Chicago Property Valuation Analysis

Justin Cox, Will Dibb, Taylor Rasley, Taylor Williams, Zhongyi Zhang

Agenda

1. Overview & Objective
2. Data Sources
3. ETL Model
4. Data Processing & Analysis
5. RDBMS
6. Data Visualization
7. Limitations & Lessons
8. Alternative Databases
9. Discussion & Next Steps

Overview

- Model development for property value comparison of Chicago areas
- Various livability considerations to identify areas where property values might be overvalued or undervalued.
- Objective: identifying these areas where property is out of alignment with livability provides potential investment opportunities and consumer information

Data Sources

CHICAGO DATA PORTAL

- Chicago Zip Geospatial Map
- Active Business Licenses
- Crime
- Community Assets:
    - Grocery Stores
    - Schools

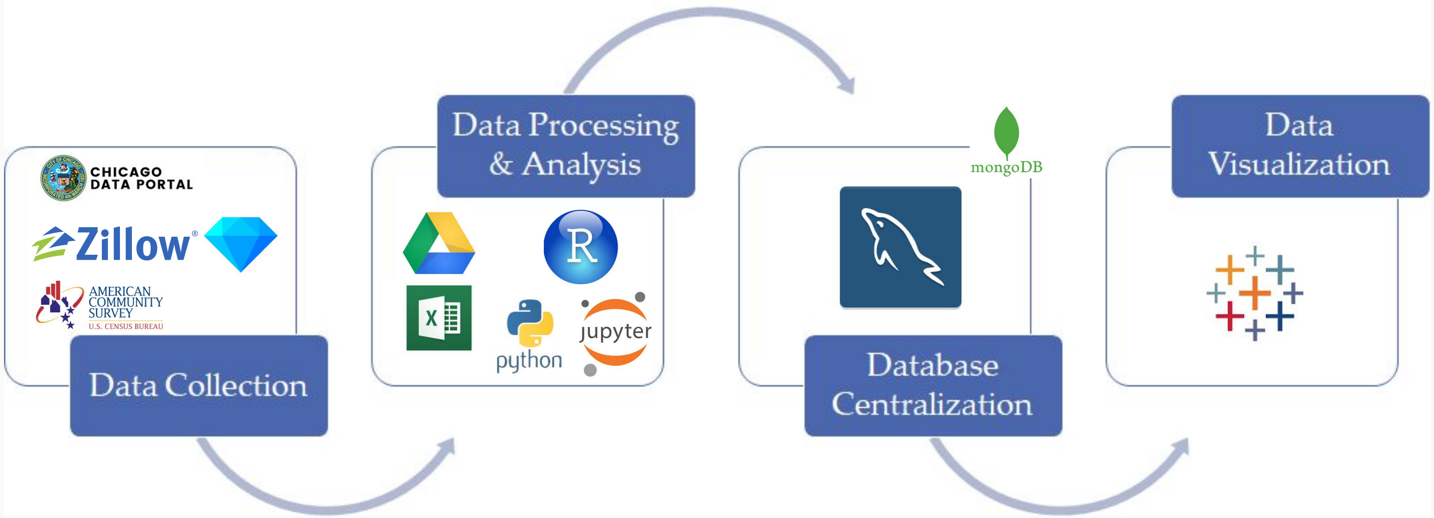
Zillow

- Property Values

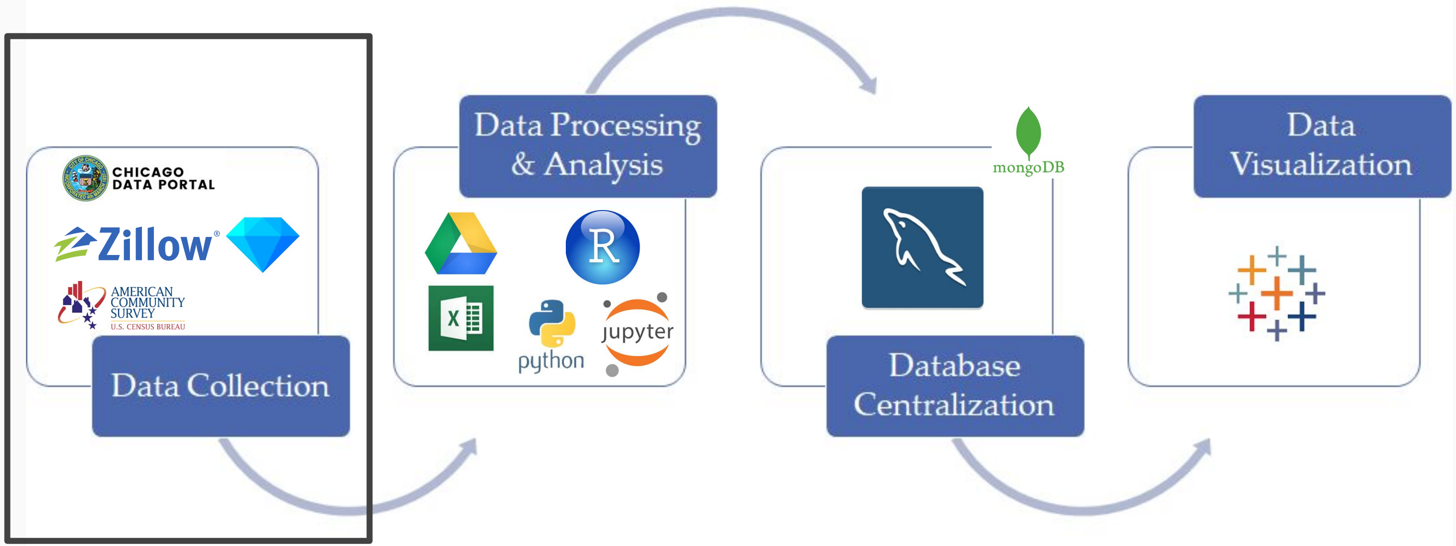
AMERICAN COMMUNITY SURVEY
U.S. CENSUS BUREAU

- Population (Zip)

Data Pipeline

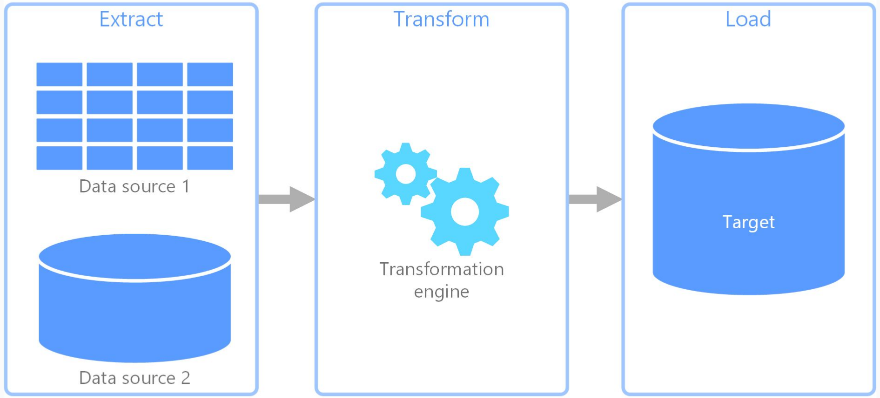


Data Pipeline: Collection

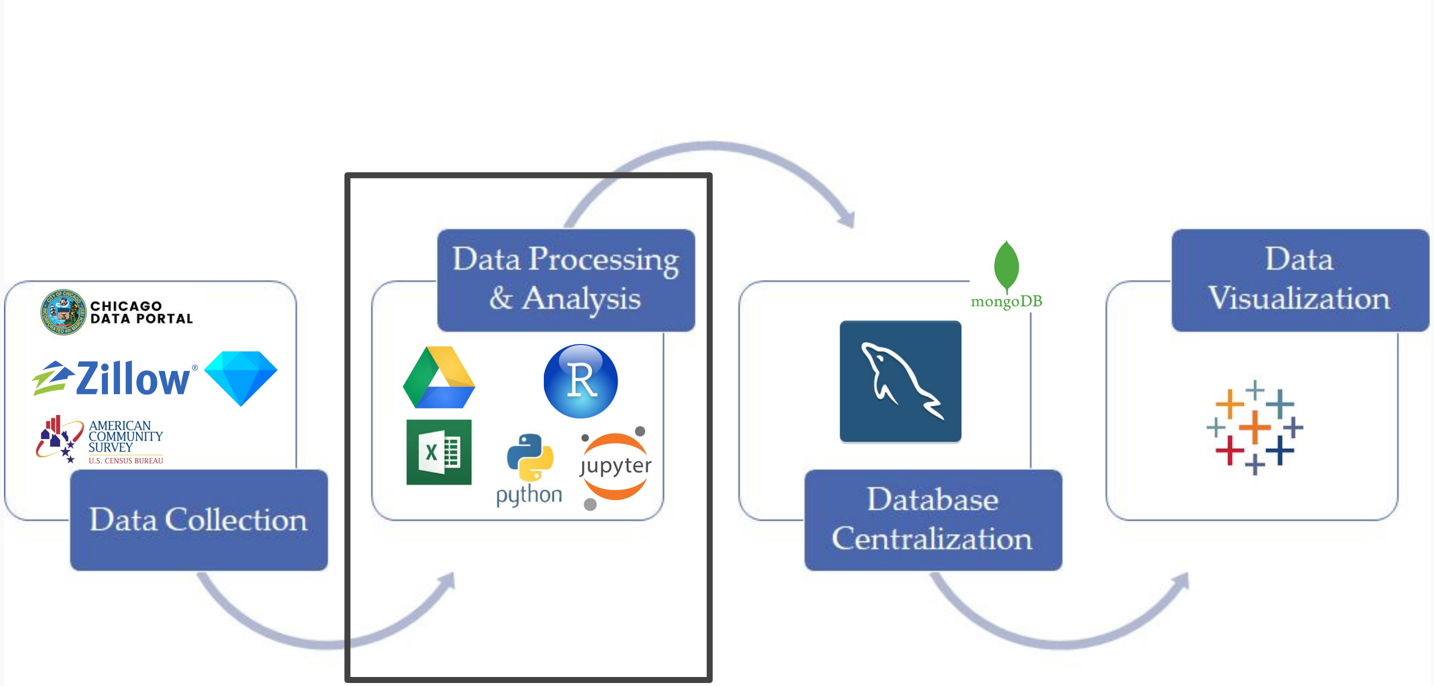


ETL

- **Extract**
  - OpenRefine used to retrieve current data from Zillow API
  - Additional data sources identified and static CSV files aggregated
- **Transform**
  - R and Python used to read, join, and clean tables, select and transform features
- **Load**
  - R used to export clean CSV files into MySQL path and imported into SQL database



Data Pipeline: Processing & Analysis



Scoring & Analysis

- **Scoring**
  - As part of feature transformation and RDBMS model, zip level scoring systems were developed and implemented for each respective data table
- **Analysis**
  - Multiple regression analyses demonstrated association between estimated property values and active business licenses (p-value = 0.025)
  - Weights for features were as anticipated, but not significant p-values (e.g. groceries p-value = 0.079)

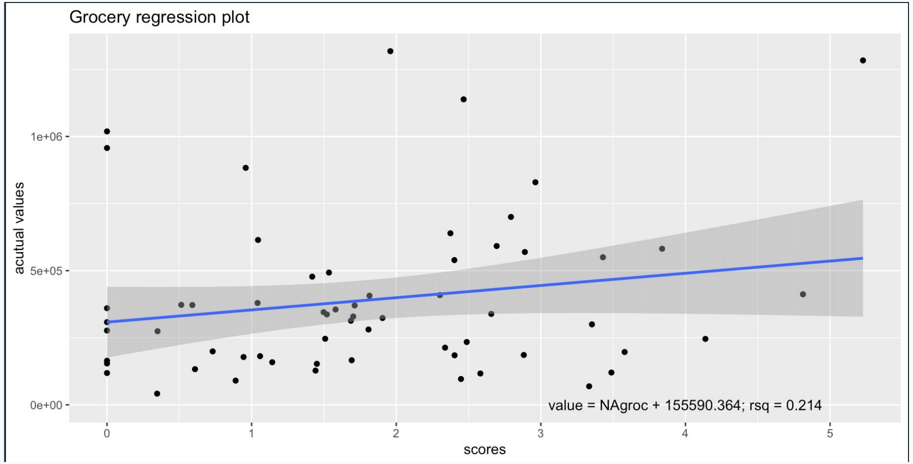
Multiple Linear Regression Result

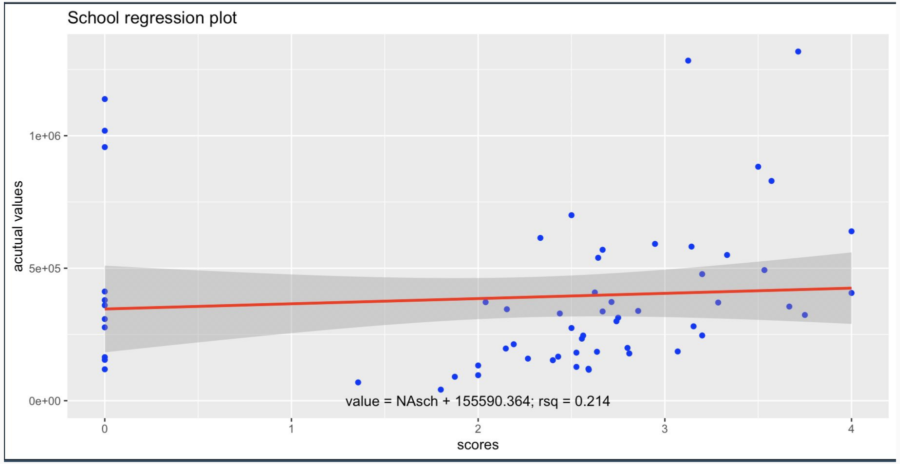
P-value < 0.05 reject null hypothesis
Negative - Lower crime scores - higher property values

(Intercept)	count_groceries	crime_score	business_score	school_score
155590.36	54359.83	-31375.13	20857.46	41326.09

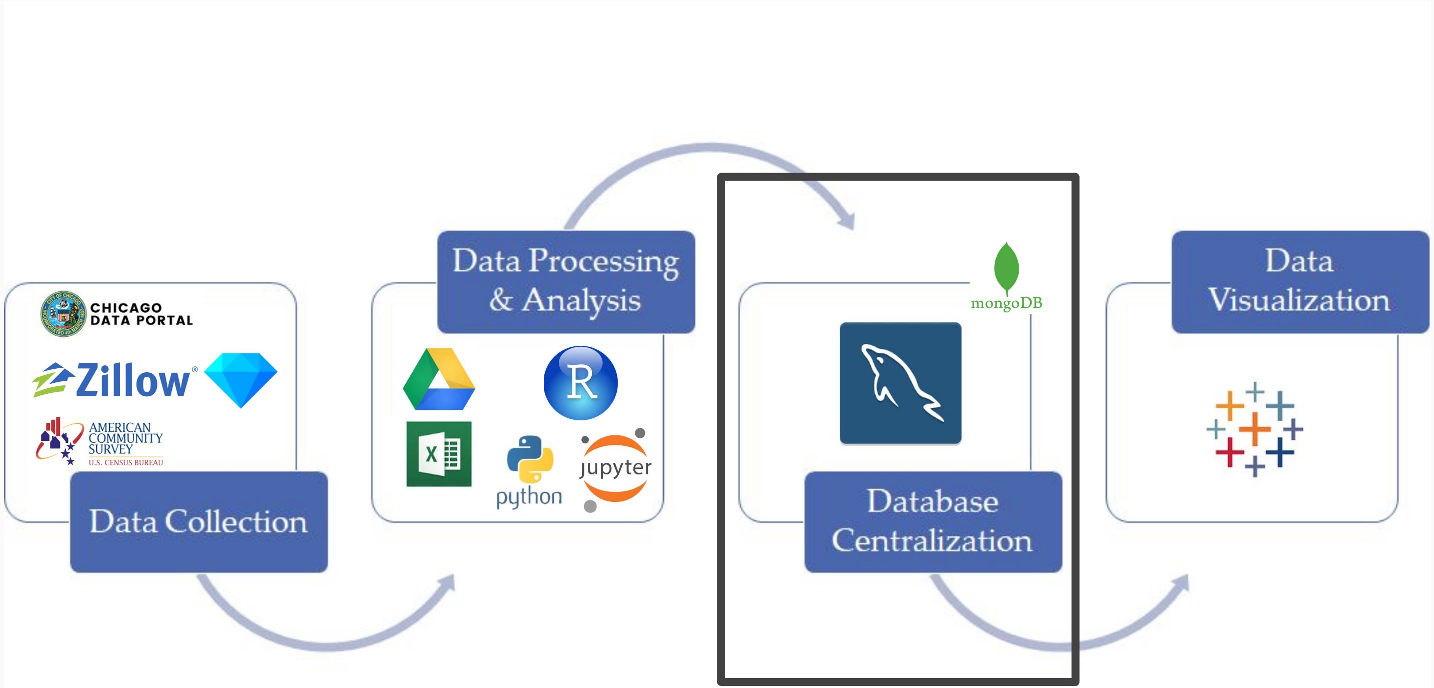
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	155590	90178	1.725	0.0900	.
count_groceries	54360	30413	1.787	0.0793	.
crime_score	-31375	34928	-0.898	0.3729	
business_score	20858	9052	2.304	0.0249	*
school_score	41326	33962	1.217	0.2288	





Data Pipeline: Database Development



RDBMS

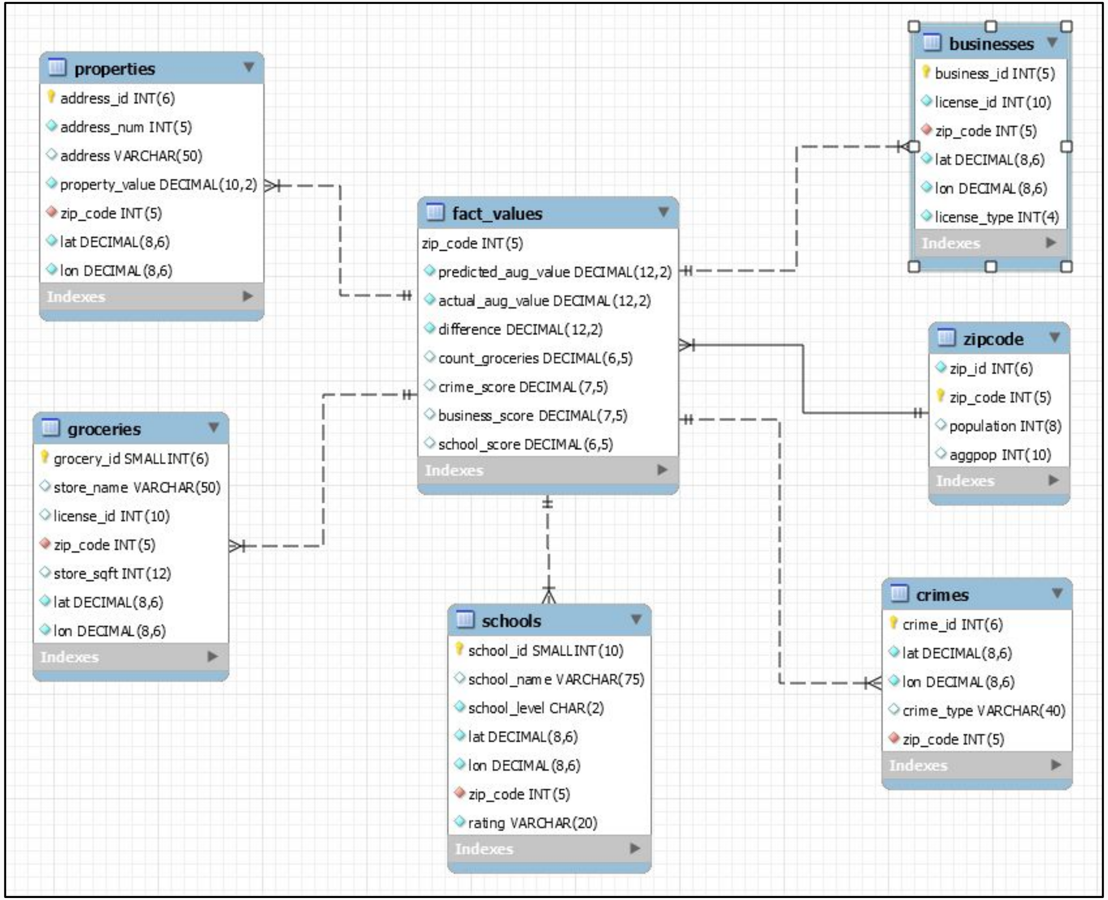
- **DDL**
  - SQL data definition language written in MySQL workbench to create defined tables and relationships
  - Reverse engineered schema for star model with central fact table for property valuations and livability index scores
- **Data Import**
  - SQL CSV import language written to populate generated relational database model
- **MongoDB Use Case** - NoSQL document database MongoDB use case also generated

Entity Relationship Diagram - Star Model

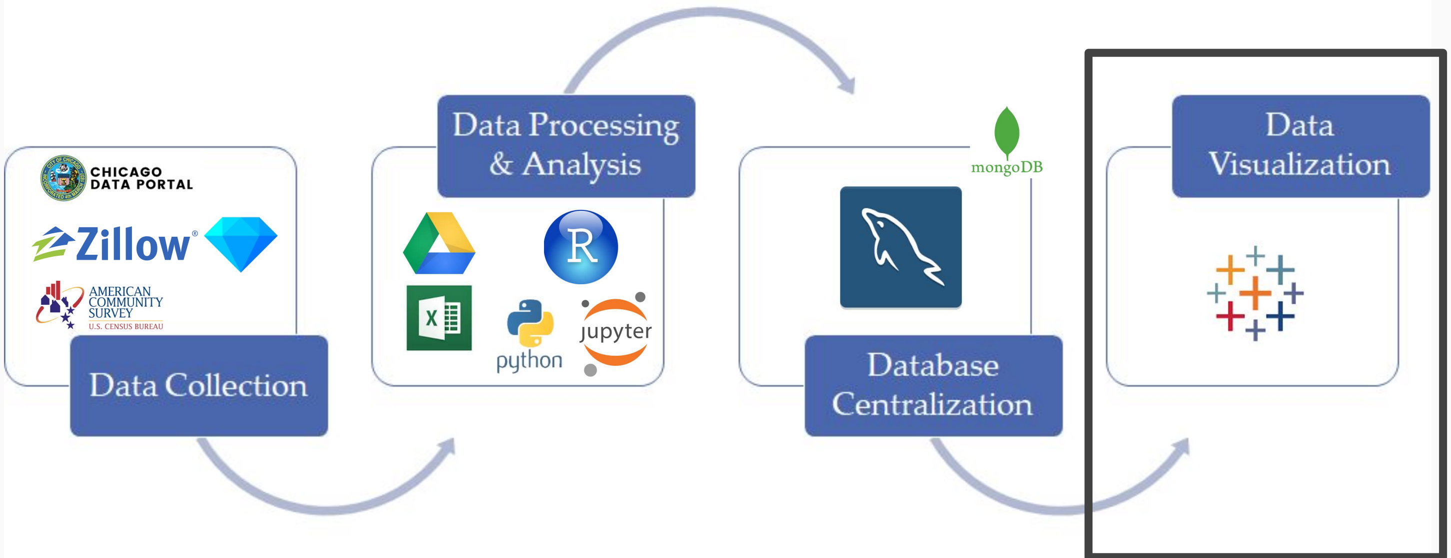
Dimensional Design Considerations

1. Modeling zip code area value based on livability considerations against actual average property values
2. Granularity - Zip Code
3. Dimensions - Each livability factor that is considered for livability
4. Facts - scores derived from dimension data for each zip code as well as model predicted scores

Additional Considerations:
-Dimension normalization not required
-Updating dimensions requiring bulk reload



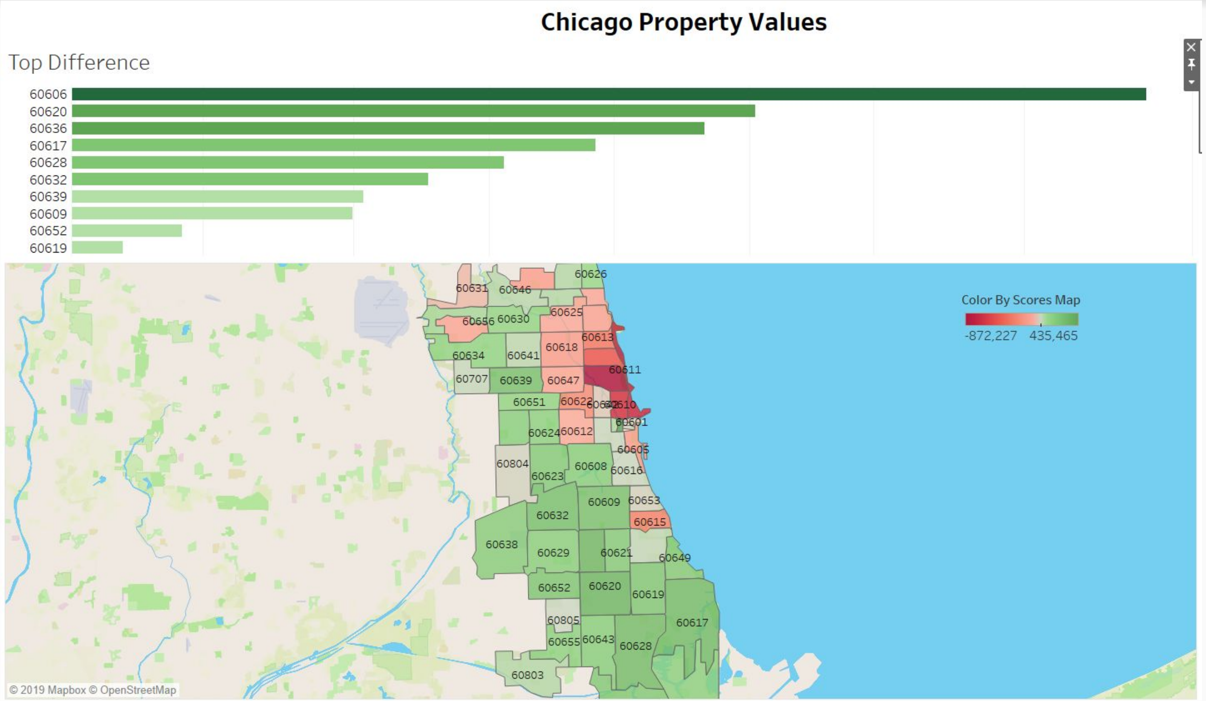
Data Pipeline: Dashboard Visualization



Data Visualization & BI

- **MySQL Server Connection**
    - RDBMS extracted into Tableau dashboard workbook
- **Reports & Dashboard**
    - Primary data visualization is an interactive heat map for zip codes for undervalued or overvalued
    - Additional reports allow data visualization for specific dimensions as well as zip-level reference information

Tableau Dashboard



Scoring Considerations & Limitations

- **Property Values**
  - Does not account for property type (apartment vs single family home)
- **Grocery Stores**
  - Excluded “Liquor” stores
  - Scored based on square footage
- **Schools**
  - Scored based on School Quality Rating Policy (only public schools)
- **Crime**
  - Scored crime types using heuristics (violence, severity)
- **Businesses**
  - Binary score for positive and negative value adding businesses

Data Limitations & Lessons Learned

Limitations from approach:

- Dimension scoring was more heuristic than scientific
- Scores don’t account for neighboring areas
- Point in time snapshot of data
  - Reliant on infrequently refreshed static data sources

Lessons learned:

- Zip codes change
- More factors could have been considered:
  - CTA/Transit
  - Parks & Attractions
  - TIFs
- Geospatial data has many potential identifiers which makes getting consistent crosswalks a challenge

Document Database Considerations

- Analytical nature of use case (as opposed to OLTP) aligns with document database
- Flexible schema beneficial for evolving data set/model
- Scripts for loading/cleaning data would need to change (for example references to JOINS)
- Scaling of analysis beyond Chicago would be more economically feasible

```
// We can provide the appropriate zip code with our predicted and actual value of the property there
// if a customer looks for a specific zip code with its property value by expecting:
// 1. the business score higher than 1,
// 2. total number of crimes lower than 1500 with # of violent crimes lower than 800
// 3. more than 3 schools around the property
// 4. more than 2 grocery stores in this zip code with a total store area higher than 10,000 sqft
db.property0.find(
    {$and: [
        {"# stores": {$gt: 2}}, {"total store sqft": {$gt: 10000}},
        {"# of schools": {$gt: 3}}, {"# crimes": {$lt:1500}},
        {"# violent crimes": {$lt: 800}}, {"business_score": {$gt: 1.00}}
    ]}).projection({"_id":0, "Zip_code":1, "predicted_aug_value":1, "actual_aug_value":1})
```

property0 0.010 s 4 Docs 20

	Zip_code	predicted_aug_value	actual_aug_value
1	60,607 (60.6K)	557,157.246 (0.56M)	549,911.476 (0.55M)
2	60,647 (60.6K)	422,136.14 (0.42M)	539,363.064 (0.54M)
3	60,657 (60.7K)	383,695.158 (0.38M)	883,022.539 (0.88M)
4	60,659 (60.7K)	409,604.906 (0.41M)	370,432.701 (0.37M)

mongoDB

Graph Database Considerations

- Graph Compute Engine aligns for OLAP
- Different factor nodes would have relationship “in” zip code
- Limited relationship types (i.e. school is in zip code) included so far limit added benefit of graph database but additional relationships identified would increase benefit of using graph database
- Zip code nodes would allow for easy identification of associated factors

Scope for Improvement & Next Steps

Process:

- Utilize additional tools in cloud platform
- Automate or further streamline procedures for initial data collections via web scraping and interval static file collections from respective sources

Analysis:

- Enhance scoring model for livability index values with additional data sources and context
- Include more factors such as: traffic congestion, public transit proximity, environmental quality
- Update DDL/data pipeline to support time variant analysis

Questions