Kimball Lifecycle Proposal

CIS 9440 - Data Warehousing for Analytics

Final Project Milestone 1

Group Number - 21

Student(s) - Leah Lu, Porfirio Mohabir, Eric Sedaghat, Tori Torsaengtong, Tshering Sherpa

This Proposal is the beginning of your semester-long Final Project. The goal of the project is to develop a working Data Warehouse using a commercial database management system. Your project will use data from at least 2 sources, dimensionally model the data inside your Data Warehouse, and connect to a Business Intelligence application to produce valuable, actionable insights.

For motivation on project ideas, search for public datasets that interest you and your group. Then, think about how these datasets (maybe combined with other datasets) could address a problem or opportunity. Below are just a few (of many) public data sources:

1. Kaggle
2. NYC Open Data
3. Opendata.gov
4. Gapminder
5. Zillow
6. NOAA Climate Data
7. Google’s public datasets

**Data Warehouse Project Title:**

The Effect of family income on property value in New York

**Motivation for Project idea:**

As a real estate investment firm we will identify the relationship between income level and property value so that we will be better able to find areas with average and median income on the rise which will signify undervalued properties for our firm to invest in.

**Description of the issues or opportunities the project will address:**

Identify how change in income level in a particular area in turn changes the property value so that we can better identify properties that may increase in value in the near future based on an influx of job in a certain area(i.e. Amazon’s New Headquarters moving into an area).

**Business Justification:**

High-level Business Initiative:

Properly account for the increase/decrease in property value due to income level change.

BI Sponsors and Stakeholders (who will own this project?)

Real Estate Agency

Real Estate Investment Firms

Anyone who own property in New York

What’s the Business Value?

By accurately identifying the relationship between change in income and change in property value we will be able to look for areas where income level is on the rise and invest there if we find a significant link between income level and property value.

How long will this take? How much will this cost?

* The project will take 3 months
* The cost will be time spent by 5 workers designing and implementing a Data Warehouse
* Additionally whatever it costs to house our data

**Technical Justification:**

Which data sources do we already have for this project?

Dataset 1: Property value in New york by zip code

<https://www.zillow.com/research/data/>

Dataset 2: Median income

<https://data.cccnewyork.org/data/table/66/median-incomes#66/107/22/a/a>

Dataset 3: Real estate listings by zip code

<https://www.kaggle.com/waragones/us-real-estate-listings-by-zip-code>

What new data sources do we need (if any)?

Dataset 1: More granular data in median and mean income levels per year broken down by zipcode

Is the data we have conformed, consistent, and current? (data quality)

* Consistent: Our data we have is consistent.
* Conformed: Our data we have is conformed, however, if we need additional median income data after the year 2017, the data is not conformed.
* Current: Our data is current. We have the data by year which will be used to predict changes in property values year over year.

What technical skills will we need to complete this project?

* We must be able to design a data warehouse that will can hold all the relevant information we have gathered
* Implementation of data warehouse design
* Basic level of data modeling skills to identify the relationship between income level and property value

Will we need any new types of technologies?

We will need a Data Warehousing tool like AWS/BigQuery. In addition, we will use programs capable of modeling the relationship between income level and property value(Python, R, Stata)

**Key Performance Indicators (KPI’s) your Data Warehouse will display:**

1. Property Value
2. Median Income Level
3. Mean Income Level
4. Change in property value/change in income
5. Average Property value/ income
6. Average price reduction count per year
7. Change in Income
8. Change in Property Value
9. Min property value
10. Max Property Value
11. Min Income Level
12. Max Income Level
13. Days on market
14. Average List Price
15. Average Sales Price
16. Year-over-Year (YoY) Variance of Average Sold Price
17. Revenue Growth