Project 2: ETL

FitBit Data

After conducting research on Kaggle, our team decided to perform ETL on FitBit data. The FitBit Data folder we found had 13 files of various types of data collected for about 30 individuals. We reviewed all 13 csv files and decided to perform ETL on the Activity Data and Sleep Data. We chose these two data sets since we thought daily activity data and daily sleep data could lead to an interesting analysis of how daily activity could affect daily sleep and vice versa. All other 11 datasets included in the FitBit folder are interesting but not in scope.

Please see below for our step by step thought process while performing ETL on our FitBit datasets.

Extract:

1. Downloaded CSV files of both the activity data and sleep data from Kaggle to study which parts of the dataset would be most valuable.
   1. URL: <https://www.kaggle.com/arashnic/fitbit>
      1. Dataset 1: dailyActivity\_merged
      2. Dataset 2: sleepDay\_merged
2. Used Pandas in a Jupyter Notebook to pull both datasets from into a dataframe to more easily visualize and perform data cleanup.

Transform:

1. Checked data types of each column in the both datasets
2. Decided how and on what to merge the datasets, and began to clean the datasets individually to perform the merge
   1. Decided to merge on date and Individual ID columns since both were common on the Activity Dataset and Sleep Dataset
   2. Noticed that the SleepDay column in the Sleep Dataset contained multiple lines of data in one cell (date, time, am/pm)
   3. Utilized “split” functionality in pandas to split up all three components of the column into three separate columns called “date”, “time”, and “am/pm”
3. Merged the Datasets to create larger combined dataframe
   1. Merged on two different columns: “ID” and “Activity Date” / “date” since these were the two consistent columns across both datasets. Merging on these column would provide a seamless analysis between the Activity Data and Sleep Data.
   2. Performed an inner join to ensure that only data that was on both datasets were included in the new, combined dataframe
4. Checked the names and use of all columns in the newly created dataframe and began to drop columns that did not have significance to the analysis we hope can be performed from this data (for example: dropped the “date” column originally a part of the Sleep Dataset to avoid duplicate columns)
5. Verified the data types in all columns once more to ensure the load into the SQL Database would be smooth

Load:

1. Utilized the “fitbit\_clean.to\_sql('fitbit\_2', engine)” code to load our clean FitBit data into a relational database (SQL database (Postgres)) ready for others to use and analyze.
2. Double checked the formatting of all columns in the table to ensure that all data had been accurately transferred over and that the columns had the correct format