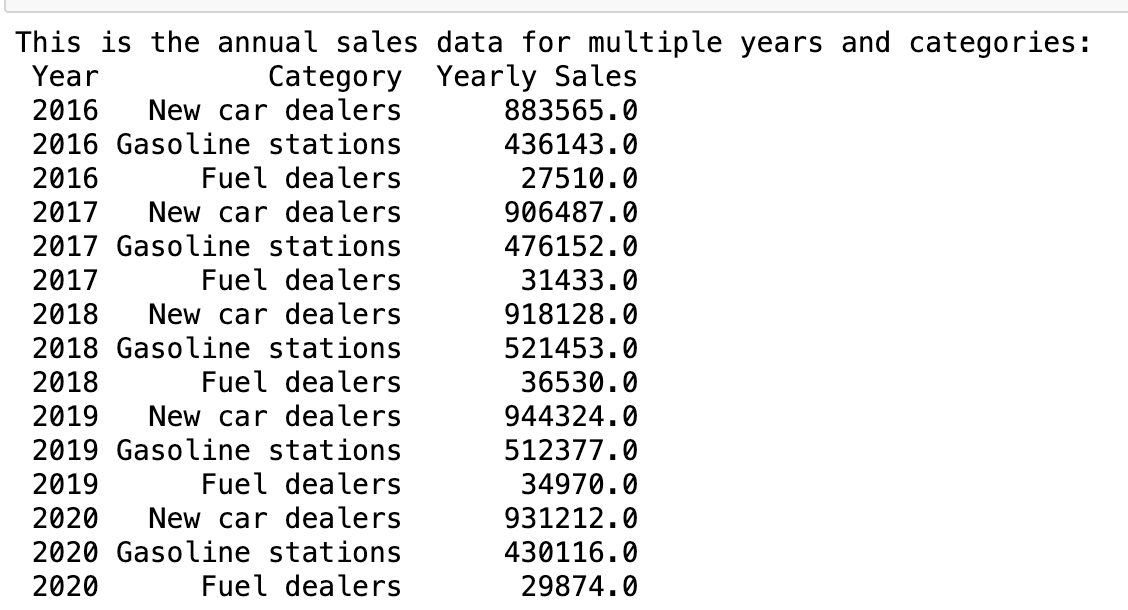
Then I decided to focus on total sales for only the MRTS category, “New car dealers”, for the five year period 2016 through 2020:



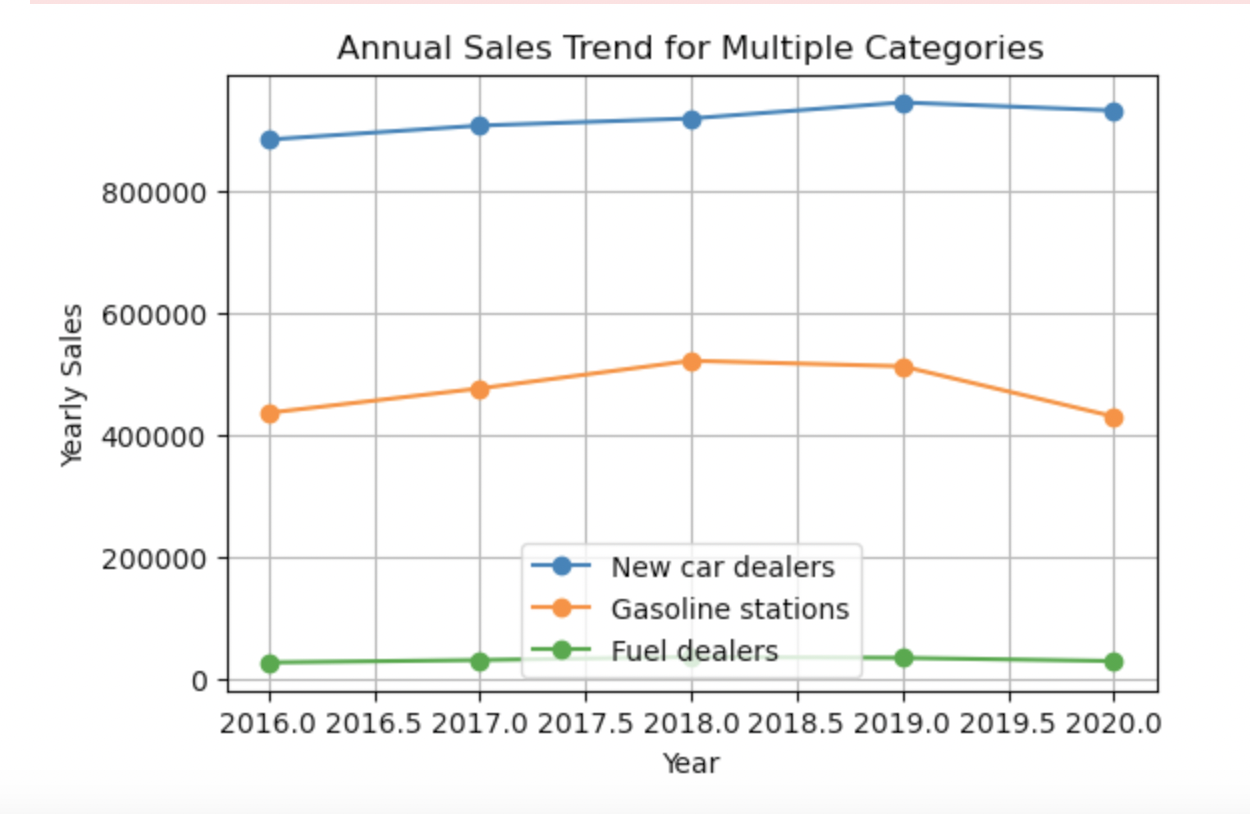
I also plotted this data:



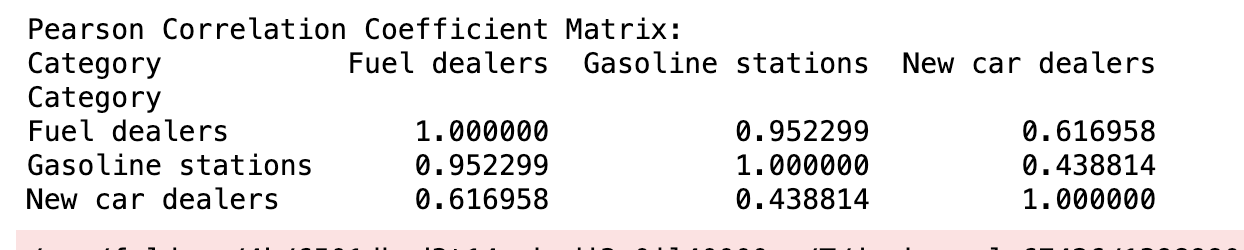
The next step was to introduce more MRTS categories and track them over time. I picked two related industries to “New car dealers”, namely “Gasoline stations” and “Fuel dealers”. As you can see, I ran a query to obtain each category’s total sales from 2016 through 2020.



I also plotted this data:



Finally, I decided to determine if these industries *really are* correlated. To that end, I calculated the Pearson Correlation Coefficient and created a heatmap of the results using Seaborn.



As you can see, there is indeed correlation among these three industries.

