SIADS 591 Milestone I project proposal

Instructions: please make a copy of this template file and replace ***this*** line with your proposal title.

## Team members: John Zupan (jtzupan), Stephen Nguyen (stphn)

### 1. Summarize your proposed project in a few sentences.

#### What is your proposed project and why are you proposing it?

#### What are the question(s) you want to answer, or goal you want to achieve?

|  |
| --- |
| We are interested in investigating the relationship between crime (arrests) and real estate prices in New York City. We know many factors influence the price of a home, from the quality of schools to the size of homes to the supply/demand in the market. We are interested in the specific role arrest and crime play in the home values in that area.  The specific questions we are answering are:   * Is there correlation between the number of arrests in a zip code and the average home price in that zip code? * Are arrests for certain classes of crimes that have a higher correlation with home prices than others? |

### 2. Describe your primary dataset. How was it collected and how will you access it? Please share what features in the dataset are relevant to your topic. At a minimum, include the following information:

#### Short description (i.e., 1-3 sentences) of its key features

#### Estimated size (in records and/or bytes)

#### Location (give the URL or other access method)

#### Format (CSV, JSON, etc.)

#### Access method (download, web scraping, API, etc.)

#### The dataset **MUST** be publicly available to all members of the class (students, instructors, course support personnel, etc.). You may not use a proprietary dataset for this project.

|  |
| --- |
| [Please use this space for your response. You may expand or contract this box as needed.] |

### 3. Describe your secondary dataset. How was it collected and how will you access it? Please share what features in the dataset are relevant to your topic and describe the data types you’re expecting. At a minimum, include the following information:

#### Short description (i.e., 1-3 sentences) of its key features

#### Estimated size (in records and/or bytes)

#### Location (give the URL or other access method)

#### Format (CSV, JSON, etc.)

#### Access method (download, web scraping, API, etc.)

#### The dataset **MUST** be publicly available to all members of the class (students, instructors, course support personnel, etc.). You may not use a proprietary dataset for this project.

#### **Please note**: the two datasets should have different features/columns and/or different access methods, like using csv, JSON files, retrieving from API or scraping from the web etc. Different time periods, for example, with the same features/columns could not be considered as different datasets. Remember, the focus of the project in this Milestone course is to give you the opportunity to practice your data manipulation skills, so feel free to challenge yourself. If you're not sure your data sets are "different enough" go ahead and ask in the #siads591\_xxxx\_001\_project Slack channel.

|  |
| --- |
| Short description  * The data set includes every arrest in New York City by the New York Police department between 2006 and the end of 2019. The data set includes the date of the arrest, metadata about the crime (description, actual law broken, offense level), metadata about the perpetrator (age group, sex, race) and geographic information (lat/long of arrest and borough).  Estimated size (in records and/or bytes)  * Rows: 5,012,956 * Size:  Location: https://data.cityofnewyork.us/Public-Safety/NYPD-Arrests-Data-Historic-/8h9b-rp9uFormat (CSV, JSON, etc.)  * CSV  Access method  * There is an API that can be used to access the data set, but we will be downloading the complete CSV as our initial data set |

### 4. [ ] Please check this box to confirm that your primary and secondary datasets are accessible and available to your classmates and the instructional team.

### 5. How will you join your primary and secondary datasets? What challenges, if any, do you anticipate?

|  |
| --- |
| [Please use this space for your response. You may expand or contract this box as needed.] |

### 6. Describe any analyses you plan to undertake. For each, please give the technique or approach and briefly explain what you expect to learn from it.

|  |
| --- |
| The goal of our analysis is to understand any relationship between the number of crimes (measured indirectly via arrests) and home prices in New York City zip codes. We also plan to investigate if more arrests for certain types of crimes have a higher correlation to home prices. These questions can be by simply calculating and looking at the correlation coefficient, r, and r-squared between the counts of arrests and home prices. Several things to be cognizant of: we are identifying correlation, not causation. While arrests may or may not be correlated to home prices, there are also many factors that could account for differences in home prices like quality of education or demand in that zip code. While out of scope for this project, understanding exactly why more crime (higher arrests) is correlated with home prices would be an interesting extension. Additionally, looking at the effect of every type of crime on home prices may take us down the path of p-hacking and increase the chance of incorrectly identifying a type of crime that has a statistically significant correlation to home prices. |

### 

### 7. Describe in 1-3 sentences at least one data visualization you plan to create. Include the chart type (e.g. bar chart, scatterplot, SPLOM, etc.) as well as the variables (features) you intend to plot.

|  |
| --- |
| [Please use this space for your response. You may expand or contract this box as needed.] |

### 8. Does your choice of data raise any ethical issues? If so, briefly describe the concern and how you plan to mitigate it.

|  |
| --- |
| Home price data does not raise any obvious ethical question. While it is common knowledge that home sales are public records, including the sales price, it could be argued that the buyer has no choice in the matter and that they could not possibly know every way in which the data may be used. However, this is mitigated by the fact that the data we are using, from Zillow, is aggregated to the zip code level, with no personally identifiable information.  The second data set consists of every arrest by the NYC Police in New York for the past 14 years. This data has higher potential for ethical issues. To begin with, while each arrest is assigned a random identification number, it may be possible to trace these back to the actual arrests and the individuals who were arrested. Next, is revealed during NYC’s “broken windows” policing experiment, more people will get arrested in areas where police patrol more often. Therefore, the data, arrest counts, could potentially reveal more about the police patrol patterns than where crimes are actually occurring. While we have no plans to segment the arrest data beyond where the arrest occurred and what type of crime the arrest was for, the presence demographic data (race, age, gender) is a reminder that we should carefully scrutinize any analysis using these fields.  Finally, if there is some relationship between home prices and crime, our analysis could potentially reveal the “cumulative disadvantage.” High crime leads to low home prices due to lack of demand, leads to little neighborhood investment and opportunity, leading to crime and so on. If our analysis was going to be used to make decisions about where the city should invest resources, the ethical implications may be more important, but our analysis is only attempting to identify any correlation between home prices and arrests. |

### 

### 9. Indicate the contribution that each team member will make to the project.

|  |
| --- |
| [Please use this space for your response. You may expand or contract this box as needed.] |

### 