1. **Executive summary**
2. **Business case and objective(s)**

One of the major concerns while buying or renting a place to stay is safety of the area especially in high rise cities like Chicago. Our team, the Real Estate Vigilantes are building a recommendation system, that would help the residents of Chicago find their very own SAFE home.

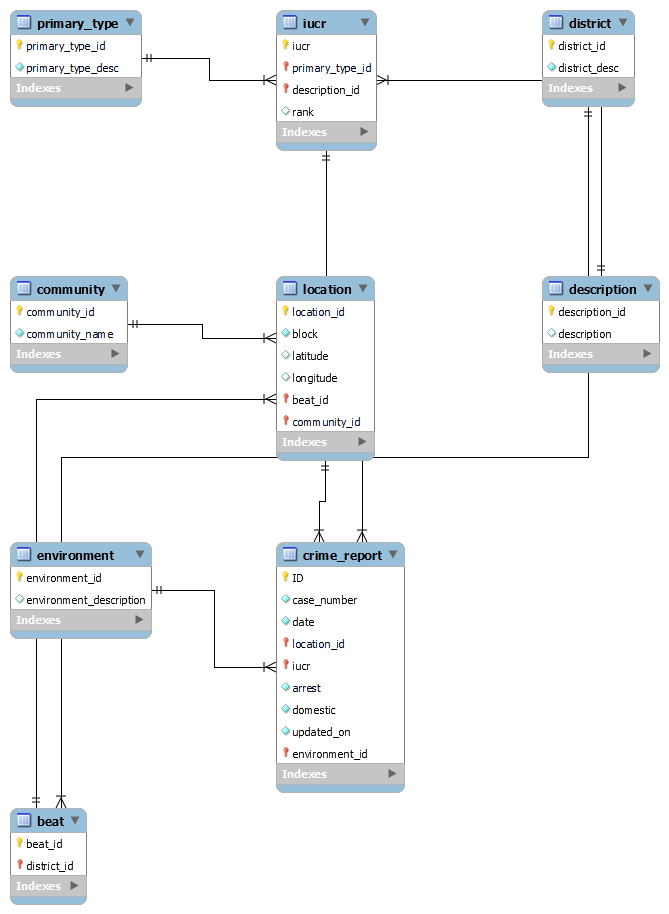
1. Objectives:
   * Collect, prepare and analyze the crime data
   * Set up the data on Azure, connect with MySQL and Python
   * Create a model to score the areas based on severity of crime on a scale of 0-100.
2. **The problem to be solved and datasets you plan on using**

Calculate and map a safety score to each beat and district based on the crime.

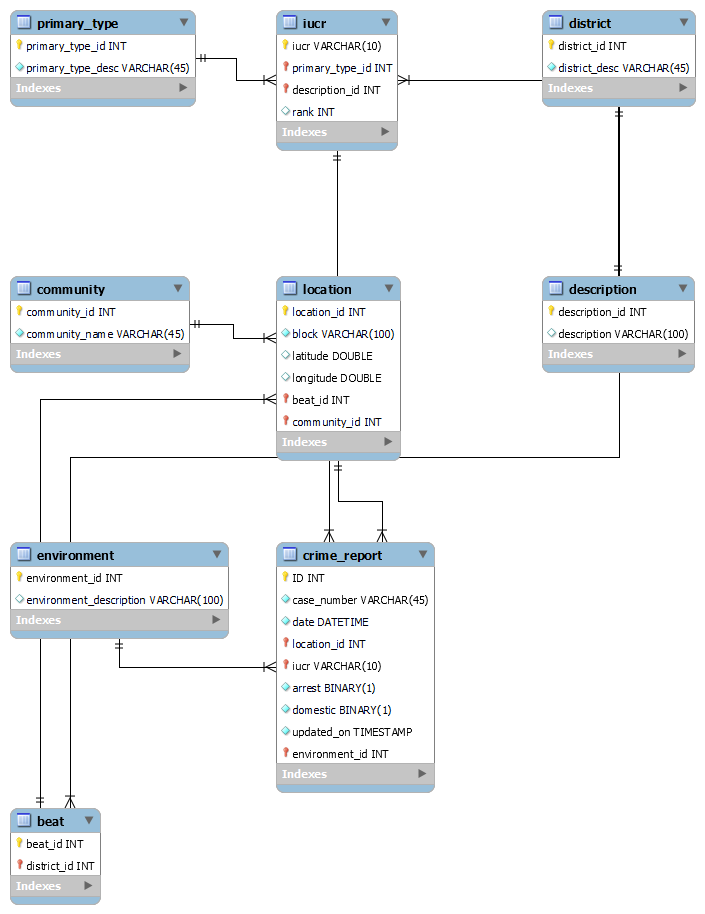
The data set used was the Crime Statistics provided on the Chicago Data Portal from 2001 to present which is a record of the crimes and their details: [Crime Statistics for Chicago](https://catalog.data.gov/dataset/crimes-2001-to-present)

1. **Data Models**
2. After manually ranking the crimes in the IUCR table, they are mapped to the crimes by beat and district.
3. The sum of the rank is then normalized between 0-1 by beat and district for day and night.
4. The following formula is then used to calculate the safety score with more weightage given to day scores as compared to night scores as we believed the crimes during the day are more significant in nature like robberies which take place when an owner is generally not present at home.
5. **Conceptual, logical, and physical data models**

* **Conceptual model**
* **Logical model**

****

* **Physical data model**

****

1. **Relational and Dimensional, (and if applicable No-SQL/Document, and Graph), and rationale behind using one over the other**

(Relational model is provided in point 5). As the purpose of the project is to calculate the scores of the district and beat, we saw that it involved just about 2-3 joins for analysis and the performance wasn’t affected by these when compared to a dimensional model. Therefore, we decided to stick with a relational model.

1. **The model should contain at least 5 tables (more is fine), which are completely Normalized till 3rd normal (if applicable) form and ER diagram provided in the document**

ER diagram provided in point 5, under logical model which is the normalized data till 3NF.

1. **Any denormalization required in the physical model, if yes, why (provide rationale for doing so)**
2. **Data Profiling**
3. **Perform high level data profiling and cleaning, document any observations along with sample values such as outliers, data quality anomalies, any aggregations already applied in the data set or you plan on calculating**
4. **Methodology and various tools used in the process**
5. **Evaluation and rationale behind using a certain Methodology, tool, and technology**
6. **Automation methodology (if any, or recommended) for the End to End pipeline.**
7. **Insights**
8. **Reports and Dashboards**
9. **Any recommendations and lessons learned**
10. **References**