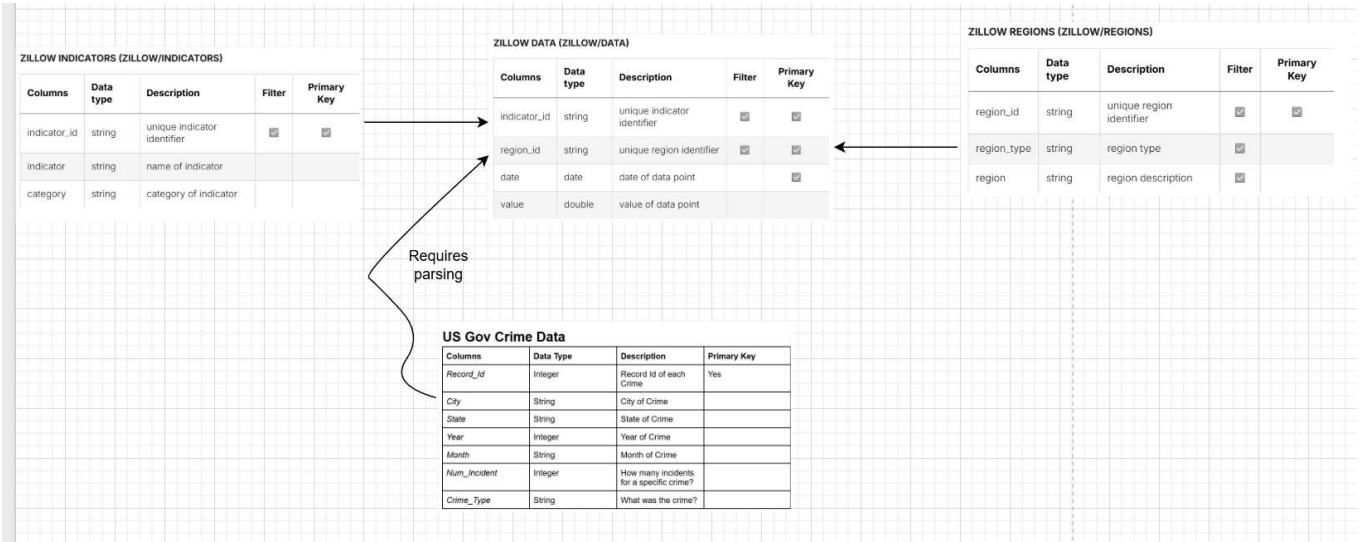
Overview:
We want to analyze how crime rates affect the housing prices of an area.

Research Question: How does crime in specific locations affect real estate market prices?

Team: We will work together and communicate to ensure that no one person carries a majority of the burden. We will play to our strengths where Jaitin will work on the more explanatory parts of the project (Modules 1-2, 4-5, 11-15) while Zach will work more on the technical parts of the project (Modules 3, 6-10). If one of the partners needs help then the other will help out as needed.

Datasets: One of our sources is the nasdaq database that contains Zillow real estate information and it contains three tables. The first table contains the value of the house and the date of when the value was created. It also has an indicator id and a region id which act as foreign keys to link with the other tables. Another table from this database contains data about the indicators of the houses and also contains a unique id for each indicator. The last table contains data about the specific region where the data comes from. Some are more specific where the whole zip code, city, county, and state are listed while some others only list the state where the house is located. This region table also contains a unique id for each region. The other source comes from kaggle that was collected from the US government crime department. It contains the city/state, when, the weapon involved, the victim's relationship to the perpetrator, and some data about the victim/perpetrator themselves. We will link all these tables together to further analyze them.

- Zillow: https://data.nasdaq.com/databases/ZILLOW
- Crime: https://www.kaggle.com/datasets/mrayushagrawal/us-crime-dataset

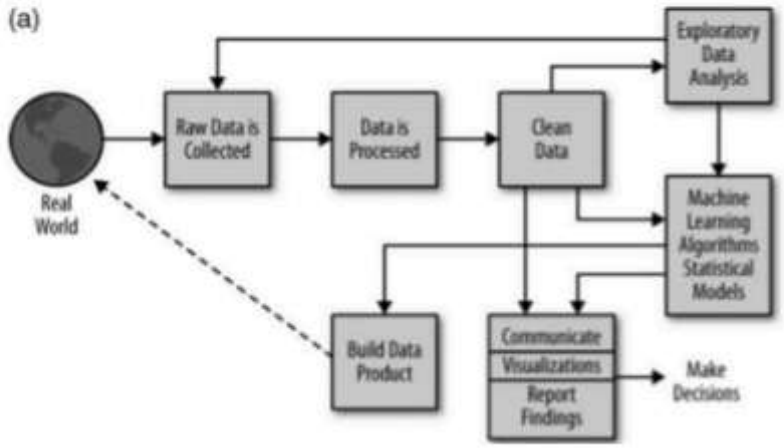


Timeline:

1. Understand Data/Schema design: Zach and Jaitin
    a. Understand all object types and identify potential problems
        i. Ex: Region_Id for Zillow data has cities incorporated while Crime dataset only has cities
            1. Solution: break down region Id to cities to use merge datasets(INNER JOIN) later
    b. Understand Schemas: we have basic schema understanding, but might need to add more data if this data isn't enough
        i. Ex: maybe education data is helpful to look at school ratings. Could we link a dataset with school ratings to the Region_Id foreign key?
2. Data Cleaning: Zach mainly, Jaitin helps point out the problem, Zach cleans it
    a. Based on previous step, clean the data and/or research other information to potentially add
        i. Ex: crime data may not be enough, we might need city population to look at crime per capita vs. total crime
    b. Missing/Null values: Zach will clean these up(Region_Id types are off for example)
3. Analysis and Visualization: mostly Jaitin
    a. Analysis: what is the crime per capita in certain cities? Which ones are the highest?
    b. Visuals: boxplot showing distribution of crime, showing distribution of real estate pricing. Bar graph showing median pricing by certain indicators like sizing of house, education ratings, etc.(based on indicator variable)
    c. Generally: many different questions we could ask that could be answered with analysis through reports and visualizations. As we develop these questions, we can be more confident and specific on which visuals to showcase.
4. Workflow and Automation: Zach and Jaitin
    a. Automate raw to clean data: automatically clean data, Zach will probably do this since he cleaned the data
    b. Automate visuals/reports: Jaitin will do most of this, but Zach may help as well
        i. Uncertain: not known exactly how this will work because it is dependent on the previous steps
5. Metadata Analysis: Jaitin and Zach
    a. Look at creation dates, authors, time of creation and implement this in final submission
6. Interim Status Report: Zach and Jaitin
    a. Explain progress, updated timeline, changes to our project, etc.

Lifecycle

Initially for this project, we did not formulate a specific research question; however, Zach and I knew we wanted to look at real estate data and see key variables that affect pricing. In order to do this, we researched important factors impacting pricings, specifically in certain locations(recorded by cities or zip codes). For example, we found school ratings, amenities like restaurants, and crime were the most important factors relating to the location. Obviously, there are many economic factors like interest rates, property taxes by state, and general financials relating to the housing market; for the sake of this project, we chose to focus on crime, and see how it correlates with the housing market. Next, Zach inspected the data looking for primary keys within each data set. Columns such as region_id and indicator_id are important to note for merging data sets to compare more variables. To clean the data set, we will have to parse certain strings or replace them in order to ensure foreign keys have similar object types and interpretable values. Once the data is cleaned, we hope to do some exploratory data analysis, creating basic visualizations to showcase statistics/trends of the data set. Zach may go into further depth with machine learning(ex: backward/forward feature selection with linear regression). The main goal is once we have our visualizations, reports and machine learning, I would create an algorithm to parse/clean the data automatically, so that these reports and visualizations could be generated automatically if/when new data is added in the future.



Ethics and Licensing

Our dataset looking at crime per city was obtained by the U.S. government. There are no privacy/confidentiality laws applying to this data, particularly because it is informational data about crime rates in cities. Copyright laws only apply to musical, pictorial, literary, and architectural works along with other “creative” works. For the Zillow datasets, Zach and I are allowed to use it for “personal” or “business” purposes, creating graphs, dashboards, and statistical summaries, as long as we cite Zillow for all of our reports/visualizations. However, we are unable to redistribute real estate listings like pricing info publicly on our own without approval from Zillow. If we use an API from Zillow, we can obtain the data but cannot perform

automated queries like webscraping to obtain the data, because this would give us the ability to obtain other zillow data(outside of this dataset) that may be prohibited.

Storage

All of our three tables for our Zillow data as well as the crime table is tabular, meaning all of this data will be stored within Excel files on OneDrive. In this folder, each dataset will be named from its source, primary key and type of data(all of them are xlsx). For example, our Zillow Indicators table would be named, Zillow_Indicator.xlsx, while our crime data would be UsGov_Crime.xlsx.

Workflow Automation, Reproducibility, & Metadata

Potentially, we could use Snakemake to automate cleaning, analysis, and visualizations. For reproducibility, this comes down to being disciplined on commenting and formatting our code, categorizing reports/visuals that are helpful and unhelpful, and weighting the importance of our analysis while communicating this to an outsider user. We have metadata on all four datasets explaining the creator, creation date, and general information on the datasets. More description and explanation of this will come as we work on the project.

Constraints

One constraint is that the crime dataset is only from 1980-2014 so the data is a little outdated and might be hard to relate to current times since the real estate market is ever changing. Another constraint is that the crime dataset only contains murder or manslaughter data, therefore, there isn't much diversity with the crime data.

Gaps

One gap is that the region data from the database is all in different formats, so the data will need to be parsed and cleaned to just get the city itself in order to merge the table with the crime dataset.