Extraction, Transformation, and Load Technical Report

“Fiscal Heartbreak”

**Team Members:**

Joseph Ayala

Andrew Behrman

Sean Findlay

Michael Fox

Michael Hankinson

09-12-2019

**TABLE OF CONTENTS**

[**Summary**](#_heading=h.30j0zll) **3**

[**Scope**](#_heading=h.3znysh7) **3**

[**1.2.1 Marriage and Divorce Statistics**](#_heading=h.ood4currpyt4) **3**

[**1.2.2 Debt Statistics**](#_heading=h.nrv094wwkk6h) **4**

[**Technologies and resource contributions**](#_heading=h.tyjcwt) **4**

[**Definitions, Acronyms and Abbreviations**](#_heading=h.1t3h5sf) **5**

[**Data Import/Extract Sources and Method**](#_heading=h.2s8eyo1) **6**

[**Data Acquisition**](#_heading=h.3rdcrjn) **6**

[Marriage Data Acquisition](#_heading=h.jgbzxcrdgvyx) 6

[Debt-to-Income Data Acquisition](#_heading=h.kc9ii9kei8m) 6

[**Data Transform**](#_heading=h.lnxbz9) **6**

[Marriage Data Transform](#_heading=h.dkc7ao4m4g2e) 7

[Debt-to-Income Data Transform](#_heading=h.q91ewwd98v8y) 7

[**Data Integrity**](#_heading=h.1ksv4uv) **7**

[Marriage Data Integrity](#_heading=h.bjs1rvlx0zcj) 7

[Debt-to-Income Data Acquisition](#_heading=h.jx5rzu66etpd) 7

[**Data Refresh Frequency**](#_heading=h.2jxsxqh) **7**

[**Data Security**](#_heading=h.3j2qqm3) **8**

[**Data Loading and Availability**](#_heading=h.knhiwdag34bm) **8**

[**Data Quality**](#_heading=h.ifl1ikhrn9ew) **11**

**1. INTRODUCTION**

*The purpose of the Extraction, Transformation, and Load (ETL) Technical Report is to capture details that pertain specifically to the ETL portion of the data pipeline that is to be used in the Fiscal Heartbreak project. It will detail the data sources, the methods employed to acquire the data, any transformations necessary, and the resulting target tables now available for analysis.*

# Summary

“Fiscal Heartbreak” will analyze if the US fiscal policy has an impact on the country’s family units. More specifically, trends of consumer debt will be compared to trends of divorces and marriages by county and by year over a 9-year span.

**Preliminary Hypothesis:** *Higher levels of consumer debt lead to higher rates of divorce and lower rates of marriage.*

**Null Hypothesis:** *There is no relationship between consumer debt and divorce rates nor marriage rates.*

# Scope

Two primary data sources will be leveraged for this project: i) Marriage Statistics from American Community Survey - 5-year Estimates (ACS5) at factfinder.census.gov and ii) Debt Statistics from EFA Household Debt at federalreserve.gov

# 1.2.1 Marriage and Divorce Statistics

From census.gov, the ACS 5-year table S1201 provides statistics about the percentage population married and divorced by county and by year.

Dimensions:

* County Name
* State Name
* Year

Metrics:

* Population
* Married Percentage
* Married Error Range
* Divorced Percentage
* Divorced Error Range
* Separated Percentage
* Separated Error Range
* Never Married Percentage
* Never Married Error Range

# 1.2.2 Debt Statistics

From federalreserve.gov, the EFA Household Debt table provides Debt to Income ratio ranges by county and by year. Detailed information here: <https://www.federalreserve.gov/econres/notes/feds-notes/household-debt-to-income-ratios-in-the-enhanced-financial-accounts-20180109.htm>

Dimensions:

* County Name
* State Name
* Year
* Quarter

Metrics:

* Debt to Income Ratio Low
* Debt to Income Ratio High

**1.2.3 Out of Scope**

The following collections of data are out of scope for this phase of the project, but could be interesting to analyze in future phases:

* Federal Reserve Interest Rates
* Prime Lending Rates
* Unemployment Rates
* Inflation Rates
* Minimum Wage Trends
* Average Household Income
* Average Household Debt by Category (i.e., Mortgage, Car, Credit Card)
* Number of Dependents per Household
* Marriage and Divorce Data by Age group
* Marriage and Divorce Data by Race

# Technologies and resource contributions

Python and PostgreSQL will be the primary tools leveraged to complete the ETL processes for “Fiscal Heartbreak”. The core python libraries that will be leveraged are: Pandas and SQLAlchemy.

The workload has been divided across the team resources as follows:

|  |  |
| --- | --- |
| ***Task*** | ***Owner*** |
| *Target Data Model and Design* | ***Mike Fox*** |
| *Extract and Transform Debt Data* | ***Joseph Ayala*** |
| *Extract and Transform Marriage Data* | ***Michael Hankinson*** |
| *Database Load* | ***Sean Findlay*** |
| *Create sample database view* | ***Andrew Behrman*** |
| *Repository Admin* | ***Joseph Ayala*** |
| *Validation/QA* | ***Andrew Behrman*** |
| *Document Owner* | ***Andrew Behrman*** |

# Definitions, Acronyms and Abbreviations

* **ETL**: Extract, Transform and Load. Describes the 3 primary stages of moving data from source to destination in preparation for analysis.
* **ACS5**: American Community Survey 5-year estimates. Statistics collected over a rolling 5-year window across communities in the USA from a sample of the population.
* **DTI:** Debt to Income.Ratio of Household Debt to Household Income based on Federal Reserve data sourced from: FRBNY Consumer Credit Panel/Equifax Data for household debt, and the Bureau of Labor Statistics for household income
* **Married Percentage**: Percentage of population currently married and living together at the time of survey based on 5-year estimates from American Community Survey
* **Divorced Percentage**: Percentage of population currently divorced at the time of survey based on 5-year estimates from American Community Survey
* **Separated Percentage**: Percentage of population currently separated, but still married at the time of survey based on 5-year estimates from American Community Survey
* **Never Married**: Percentage of population never married based on 5-year estimates from American Community Survey

**2. ETL DETAILS**

*This section outlines a more detailed description of the processes utilized/proposed to achieve the objectives of this initiative.*

# Data Import/Extract Sources and Method

#### Marriage Import/Extract Sources and Method

Several websites were investigated to find a good source of marital status data. The website ultimately selected was chosen due to its granularity of data. The CSV’s were readily available and used in contrast to the API’s which had instructions that had bad links The columns chosen were all percentages of marital statuses.

#### Debt Import/Extract Sources and Method

The data on debt-to-income came from the Federal Reserve website. It was presented with an interactive graphic of a map of the United States, but there was also a CSV download available. We chose the county-level dataset, which contained five columns for each record. We used the year, area fips code, low end ratio, and high end ratio. We dropped the quarter column after filtering out only the first quarter rows.

To attach the county names to their respective data point, we downloaded a separate table that listed each county next to its fips code and merged the table with the data set in Pandas.

# Data Acquisition

#### Marriage Data Acquisition

While there are several demographic choices including race, the data chosen is for the whole population, per county, per year, per marital status. The percentages of married, divorced, widowed, separated, and never married add up to approximately 100 percent but may differ due to rounding errors. There are geography choices as well - the county level was chosen to provide granularity. The CSV files are created annually with about a 2 year delay.

An API is available to obtain marriage data but CSV’s were used due to ease of availability.

#### Debt-to-Income Data Acquisition

The Federal Reserve official website provided the median household debt-to-income ratio. Geographically, the data was broken down by state, county, and metropolitan area. The state level didn’t provide enough data sets, and the metropolitan level didn’t cover the entire country. The data sets covered years 1999 to 2018 and broke down each year into quarters with the low end and high end ratio for each quarter.

Data was available for download in CSV format from the website and read using the Pandas library.

# Data Transform

#### Marriage Data Transform

Allowances were made due to the CSV containing two headers, the second header was dropped. The county and state column was split into separate columns to allow more flexibility with the data. As each spreadsheet was for a given year, the year for each CSV had to be inserted manually into the panda dataframe. Each year was then appended into one dataframe to allow easier loading.

Please note the encoding for the CSV is ISO-8859-1.

As far as transforms go, the marriage data transform was straightforward.

#### Debt-to-Income Data Transform

After the data was extracted into a Pandas dataframe, the data was cleaned. All NaN values were removed. Columns were renamed to remove spacing. The biggest transformation incorporated a second table that had a list of the county names next to their respective fips code. The debt-to-income data did not include the names of the county so a merge was required. The county names were also joined by the state names in the same column so they were split into two in the process.

# Data Integrity

#### Marriage Data Integrity

The marriage data chosen is in 5 year increments year by year. The 1 year increments data that was available and would have been preferred was not used due to too many null values. The county within state geography was chosen additionally to allow more data points.

#### Debt-to-Income Data Acquisition

The data appears to be updated with each new year’s data for each region; however, the data presented would not change. The data did have a significant amount of NaN values, which removed many rows, but there were still thousands of rows available to work with.

# Data Refresh Frequency

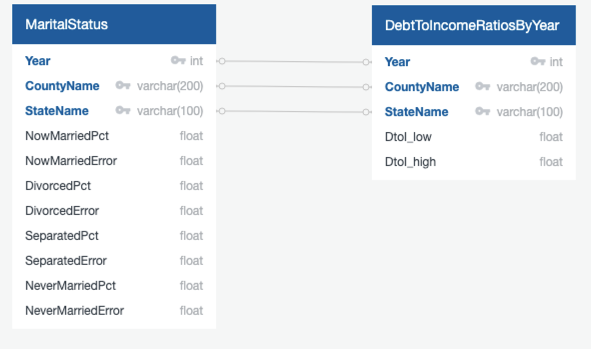
This ETL process was executed once in order to prepare for Fiscal Heartbreak analysis. However it can be re-run on an annual basis as the two underlying data sources are updated annually.

# Data Security

All of the data used in this project is available in the public domain and thus no security or data protection mechanisms are required. Specific auditing of the ETL process is also not required, however Transaction and Audit Logs native to the PostgreSQL database can provide insight into any data load activity.

# Data Loading and Availability

The following denormalized schema was designed to store the final data for analysis:



Once the schema was designed, the following SQL script was generated to create the tables within a PostgreSQL database. This SQL can be re-used to create the target tables in any PostgreSQL database. It is important to note that these tables should not exist yet in the database when the SQL is executed. If tables with the same names do exist and it is desired to remove them and re-create empty tables, use the “DROP TABLE <tablename> CASCADE” command to remove each table first.

CREATE TABLE "MaritalStatus" (

"Year" int NOT NULL,

"CountyName" varchar(200) NOT NULL,

"StateName" varchar(100) NOT NULL,

"NowMarriedPct" float NOT NULL,

"NowMarriedError" float NOT NULL,

"DivorcedPct" float NOT NULL,

"DivorcedError" float NOT NULL,

"SeparatedPct" float NOT NULL,

"SeparatedError" float NOT NULL,

"NeverMarriedPct" float NOT NULL,

"NeverMarriedError" float NOT NULL,

CONSTRAINT "pk\_MaritalStatus" PRIMARY KEY (

"Year","CountyName","StateName"

)

);

CREATE TABLE "DebtToIncomeRatiosByYear" (

"Year" int NOT NULL,

"CountyName" varchar(200) NOT NULL,

"StateName" varchar(100) NOT NULL,

"DtoI\_low" float NOT NULL,

"DtoI\_high" float NOT NULL,

CONSTRAINT "pk\_DebtToIncomeRatiosByYear" PRIMARY KEY (

"Year","CountyName","StateName"

)

);

-- REMOVING the FK relationship as the data is denormalized and there is no guarantee

-- that data exists for every county equally in both fact tables. The two tables should

-- be inner joined during analysis to properly evaluate marriage stats compared to debt

-- stats in counties where both stats exist

-- ALTER TABLE "MaritalStatus" ADD CONSTRAINT

-- "fk\_MaritalStatus\_Year\_CountyName\_StateName" --FOREIGN KEY("Year",

-- "CountyName", "StateName")

-- REFERENCES "DebtToIncomeRatiosByYear" ("Year", "CountyName",

-- "StateName");

With the tables created and the data cleanly staged in pandas dataframes, the data load itself becomes straightforward utilizing the pandas.dataframe.to\_sql() method call

**3. DATA QUALITY**

*This section provides an overview of the user-acceptance testing process completed for this project along with the associated outcomes from the test work completed. Practical steps for the client are provided below to ensure satisfactory data quality performance.*

# 3.1 Data Quality

The following user acceptance tests were performed to ensure data quality:

* **Debt-to-Income Ratio Boundaries** – ensured that no lower debt-to-income ratio bands exceeded upper debt-to-income ratio band in any given row in order to confirm data reliability and accuracy. This was confirmed by visualizing our debt-to-income ratio differential (upper-band minus lower-band) in ascending order:

**A screenshot of a cell phone

Description automatically generated**

**Sum of Marriage Statistics** – Made sure that the sum of overall marriage statistic percentages tied out to ~100% to ensure that data found in each record was complete and accurate. The extrema of the total percentages calculated ranged from 99.8% to 100.2%, which was identified as being caused by insignificant rounding adjustments within the original datasets. Sample python command can be found below:

**A screenshot of a cell phone

Description automatically generated**

* **Dataset Size and Reliability** – Lastly, we gained comfort with our dataset’s reliability via its overall size (24,253 rows) as well as the overall dataset’s diversity by state (all 50 states) and county (1,748 total counties); select statements from sql table below:
* select count(distinct "CountyName") from fiscal\_heartbreak\_analysis
* select count(distinct "StateName") from fiscal\_heartbreak\_analysis

# Appendix: Example Dataset Table View Command and Illustrative Screenshot

create view fiscal\_heartbreak\_final\_analysis as

select dti."Year", dti."CountyName", dti."StateName", dti."DtoI\_low", dti."DtoI\_high",mar."NowMarriedPct",mar."NowMarriedError",mar."DivorcedPct",mar."DivorcedError",mar."SeparatedPct",mar."SeparatedError",mar."NeverMarriedPct",mar."NeverMarriedError"

from "DebtToIncomeRatiosByYear" as dti

inner join "MaritalStatus" as mar on dti."Year"=mar."Year" AND

dti."CountyName"=mar."CountyName" AND

dti."StateName"=mar."StateName"

A screenshot of a cell phone

Description automatically generated