Outliers

Please refer to <https://github.com/Miika0320/dataScience.git> under util to find the python code.

Using the neighbourhoods, dates, number of crimes, average rent, number of rentals, and the computed safety score, our outliers were computed using sklearn.svm.

A screenshot of a graph

Description automatically generated

From the outputted ouliers.csv file we can determine that the data was biased toward Sandy Hill (id 47) as we gathered much of our rental data from student housing which has the on-campus bias of being in Sandy Hill. The most dangerous neighbourhood based on our calculated safety scores was Trend-Arlington (id 65) followed closely by Stittsville (id 72). The safest neighbourhoods were Elmvale - Eastway - Riverview - Riverview Park West (id 52) followed by Iris (id 32). Our biggest outlier for number of crimes was Stittsville as within 2016-2017 of the approximate 62000 crimes that occurred, 61000 approximated were in Stitsville, however that is only an outlier based on date. For our entire data mart, Sandy Hill was an outlier for most crimes, as well as most rentals. Rothwell Heights - Beacon Hill North (id 73) was the highest outlier for rent and Merivale Gardens - Grenfell Glen - Pineglen - Country Place (id 67) was the lowest. What we can improve on with this data is to collect more data to have less bias in concentrated areas as well as ass other factors in such as proximity to high traffic areas when calculating the safety scores.