ECON 8320 – Tools for Data Analysis
Assignment 7 [25 points]

Data Frames and SQL allow us to store our data very efficiently on a database, and to quickly retrieve and work with that data to solve real research questions. In this lab, we will work with raw data to practice importing data into Data Frames, and use SQL (via pandasql) to clean that data.

1. Import the data from assignment7.csv into a dataframe
2. Using pandasql, create an aggregated dataset, taking averages over the area variable
3. Which area code has the highest value for churn?
4. Without using pandasql, reduce the dataset to the following data:
    - Accounts in California (CA)
    - Accounts with accLen greater than 100
    - Only keep the area, vmMessages, dayCalls, eveCalls, nightCalls, intlCalls, and churn columns
    - Create a new column called allCalls, which is the sum of dayCalls, nightCalls, and intlCalls
    - Restric the dataset one more time to only contain the area, vmMessages, allCalls, and churn columns
5. Print the head of all three datasets, so that I can verify your results