**ISQS 7339**

**Prescriptive Analytics**

**Team 3**

**Project Deliverable – Final**

**Professor:**

Dr. Zhangxi Lin

**Team Members:**

Tiju Mathew George (R11287760)

Praveena Kathera (R11372933)

Fazil Sheik Mohammed (R11360243)

Prathamesh Tarkar (R11357607)

**11/02/2015**

Table of Contents

[1 Motivation: 3](#_Toc434178595)

[2 Objective: 3](#_Toc434178596)

[3 Project Overview: 3](#_Toc434178597)

[4 Data Set Description: 4](#_Toc434178598)

[4.1 Data Availability 4](#_Toc434178599)

[4.2 Provided Data: 4](#_Toc434178600)

[4.3 Descriptions: 4](#_Toc434178601)

[4.4 Data Quality: 9](#_Toc434178602)

[4.5 Data Pre Processing Tasks 9](#_Toc434178603)

[4.6 Data Exploration: 10](#_Toc434178604)

[4.6.1 Average Dropout vs Crime rate grouped by City 10](#_Toc434178605)

[4.6.2 Average of % below poverty vs Crimerate grouping City 11](#_Toc434178606)

[4.6.3 Normalincome Vs Crime rate grouping City 11](#_Toc434178607)

[4.6.4 Unemploymennt rate vs Crime rate grouping City 12](#_Toc434178608)

[4.6.5 Cloud cover vs Crime rate grouping City 13](#_Toc434178609)

[4.6.6 Crime rate vs Day: 13](#_Toc434178610)

[4.6.7 Crime rate vs Stormclass grouping by City 14](#_Toc434178611)

[4.7 Data Modeling Outputs 15](#_Toc434178612)

# Motivation:

The motivation of our project is to identify the crime rate of a particular city with data available from five different cities of US. Crime rate is determined with the help of various factors identified. We are trying to identify those factors which will help our police force to deploy in event of contingency plan. A model is being built which will relate all the relevant variables and help us to predict the crime numbers by crime type by scoring the scenarios for the future.

# Objective:

The objective of the project is to come to an inference on those factors like weather, celestial phenomenon, demographic mix, etc. which would be significant in determining the occurrence of crime. This would help the police department in designing their working and deployment pattern.

We went on to define our objective in terms of below aspects:

* **Police intervention based on the big data**: We have tried to indentify the natural factors and data from various aspects to decide when and how quickly the police can act to the situation.
* **Effect of day-light saving related to crime-rate:** Also we tried to check how day-light saving has its effect on the crime-rate, whether to suggest the police to keep a track of the prominent occurrence of crime in this period.

We have provided several sets of data related to crime, celestial phenomenon, weather, population, daylight etc. for achieving the goal. We need to analyze the various factors of the data dependent on the crime rate and on the basis of tat design a future model that can be applied to other cities as well.

The model is to build and predict the crime occurrences on the basis of time, period and places with the help of available variables and help the police department in handling and deploying the forces.

**We intend to achieve the following objectives in our analysis-**

1. **Perform prediction of the crime rate for a given year in the future.**
2. **Find out the significant variables used to determine the crime rate using a predictive analysis technique.**
3. **Determining whether daylight saving plays a role in affecting the crime rate.**
4. **Determine whether temperature plays a role in affecting the crime rate for a particular time in the year and for a given city.**

# Project Overview:

The hypothetical Techno heat (police resource & Deployment) center is using the previous study references to expand the efforts to look at the dynamics & specific elements of natural phenomenon on crime. Using the datasets provided, we are combining all the datasets and analyzing the datasets to understand how different factors such as unemployment rate, poverty rate, storm class, income level, cloud cover etc. affect the crime rate.

Our main goal in this project is to separate out the most significant factors that are affecting the crime rate and are combined into a single dataset which is done by building a model. In our project we are provided with daily crime statistics from 5 cities over the years from 2005-2012. We create different model to analyze the effect of different factors on crime rate. The interesting aspect of answering these questions would help police and criminal justice employees to get a better estimate of happening of particular crime type.

The crime data is provided at the census tract granularity by overlapping the reporting Crime Area data with the census tract maps; thus, some Crime Areas have reported crime data for a portion of the census tract, and other census tracts are completely covered with reporting Crime Areas. Thus it is important to decide which census tracts to include in your analysis and any imputation methods you might use to merge data provided by census tract with data provided by Crime Area.

# Data Set Description:

## Data Availability

The data sets are provided by the SAS Data Analytics Shootout 2014 project. The data lists the occurrence of crimes and the values of factors that may have favored or inhibited these crimes. The cities covered in this project include Atlanta, Chicago, Denver, Houston and Sacramento. This document lists the datasets provided to accompany the 2014 SAS Shootout Competition. Each dataset is described for content, the dataset names are provided and data elements are described.

## Provided Data:

* CRIME.sas7bdat
* POPULATION.sas7bdat
* UNEMPLOYMENT.sas7bdat
* SOCIOECONOMIC.sas7bdat
* HEATING\_COOLING\_DAYS.sas7bdat
* CALENDAR\_DAYS.sas7bdat
* SOLAR\_LUNAR.sas7bdat
* STORMS.sas7bdat
* WEATHER.sas7bdat

## Descriptions:

**1. CRIME.sas7bdat:** Crimes which occurred in Atlanta, Chicago, Denver, Houston, and Sacramento. The crimes are by day and hour, made available by the individual city police departments. Separate tables for each city were combined to make this dataset. Contains 12 columns and 1,739,055 rows.

**Crime data included for the following years:**

Atlanta: 2009 ‐ 2012

Chicago: 2005 - 2012

Denver: 2008 - 2012

Houston: 2008 - 2012

Sacramento: 2005 - 2012

* **Crime type:** The category of crime as reported by the FBI. Includes: assault, auto theft, burglary, homicide, larceny, robbery, sexual assault, theft.
* **ID:** Unique crime ID
* **City:** City where the crime occurred. Includes: Atlanta, Chicago, Denver, Houston, and Sacramento.
* **Date:** The date the crime occurred, in SAS date format.
* **Year:** the four-digit year

**•**  **Month:** numeric month

**•**  **Date\_num:** the numeric date of the month

**•**  **Day:** the three‐character day of the week (Mon, Tue, etc.)

**•**  **Hour:** the hour of the day, from 0 to 24, when the crime occurred

* **CrimeArea:** the police designated reporting area where the crime occurred. Each police department has a number of designated patrols areas. Some examples of police designated areas are “precincts”, “beats”, “districts”, or “neighborhoods”.
* **Tract\_ID:** The census tract in which the crime occurred, geographically calculated from the latitude/longitude given in the FBI crime data.
* **Area\_Ratio:** the fraction of the tract area that falls in the associated ‘Crime Area’. Each crime occurrence is associated with one ‘CrimeArea’ and one census tract.

**2. POPULATION.sas7bdat:** contains the population data by gender by age group by census tract from the 2000 and 2010 censuses. Census tracts from 2000 have been geographically transformed to 2010 census tracts. Contains 7 columns and 1,432,524 rows.

* **Year:** the four-digit year (2000 or 2010)
* **Gender:** male or female
* **Age group:** the age groups as reported by the Census Bureau. Includes: under\_1, 1, 2, 3...99, 100\_to\_104, 105\_to\_109, 110\_and\_over.
* **Tract\_ID:** the ID of the Census Tract.
* **City:** City of population record.
* **Population:** the population of people in the census tract, by gender and age group3.
* **Census tract area:** area of land in the census tract in square miles.

**3. UNEMPLOYMENT.sas7bdat:** contains the unemployment rate by city by month for the years 2005 to 2012. Taken from the US Bureau of Labor Statistics. Contains 5 columns and 448,704 rows.

* **Cityname:** Includes: Atlanta, GA; Chicago, IL; Denver, CO; Houston, TX; Sacramento, CA.
* **Tract\_ID:** The ID of the Census Tract.
* **Year:** four-digit year, from 2005 to 2012.
* **Month:** numeric month.
* **Unemployment\_rate:** percent unemployment. Value calculated using the formula: (Number of unemployed people 25 years and older)/ (Number of people 25 years and older)

**4. SOCIOECONOMIC.sas7bdat:** contains the yearly dropout rate and normalized income by census tract. Calculated from the Census Bureau’s American Community Survey estimates for the years 2005 to 2012. Contains 6 columns and 37,392 rows.

* **City:** Includes: Atlanta, GA; Chicago, IL; Denver, CO; Houston, TX; Sacramento, CA.
* **Tract\_ID:** the census tract ID.
* **Year:** the four-digit year, from 2005 to 2012.
* **Normal income:** normalized income, calculated as median household income divided by median household value.
* **Dropout rate:** the proportion of people over age 25 who did not complete high school.
* **Pct\_below\_poverty:** percent of population living below the poverty line.

**5. HEATING\_COOLING\_DAYS.sas7bdat:** by city for years 2008 to 2012. From the US National Weather Service: “To calculate the heating degree days for a particular day, find the day's average temperature by adding the day's high and low temperatures and dividing by two. If the number is above 65, there are no heating degree days that day. If the number is less than 65, subtract it from 65 to find the number of heating degree days. For example, if the day's high temperature is 60 and the low is 40, the average temperature is 50 degrees. 65 minus 50 is 15 heating degree days. Cooling degree days are also based on the day's average minus 65. They relate the day's temperature to the energy demands of air conditioning. For example, if the day's high is 90 and the day's low is 70, the day's average is 80. 80 minus 65 is 15 cooling degree days.”

([http://www.erh.noaa.gov/cle/climate/info/degreedays.html).](http://www.erh.noaa.gov/cle/climate/info/degreedays.html) Contains 6 columns and 600 rows.

* **City:** includes: Atlanta, GA; Chicago, IL; Denver, CO; Houston, TX; Sacramento, CA
* **Year:** four digit year from 2008 to 2012
* **Month:** name of month
* **Degree\_day:** indicates whether the record is a ‘heating’ or ‘cooling’ day
* **Month\_Total**: Summation of monthly total
* **Mon\_dev\_from\_norm:** The difference of the period total from the normal (1971-2000) period total. Negative values indicate less degree days than normal. WEEK (or MONTH) DEV FROM NORM

**6. CALENDAR\_DAYS.sas7bdat:** includes calendar, holiday, and school day data. Contains 7 columns and 7,312 rows.

* **Date:** SAS date
* **Week ID:** ID for each week
* **Day of week:** the numeric value of each weekday (Sun=1, Mon=2…)
* **Weekdays:** three-character day of week
* **Holiday name:** if a day is a holiday, the name of the holiday is indicated
* **Holiday week:** indicates the week surrounding Christmas, New Year’s Day, and Thanksgiving Day, giving the name of the holiday that occurs during that week
* **School day:** binary column indicating whether a day is a school day

**7. SOLAR\_LUNAR.sas7bdat:** includes the daily solar, lunar, and daylight savings time data for each city from 2005 to 2012. Includes 8 columns and 14,016 rows.

* **Date:** SAS date.
* **City:** includes Atlanta, Chicago, Denver, Houston, Sacramento
* **Full\_moon:** binary column indicating whether the full moon began on that date. The full moon occurs on about a 29.5305 day cycle
* **Days\_past\_full\_moon:** zero on days when there is a full moon and increments up daily to the next full moon
* **Full\_moon\_group:** binary column indicating days which occur around a full moon. The value is one on days 26, 27, 28, 29, 0, 1, 2, 3, 4 of days past full moon. Not all cycles have the same number of days in the full moon group.
* **Standard\_sunrise:** the standard time of sunrise by city by date (not accounting for daylight savings time)
* **Standard\_sunset:** the standard time of sunset by city by date (not accounting for daylight savings time)
* **Daylight\_Savings\_Time:** binary column indicating whether Daylight Savings Time (DST) is in effect. When DST is in effect, one hour is added to the standard time. DST goes into effect according to the following rules:

**Daylight Savings Time Rules for the United States:**

* **1990 – 2006:** 1st Sunday in April until last Sun in Oct
* **2007 – Current:** 2nd Sunday in Mar until 1st Sunday in Nov

**8. STORMS.sas7bdat:** includes selected significant storm events, by city, which occurred from

2005 to 2012. Contains 8 columns and 627 rows.

* **City:** Includes Atlanta, Chicago, Denver, Houston, and Sacramento.
* **Storm\_class:** The event type, grouped. Includes: cold, drought, flood, hail, heat, heavy rain, High wind, lightning, winter storm
* **Begin\_year\_month:** the year and month of the beginning of the storm event.
* **Begin\_day:** day the storm event began.
* **Begin\_time:** time the storm event began
* **End\_year\_month:** the year and month of the end of the storm event
* **End\_day:** day the storm event ended.
* **End\_time:** time the storm event ended.

**9. WEATHER.sas7bdat:** includes the daily weather, by city for years 2005 to 2012. Contains 10 columns and 14,610 rows.

* **Date:** SAS date
* **City:** Includes Atlanta, Chicago, Denver, Houston, and Sacramento
* **Year:** four-digit year
* **Month:** numeric month value
* **Day:** day of month
* **Cloud\_cover:** median fraction of sky covered by clouds
* **Precipitation:** the liquid quantity of precipitation (in mm) over the course of the day
* **Pressure:** the mean altimeter setting (in mbar) given in the weather reports of the day.
* **Temperature\_high:** highest recorded temperature of the day, in degrees Celsius
* **Temperature\_low:** lowest recorded temperature of the day, in degrees Celsius.

## Data Quality:

The data quality for each of the provided datasets is as described below:

1. The CRIME\_V2 dataset has 130 missing values for the tract\_id attribute and 84 missing values for the CrimeArea attribute.

2. The CRIME\_V2, WEATHER datasets has multiple attributes relating to date hierarchy (attribute date already has the year, month and date of occurrence of a particular crime).

3. The UNEMPLOYMENT, HEATING\_COOLING\_DAYS datasets has two attributes year and month.

4. The date attribute is not consistent across all the available datasets.

5. The tract\_id attribute is also inconsistent across given data sets.

## Data Pre Processing Tasks

Data preprocessing is a technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing.

Data preprocessing is used database-driven applications such as customer relationship management and rule-based applications (like neural networks).

We are doing data preprocessing to do

* Data cleaning
* Data transformation
* Data separation

**1. Data Cleaning:** There were missing values in the Weather & the CRIME\_V2 dataset were imputed.

**2. Data Transformation:** When we combined the data set into the Crime data set we got some missing values. For non-holiday days we replayed the missing values with "Not Holidays".

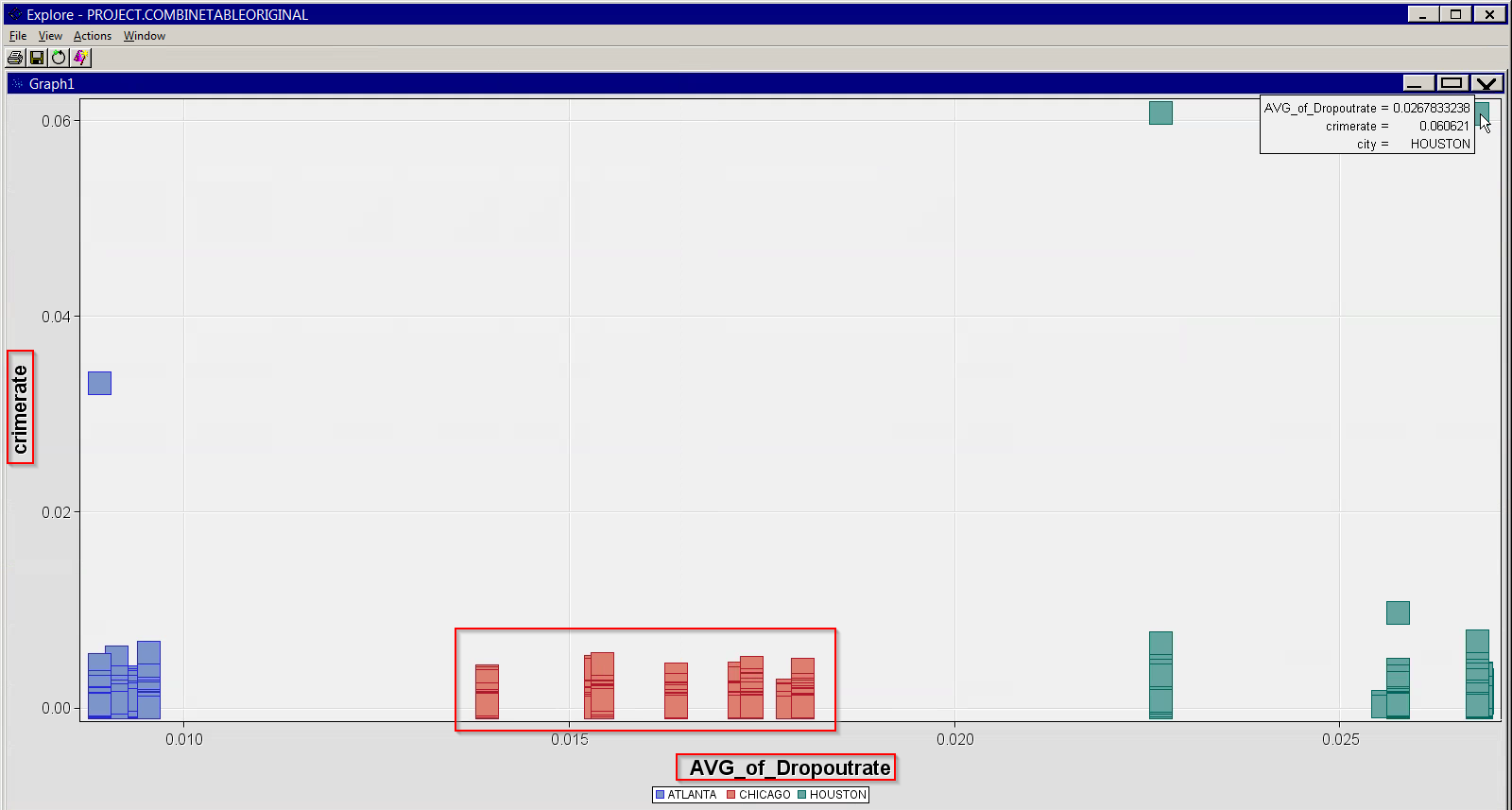
**3. Data Separation:** In storm dataset the Date was not in proper format. For e.g. In storm dataset, the date was jumbled up as "Begin\_YearMonth". As a solution we separated "Begin\_YearMonth" into "Year" & "Month".

## Data Exploration:

After successfully carrying out the data cleansing and preprocessing tasks, we finally obtained the data exploration task. The results of various explorations are described below:

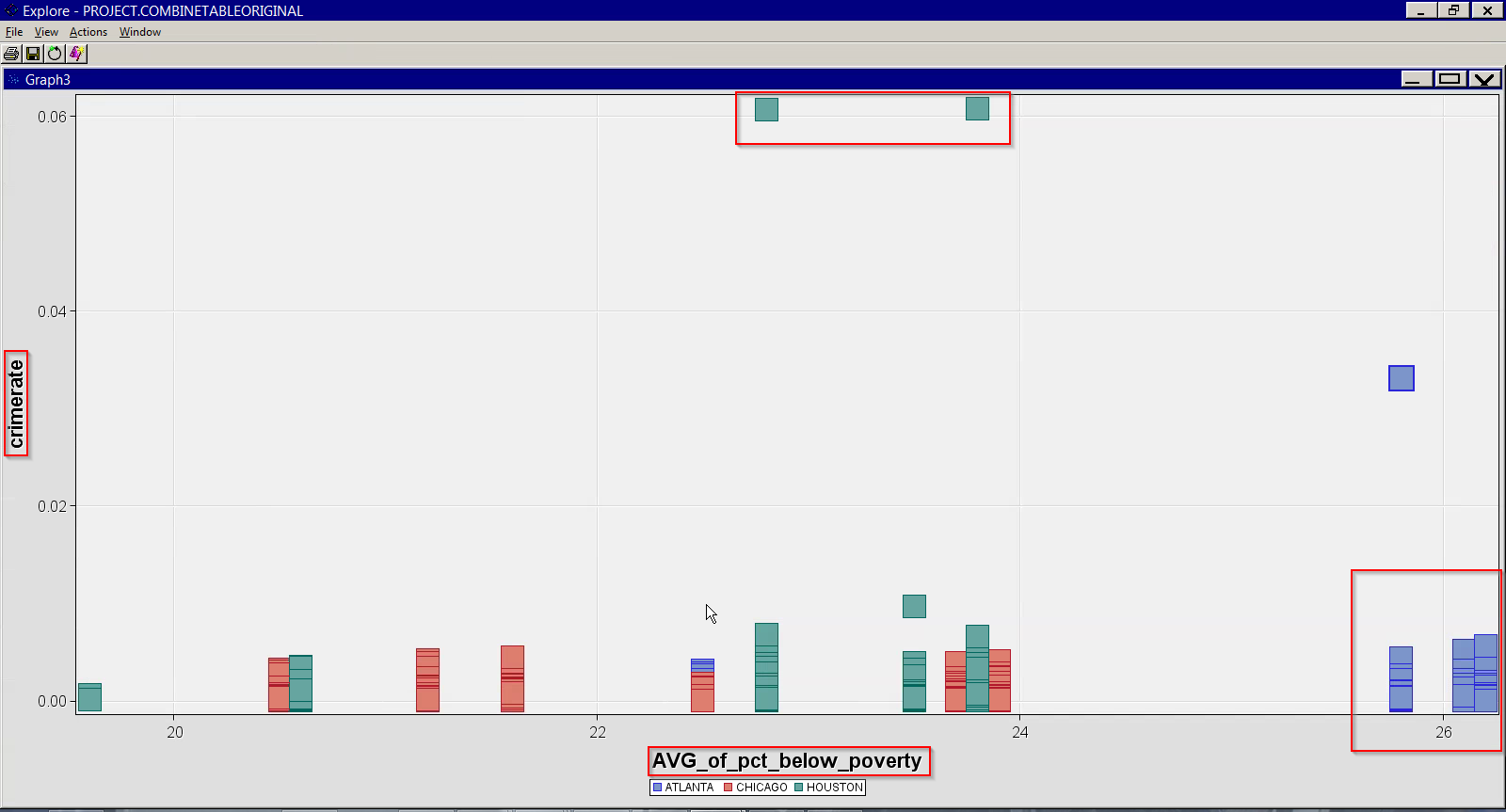
Here we studied the relationship of the Crime with each individual factors.

### Average Dropout vs Crime rate grouped by City

****

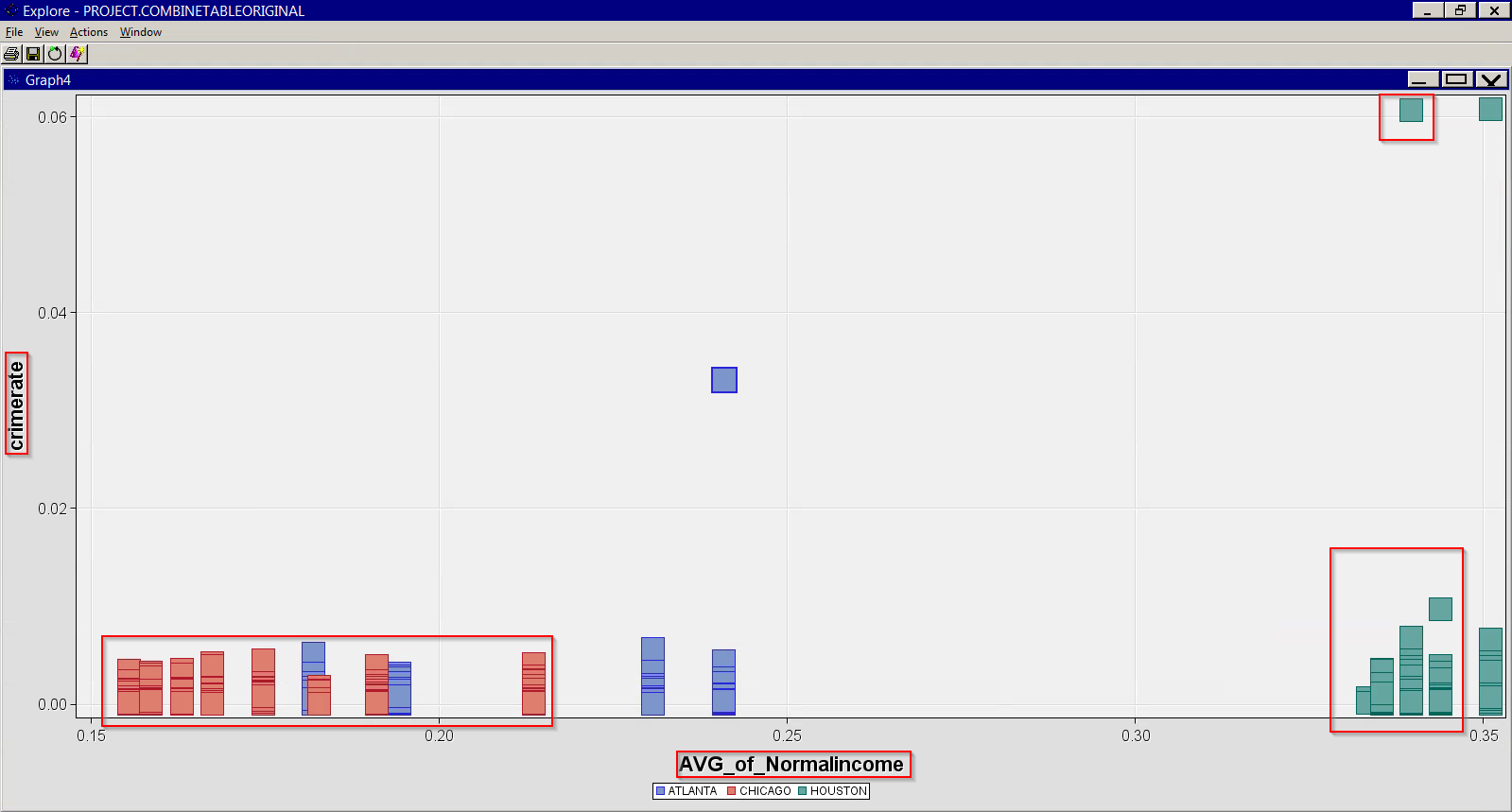
Houston has the highest dropout rate and is also highly incident of crimes as compared to Atlanta and Chicago.

### Average of % below poverty vs Crimerate grouping City

****

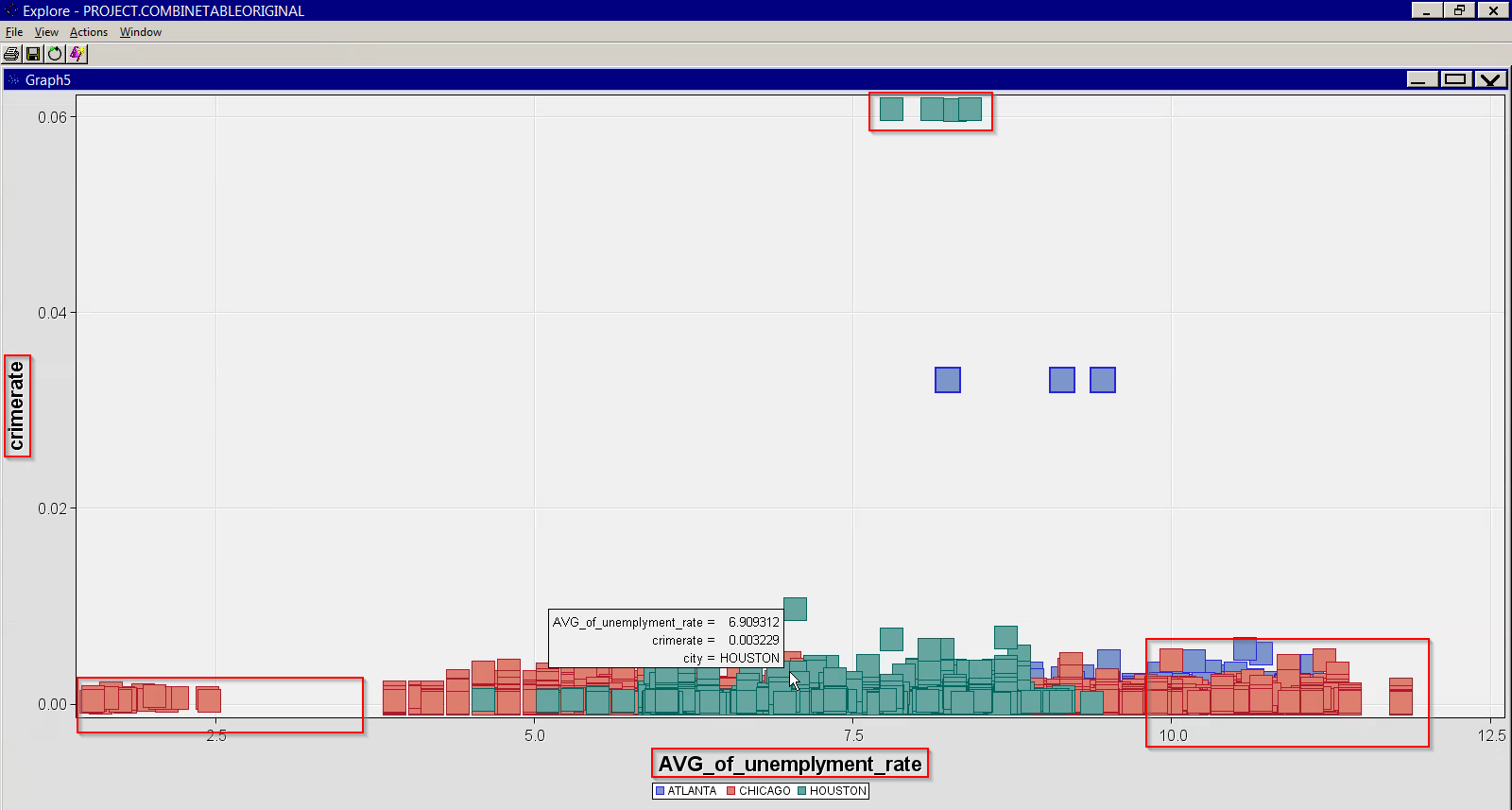
From the above graphs we can see that the crimerate in Chicago is low even though they have a good amount of dropout and percent below poverty rate. We can say that the police are most effective in Chicago compared to Atlanta and Houston.

### Normalincome Vs Crime rate grouping City

****

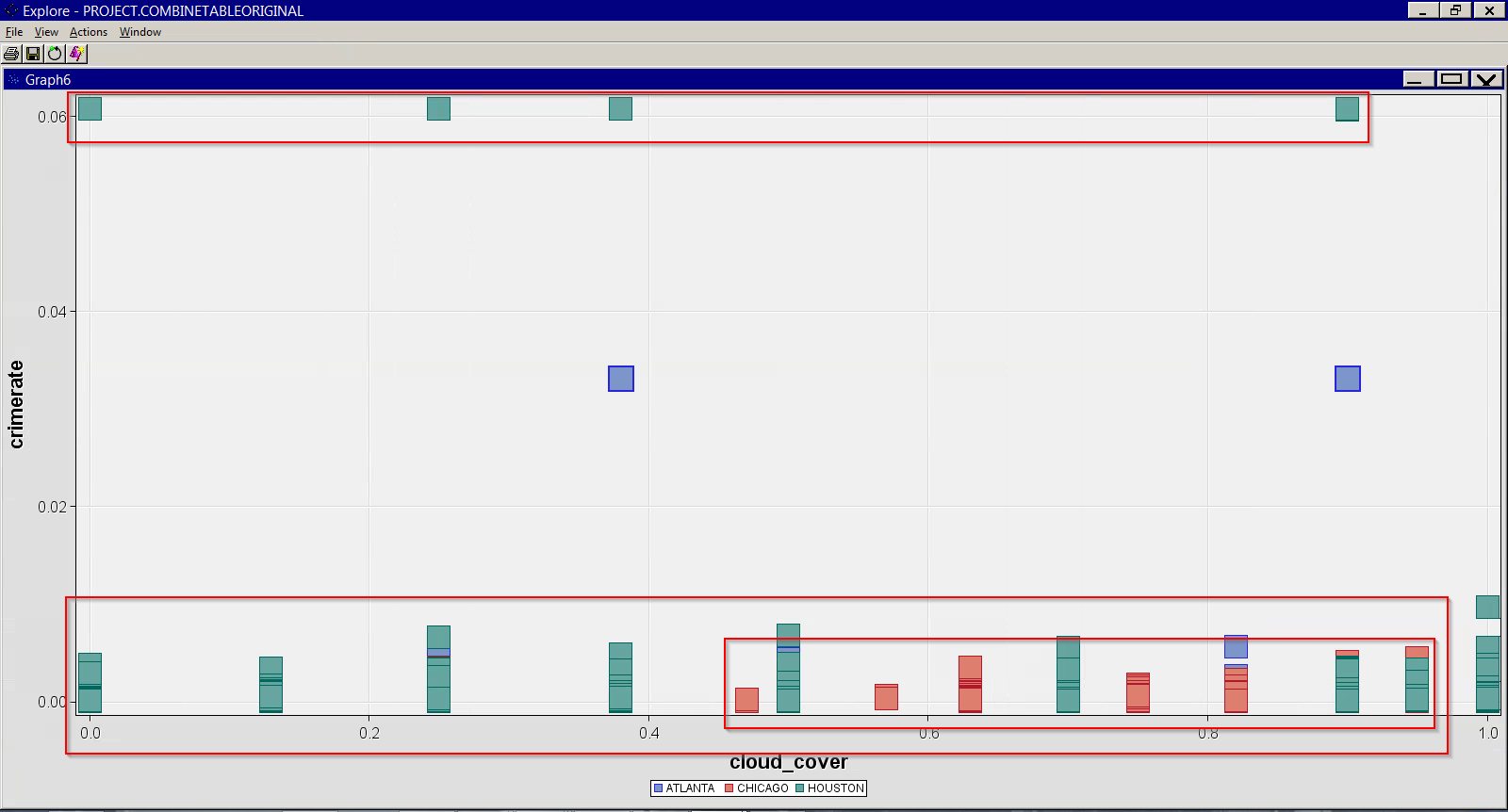
As we can see here the normal income is high in Houston and yet the crime rate is high. At the same time we can see that Chicago has low normal income and yet the crime rate is low.

### Unemploymennt rate vs Crime rate grouping City

****

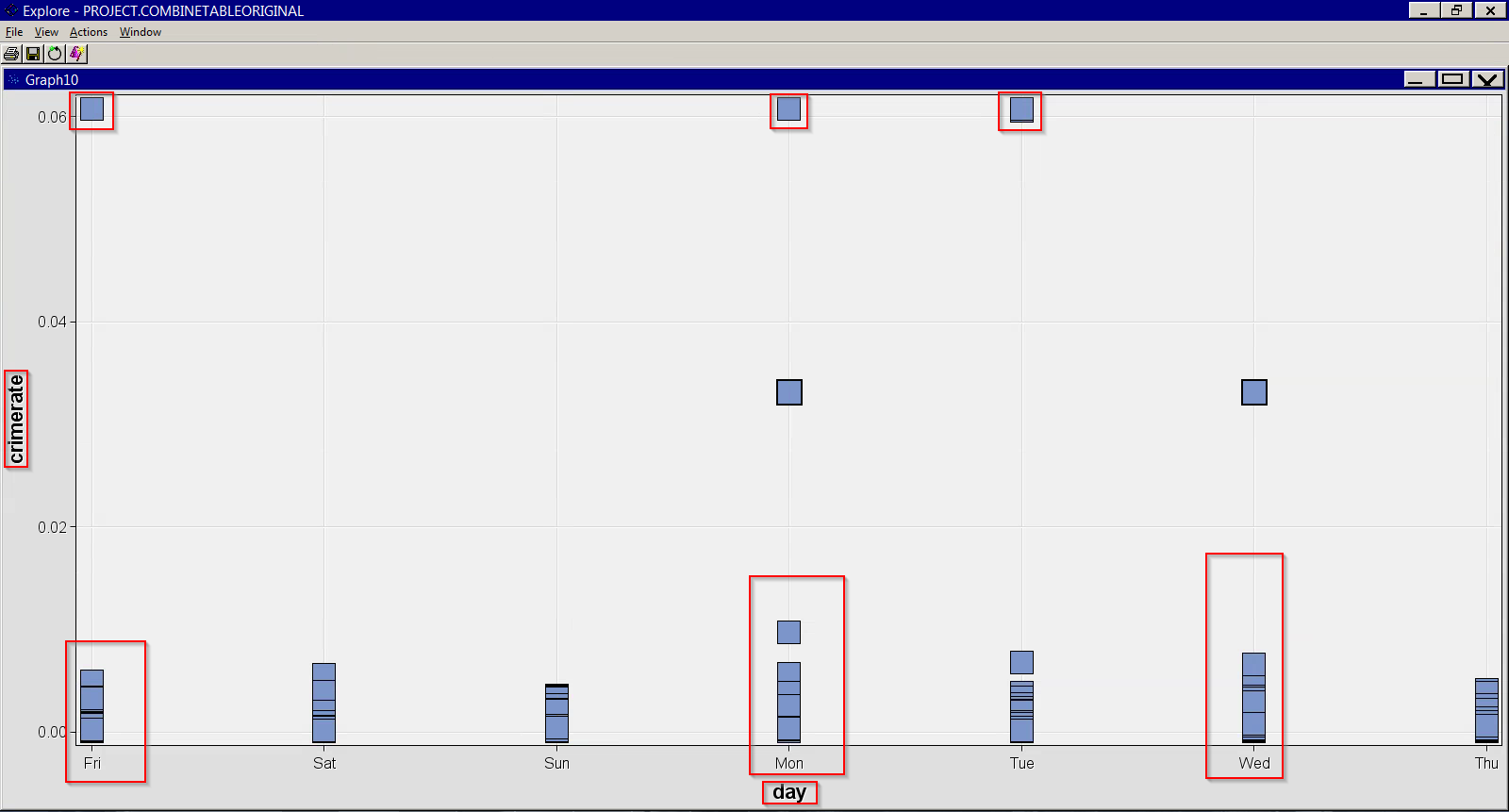
Here we can see that Chicago has unemployment rate ranging from lowest and highest and yet the crime rate is low. But we can see that though Houston has less unemployment rate compared to Chicago, crime rate is still high in Houston.

### Cloud cover vs Crime rate grouping City

****

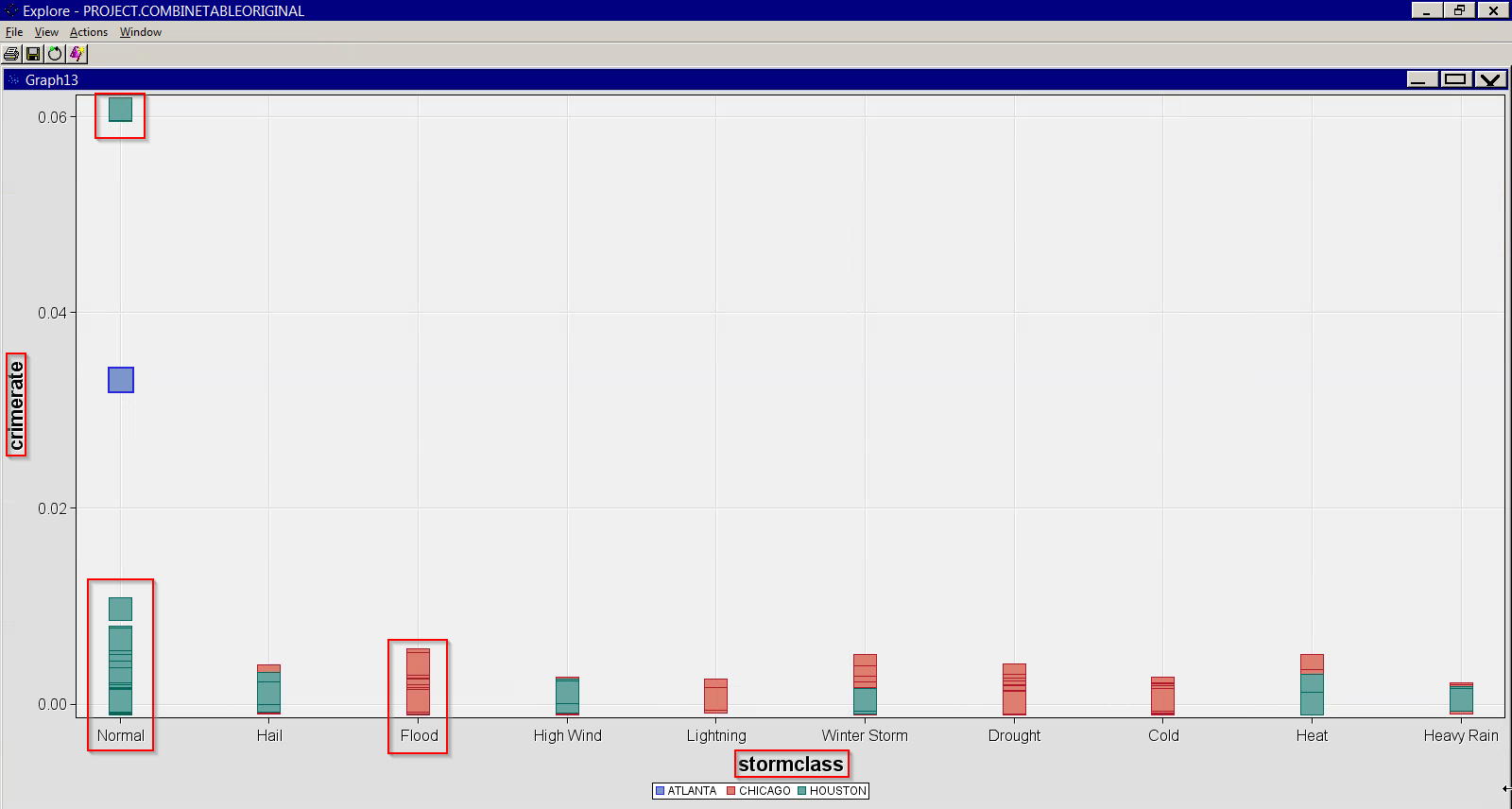
Houston has cloud cover ranging from lowest to highest and they have the highest crime rate. But we can also see that Chicago has decent cloud cover and yet the crime rate is less.

### Crime rate vs Day:

****

The crime rate is more on Friday, Monday and Wednesday in all the cities. These days we need more deployment of police. We can also group specific to crime area which needs more attention.

### Crime rate vs Stormclass grouping by City

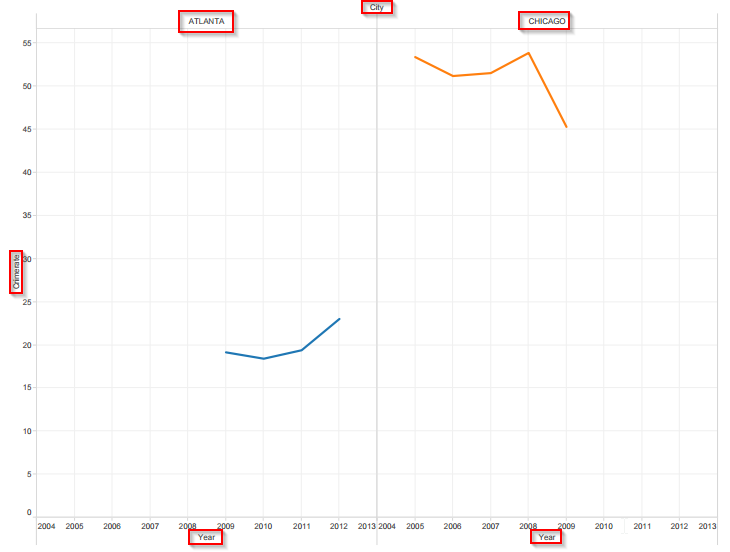
****

Here we can see that crime rate is more in Houston during normal days rather than any other storm class. We can also see that during flood storm class, crime rate is high in Chicago.

## Data Modeling Outputs

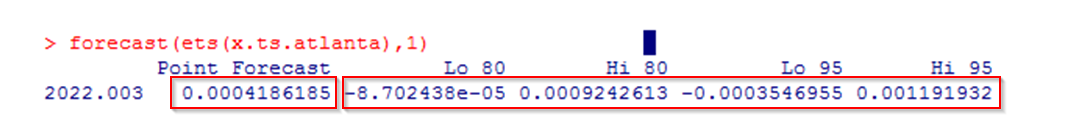
Through this project we are trying to answer the following questions of value for business.

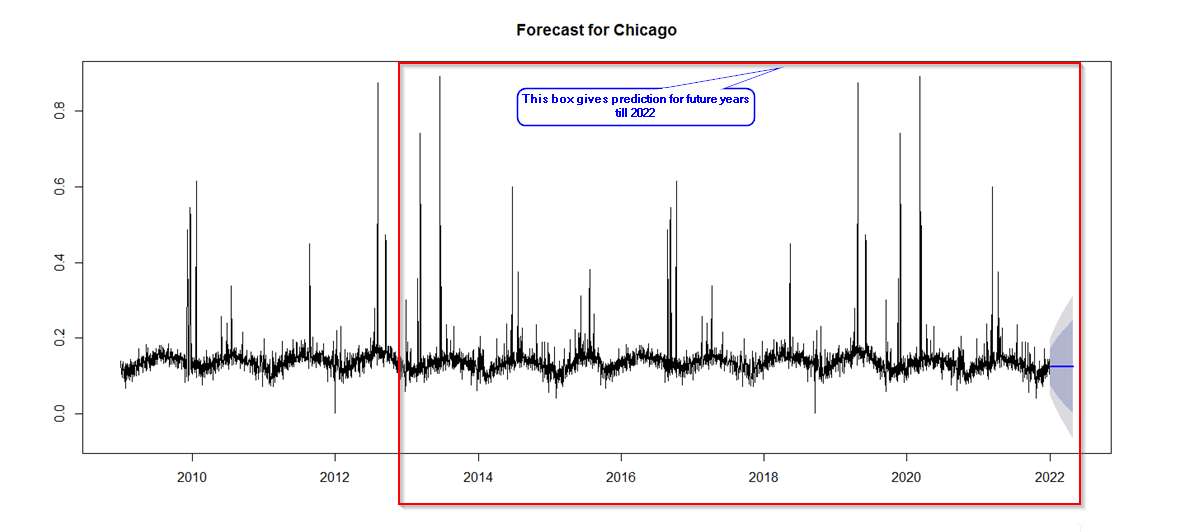
**1) Predicting crime-rate for cities for the year 2022.**

For this problem, we used the given dataset to form the trend for existing data. This trend shows how crime-rate varies each year for different cities for given data. Below is the screenshot for the same using Tableau Dashboard.

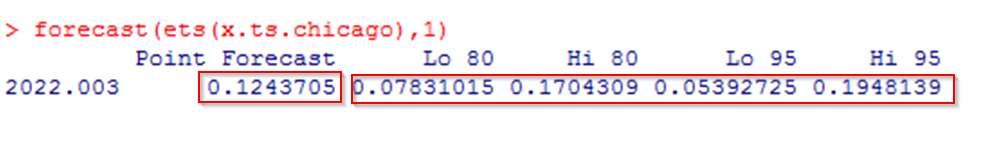
Now we try to predict the crime-rate for these cities for the year 2022 using R Studio.

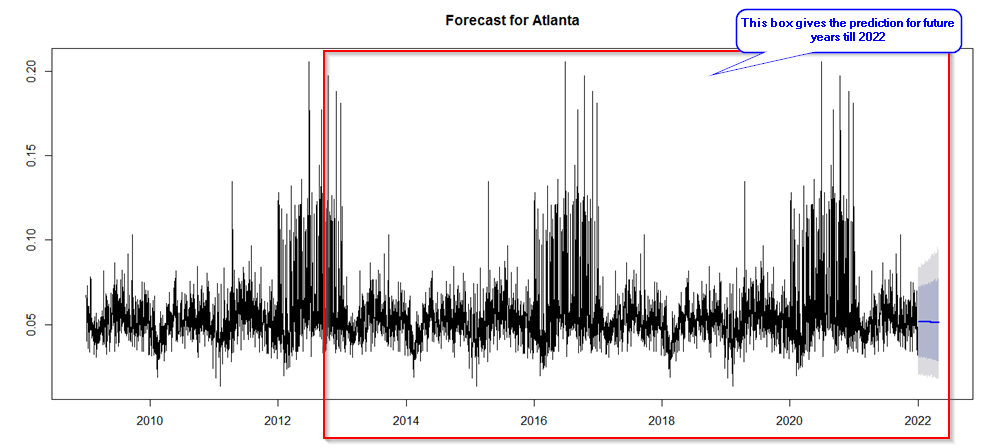
i) **For Chicago,** we had data from 2009 to 2012. Since we wanted to estimate the crime-rate for the year 2022, which is 10 years in future, we had to assume the replication of same data. We **observed 80% prediction limits and 95% prediction limits for the predicted crime-rate which lies around 0.12**

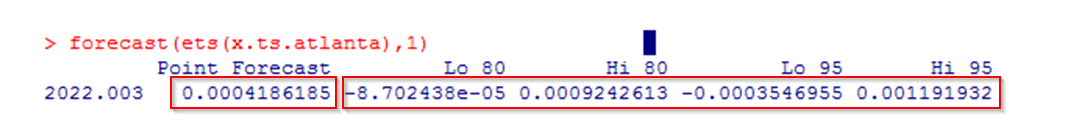




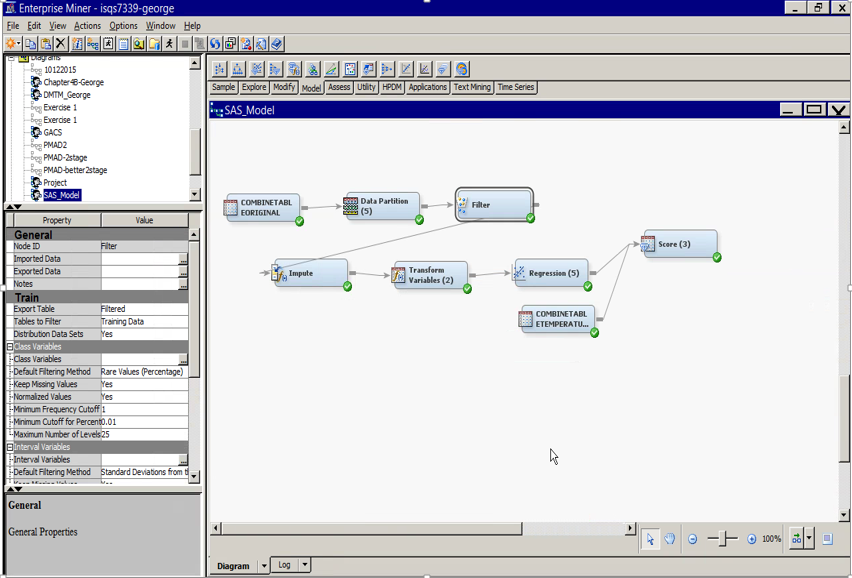
Using forecast function in R we **predicted the crime-rate for the year 2022 with 80% prediction interval and 95% prediction interval**. Below is the screenshot of the same.



ii) **For Atlanta,** we had data from 2009 to 2012. Since we wanted to estimate the crime-rate for the year 2022, which is 10 years in future, we had to assume the replication of same data**. We observed 80% prediction limits and 95% prediction limits for the predicted crime-rate which lies around 0.004**. Below is the screenshot for the same.

Using forecast function in R we predicted the crime-rate for the **year 2022 with 80% prediction interval and 95% prediction interval.** Below is the screenshot for the same.

**2) Determine the significant inputs that help you to predict crime-rate**



We used the following nodes for processing the data in the process flow.

**A. Filter Node:**

We used the filter node to restrict the outliers. For e.g. In our case in the Crime dataset there were missing crime areas and trackID were missing, etc.

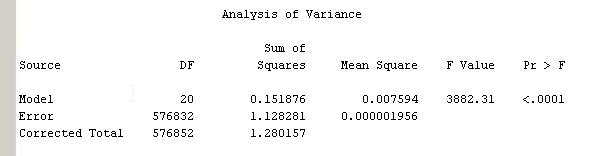
**B. Impute Node:**

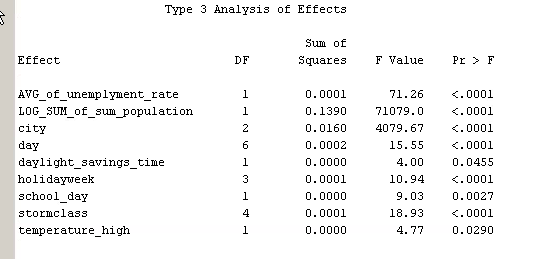
We used impute node to replace the missing values. For e.g. in our case when we combined the Crime dataset and the Calendar Dataset, the combined dataset contained many missing values. So we replaced these missing values with "Not Holiday". Additionally, in the weather dataset, the pressure data was missing.

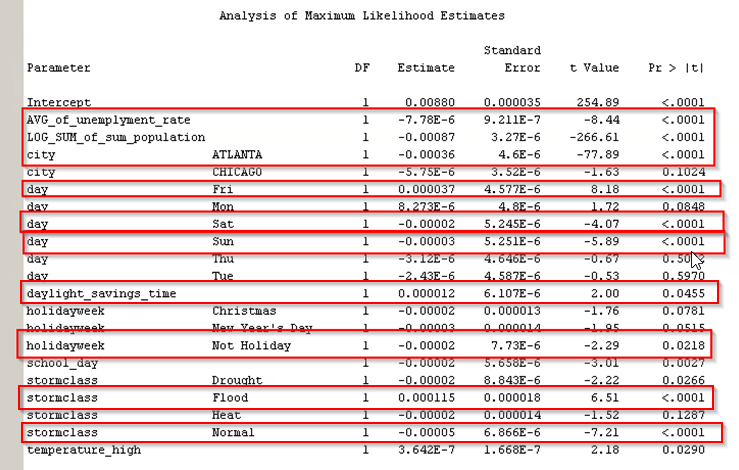
**C. Transform Node:**

We used transform node to transform the original data into a new one. For e.g. In our case the range of variables like population & precipitation is too large. To avoid incorrect interpretation of the outcomes, we applied this node.

**As we can see from the below Analysis of Variance, we obtain a significant value for the above regression model.**



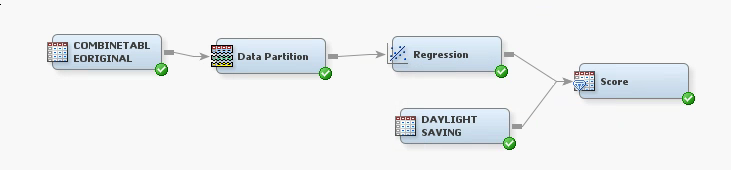




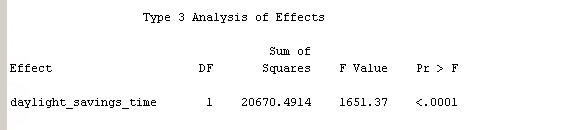
When we do a regression by considering all possible and relevant inputs, we find that **AVG\_of\_unemployment\_rate, LOG\_SUM\_of\_sum\_population, city –Atlanta, city-Chicago, day- Fri, Sat, Sun, daylight\_savings\_time, holidayweek-Not holiday, Stormclass- Flood and Normal are significant since they have a p-value less than 0.05**

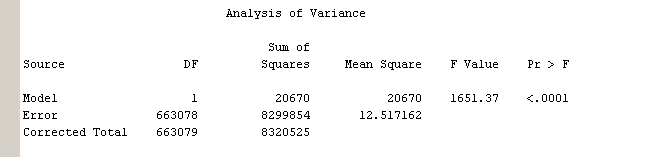
**3) What is the effect of Daylight Savings on crime-rate? And what is the effect on crime-rate when no daylight saving is applied?**

**Regression Model for Daylight saving**

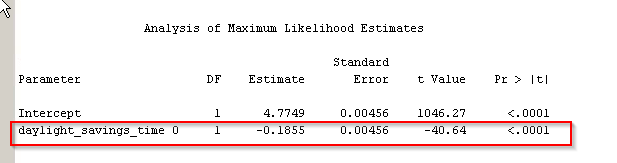


**The Type 3 Analysis & Analysis of Variance shows the significance of using daylight\_savings\_time for the above regression model.**



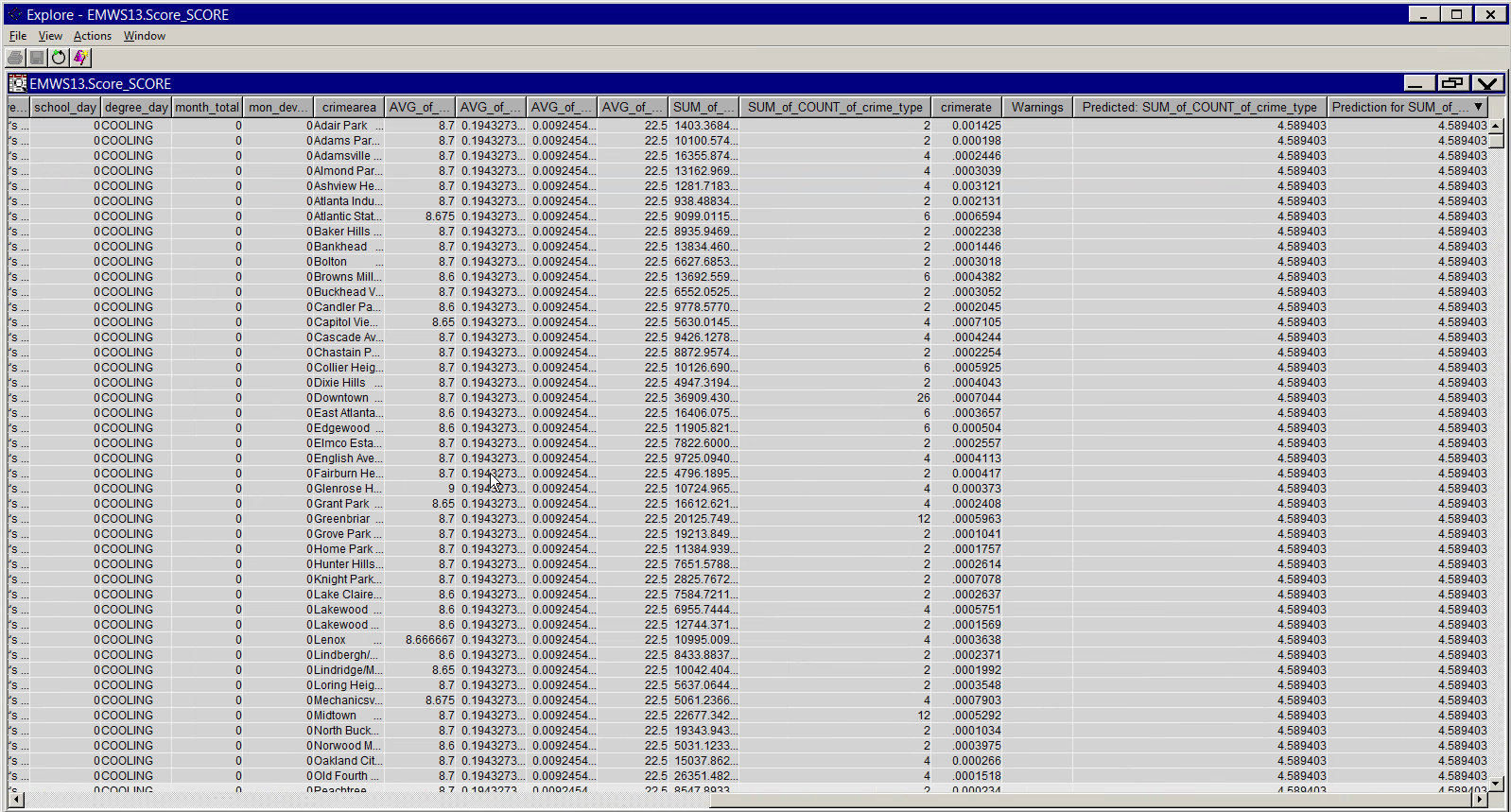


The below screenshot shows that **as we move daylight\_saving\_time from 1 to 0, we see that crimerate target decreases which is indicated by the negative sign of the parameter estimate**.



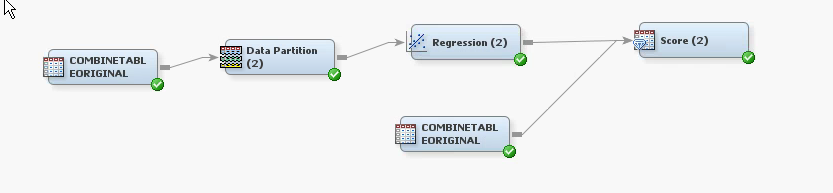
Consequently, **when the daylight savings are implemented (Daylight savings = 1), the crime rate increases.**

Below the screenshot for the scored dataset using Regression Model for Daylight Saving

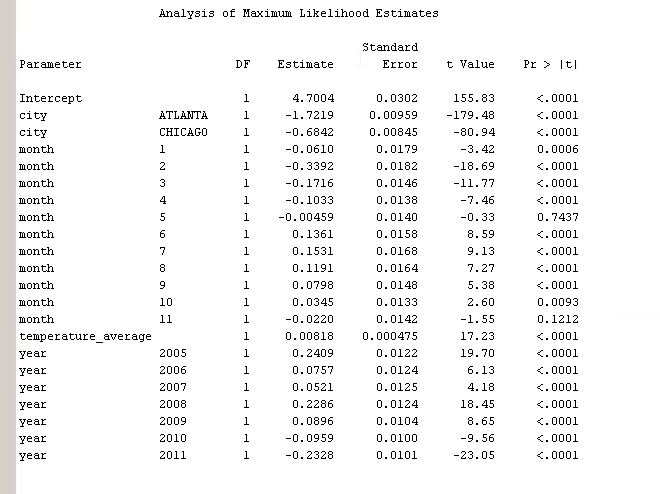


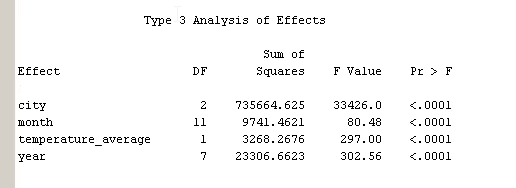
**4) What are the effects of average temperature with respect to a particular city and a particular timing in the year for crime-rate?**

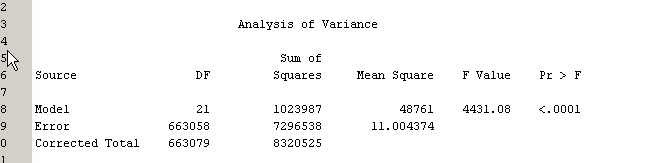
**Regression model using inputs City, temperature average, Month and year for the target Crime Rate.**



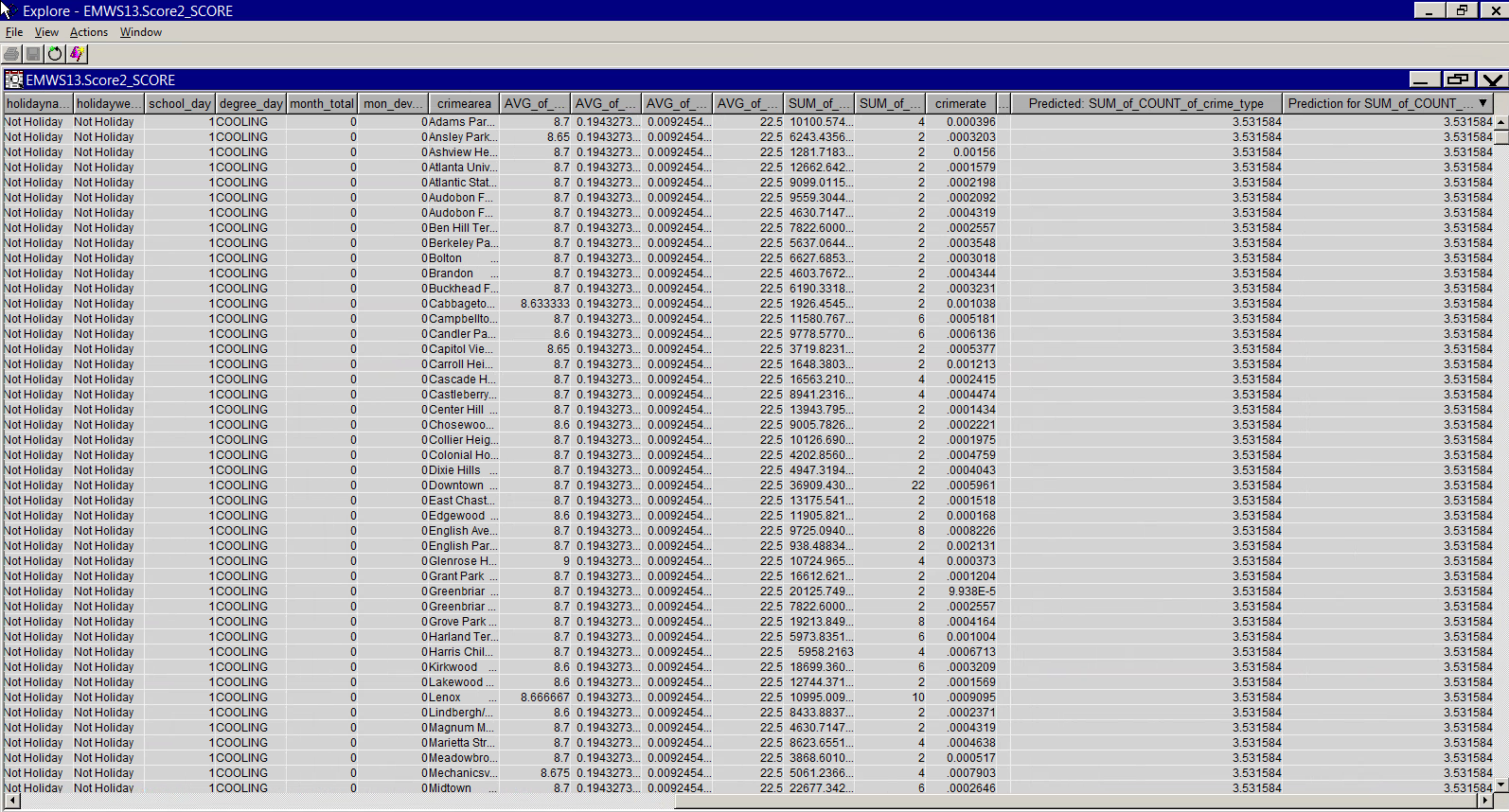
We see **significant relationship between month, temperature\_average, city and year for the target Crime Area as indicated by the Analysis of Variance, Type 3 Analysis and Maximum Likelihood estimates**.







Subsequently, we obtained a scored data for the above Regression Model giving predicted values for target crime rate.



**Conclusions-**

1. **For Chicago, in the year 2022, using 80% and 95% prediction limits, we predict that there is 80% and 95% chance respectively that the crime rate will be around 0.1243**
2. **For Atlanta, in the year 2022, using 80% and 95% prediction limits, we predict that there is 80% and 95% chance respectively that the crime rate will be around 0.0004**
3. **We have a significant model for regressing crime rate based upon Average of Unemployment Rate, Sum of the population, city, day, daylight savings time, holiday week and storm class.**
4. **We see that the crime rate increase as daylight time savings are implemented.**
5. **Crime rate significantly decreases during Spring time of the year depending upon the city and the temperature.**
6. **This regression models obtained can be used in scoring new crime data for the target crime rate.**