ETL Project

**Data Analytics and Visualization**

**Rice University**

Dr. Sailendra Mahapatra

Data Sources:

<https://www.recenter.tamu.edu/data/housing-activity#!/activity/County/>

<https://www.bls.gov/data/>

Methodology:

The project is attempted to create a linear data base of storing diverse data for TX counties which can be further populated with linear information.

The housing price details (Sales, Dollar Volume, Average Price, Median Price, Total Lisings Months, Inventory) are downloaded from TAMU web page for the Texas counties (Anderson & Angelina) for the year 2011-2018 (removing /n, etc) in csv format and re-formatted in excel. Python, python library pandas, csv reader, json, request and pymongo, etc are used to create a data base collection (db.housing\_price) in Mongo after cleaning the data.

Data for annual wages are retrieved from BLS.gov site for these counties by API call.

The BLS data are uploaded to Mongo data base as series (collections).

The data base is chosen as a linear series as it is envisaged more series like population, family members, employment status of family members, crime statistics, county development activities can be further added (which may or may not be related to each other).

Schema:

1. County

:

"Anderson"

1. Date

:

2011

1. Sales

:

21

1. Dollar volume \_3

:

2283718

1. Average Price\_4

:

108724

1. Median Price\_5

:

86696

1. Total Listings\_6

:

396

1. Months Inventory\_7

:

24

1. \_id

:

5c78b0c0ddc8f4bf5488f004

1. County

:

"Anderson"

1. Date

:

"2017"

1. Annual\_Income

:

"42335"

Hypothetical use

Relevant queries can be developed in any programming languages to selectively retrieve data for counties. For example, is housing price is changing based on annual income or is number of crimes are less for county with higher housing prices, etc.