Katie Wood, Data Analyst

This Github repository contains one Python code sample, entitled `New York CitiBike data ETL + analysis.ipynb`, which runs an ETL pipeline on the New York CitiBike data set. The code then performs K-means clustering to identify groups of similar bike trips. Key results are displayed below:

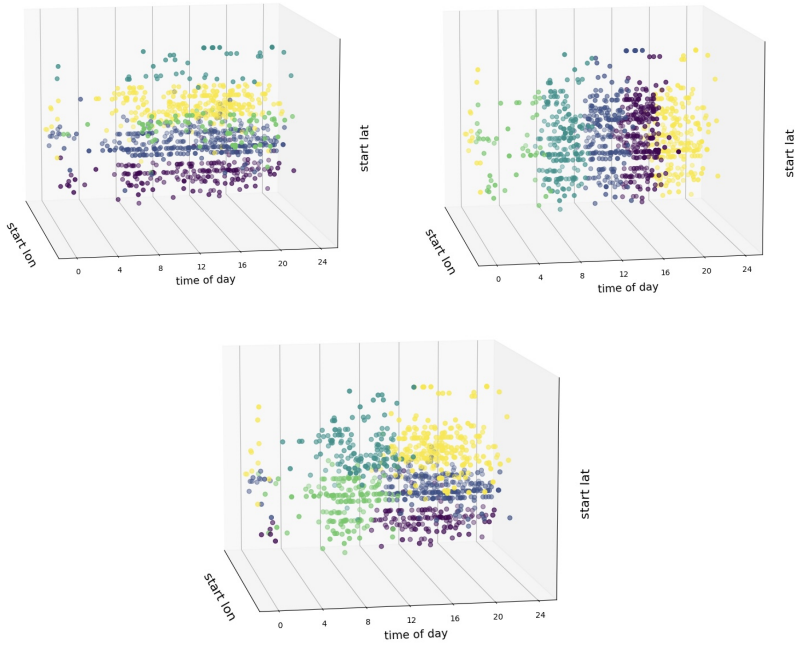

Figure 1: By tuning one parameter, we can discover geographic clusters (above left), temporal clusters (above right), and hybrid geo-temporal clusters (below center).

The repository also contains one SQL code sample, entitled `Internet Movie Database SQL queries.sql`, which analyzes the Internet Movie Database (IMDB). Key results are displayed below:

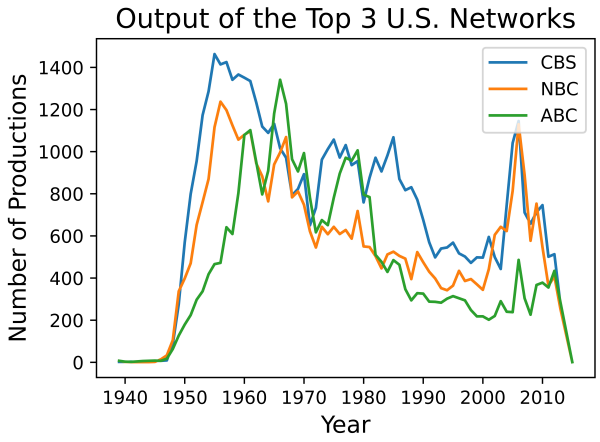

Figure 2: A sudden spike in the output of the top 3 U.S. television networks occurs just after 1950. Output then declines steadily for the next five decades, spiking again just before 2010.