Timeline

1. <u>Understand Data/Schema design:</u> Jaitin
    a. Understand all object types and identify potential problems
        i. <u>Ex:</u> Region_Id for Zillow data has cities incorporated while Crime dataset only has cities, Zillow Region_ID has weird syntax with super long names
        ii. <u>Ex 2:</u> Indicator_ID for Zillow data is hard to interpret and requires more work on what it means. What we know is the indicators represent what the house value is(is it one bedroom, two bedroom); however, the strings are very long for the indicators, which requires further parsing and mapping to interpret each specific indicator.
        iii. <u>STATUS:</u> Completed by Jaitin in interpreting Region_ID, In progress to interpret Indicator_ID(much work needed). This is the first priority.
    b. <u>Understand Schemas</u>: we have basic schema understanding(look at the diagram below), but might need to add more data if this data isn't enough
        i. <u>Ex:</u> maybe education data is helpful to look at school ratings. Could we link a dataset with school ratings to the Region_Id foreign key?
            1. This was our question from the last report; however, this may not be feasible as it requires lots of cleaning and tremendous amounts of data beyond the scope of this project.
2. <u>Data Cleaning:</u> Zach and Jaitin
    a. Based on our understanding of schema and object types of each variable/column, we developed two parts to clean
    b. <u>Part One(Zach):</u> clean the Zillow Data Region_ID by parsing the super long string to capture only the city and state.
        i. <u>STATUS:</u> completed by Zach
    c. <u>Part Two(Jaitin):</u> based on how we interpret the Indicator_ID, we have to map out descriptions and/or "interpretable outputs" of each input value from the Indicator_ID. Based on that, we merge these new outputs into the dataset, allowing us to see what type of home each datapoint refers to.
        i. <u>STATUS:</u> In-Progress by Jaitin. This will take a lot of working since there are over 30 different types of Indicator_IDs. Lots of mapping work needed. Second priority as Jaitin will complete after first priority is done.

3. <u>Analysis and Visualization:</u> Jaitin
    a. <u>Analysis:</u> break down analysis questions looking at crime based on certain cities in our dataset along with value of our houses based on the type of home(indicator_id) in a specific city/state
    b. <u>Part 1:</u> creating analysis questions and generating visuals like bar graphs and boxplots to look at crime and house value by city state
        i. <u>Status:</u> Completed by Jaitin. We might add more analysis after we complete Part two from this section.

        c. Part 2: creating analysis questions and generating visuals to look at the home value sectioned off by specific indicators, to then compare them by city/state. Then we could look at the crime data and see if this correlates with home value.
            i. Status: InComplete by Jaitin. Requires Part Two from the Data Cleaning section to be finished. This will be completed after part two from data cleaning is finished, meaning it is third priority.

4. Workflow and Automation: Zach and Jaitin
    a. Checksums and data integrity assurance: we must make sure certain columns exist and there are enough in the raw data sets to ensure the workflow is possible to enter
        i. Status: Incomplete by Jaitin. This will be completed after everything else once the cleaning process is set in stone, so it is the fourth priority.
    b. Automate raw to clean data: automatically clean data, Zach will probably do this since he cleaned the data
        i. Status: Incomplete by Zach as we have to finalize all the ins and outs/details about how we clean the data. This will be completed once check sums and data integrity is managed. This is the final step.

DATA ORGANIZATION

Regions are broken down under the following types:

Region	Region Type
State	state
County	county
Metro area & USA	metro
City	city
Neighborhood	neigh
Zip Code	zip

Region Type is the field used as a filter in the REGIONS table. Note that data for entire USA is categorized under the *metro* region type. The USA region code will always be **102001**.

This product can be accessed via Nasdaq Data Link's Tables API.

- There are three tables included in this product, as listed below.
- Each table has its own Table code, shown in the second column below.

Table	Table Code	Table Description
Zillow Data	ZILLOW/DATA	Values for all indicators
Zillow Indicators	ZILLOW/INDICATORS	Names and IDs of all indicators
Zillow Regions	ZILLOW/REGIONS	Names and IDs of all regions

ZILLOW DATA (ZILLOW/DATA)

Columns	Data type	Description	Filter	Primary Key
indicator_id	string	unique indicator identifier	✅	✅
region_id	string	unique region identifier	✅	✅
date	date	date of data point		✅
value	double	value of data point		

ZILLOW INDICATORS (ZILLOW/INDICATORS)

Columns	Data type	Description	Filter	Primary Key
indicator_id	string	unique indicator identifier	✅	✅
indicator	string	name of indicator		
category	string	category of indicator		

ZILLOW REGIONS (ZILLOW/REGIONS)

Columns	Data type	Description	Filter	Primary Key
region_id	string	unique region identifier	✅	✅
region_type	string	region type	✅	
region	string	region description	✅	

1. <u>Understanding Data/Schema Design</u>

The Zillow website did a great job of explaining the schema of the three housing datasets. Particular variables of interest are *Region_Id* to look at the location of houses and *Indicator_Id* which indicates specific information about each house. Region_id is broken up into many parts, relating to zip code and county. Since our crime data set has crime data only linked to the city and state, we extracted the city and state from *Region_Id* so that we could match the crime data with the Zillow data. *Indicator_Id* looks as the information regarding the house such as how many bedrooms it has and listing information. This is important because these factors greatly affect the value of a house. Moving on, Zach and I used crime data based on the *Region_Id* in order to see if home value is related to crime. Our crime dataset has many different variables looking at all the crimes from 1980-2014 that occurred in a specific city and state. Our research question specifically is looking at crime's effect on home values, so for simplicity we counted all the crime incidents from 1980-2014 and used that number as a basis for crime over the last 35 years. One could argue that total crime incidents from 1980-2014 is not the most accurate measure of crime in a specific city, but it is the best we could do with our dataset. Now that we understand our schema for the housing data while also developing a crime value to judge crime per city, we decided to create a *final_df* that merges all the important variables from the housing data along with the total_incidents(measuring total crime incidents from 1980-2014). In our *final_df* dataframe, we look at the variables of *Region_Id* to specify house location, *Indicator_Id* to specify information about the house, date of the listing, house value in dollars, and *total_incidents* of crime from 1980-2014.

2. <u>Data Cleaning</u>

To clean the data, I (Zach) had to merge the region, the crime dataset, and the home value dataset all together. For the region dataset all the formats for how the region is stored is different depending on the region type. For example, the regions with the zip region type included the zip code, the greater area, the city, state and county.

	region_id	region_type	region	city	state
0	96208	zip	90706;CA;Los Angeles-Long Beach-Anaheim, CA;Be...	Bellflower	CA
72	95315	zip	87121;NM;Albuquerque, NM;Albuquerque;Bernalill...	Albuquerque	NM
73	91325	zip	76244;TX;Dallas-Fort Worth-Arlington, TX;Fort ...	Fort Worth	TX
74	91732	zip	77083;TX;Houston-The Woodlands-Sugar Land, TX;...	Houston	TX
75	61616	zip	10002;NY;New York-Newark-Jersey City, NY-NJ-PA...	New York	NY
...	...	...	...	...	...
89299	92436	zip	78402;TX;Corpus Christi, TX;Corpus Christi;Nue...	Corpus Christi	TX
89300	59107	zip	3293;NH;Lebanon, NH-VT;Woodstock;Grafton County	Woodstock	NH
89301	94027	zip	83025;WY;Jackson, WY-ID;Wilson;Teton County	Wilson	WY
89302	58463	zip	1718;MA;Boston-Cambridge-Newton, MA-NH;Acton;M...	Acton	MA
89303	75653	zip	40363;KY;nan;nan;Owen County	nan	KY

Furthermore the region types with the metro region type included only the city and state so I just renamed that column.

	region_id	region_type	region
1	394415	metro	Bridgeport, CT
2	394653	metro	Greenville, SC
3	394312	metro	Albuquerque, NM
4	394357	metro	Bakersfield, CA
5	394308	metro	Albany, NY
...	...	...	...
89082	753924	metro	Urban Honolulu, HI
89083	395169	metro	Tulsa, OK
89084	394619	metro	Fresno, CA
89085	395238	metro	Worcester, MA
89086	394938	metro	Omaha, NE

	region_id	region_type	city_state
1	394415	metro	Bridgeport, CT
2	394653	metro	Greenville, SC
3	394312	metro	Albuquerque, NM
4	394357	metro	Bakersfield, CA
5	394308	metro	Albany, NY
...	...	...	...
89082	753924	metro	Urban Honolulu, HI
89083	395169	metro	Tulsa, OK
89084	394619	metro	Fresno, CA
89085	395238	metro	Worcester, MA
89086	394938	metro	Omaha, NE

And there were some that only included the state so I dropped those because they didn't have any data about the city.

Since all the formats were different I split the larger regions dataset into subsets based on their region type. I then split the long string and parsed the data for each individual region type to only get the city and state data. I then remerged all the subsets together to get the larger dataset with the city and state as separate columns.

	region_id	region_type	region	city	state	city_state
0	96208	zip	90706;CA;Los Angeles-Long Beach-Anaheim, CA;Be...	Bellflower	CA	NaN
1	95315	zip	87121;NM;Albuquerque, NM;Albuquerque;Bernalill...	Albuquerque	NM	NaN
2	91325	zip	76244;TX;Dallas-Fort Worth-Arlington, TX;Fort ...	Fort Worth	TX	NaN
3	91732	zip	77083;TX;Houston-The Woodlands-Sugar Land, TX;...	Houston	TX	NaN
4	61616	zip	10002;NY;New York-Newark-Jersey City, NY-NJ-PA...	New York	NY	NaN
...	...	...	...	...	...	...
81192	2400	county	Manistee County;MI;nan	Manistee County	MI	NaN
81193	2353	county	Humboldt County;NV;Winnemucca, NV	Humboldt County	NV	NaN
81194	1077	county	Hardin County;IA;nan	Hardin County	IA	NaN
81195	263	county	Otoe County;NE;nan	Otoe County	NE	NaN
81196	131	county	Clay County;TN;nan	Clay County	TN	NaN

I then merged this cleaned dataset with the zillow home values dataset using the region id because that is the attribute they had in common.

```
merged_df_clean = pd.merge(data_2014, check_merge_combine.drop(["region_type","region"], axis=1), on="region_id", how="inner")
merged_df_clean["date"] = pd.to_datetime(merged_df_clean["date"])
merged_df_clean
```

✓ 3.6s

	indicator_id	region_id	date	value	city	state	city_state
0	ZATT	1146	2014-03-31	288076.000000	Lehigh County	PA	NaN
1	ZATT	1146	2014-04-30	289159.000000	Lehigh County	PA	NaN
2	ZATT	1146	2014-01-31	296171.523003	Lehigh County	PA	NaN
3	ZATT	1146	2014-02-28	299399.064690	Lehigh County	PA	NaN
4	ZATT	1146	2014-05-31	313516.092252	Lehigh County	PA	NaN
...	...	...	...	...	...	...	...
6496380	Z4BR	95019	2014-09-30	189555.079592	Sierra Vista	AZ	NaN
6496381	Z4BR	95019	2014-10-31	188631.870053	Sierra Vista	AZ	NaN
6496382	Z4BR	95019	2014-11-30	188488.806946	Sierra Vista	AZ	NaN
6496383	Z4BR	95019	2014-12-31	188611.779814	Sierra Vista	AZ	NaN
6496384	Z4BR	95019	2014-01-31	194980.837292	Sierra Vista	AZ	NaN

I noticed in the crime dataset that the state was fully listed out whereas in the zillow dataset it was only the abbreviation. Because of this I had to convert the full state names to their abbreviations so that the naming convention is consistent which helps when merging the datasets.

In both datasets I then created a new column that combined the city and state so that I could merge the two datasets together. I then merged the dataset with the combined region and house data with the indicator dataset to create one giant zillow dataset with all the variables that we will be using for further analysis.

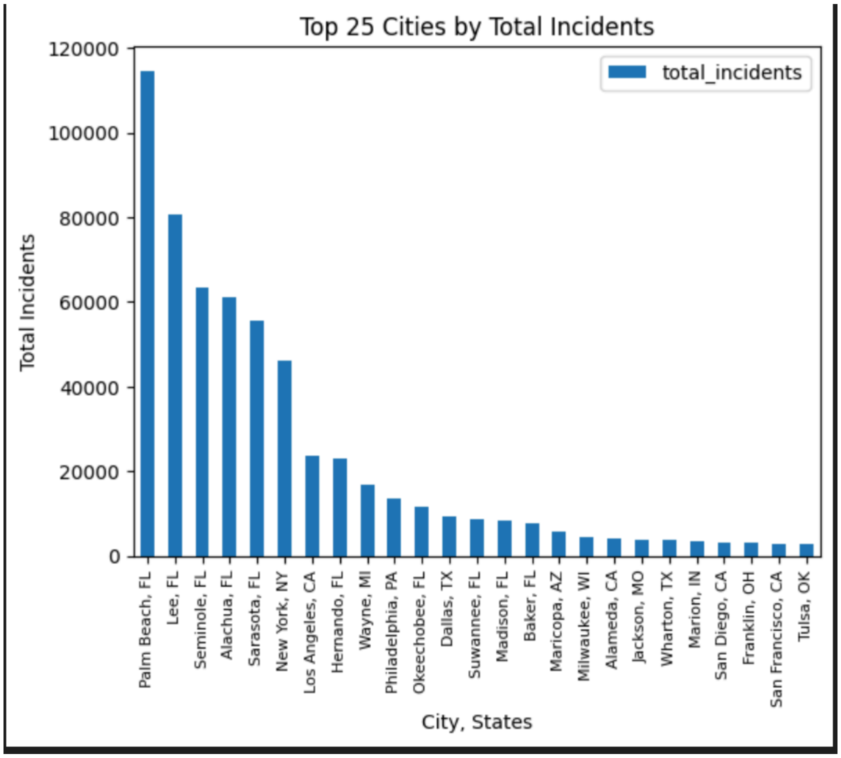
	indicator_id	region_id	date	value	city	state	city_state	indicator	category
0	ZATT	1146	2014-03-31	288076.000000	Lehigh County	PA	Lehigh County, PA	ZHVI All Homes- Top Tier Time Series ($)	Home values
1	ZATT	1146	2014-04-30	289159.000000	Lehigh County	PA	Lehigh County, PA	ZHVI All Homes- Top Tier Time Series ($)	Home values
2	ZATT	1146	2014-01-31	296171.523003	Lehigh County	PA	Lehigh County, PA	ZHVI All Homes- Top Tier Time Series ($)	Home values
3	ZATT	1146	2014-02-28	299399.064690	Lehigh County	PA	Lehigh County, PA	ZHVI All Homes- Top Tier Time Series ($)	Home values
4	ZATT	1146	2014-05-31	313516.092252	Lehigh County	PA	Lehigh County, PA	ZHVI All Homes- Top Tier Time Series ($)	Home values
...	...	...	...	...	...	...	...	...	...
6496380	RSSA	71340	2014-06-30	775.000000	Augusta	GA	Augusta, GA	ZORI (Smoothed, Seasonally Adjusted): All Home...	Rentals
6496381	RSSA	71340	2014-07-31	769.000000	Augusta	GA	Augusta, GA	ZORI (Smoothed, Seasonally Adjusted): All Home...	Rentals
6496382	RSSA	71340	2014-09-30	756.000000	Augusta	GA	Augusta, GA	ZORI (Smoothed, Seasonally Adjusted): All Home...	Rentals
6496383	RSSA	71340	2014-10-31	750.000000	Augusta	GA	Augusta, GA	ZORI (Smoothed, Seasonally Adjusted): All Home...	Rentals
6496384	RSSA	71340	2014-11-30	743.000000	Augusta	GA	Augusta, GA	ZORI (Smoothed, Seasonally Adjusted): All Home...	Rentals

In addition, I summed up the total number of crimes in each city-state pairing in the dataset so that we can do further analysis about how the total number of incidents affects the price of a house. I then combined this data with the combined zillow dataset to get the final dataset that we will be using.

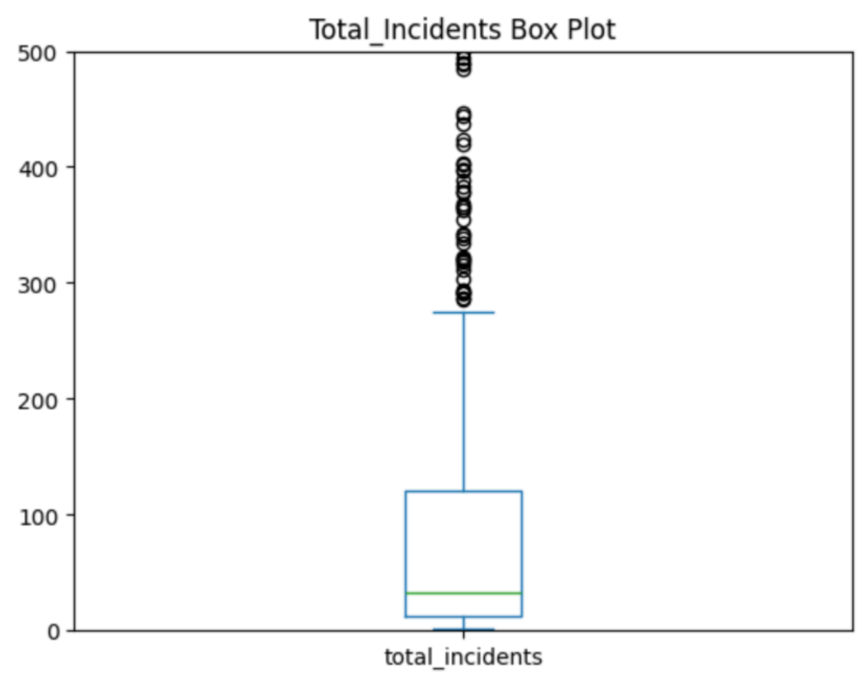
	indicator_id	region_id	date	value	city	state	city_state	indicator	category	total_incidents
0	ZATT	52334	2014-01-31	173491.0	Greenwood	SC	Greenwood, SC	ZHVI All Homes- Top Tier Time Series ($)	Home values	150
1	ZATT	52334	2014-02-28	174812.0	Greenwood	SC	Greenwood, SC	ZHVI All Homes- Top Tier Time Series ($)	Home values	150
2	ZATT	52334	2014-03-31	177605.0	Greenwood	SC	Greenwood, SC	ZHVI All Homes- Top Tier Time Series ($)	Home values	150
3	ZATT	52334	2014-04-30	178458.0	Greenwood	SC	Greenwood, SC	ZHVI All Homes- Top Tier Time Series ($)	Home values	150
4	ZATT	52334	2014-05-31	178872.0	Greenwood	SC	Greenwood, SC	ZHVI All Homes- Top Tier Time Series ($)	Home values	150
...	...	...	...	...	...	...	...	...	...	...
284551	Z4BR	92159	2014-08-31	83411.0	Victoria	TX	Victoria, TX	ZHVI 4-Bedroom Time Series ($)	Home values	190
284552	Z4BR	92159	2014-09-30	83676.0	Victoria	TX	Victoria, TX	ZHVI 4-Bedroom Time Series ($)	Home values	190
284553	Z4BR	92159	2014-10-31	84146.0	Victoria	TX	Victoria, TX	ZHVI 4-Bedroom Time Series ($)	Home values	190
284554	Z4BR	92159	2014-11-30	84545.0	Victoria	TX	Victoria, TX	ZHVI 4-Bedroom Time Series ($)	Home values	190
284555	Z4BR	92159	2014-12-31	84925.0	Victoria	TX	Victoria, TX	ZHVI 4-Bedroom Time Series ($)	Home values	190

For the zillow dataset we are only looking at the values from 2014 because those are the most relevant to us today. As the housing market is always changing we want to look at the data that would be most relevant to a home buyer today.

3. Analysis and Visualization



With the *final_df*, we have access to all the homes for each city within each state from 2014 specifically along with the *total_incidents* of crime data from 1980-2014. In total we have 886 unique cities with homes. First, we looked at the top 25 cities with the most *total_incidents* in a bar graph. We saw the top five were all cities from Florida. Some of these cities are fairly safe despite them having high *total_incidents.* This attests to potential inaccuracies with our *total_incidents* variable in assessing crime in cities. Additionally, there could be errors within the data, which would require further checks. Nonetheless, this is the best we had to measure crime, so we did the best we could.



Next, I created a boxplot to look at the *total_incidents* distribution. I noticed it was greatly right skewed with the median being below 100, but having significant outliers in the tens of thousands. This goes to show the *total_incidents* variable might have many errors that need cleaning. We may need to also do more inspecting on the crime dataset.