



IS424

Data Mining and Business Analytics

Term 2 Group Project

Crime Data Analysis

Professor: FeiDa Zhu

George He Shuxian

Varun Ponnuru

Eric Chow

Table of Contents

INTRODUCTION	3
DATA DESCRIPTION	4
DATA PREPROCESSING	5
TOOLS EMPLOYED	6
SAS ENTERPRISE MINER	6
MANIFOLD	6
PROJECT OBJECTIVE	6
RESULTS	8
PRELIMINARY ANALYSIS	8
LOCATION BASED	8
TIME BASED	9
CRIME TYPE BASED	10
ANALYSIS	10
CLASSIFICATION: DECISION TREE ANALYSIS	10
CLUSTERING: GEOSPATIAL DATA ANALYSIS	12
ASSOCIATION ANALYSIS: CRIME DATE FREQUENCY ANALYSIS	14
LIMITATIONS	15
CONCLUSION	16

Introduction

Singapore has built a reputation for being a very safe city. The island state has one of the lowest crime rates in the world. However, as a recent Singapore Police campaign has articulated – “Low Crime doesn’t mean No Crime”. With a view of generating results to help the Singapore Police Force tackle crime effectively, we mined a set of crime data. Our findings are presented in this report.

From our preliminary analysis we found that a large number of crimes occurred in residential areas. This is pertinent as Singapore has a very high population density and a large proportion of the population live in government housing. Therefore, besides gearing our results towards the Singapore Police Force, we sought to help the Singapore Housing Development Board as well.

Due to the sensitive nature of crime data, our group was given a censored version of the dataset. Our version was wiped of any sensitive details, which limited our analysis, but was a necessary step.

We have attempted to mine the data for useful and relevant patterns, to substantiate our suggestions to the Singapore Police Force and Housing Development Board. We encountered limitations along the way, which are discussed later. We also brainstormed ideal circumstances in terms of data available to us, to construct an ideal crime prediction model.

Data Description

The dataset used, was provided to us, by Prof. Kam Ting Seong of the Singapore Management University. The data is primarily geospatial data with a few additional attributes describing the type of crime. To be systematic, we segmented the data into two types – record and ordered. The record data was represented by descriptions of the crime including the time and location of occurrence. The ordered data was represented by geospatial data that allowed us to plot the crimes on a map of Singapore, for a visual overview of the data.

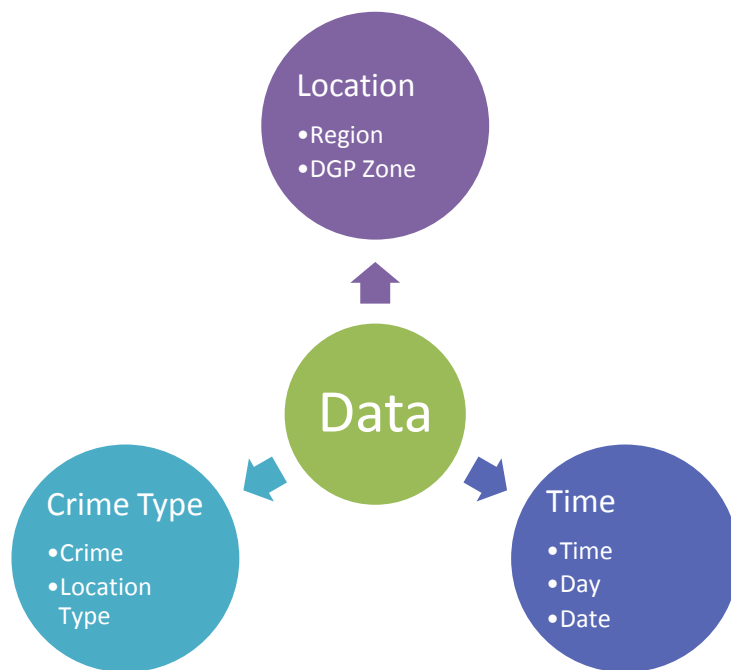
The dataset contained a total of over 5,000 records, ensuring an appropriately large size to mine. The dataset contained the information about the crimes that occurred in Singapore in the year 2000. The attributes are as follows:

Attribute Name	Description	Data Type
First Date	The date the crime started	Interval
LastDate	The date the crime record was closed	Interval
FirstTime	The time the crime start	Interval
LastTime	The time the crime record was closed	Interval
Crime	Type of crime	Ordinal
Location Type	The type of Location at which the crime occurred	Ordinal
FirstDay	The day of the week the crime started	Interval
Last Day	The day of the week the crime record was closed	Interval
DGPZ Name	The Development Guide Plan (DGP) zone in which the crime occurred	Ordinal
Region	The region of crime scene	Ordinal

To expand our analysis, we generated another few fields from the data, namely:

Attribute Name	Description	Data Type
First Time Of Day	The period of time during which the crime occurred. E.g. Night, Morning, etc.	Ordinal
Holiday	Whether the FirstDate was a public holiday or not	Binary

From a preliminary examination of the data attributes, we divided the attributes into three main types:



Given any set of data, there are endless combinations of attributes that one can use to represent the data.

This division helped us categorize the data for analysis. By asking ourselves how best to combine two segments to present relevant results, we were able to narrow down our results.

Data Preprocessing

The data required minor preprocessing. While scanning the data, we found certain fields that were recorded in 2001. We removed this data as it would make the analysis a little inconsistent.

The time of occurrence of the crime was discretized to allow for analysis by grouping them into common categories such as Late Night, Evening, and Morning to name a few. The discretization allowed for a more general analysis of the time of crime.

The dates were binarized for the purpose of differentiating public holidays. An analysis was performed using the binary holiday data, which would have been tedious otherwise.

Tools Employed

We have attempted to visualize the data using SAS Enterprise Miner and Manifold. SAS Enterprise Miner was used to generate graphs to compare different aspects of the data. Manifold was used to plot the geospatial aspects of the data on a map of Singapore.

SAS Enterprise Miner

SAS Enterprise Miner is a professional data mining software, capable of performing a number of data mining operations on large sets of data. The SAS suite is capable of exporting datasets from external files, which was employed in our project. Enterprise Miner offers a number of different views of a dataset, which help immensely in visual analysis.

The SAS Data Mining suite however, doesn't come without drawbacks. Being a very complex suite of software, it is often prone to cross program errors that are not straightforward to debug. Due to the complexity, it is also difficult to master its functions in a limited amount of time. The team faced a steep learning curve when it came to employing SAS effectively. Nonetheless, the tool's capabilities overshadowed its snags, proving useful to our analysis.

Manifold

Manifold is a Geographical Information System (GIS) package, used to represent and plot geospatial data on graphic maps. The software supports multiple layers of geospatial data as well as the ability to plot different types of co-ordinates on the same map.

Manifold is useful in data mining as it allows for another layer of analysis, into geospatial data. The software is capable of handling a large number of data points, allowing it to be used in conjunction with other data mining software.

We have employed Manifold to generate a map of Singapore with the different crime locations plotted. This aids tremendously in visually identifying clusters and patterns. Humans are more equipped to identify patterns and Manifold allows us to do that with geospatial data.

Project Objective

The central question asked in the project is what information can be generated to help improve the coverage and efficacy of the Singapore Police Force. Our two pronged analysis of record and ordered data allowed us to address these two aspects of the police force:

- *Police station coverage* - Through our map-based observations, we were able to identify blind spots in the police station's coverage radii.
- *Police manpower allocation and patrol frequency* - From data visualization, we have identified categorical hotspots where crime has historically been high.

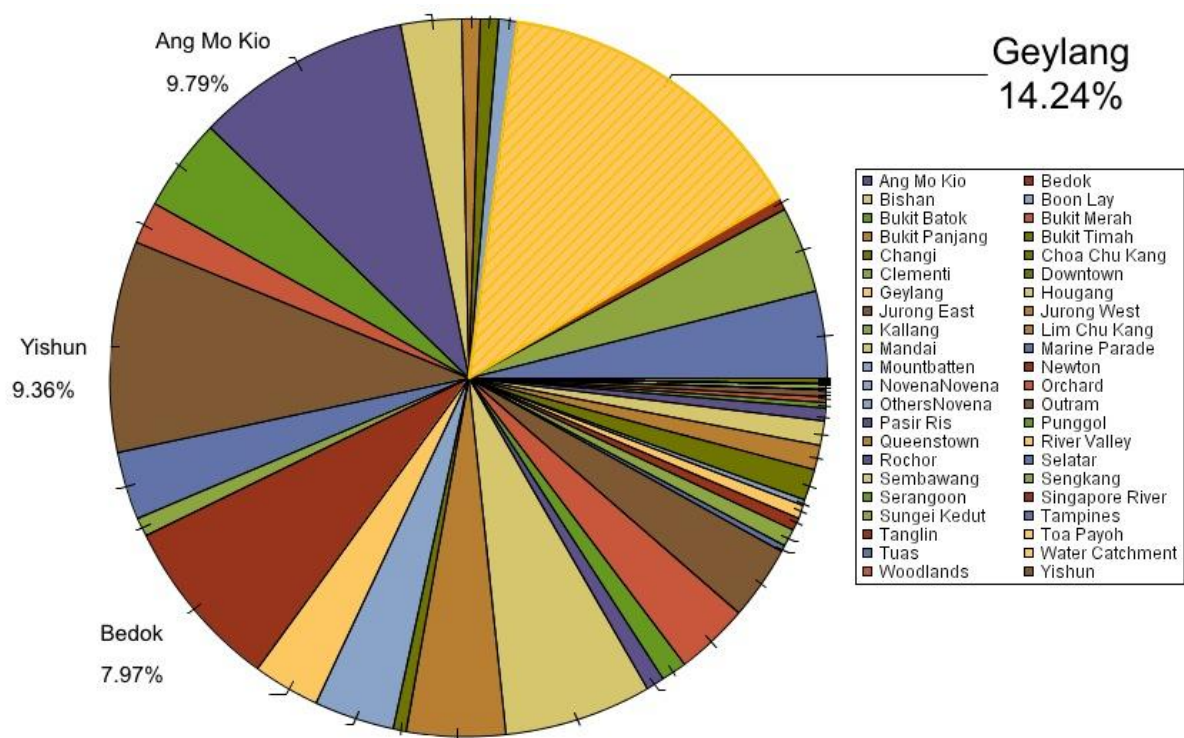
Given Singapore's limited police force and land area, it is pertinent to ensure that staffing is done appropriately and police stations are located effectively close to the crime scene to improve efficiency. To tackle crime on an everyday basis, the frequency of patrols could also be increased during times when crimes have historically been high.

Results

Preliminary Analysis

We used SAS Enterprise Miner to conduct a preliminary analysis of the data. We split the analysis to focus on one category of data at a time. The analysis proved extremely useful as it quickly identified hotspots in each category. Just the knowledge of these hotspots, provides impetus to organize manpower and patrols in a more effective manner.

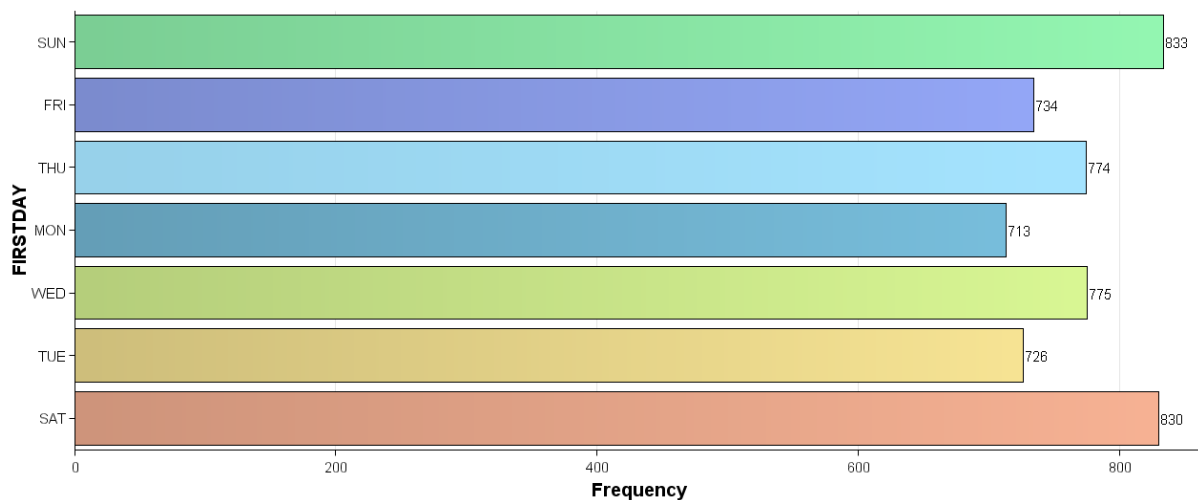
Location Based



From the above chart we can see that Geylang has the highest level of crime amongst all the DGP zones. We can therefore recommend that the Geylang police stations be allocated more manpower. Additionally the frequency of police patrols in Geylang could be increased so as to increase the probability that a police officer is in the vicinity when a crime occurs.

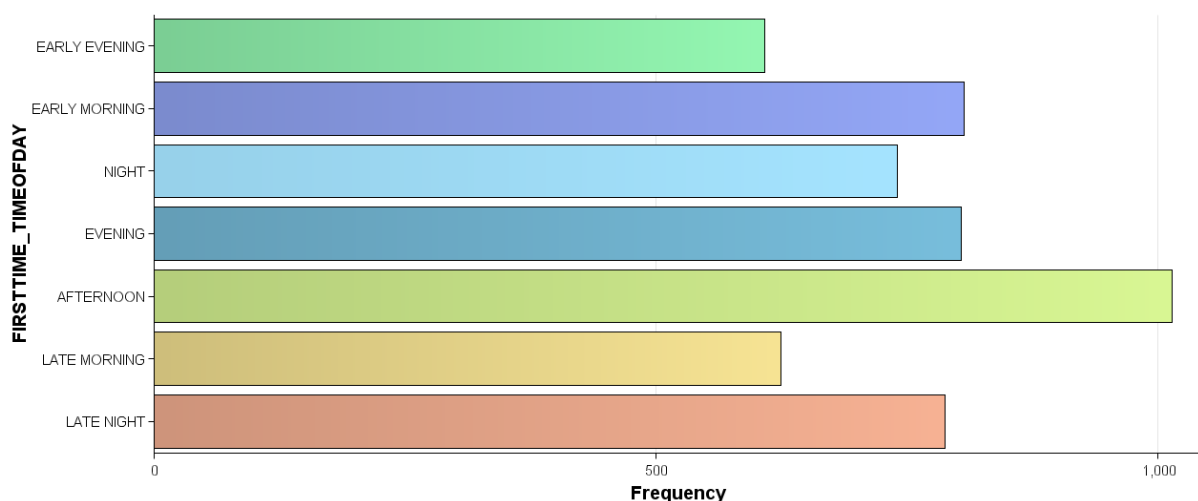
The DGP zone data is very important as it also highlights the problem of crime to urban planners. The rate of crime in a particular area is dependent on a number of factors including type of housing and overall demographics, which urban planners need to address to alleviate crime.

Time Based



We can see from the above graph that the maximum number of crimes occur over the weekend. The frequency of crimes over the weekend is nearly 12% greater than the average crimes occurring on weekdays.

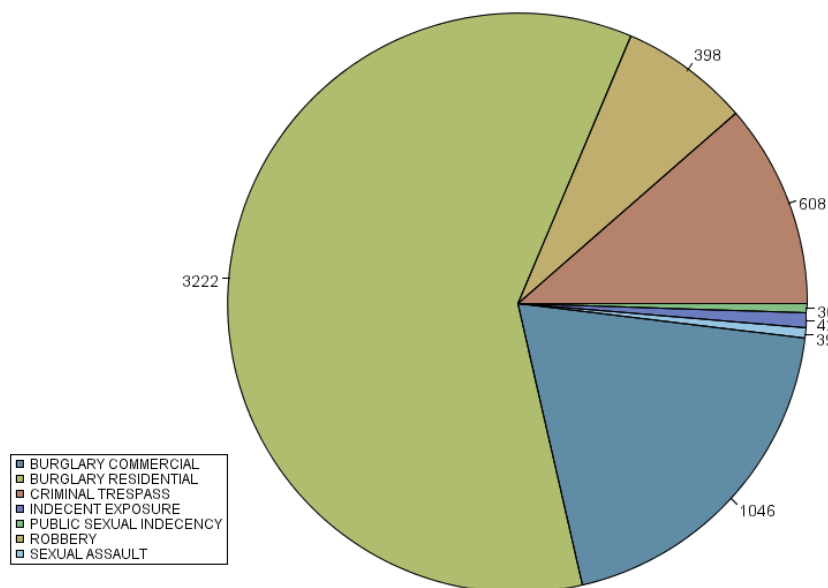
Considering that the largest segment of crimes is household burglaries, it is a popularly believed hypothesis that the crime rate increases over the weekend. We can postulate that this is because most people tend to go out at night on weekends, therefore tempting burglars. However, this hypothesis could be disproven by the next set of data, which shows the frequency of crimes by time of day.



We see that the largest number of crimes occur in the afternoons which leads us to question the popular notion that most crime occurs after dark. This is a very useful observation for the police force, as it goes contrary to popular belief.

This would imply that more police patrols and shift-based staff are needed in the afternoon as compared to late at night.

Crime Type Based



We see from the above chart that the most prevalent type of crime in Singapore is residential burglaries. A small majority of Singapore's land area is purely residential. With the density of population being very high, burglars have more opportunities to commit crimes. This is relevant to the Housing Development Board, which may be able to use the data to design campaigns to educate homeowners about security.

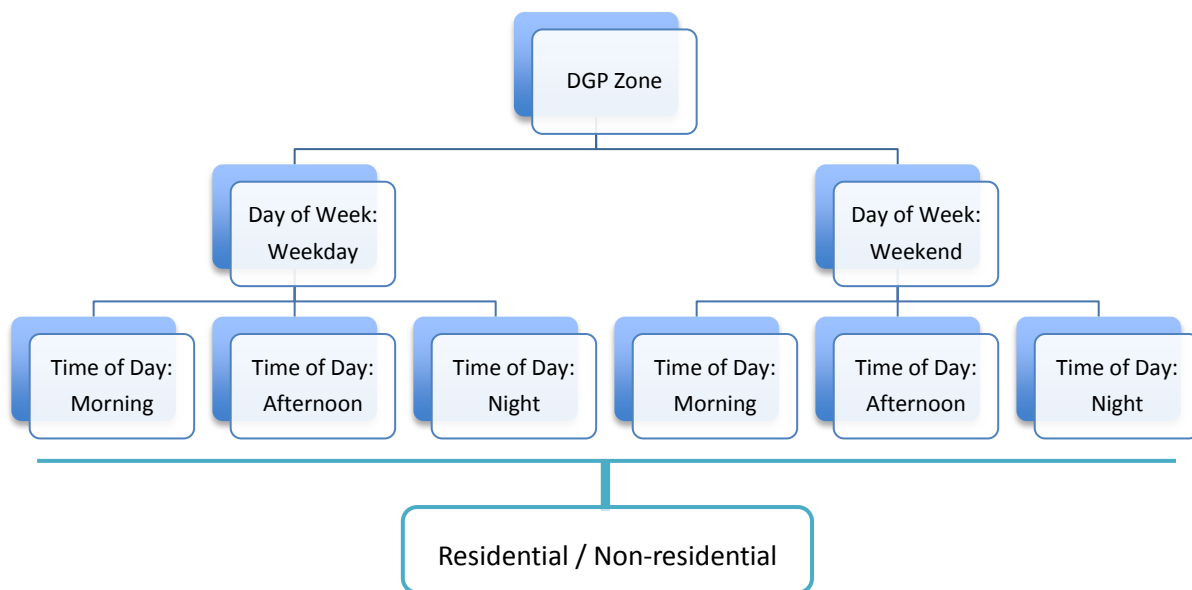
Analysis

We employed techniques learnt in class for the main analysis. We applied classification, clustering and association analysis. While the practical implementations of these techniques might be incomplete, they have been conceptually discussed below. The shortfalls of implementing them using software will be detailed in the section on Limitations.

Classification: Decision Tree Analysis

The ability to classify a potential crime accurately is immensely beneficial to the police. Using a decision tree, we can classify an anticipated crime as residential or non-residential and act accordingly. At this juncture, we recognize that due to the limited attributes that we have been provided with for the data, our model is not robust. However, we have attempted to use the resources provided to us to generate a model that will work to an acceptable level of accuracy.

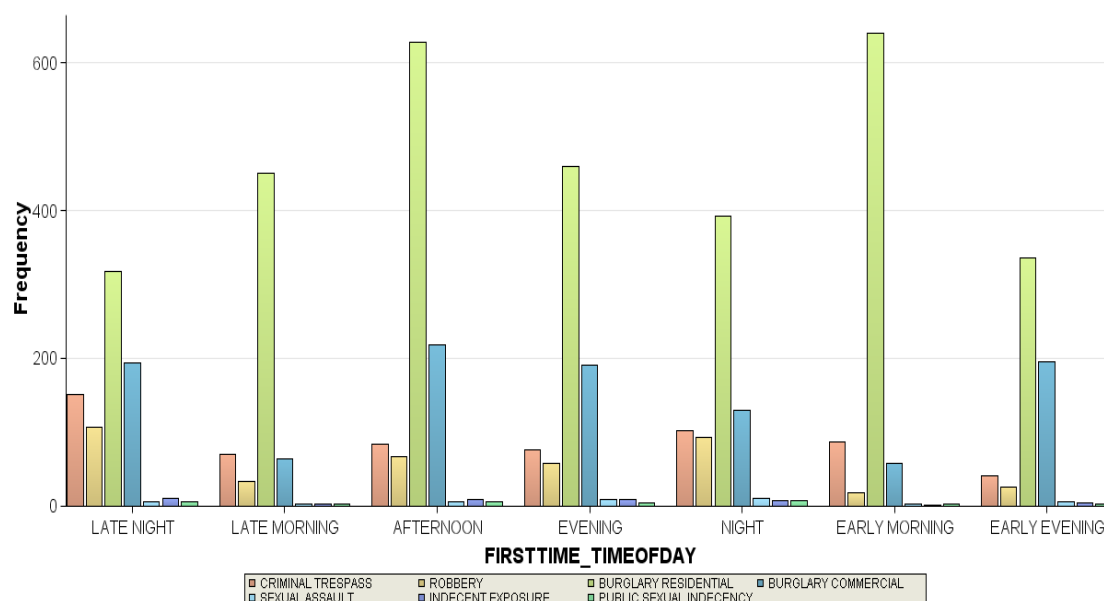
Our decision tree is as follows:



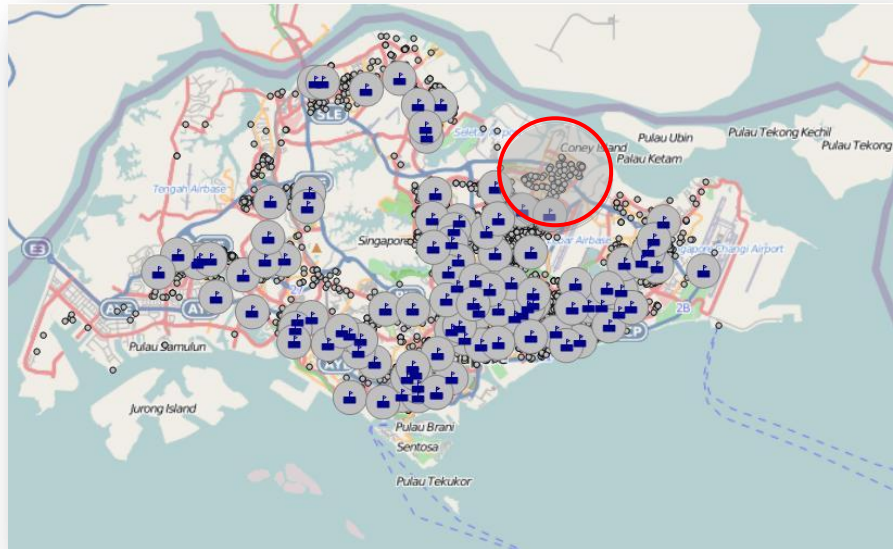
We see from our preliminary analysis that most residential crimes occur in early in the morning and in the afternoon. Also a larger proportion occurs on weekends. These trends are not correlated to the trends for non-residential crimes and hence the decision tree could prove to be accurate enough to predict this classification.

The following graphs support the assumptions used in constructing the decision tree:

1. Number of crimes by Time of Day, categorized by Type of Crime
2. Number of crimes by Day of Week, categorized by Type of Crime



Secondly, we plotted the various police stations in Singapore on the map and assumed an effective coverage radius of 1km, denoted by the circles around the stations. As we can see,



the coverage of police stations is not perfect with a few clusters being left uncovered.

In particular we focused on the cluster in the northeast region around Sengkang. This particular area has a fairly high crime rate, yet no police station coverage. Upon closer



examination, we find that the area is the Sengkang New Town. This is a township intended to replace the original Sengkang fishing village. The new town is relatively young, and hence

perhaps doesn't have a police outpost yet. However given the high rate of crime, it becomes an urgent issue for the government to setup a police station there to curb the crime as soon as possible.

Association Analysis: Crime Date Frequency Analysis

We attempted to perform an association analysis on the data. However, the SAS software that we intended to use did not work very well with our data. Due to the lack of further expertise and time, we decided to provide a possible application of association analysis, in theory.

One possibility of employing association analysis, is to analyze the frequency of crimes occurring within one week. With the data that we have, we could organize the crimes into datasets divided by weeks. For example, the first week would contain the days of the week that crimes occurred on.

With this itemset, we could mine for rules such as {Monday -> Wednesday}. It will therefore become possible to check for the reoccurrence of crime in a particular area given that a crime has already taken place there, in that week. Furthermore, we could look for an interval after which a crime reoccurs. For example, we may find that there's a nearly constant wait of two days between the occurrence and reoccurrence of crime in one area.

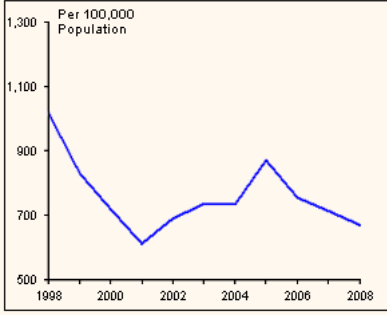
This is very useful to the police, who can forecast the day of reoccurrence and so make sure that their patrols are beefed up on that day.

The underlying mechanic supporting our theory is that the criminal world also keeps track of crimes that happen in one particular area. Burglars, for example, might be coaxed into thinking that a particular area is a soft target, if a crime has occurred there. Hence, they may lie in wait for a day or two and strike the same area again.

Limitations

While our findings were insightful, our project suffers from a few significant limitations. They are detailed below:

- *Data Attributes* –
 - The data has been censored to protect the privacy of those involved. This makes it impossible to know who committed the crime and whether there were any repeat offenders. With such information, we could perform an association analysis to determine what sort of crimes, a repeat offender may commit next.
 - The data doesn't contain information about the severity of the crime. Such a measure would be more useful to determine adequate manpower location and identify the really rowdy areas of Singapore.
 - Police station information would also be useful in adding depth to our analysis. With knowledge of the different police stations' coordinates, we could easily determine the coverage of each police station. We mitigated this concern by manually inputting the coordinates.

 - *Data Age* – The crime rate in Singapore is back at year 2000 levels – implying that overall, our analysis can be ported to the current year. However, there may have been significant policy and macroeconomic changes that underpin today's crime rate. Our data doesn't allow us to factor these in, rendering it somewhat obsolete.
- 
- | Year | Crime Rate (Per 100,000 Population) |
|------|-------------------------------------|
| 1998 | 1050 |
| 1999 | 850 |
| 2000 | 750 |
| 2001 | 650 |
| 2002 | 700 |
| 2003 | 750 |
| 2004 | 750 |
| 2005 | 850 |
| 2006 | 750 |
| 2007 | 720 |
| 2008 | 700 |
- *Inexperience with SAS* – Our team's inexperience with SAS was a large factor in determining the depth of our analysis. Given the heavy resource usage of the SAS suite, not all teammembers could even run the software. Troubleshooting was a problem as well, with SAS giving a variety of errors that the team did not know how to tackle. In hindsight, we should have tested out a number of alternatives before diving in to using SAS despite the tight time frame of the project.

Conclusion

In conclusion, our team has sought to achieve the following objectives set out:

- *Police station coverage* - Through our map-based observations, we were able to identify blind spots in the police station's coverage radii.
- *Police manpower allocation and patrol frequency* - From data visualization, we have identified categorical hotspots where crime has historically been high.

Through our analysis, we have been able to come out with the following recommendations pertaining to coverage and manpower:

1. Coverage
 - a. Build new police stations in the current coverage blindspots, starting with Sengkang New Town.
 - b. Increasing patrolling in areas where there are greater incidences of crime.
 - c. Perhaps increase the radius of effective coverage of larger police stations so as to cover the blindspots in their vicinity
2. Manpower
 - a. Increase the patrolling frequency in the afternoons
 - b. Increase the patrolling frequency over the weekends
 - c. Concentrate on high crime areas like Geylang

We would also like to recommend to the Housing Development Board to run campaigns to increase the level of personal security in HDB apartment complexes, so that residents are aware of the times and locations where crimes regularly occur.

We hope that through our analysis, we have come up with constructive suggestions to make Singapore an even safer place to live in.