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.

KPI Documentation

Our main KPIs includes the total balances of births, deaths and individual loan balances and the total loan balances across the seven years of data. We do this by aggregating the balances of each loan categories: auto loan, HELOC, credit card, and student loan per year as well as the total dollar amount for every loan and we also do the same for births and deaths categorized by race, gender, and age groups.

- KPIs including:

  Total Dollar Balance of Loans

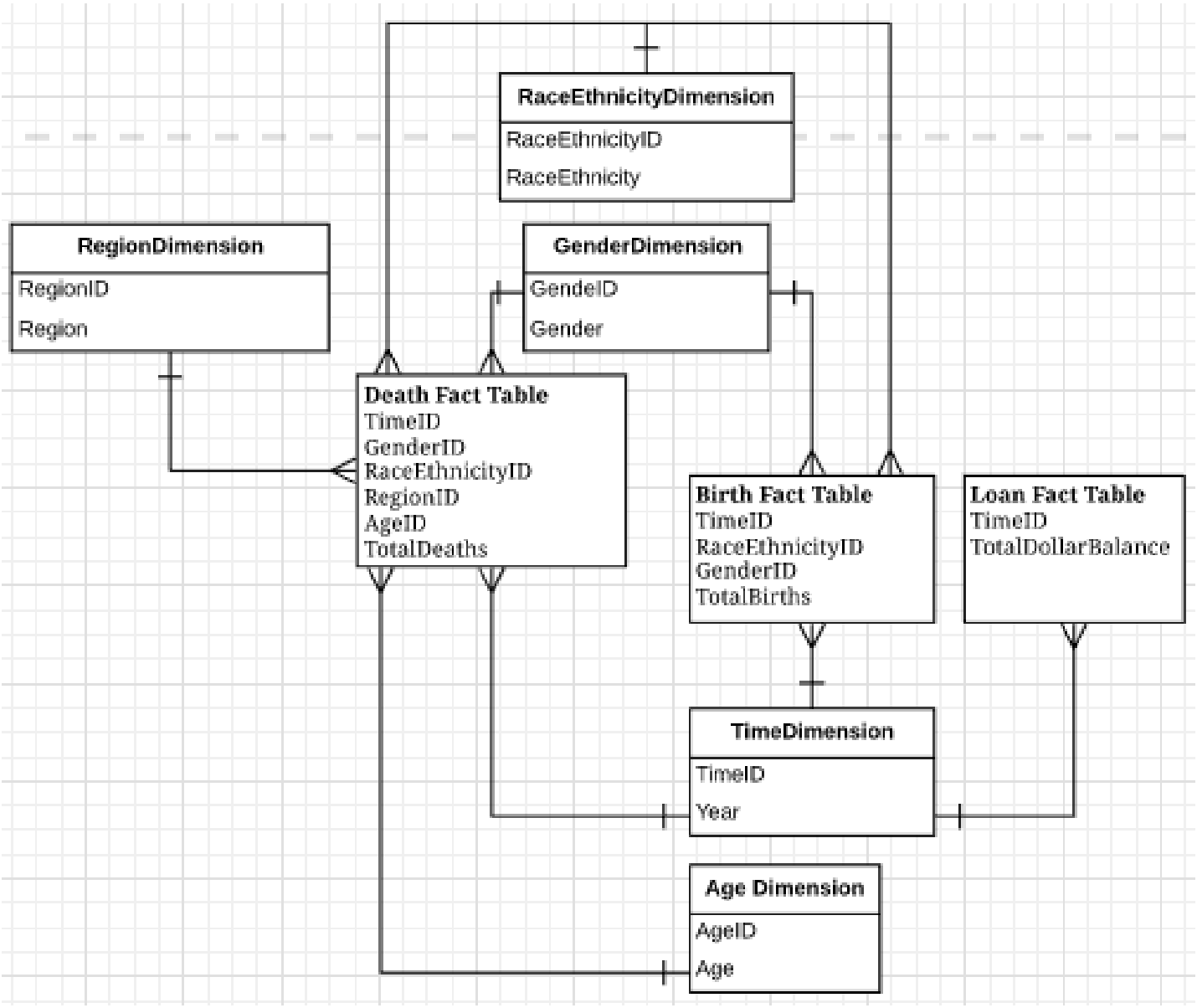
  Total Auto loan balance

  Total HELOC balance

  Total Credit Card Balance

  Total Student loan balance

Dimensions and Fact Tables



Pentaho Data Integration (ETL Process)

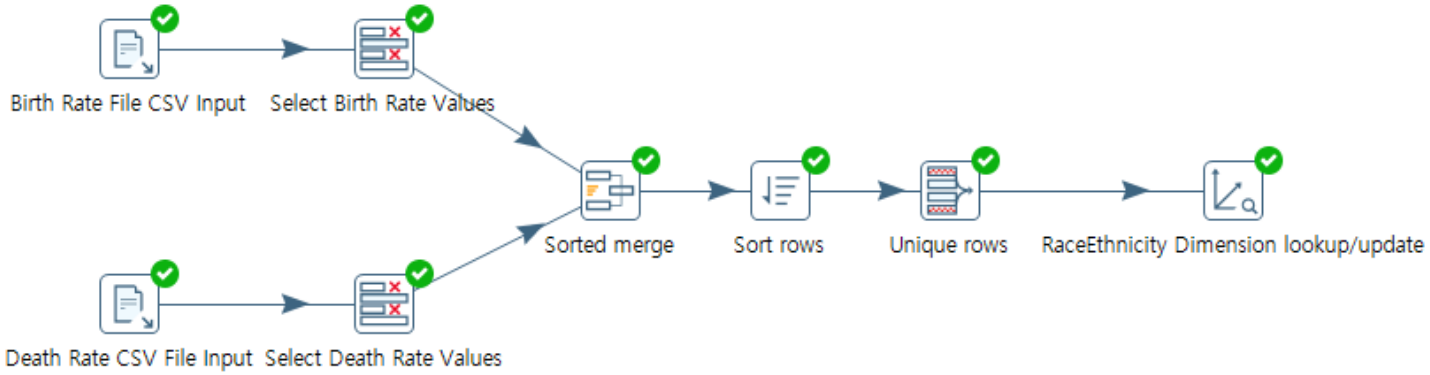
For data integration process, we used Pentaho Community Edition where we use Oracle as the target DBMS. We used Oracle DBMS because using the H2 database might not guarantee access to BI tools and Tableau. We created a total of five dimensions including Region, Race Ethnicity, Gender, Time, and Age; and three fact tables including Loan, Death, and Birth fact tables.

The following are the steps we took to produce dimensions:

Region Dimension

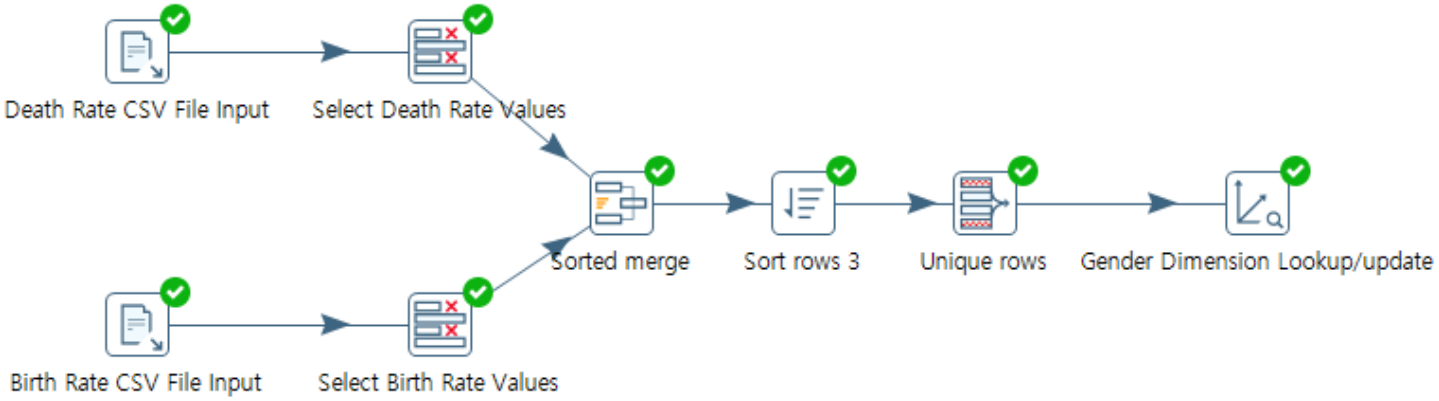
For the region dimension, we used Death Rate dataset as the input. We used the Select Values process to select RegionID and Region attributes from the dataset. Next, we used Sort Rows to sort values in ascending order. Then we used Unique Rows process to eliminate duplicate data. Finally, we used the 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



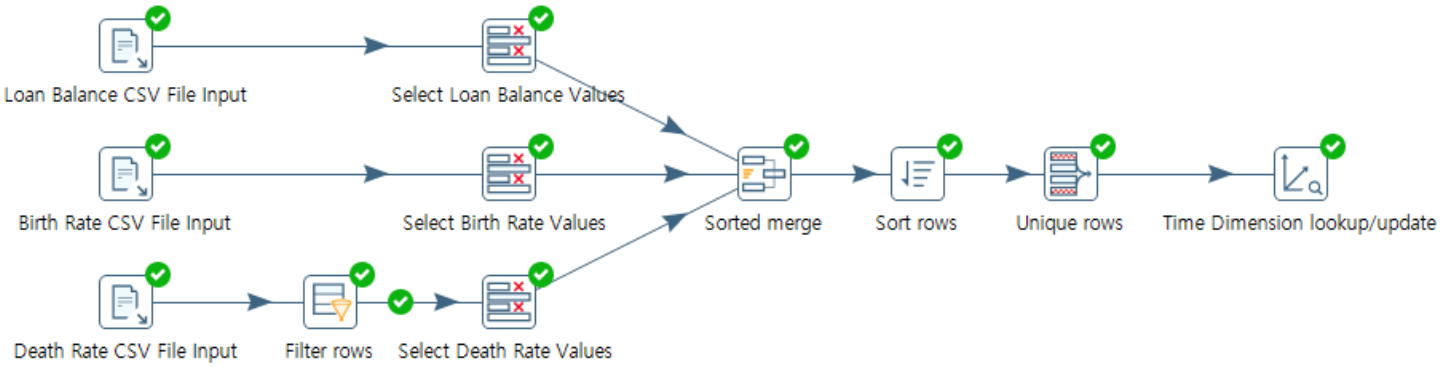
For the race-ethnicity dimension, we created a conformed dimension using two datasets where we used Birth and Death Rate datasets as inputs. In the Select Values process, we selected RaceEthnicityID and RaceEthnicity attributes. We used Sorted Merge process to merge the two datasets and Sorted Rows to 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

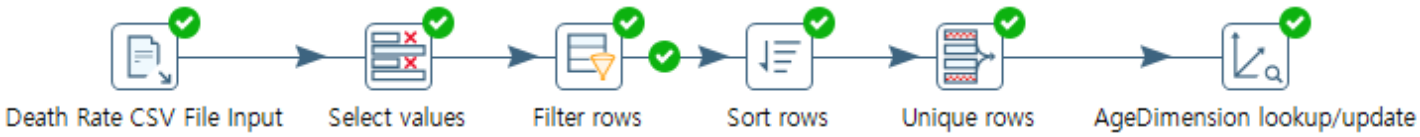


For the time dimension, we created a conformed dimension using three datasets – Loan Balance, Birth Rate, and Death Rate. For Death Rate input file, we used the Filter Rows process to filter data with the following conditions:

```
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.

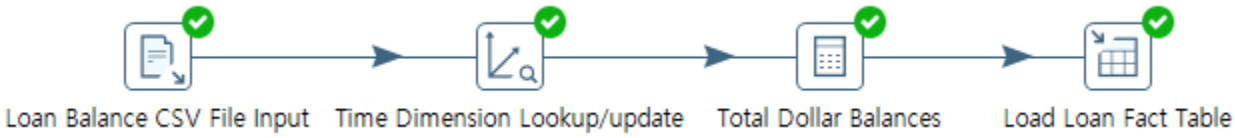
Age Dimension



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.

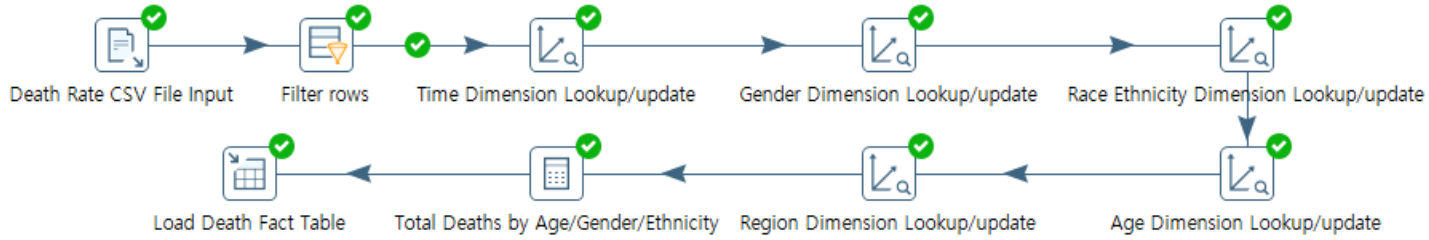
The following are the steps we took to produce fact tables:

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

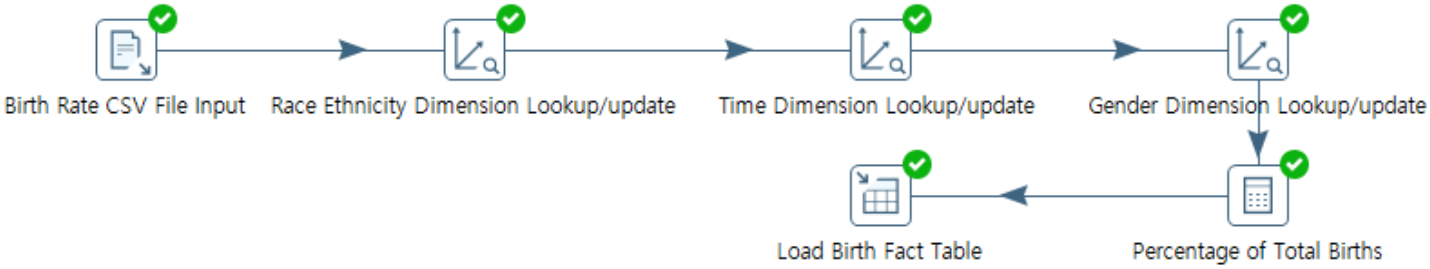


For the death fact table, we used the Death Rate dataset as the input and used Filter Rows process to filter out the following conditions:

```
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

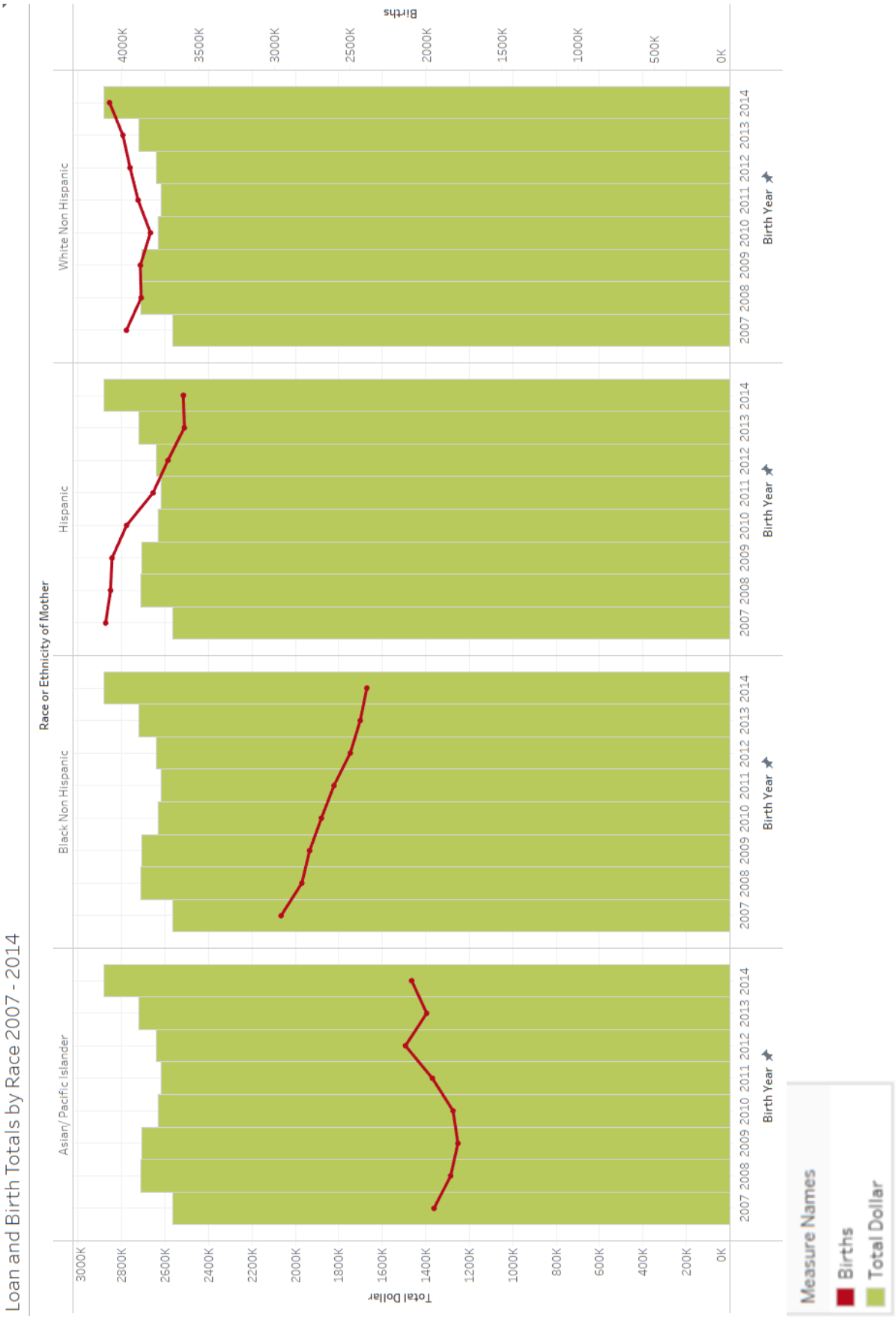


For the birth fact table, we loaded the Birth Rate dataset as the input, and we connected it to the Race, Time, and Gender dimensions. In the Calculator process, we created two new fields Total Births – which holds a constant value of the Total Births during the entire period, and the Percentage of Total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

Dashboard Application

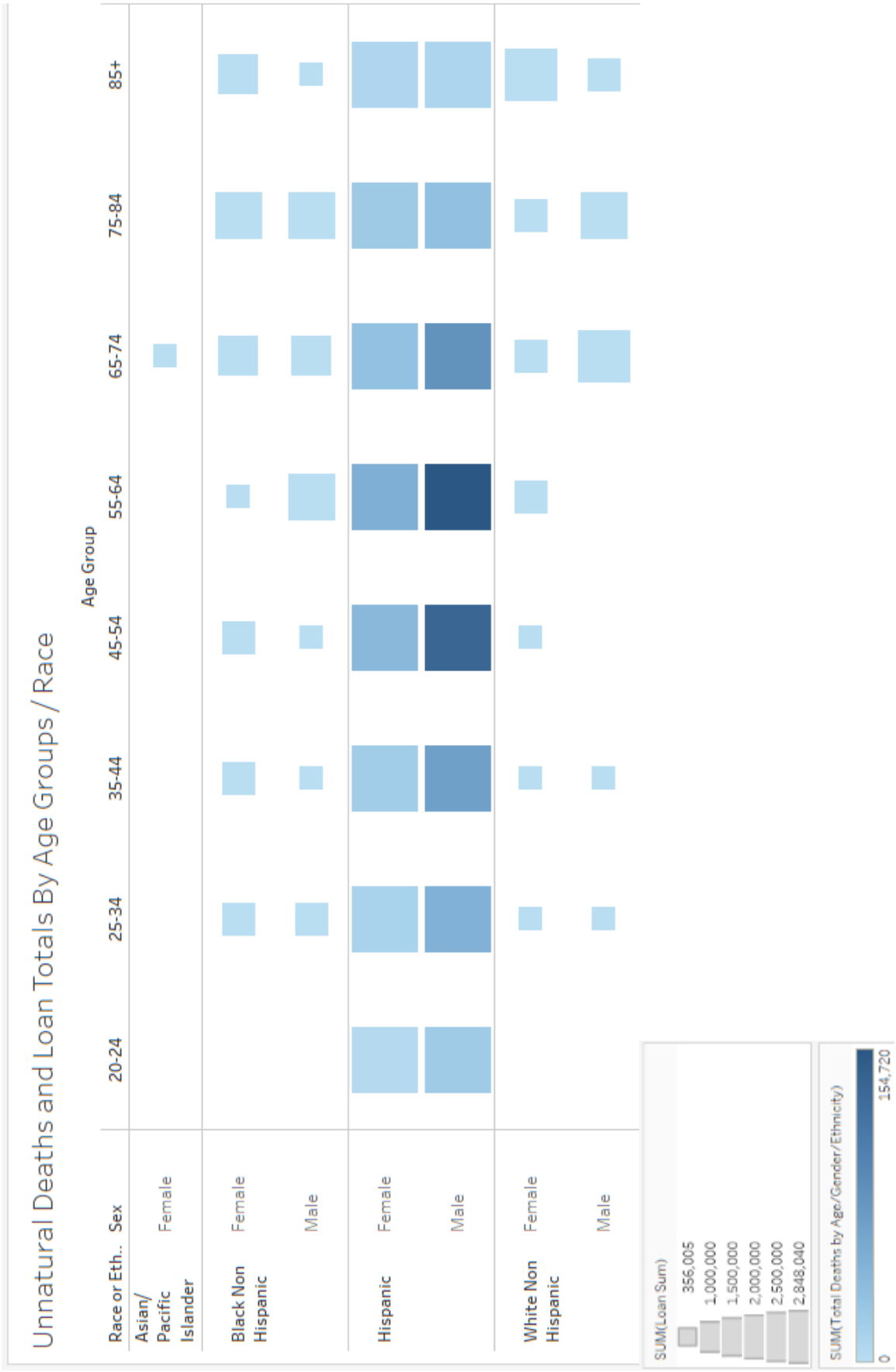
Our Dashboard containing our three graphs and charts “Unnatural Deaths and Loan Totals by Age Groups / Race”, “Total Unnatural Deaths and Total of Loans 2007 - 2014”, and “Loan and Birth Totals by Race 2007 / 2014” was created in Tableau.

Loan and Birth Totals by Race 2007 - 2014 Combination Graph



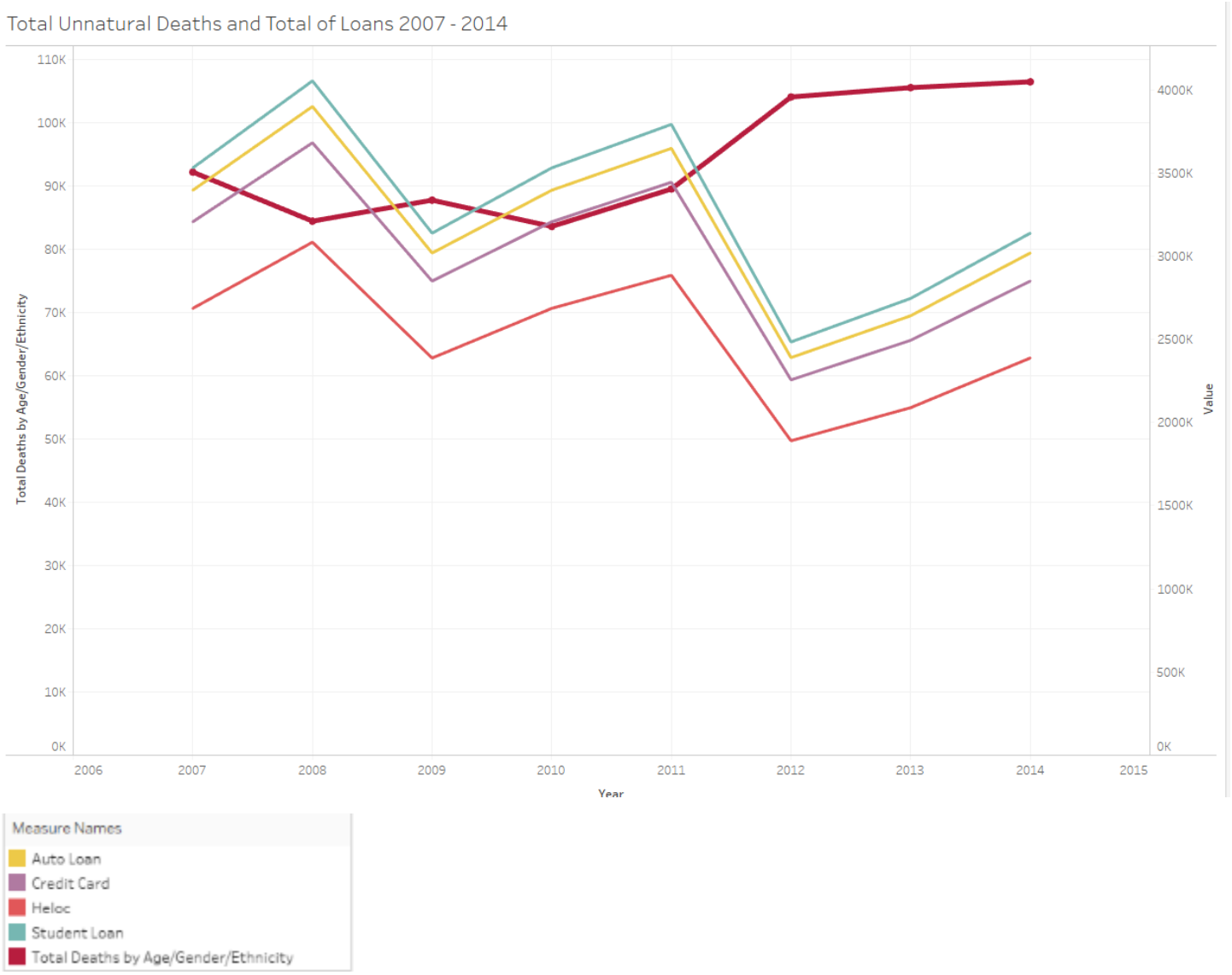
Our first visualization titled Loan and Birth Totals by Race 2007 - 2014, we decided to use a trending / correlation hybrid analysis-based graph to display our data using a combination chart of a bar graph and a line graph. There are two y-axes, the left axis contains the loan totals of each individual loan type represented by the green bars and the right axis is the total births from our birth dataset represented by the red line graph. These graphs are categorized by the race/ethnicity of the mothers. The x-axis displays the years from 2007 - 2014. From this visualization, we can see for Black Non-Hispanic and Hispanics, there is a steady drop in birth rates in relation to the increase of total loan amounts. However, it seems Asian/Pacific Islander and White Non-Hispanic people had an overall increase in birth rates from 07 - 14, but there is an initial decline.

Unnatural Deaths and Loan Totals by Age Groups / Race



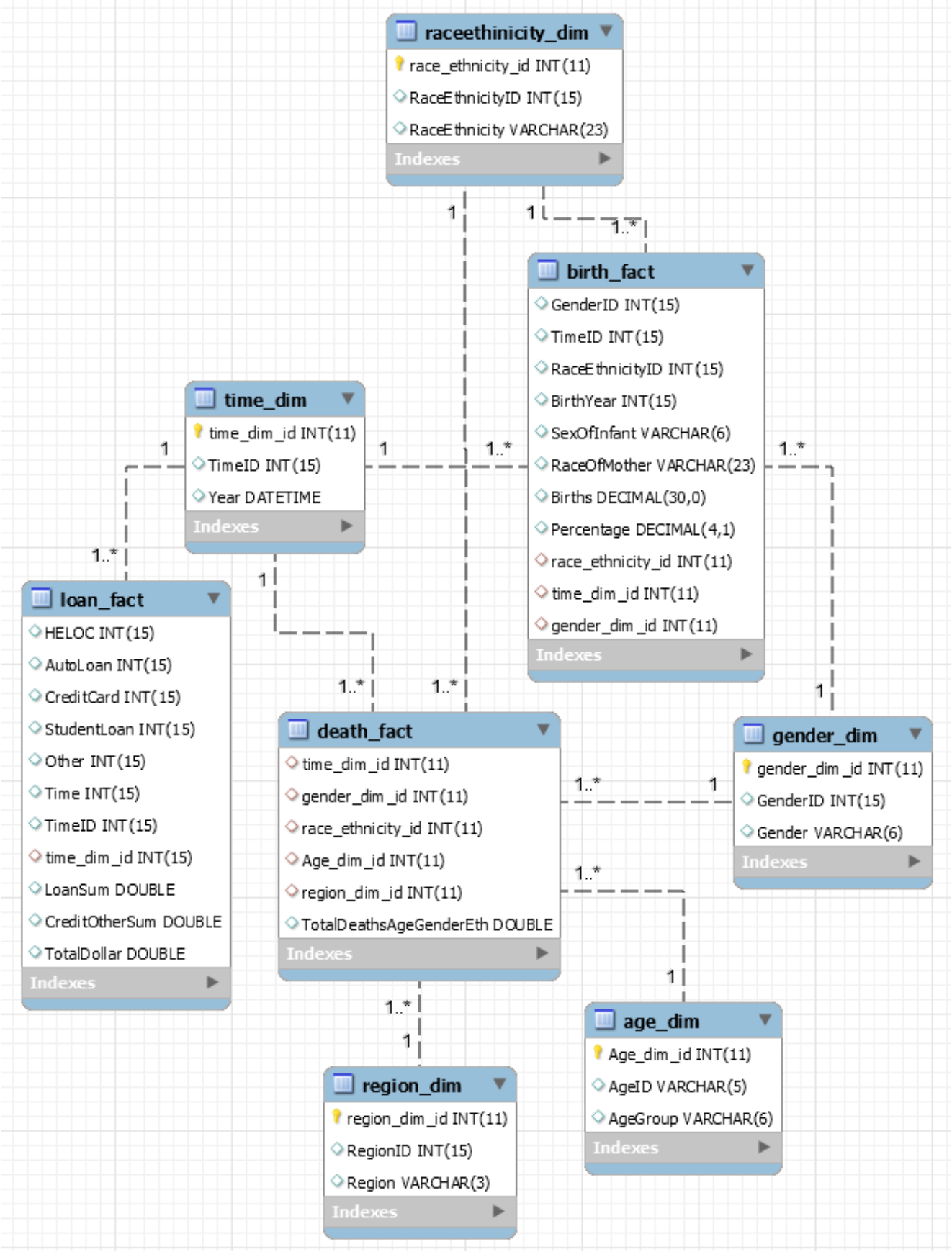
Our second visualization is a contribution analysis using a heatmap visualization that displays the total number of unnatural deaths categorized by Race/Ethnicity, Gender and Age group. The darker the color of the squares, the denser the death amount. The size of the square represents the total loans of each individual age groups separated by gender and race. According to this visualization, the age groups between 20-24 experience the least amount of deaths/loans and Asian/Pacific Islanders, apart from the age group 65-74, do not see any unnatural deaths related to loans. The more daunting display is the Hispanic group ages 45-64 that see very dark colors and abundant large squares spanning across generation representing high numbers of loans and unnatural death rates.

Total Unnatural Deaths and Total of Loans 2007 - 2014



In our final visualization, we wanted to display the separate loan totals represented by different colored lines using a trending analysis line graph. We wanted to see the relationship each loan totals have with the total death rates from unnatural causes. This graph once again has two y-axis, the left axis represents the total deaths by age/gender/ethnicity, while the right y-axis represents the total dollar amount of each loan in billions. The x-axis is the year in which the data is stored in. There doesn′t seem to be much correlation between total deaths and loan amounts until the year 2012, where each loan total jumped up significantly and we see a gradual climb in unnatural death causes. However, previous numbers do not seem to back this trend up, therefore, our findings are inconclusive, and we cannot say that an increase in loan numbers correlates to unnatural causes of death.

Physical Model



To build the physical model, we created SQL commands for each table from the dimensional model. We set the primary key for each dimension table and connected them with fact tables by assigning the primary keys in dimension tables as foreign keys in fact tables. We used the following ALTER TABLE statements to add in foreign keys into fact tables:

DEATH FACT FK ALTER STATEMENTS

```
ALTER TABLE Death_Fact ADD CONSTRAINT time_dim_id_fk FOREIGN KEY
time_dim_id REFERENCES Time_Dim(time_dim_id);
ALTER TABLE Death_Fact ADD CONSTRAINT gender_dim_id_fk FOREIGN KEY
gender_dim_id REFERENCES Gender_Dim(gender_dim_id);
ALTER TABLE Death_Fact ADD CONSTRAINT race_ethnicity_id_fk FOREIGN KEY
race_ethnicity_id REFERENCES Race_Dim(race_ethinicity_id);
ALTER TABLE Death_Fact ADD CONSTRAINT Age_dim_id_fk FOREIGN KEY Age_dim_id
REFERENCES Age_Dim(Age_dim_id);
ALTER TABLE Death_Fact ADD CONSTRAINT region_dim_id_fk FOREIGN KEY
region_dim_id REFERENCES Region_Dim(region_dim_id);
```

LOAN FACT FK ALTER STATEMENTS

```
ALTER TABLE Loan_Fact ADD CONSTRAINT time_dim_id_fk FOREIGN KEY
time_dim_id REFERENCES Time_Dim(time_dim_id);
```

BIRTH FACT FK ALTER STATEMENTS

```
ALTER TABLE Birth_Fact ADD CONSTRAINT time_dim_id_fk FOREIGN KEY
time_dim_id REFERENCES Time_Dim(time_dim_id);
ALTER TABLE Birth_Fact ADD CONSTRAINT gender_dim_id_fk FOREIGN KEY
gender_dim_id REFERENCES Gender_Dim(gender_dim_id);
```

```
ALTER TABLE Birth_Fact ADD CONSTRAINT race_ethnicity_id_fk FOREIGN KEY
race_ethnicity_id REFERENCES Race_Dim(race_ethinicity_id);
```

Conclusion

The most difficult part of the assignment was the ETL process

We were overwhelmed with the filtering steps and other tools we had to use to make sure our data was consistent throughout. A lot of our dimensional tables were low cardinality, but they were still vital and needed to be created. The easiest part of the assignment was creating the charts and graphs in Tableau.

Once we finished the ETL process, creating the graphs that helped answer some of our initial business question was satisfying and not too difficult. The majority of our group had taken the data visualization course, and one member is currently taking the class, so we were all familiar with using Tableau.

We each learned the difficulties of filtering out data and data profiling in respect to raw data. We were all so used to working with data that had already been cleaned, so we needed to take a step back and analyze the situation. In the beginning, we had not considered the data profiling step would be this grueling, so it was a great experience to learn the different ways accomplish this task.

If we had to do it over again, we would have liked to try Pentaho's visualization tool to create our graphs to take advantage of working with new software. We would also have liked to have chosen a topic that had data records that were more substantial, like we mentioned earlier, a lot of our dimension tables were low in cardinality which gave us trouble in our initial ETL process.

If the proposed benefits can be realized by the new system

In respect to our business requirement questions, I feel as though our system answered most of them. Birth rates were generally more affected by loan amounts than death rates. This knowledge and information can perhaps redirect some resources towards mental health for those who have incurred significant debt.

Final Comments

Though difficult and headache-inducing at times, working on an involved project with real datasets was extremely rewarding. We learned to utilize different tools to accomplish business task that is applicable in working environments.

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}]}}