## horizontal line



E.T.L

08/17/2019

**─**

Natalie Odgerel St

Banafsheh Vafaeenia

UCB Data Analysis and Visualization

# Overview

The purpose of this project is to gather data from different sources and transform them into a unique format and make the required changes to get the desired results and load the results into a new database.

# Goals

1. Find data from various sources.
2. Clean data to match our desired goals
3. Being able to load and transport data to a new format.

# What is ETL?

When we have a large set of data in CSV, or TSV format from different reports and we want to combine them and create on database, That is when we use ETL.

## **E**: EXTRACT

[Data extraction](https://en.wikipedia.org/wiki/Data_extraction) is finding the data from one or multiple sources.

## **T**:TRANSFORMATION

Data transformation is when we clean and change the data into a readable and proper format for the purpose of analysis.

## **L**: LOAD

Data loading is when we transfer the information we cleaned into a platform to inform our analysis on the data.

# 

# Difficulties

## Finding Data

Data sources were limited on the type of information we needed. Data is not always in the same format and sometimes has duplicates or missing information and can change the outcome of the report

## Cleaning Data

We had to make sure that all table names and column names matched in order for transformation to happen.

# Steps of this project:

We came up with the idea of finding information on the housing market and compare it to

the income per Zip Code but since we didn’t have enough time we suffice to sorting our data based on State information.

**Extract** : One source we used was the Census API and we gathered employment rate and income information based on county and State. First we had to seperate the County name from State and then set the index to state.

Our second source of data was list of states with their abbreviations we found in CSV format.

Our third source of data was a CSV list of quarterly mortgage rates from year 2000-2010. Since our Census API information was from the last census in 2010 we only needed one column in our last data set and we had to get the average of the quarterly rates to come up with a yearly mortgage rate so we had to make some changes in our excel spreadsheet for mortgage rates.

**Transform:** We had to read our CSV files and API data in jupyter notebook and make the appropriate changes to be able to merge the tables on a single column shared by all of our data sets.

**Load:** we created three tables in Pgadmin4 and imported our data from our notebook to Sql Database. Then we made the necessary joins on our tables to create one joined table of all our tables.