CIS 4400 Data Warehousing

Paramvir Singh

Andrew Ivanov

Anthony Ferraro

Ricky Vidals

NYPD Arrests

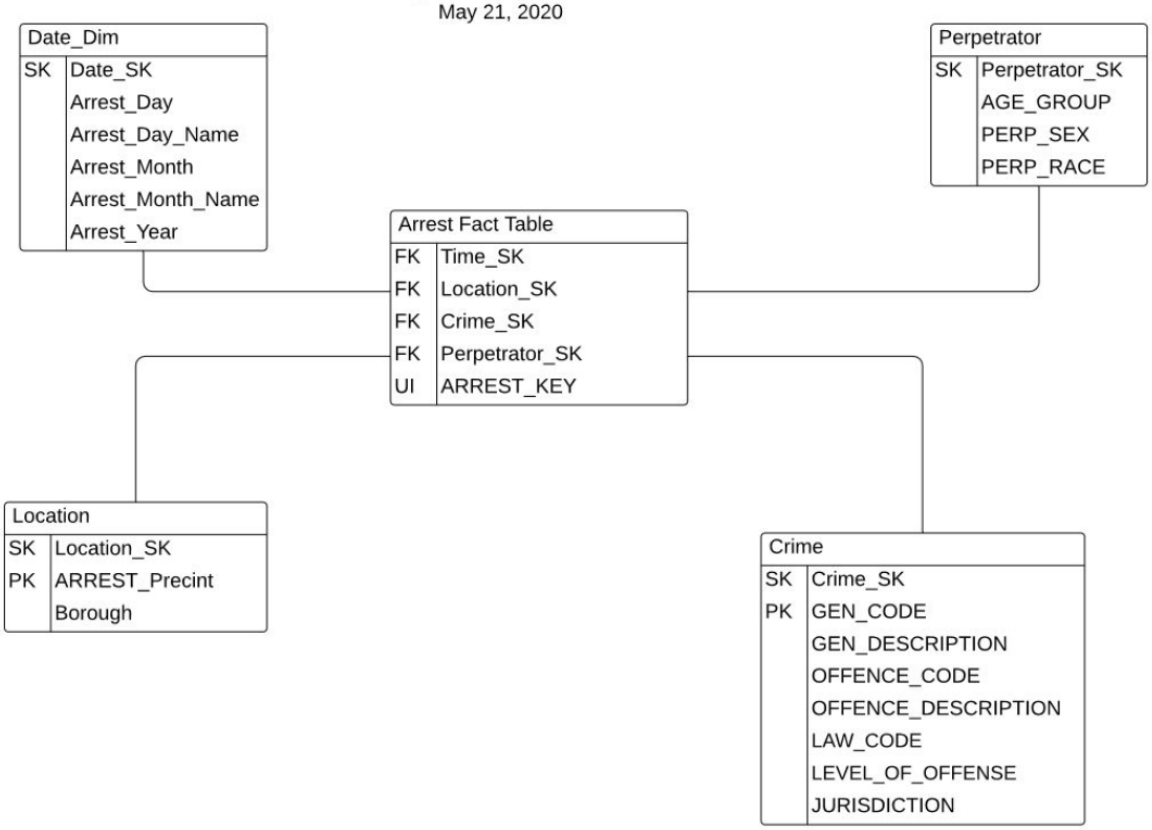
Introduction

The source data we are using for our data warehouse is the New York City the NYPD Arrest Dataset. There are over 215,000 entries in the NYPD Arrest data. The data will be used to NYPD headquarters to understand crime in New York City. The NYPD will be able to use the data warehouse to identify areas where more policing is needed and where certain types of crimes are happening more often.

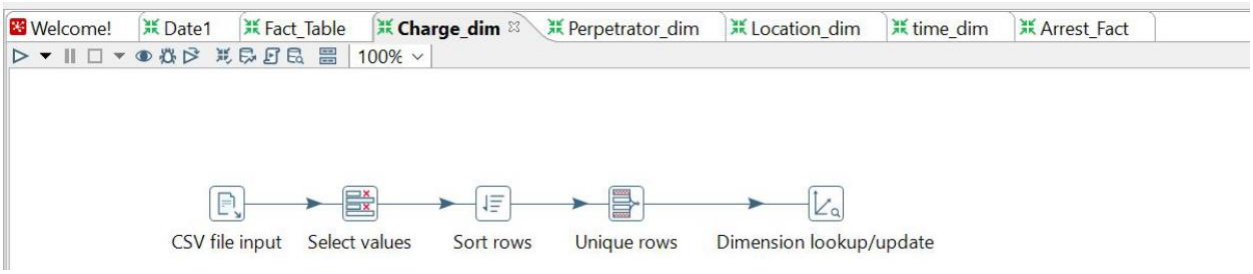
Group 4 Dimensional Model

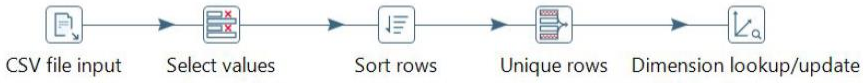
Paramvir Singh, Ricky Vidals,
Anthony Ferraro, Andrew Ivanov

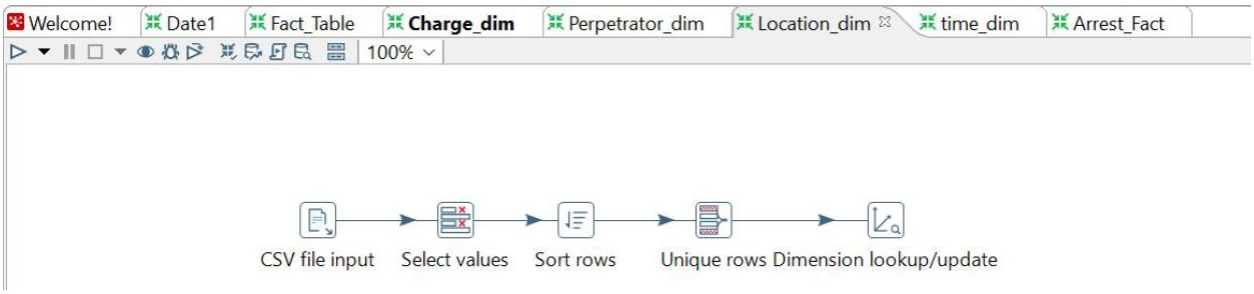
May 21, 2020

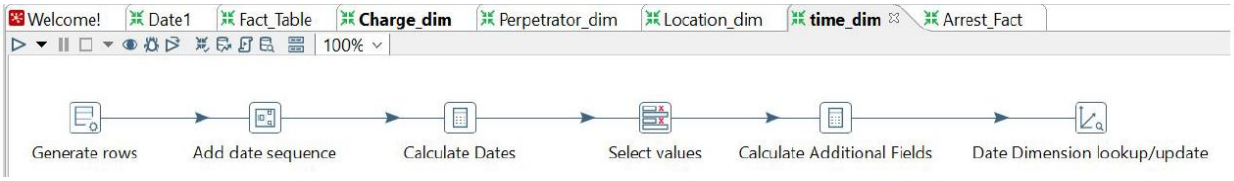


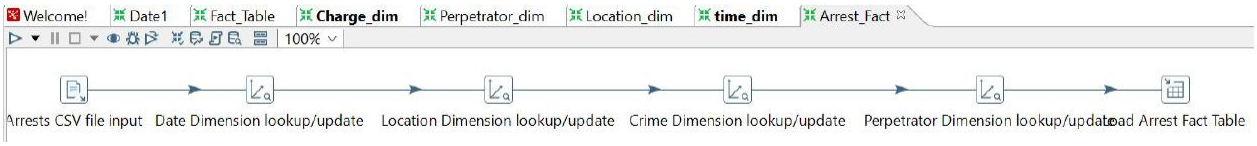
A description and screen pictures of the Pentaho ETL process





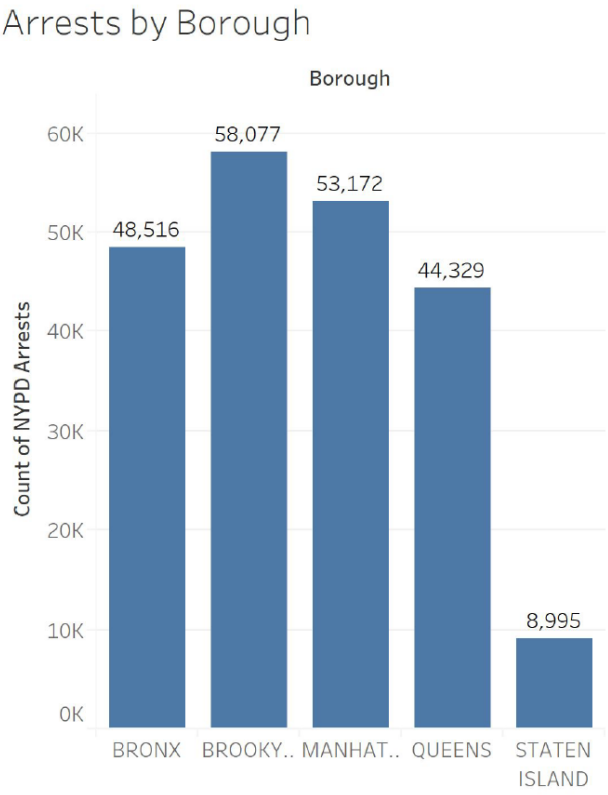




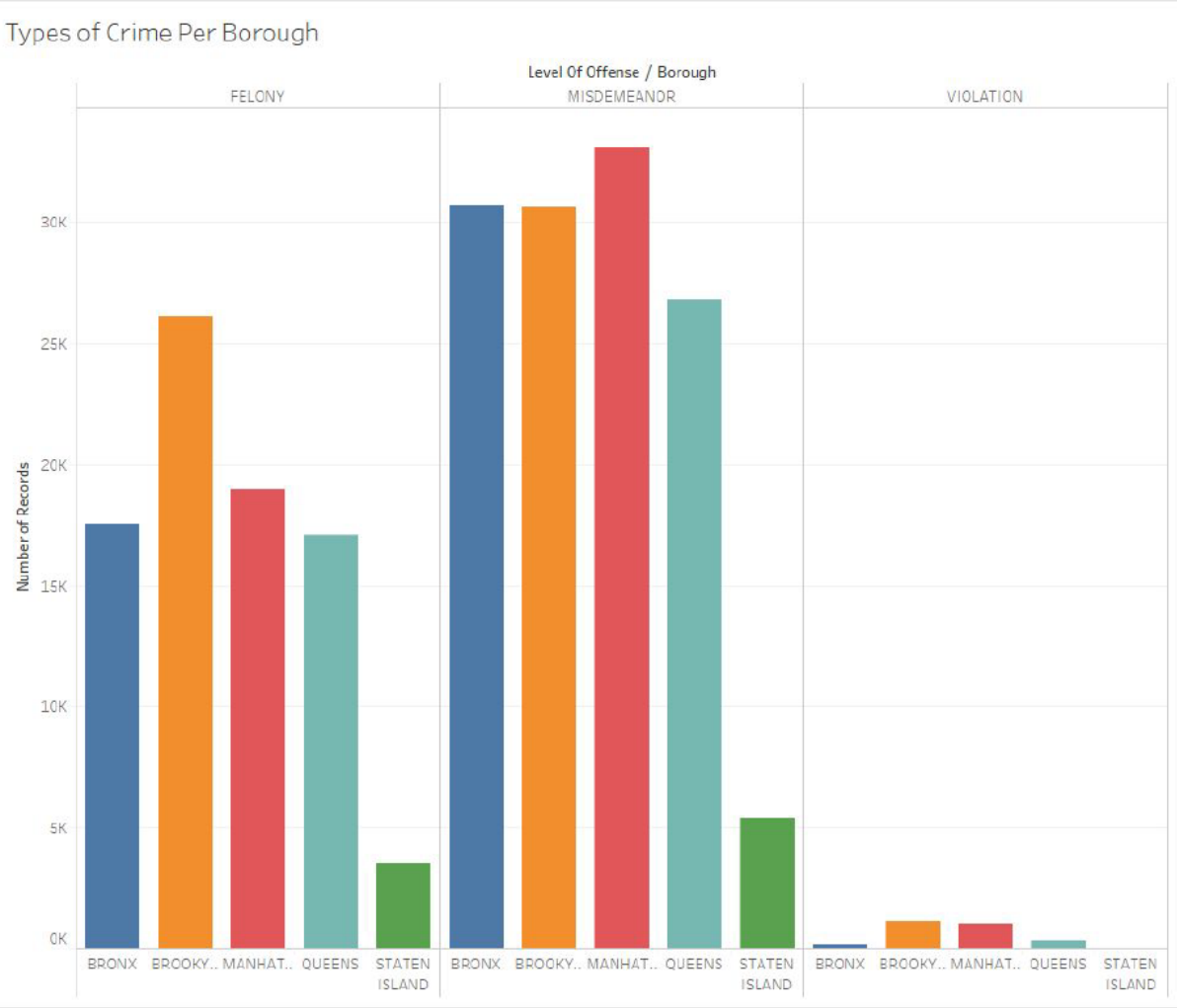


Applications and visualizations developed based on the data warehouse data.

This heat map identifies the precincts with the highest arrests per borough and compared arrests to all precincts in the city.



This bar chart highlights the differences in arrests in each borough. We can see that range of arrests in the Bronx, Brooklyn, Manhattan, and Queens is about 14,000. Staten Island has significantly less crime than the other four boroughs.



This bar plot separates the crime in each borough into three categories, felonies, arrests, and violations. Brooklyn has almost 8,000 more felony arrests than the second highest borough Manhattan. The Bronx, Brooklyn, Manhattan, and Queens are relatively close in misdemeanor arrests.

Conclusion

a) the group's experience with the project (which steps were the most difficult? Which were the easiest? what did you learn that you did not imagine you would have? if you had to do it all over again, what would you have done differently?)

The most difficult part of the project was figuring out how our source data was going to be transformed into dimensions. We had to rework the star schema a couple of times because of the time dimension. First, the original schema included two fact tables. However, the fact tables were of different granularity so that made it not possible to use two different data sources. One of the datasets was property values that were evaluated annually and the other was arrest data that was recorded daily. A solution we thought might fix this is making everything yearly. We did not get a chance to explore this because of time restraints so we ended up using only one dataset. As a result, we learned that the planning stage needs to be revisited multiple times. The easiest part of the project was getting everyone on board to help out in some way. Each of us has a different skill that benefits our project. If we had a second chance we would spend more time learning the advanced tools provided in Pentaho.

b) if the proposed benefits can be realized by the new system.

NYC would be safer because not only would the system track crime but also identify areas where policing can make a difference. There are ethical issues in using such a powerful tool and that should be addressed by the local government. Despite this drawback, we think a tool like this can prevent future crime in areas where cooperation is difficult. The data warehouse can also be used by the NYPD to identify areas in the city where additional police presence is needed to lower the amount of crimes occurring.

c) final comments and conclusions
In our final remarks, we wish to mention that this project was challenging to complete. We spent a lot of time focusing on combining our NYPD data with the real estate dataset that we had. Initially we had planned to use our real estate dataset to analyze real estate prices compared to NYPD arrest data in a particular neighborhood or zip code. After emailing with the professor last week, we decided to drop the real estate dataset and focus only on creating a data warehouse for the NYPD data.