**ETL Project**

**Group 12 team members:**

Ming Lei

Denise Person

**Objective**

Create a database that analysts can use to determine what affect the shape of the yield curve has on economic growth (represented by GDP.)

**Extraction**

Two data sets were used for this project. Monthly yield curve data for the period June 1990 to June 2019 was obtained from the U.S. Department of the Treasury. Data for the period 1970 to 2019 was sought; however, yield curve data earlier than 1990 was not available. Quarterly real and current GDP growth rates for the second quarter of 1947 to the first quarter 2019 were obtained from the U.S. Bureau of Economic Analysis. Both data sets were in csv format. Pandas was used to read the data into two separate dataframes using Jupyter Notebook.

**Transformation**

The first step of the transformation process was to delete any columns and rows that were not needed. Columns and rows deleted included summary information about the data set and NaN values. Deletion of the summary information was completed in EXCEL prior to reading the data into a Pandas dataframe. Formatting of date values were adjusted for both data sets as the dates were going to be used to join the two data sets.

**Yield Curve Data Set**

Ming Lei provide this section.

**GDP Data Set**

**Initial GDP DataFrame of CSV File**



In the GDP data set, the column headings were the quarterly date values. As the dates were going to be used to join the two data sets, the GDP dataframe was transposed so that the quarterly date values were in a column.

**Transposed GDP DataFrame**

****

The transposed GDP dataframe, which included other economic data in addition to GDP, was filtered with a new dataframe being created that included only the date, real GDP, and current GDP columns.

**Filtered GDP DataFrame**

****

**Load**

The final step was to load the data into a Database. The yield curve and GDP data were imported from their respective Pandas dataframe to populate the tables created in PostgreSql. Separate databases were created for the yield curve data and GDP data in PostgreSql. The two databases were merged to create a combined database. SQLAlchemy was used to connect to the PostgreSql database.

**Insert Combined Database**

**Summary**

The data in the database can be used in for creating various yield curve and GDP growth rate plots using Matplotlib. The analyst can use the plots to determine whether the historical data shows any patterns in the shape of the yield curve and economic growth or recessions. Regression analysis can also be performed to predict economic growth or recessions.