Loans & Birth/Death Rate Data Warehouse

Finding the Correlation between Loans and Birth/Death Rate

Group 2 Members

Victor Ou victor102496@gmail.com
Haonan Ou hn737433241@gmail.com
Soohun Han hsh804@gmail.com

CIS4400 – CMWA

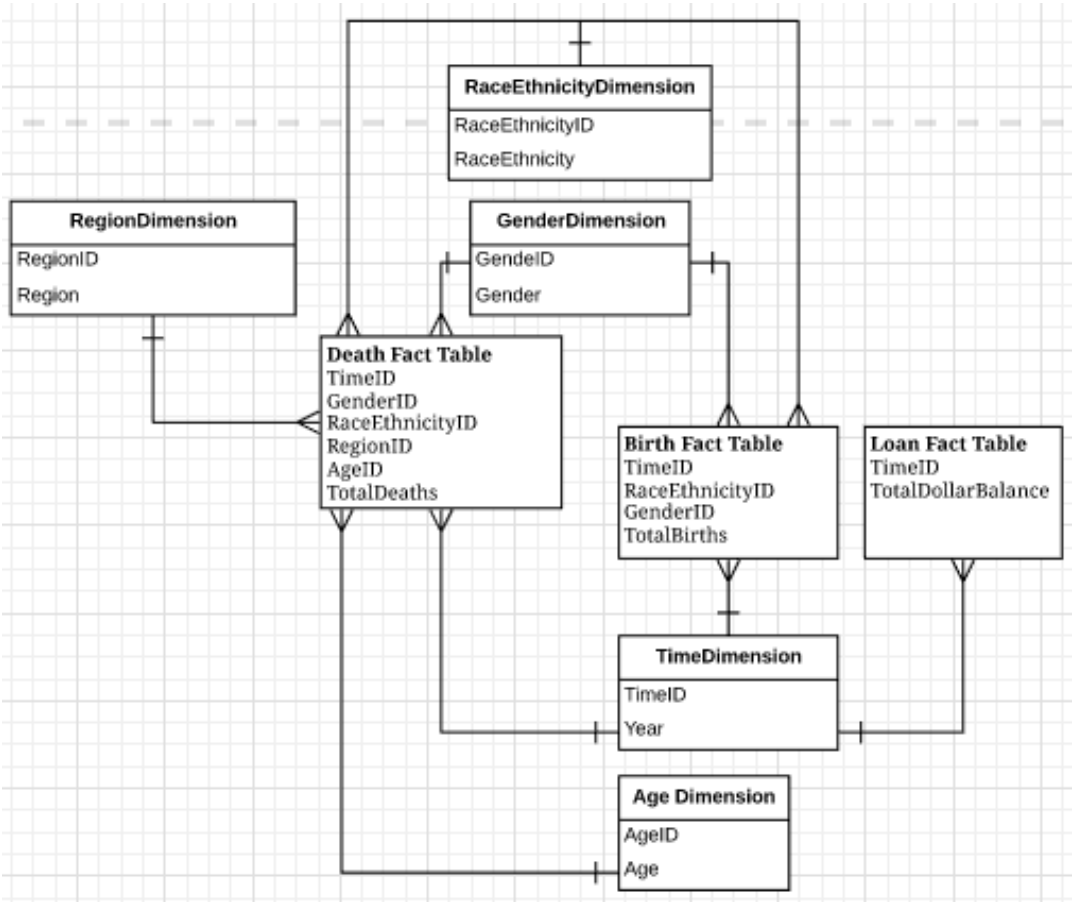
Project Description

Our project revolves around the effects loans may have on birth rates and mortality rate in New York. The economic environment over the past decade has forced many individuals and families to take out loans to either make ends meet or to invest in different opportunities. By analyzing annual birth rates with respect to continuous loan data, we hope to find trends that may help to paint an interesting narrative. Regarding the mortality rate, we want to specifically focus on deaths that occur in an unnatural manner, such as deaths relating to firearms, alcohol, and suicides. These unnatural occurrences, while morbid, may provide an insight on the effect's loans can have on an individual's mental health.

Information Needs

- Dataset source containing information on birth rates by age/race/gender spanning over years
- Dataset source containing information on mortality rates by age/race/gender spanning over years
- Continuous loan data over the span of multiple years containing data from multiple types of loans.

Dimensions and Fact Tables



Pentaho Data Integration (ETL Process)

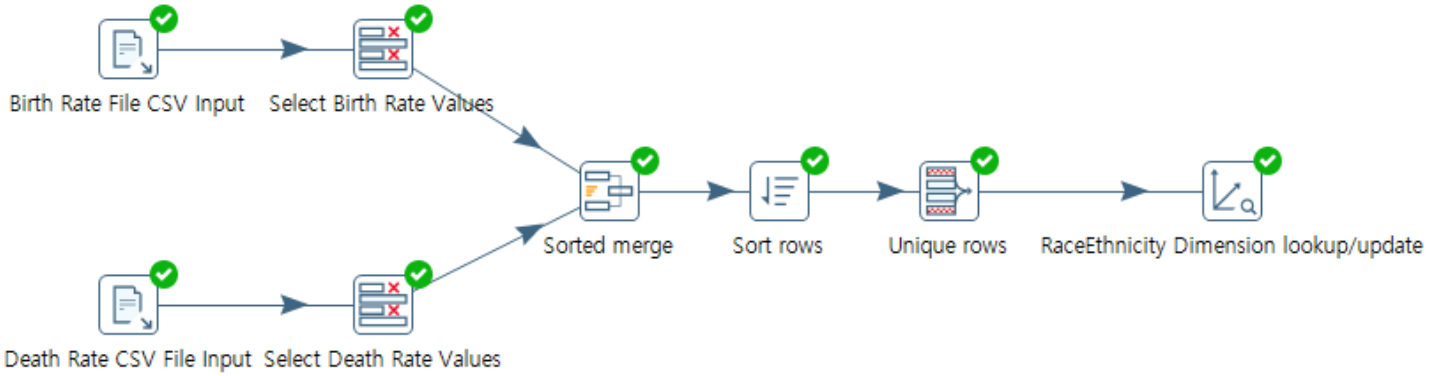
For data integration process, we used Pentaho Community Edition where we use Oracle as the target DBMS. We created a total of five dimensions including Region, Race Ethnicity, Gender, Time, and Age; and three fact tables including Loan, Death, and Birth fact tables.

The following are the steps we took to produce dimensions:

Region Dimension

For the region dimension, we used Death Rate dataset as the input. We used the Select Values process to select RegionID and Region attributes from the dataset. Next, we used Sort Rows to sort values in ascending order. Then we used Unique Rows process to eliminate duplicate data. Finally, we used the Dimension Lookup process where we set the Oracle as the target, added the surrogate key as region_dim_id with dimension type as Update, and created the dimension table named “Region Dim.”

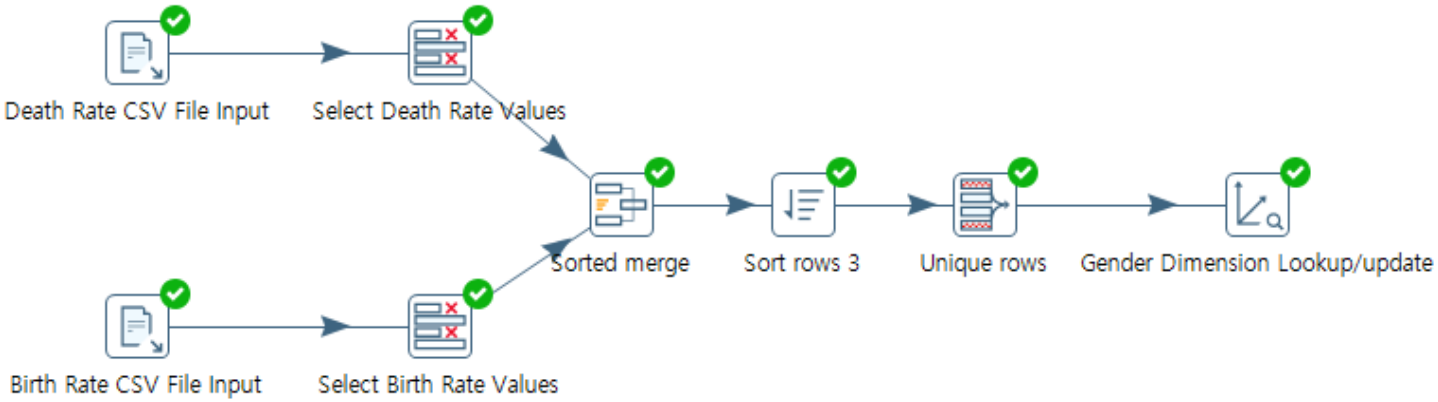
Race Ethnicity Dimension



For the race-ethnicity dimension, we created a conformed dimension using two datasets where we used Birth and Death Rate datasets as inputs. In the Select Values process, we selected RaceEthnicityID and RaceEthnicity attributes. We used Sorted Merge process to merge the two datasets and Sorted Rows to sort the data in ascending order. We used Unique Rows to get rid of duplicate data, and in Dimension Lookup process, we set

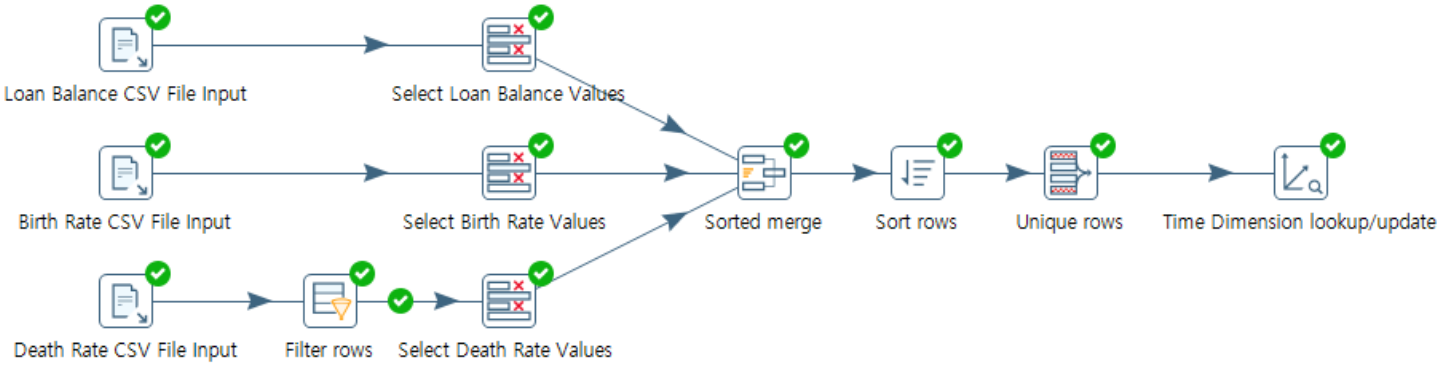
Oracle as the target system, added the surrogate key, race_ethnicity_id, as the type Update, and named the dimension table as “Race Ethnicity Dim.”

Gender Dimension



For the gender dimension, we created a conformed dimension using Birth and Death Rate datasets as inputs. We used Select Values to select GenderID and Sex attributes and merged using Sorted Merge. We sorted the data in ascending order in Sort Rows and de-duplicated the data using Unique Rows. Finally, we created a dimension table “Gender Dim” with a surrogate key gender_dim_id and type Update.

Time Dimension



For the time dimension, we created a conformed dimension using three datasets – Loan Balance, Birth Rate, and Death Rate. For Death Rate input file, we used the Filter Rows process to filter data with the following conditions:

```
TimeID >= [1]
AND
Year <= [2014]
AND
Region = [NYC]
AND
Year >= [2007]
```

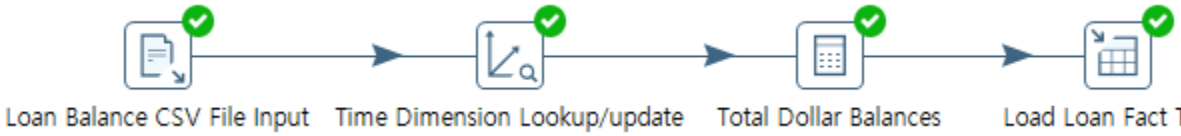
Then, we used the Select Values process to select TimeID and Year attributes from each dataset as merged in Sorted Merge. Using Sorted Rows and Unique Rows, we sorted the values in ascending order and removed duplicate values. Lastly, we created a “Time Dim” dimension table with time_dim_id as surrogate key and type Update.

Age Dimension

For the age dimension, we used Death Rate dataset as the input and selected AgeID and Age Group attributes using the Select Values process. In Filter Rows, we filtered out values where `AgeID = FALSE`. Then sorted values in ascending order and removed duplicates for AgeID values. In Dimension Lookup, we made the dimension table “Age Dim” with surrogate key age_dim_id and type Update and connected to the Oracle.

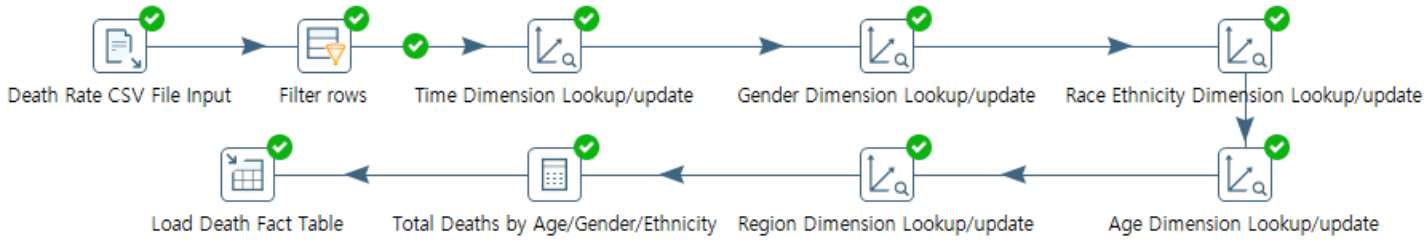
The following are the steps we took to produce fact tables:

Loan Fact Table



For the loan fact table, we used the Loan Balance dataset as the input. In Dimension Lookup process, we connected it to the time dimension. In the Calculator process, we created a LoanSum column for the sum value of HELOC, Auto, and Student Loan attributes; CreditOtherSum column for the sum of Credit Card and Other loan type attributes; and TotalDollar column where we store the sum of the LoanSum and CreditOtherSum columns. In the Table Output process, we created a fact table “Loan Fact” and set it to the Oracle target system.

Death Fact Table

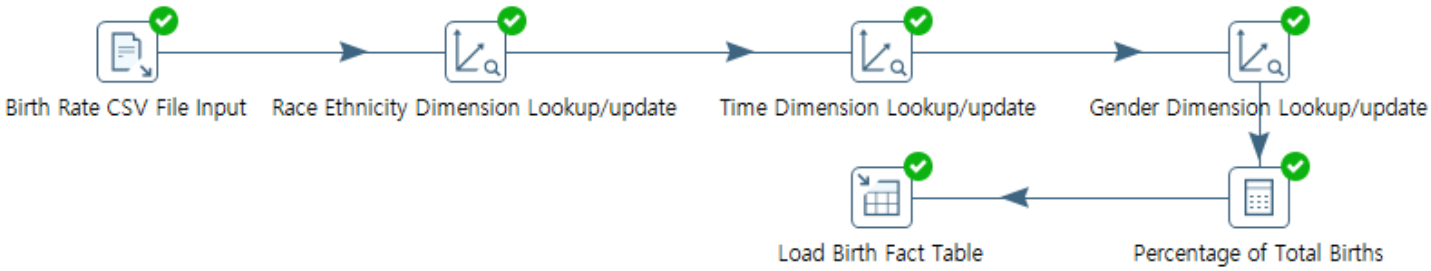


For the death fact table, we used the Death Rate dataset as the input and used Filter Rows process to filter out the following conditions:

```
NOT ( AgeID = [FALSE] )
AND
NOT ( TimeID = [FALSE] )
AND
NOT ( RegionID = [ROS] )
*Where ROS = Rest of the states (other than New York)
```

We connected it to Time, Gender, Race-Ethnicity, Age, and Region dimensions. In the Calculator process, we created a new field Total Deaths by Age/Gender/Ethnicity where we sum up values from Firearm, Alcohol-Related, and Suicide Deaths columns. Then, using the Table Output process, we load it into the fact table "Death Fact" and set the target to Oracle DBMS.

Birth Fact Table



For the birth fact table, we loaded the Birth Rate dataset as the input, and we connected it to the Race, Time, and Gender dimensions. In the Calculator process, we created two new fields Total Births – which holds a constant value of the Total Births during the entire period, and the Percentage of Total Births by Year/Sex/Race. In the Table Output process, we finalize our "Birth Fact" fact table and store it in Oracle DBMS.

Datasets Used

Debt data for years 04-14
https://data.world/finance/student-loan-debt

Percent Live Births by Infant Sex and Mother's Race/Ethnicity for New York City, 2007-2014
https://data.cityofnewyork.us/Health/Natality/wffy-3iyg

Vital_Statistics_Suicide_Deaths_by_Age-Group__Race_Ethnicity__Resident_County__Region_and_Gender__Beginning_2003
https://healthdata.gov/dataset/vital-statistics-suicide-deaths-age-group-raceethnicity-resident-county-region-and-gender-13#{view-graph:{graphOptions:{hooks:{processOffset:{},bindEvents:{}}}},graphOptions:{hooks:{processOffset:{},bindEvents:{}}},view-grid:{columnsWidth:[{column:!Firearm++Deaths,width:268},{column:!Alcohol-Related++Deaths,width:250},{column:!Suicide++Deaths,width:225}]}}