PM5 Data Warehousing

Introduction

As a crime reporting and analysis service, it's critical to provide useful insights based on collected data todate. To achieve that, we would rely on major external data sources and both identify the correlations with crime, and provide meaning predictions for our users.

The main influencer on crimes that we can hypothetically assume is the city economy which also closely connects to the population change, thus we believe it shall provide a clear picture of the reaction from crime activities. Out of curiosity, we are also interested in finding out the impact from general Seattle climate, given that being a universally fitting data set, which can potentially pose certain effect on criminal behaviour.

Datasets

1. Building Permits

   https://data.seattle.gov/Permitting/Building-Permits/76t5-zqzr

   This data is derived from sensor stations placed on bridges and surface streets within city limits. Each station has a temperature sensor that measures the temperature of the street surface and a sensor that measures the ambient air temperature at the station each second.

   From this data set, we planned to find correlation between the air temperature and the crime rates aggregated on a monthly basis.

2. Road Weather Information Stations
   https://data.seattle.gov/Transportation/Road-Weather-Information-Stations/egc4-d24i

   This data contains all building permits issued or in progress within the city of Seattle.

   Unfortunately due to the nature of economy, metrics at high levels are only aggregated at longer time spans. Therefore we moved over to the city building

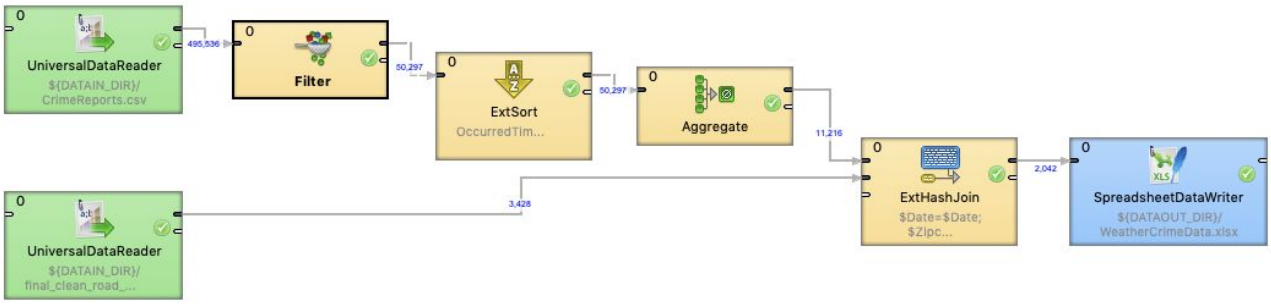
permit data which is an effective indicator of the city economy with a workable data volume.

Road Weather Information Stations

To analyze data against Seattle road weather data.

We got our weather data from Seattle government website, and based on a huge amount of temperature data resources in Seattle district, we firstly cleaned our data source before analyze the relationship between weather condition and crime reports. We select year 2017 as our research target because of its timeliness and completeness. The original data was collected in unit of minutes, in order to have rational and feasible evaluations based on our crime scenario, we aggregated the weather data in date unit and calculated the average temperature as well as the maximum and minimum value in each day.

1.ETL Workflow



The above ETL workflow includes the following components:

1) Universal Data reader

Import both the crime reports data and weather data.

2) Filter

Select the crime reports data from 2017-1-1 to 2017-12-31.

3) ExtSort

Sorts metadata by chosen representative timestamp that allows the following Aggregate to function.

4) Aggregate

Based on our analysis goal, this component aggregates both data to get the total record counts on a monthly basis.

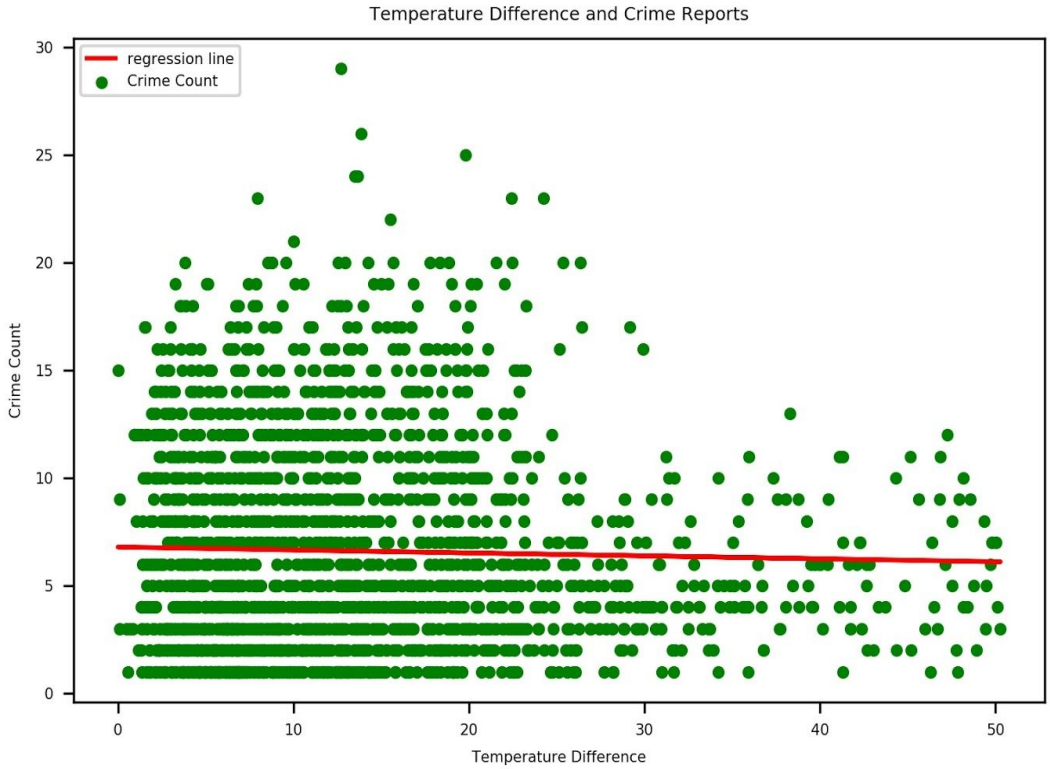
5) ExtHashJoin

Joins both data on the chosen timestamp key.

6) SpreadsheetDateWriter

Writes the joined data in a spreadsheet format.

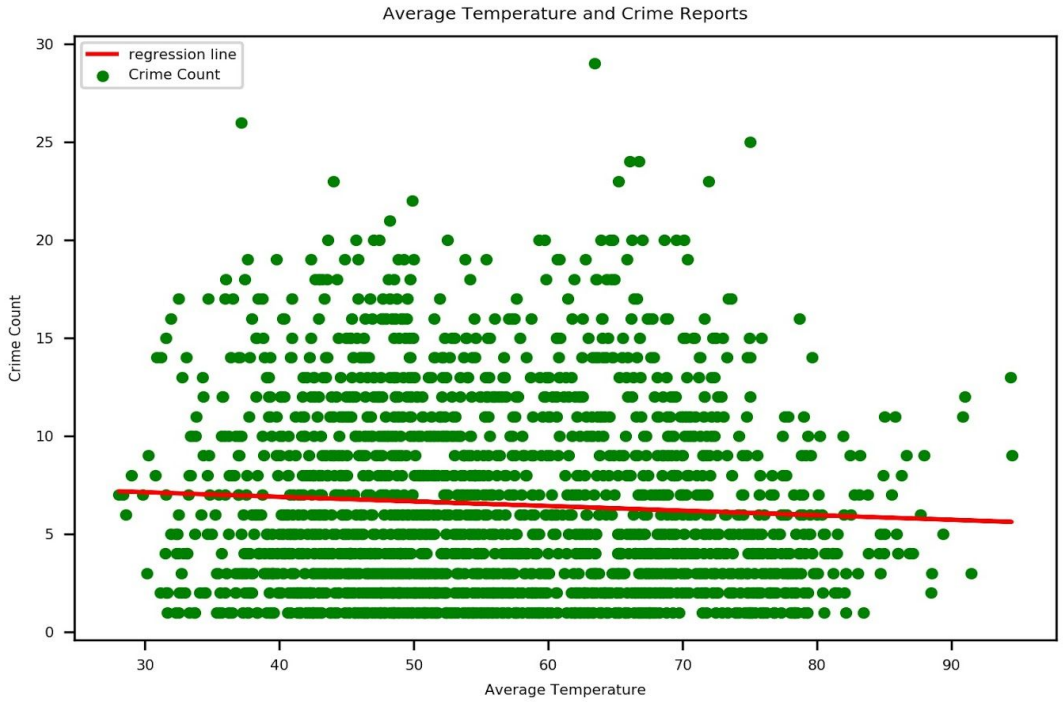
2. Analysis Result



A. The relation between temperature difference and the number of crime reports based on linear regression analysis using python

Based on our earlier hypothesis, weather condition could have influences on the frequency of crime. For example, extreme weather such that foggy and windy may reduce crime rate, for it may bring adverse impact or increase difficulty on conducting crime. Also, there would be high possibilities that robbery and theft happening in a rainy or snowy day, for the weather could cover up and wash away criminal marks so that it would be easier for criminals to escape from investigations.

The chart above shows an inverse proportional relation between temperature differences and number of crimes. With the increasing of daily temperature differences, it reflects a slightly decreasing of number of crime reports. Which means that weather is an influential factor to the frequency of crime happening in Seattle, however, the influence is minor and it will not be a determinant factor when analyzing crime rate in an area.



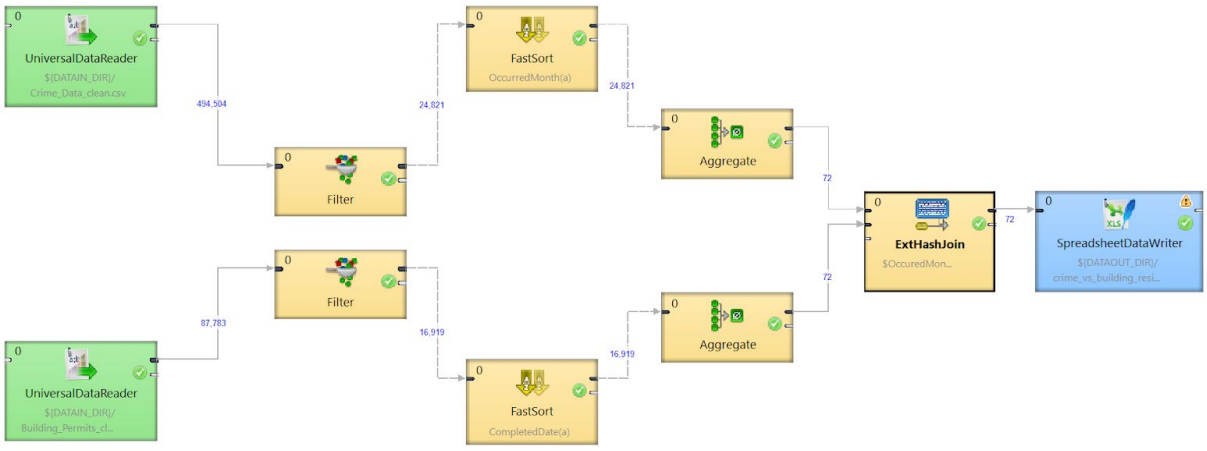
B. The relation between average temperature and the number of crime reports based on linear regression analysis using python

The graphs above represents the relationship between temperature and the number of crime reports. The red straight line in the plot shows how linear regression attempts to draw a straight line that will best minimize the residual sum of squares between the observed responses in the dataset and the responses predicted by the linear approximation. The value of correlation coefficient is near zero, which shows there is no strong linear relationship between weather conditions and crimes.

Building Permits

Before the analysis, certain level of data cleansing was performed to both clean up noises and extract the most relevant data records. For instance, not all permit records are complete in terms of registration time or project category, nor do many projects reached completion yet. Besides, we have determined that only projects that are at certain scale may have significant impact on crime activities, it's hard to believe that small scope single family retrofits and additions would have much influence.

1.ETL Workflow



The above ETL workflow includes the following components:

1) UniversalDataReader:

Imports both the crime and the building permit data, parsed original data format and extracted them into metadata.

2) Filter:

Filters both data to focus only on the period from the beginning of 2012 to the end of 2017. Then for the other two following analysis targeting on specific segments, the filter helps to only focus on either Residential or Commercial scope.

3) FastSort:

Sorts metadata by chosen representative timestamp that allows the following Aggregate to function.

4) Aggregate:

Based on our analysis goal, this component aggregates both data to get the total record counts on a monthly basis.

5) ExtHashJoin

Joins both data on the chosen timestamp key. Note that in order to bringing the permit number count to the same scale as crime rate, permit counts have been multiplied by 10 in the joiner. Nevertheless, the overall data pattern shall still serve the purpose of comparison.

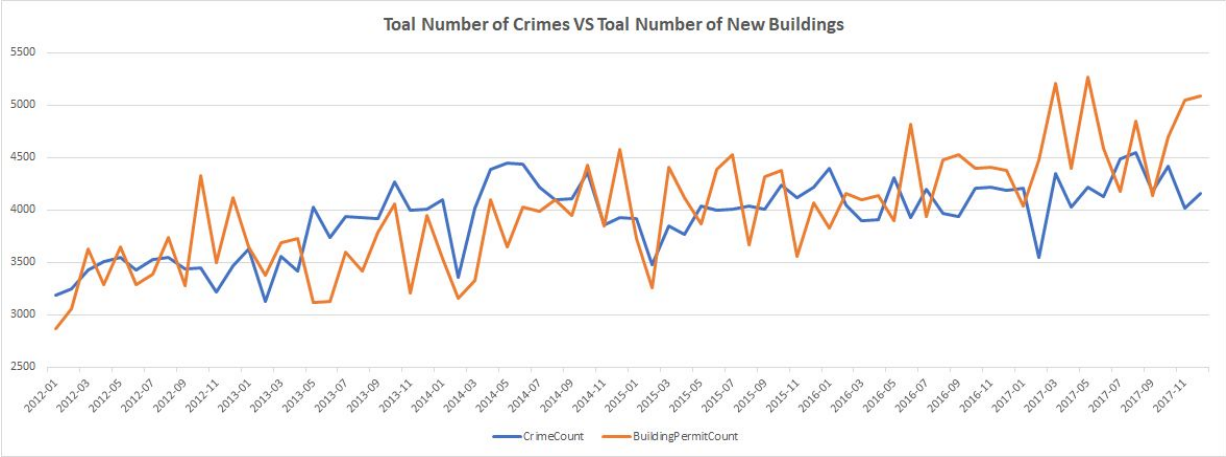
6) SpreadsheetWriter

Writes the joined data in a spreadsheet format.

2. Analysis Result

A. Total crime incident number against number of building permit, month by month

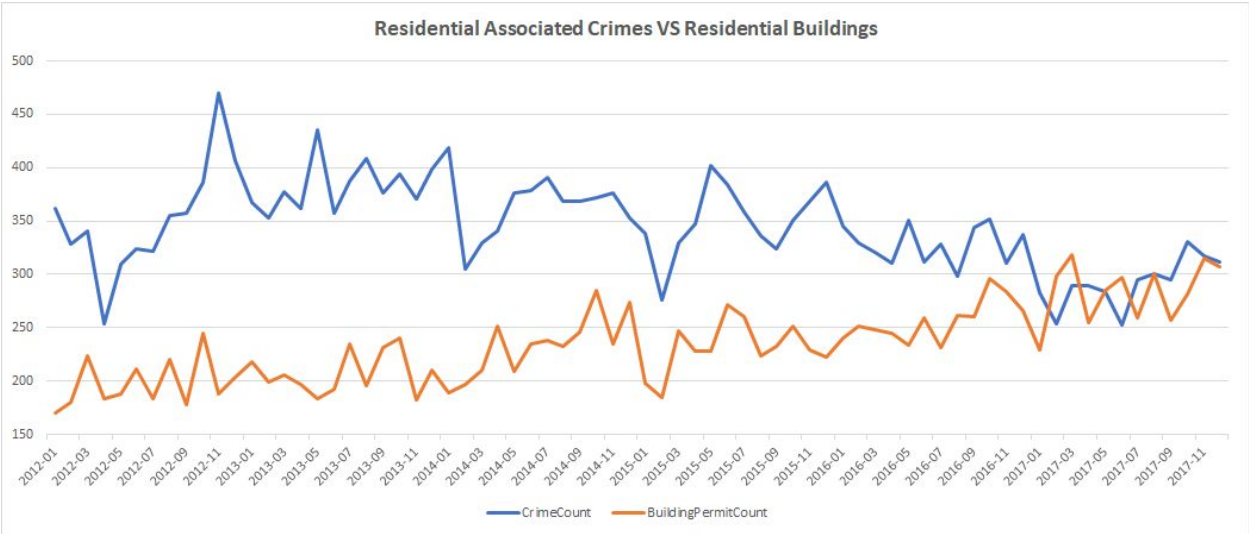
   General hypothesis for this analysis is that there should be positive correlation between the city development and the rate of crime activities, since theoretically a mass city development usually brings more residential as well as transitioning population, which fundamentally increases the rate of all sorts of social activity, including crimes.



As a result, the general trend from both dataset tightly follows each other, which very well validates our hypothesis. This finding opens up the opportunity for us to provide additional service of crime predictions on a high level through this particular perspective.

B. Total residential related crime number against residential construction permit number, month by month

General hypothesis for this analysis is that the overall trend of residential oriented crime activities should also follow the pace of residential development in the same direction.



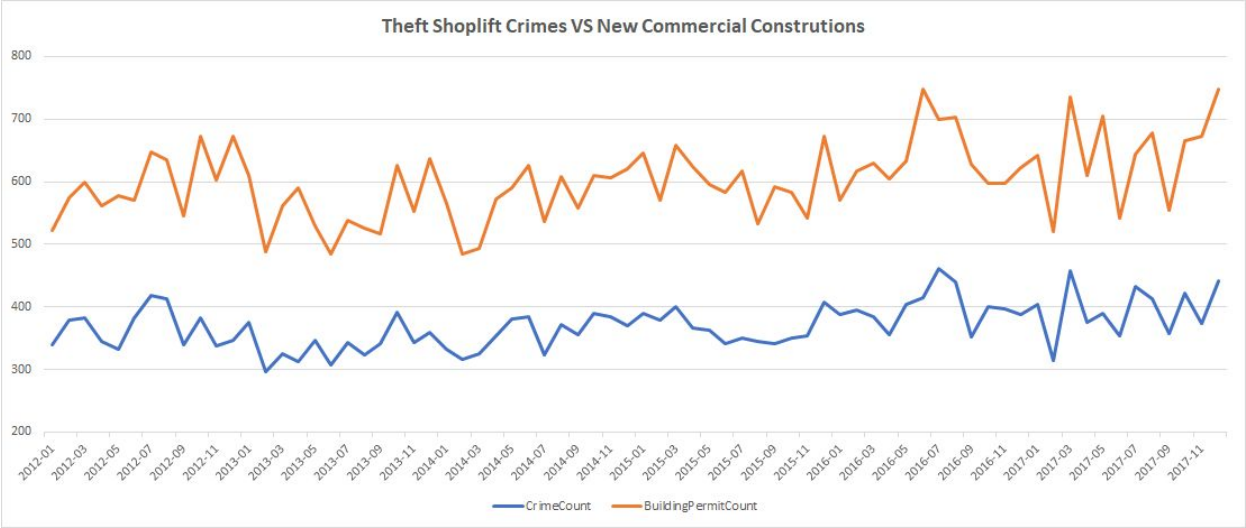
Interestingly enough, contrary to our original hypothesis, the result actually tells a different story. With a steady growth of residential development, the crime rate, despite fluctuations, actually drops through the period, and only ticks up a little bit starting mid-2017.

A couple of acceptable explanation to this result would be that:

- With the development in the residential segment, more property tax was collected and have been invested towards community safety.
- Real estate development have also brought more employment, which could also effectively drop the crime rate.

C. Commercial related crime number against commercial construction permit number, month by month

General hypothesis for this analysis is that the overall trend of commercial associated crime activities should also follow the pace of commercial development in the same direction.



The result generally agrees with our hypothesis. Possible reason for a different trend from the residential sector would be that, majority of the commercial development through recent years are commercial office buildings, and typically

those buildings are equipped with comprehensive security solutions, therefore such new constructions do not easily become new targets to potential offenders. The typical commercial oriented crime activities should still happens against shopping stores, and its rate grows following the growth of general population.