**Forecast the crime activity from Chicago crime database**

**Part 1**

**Data Source:** City of Chicago, Data Portal

**Outcome:** From the raw data file, created a data set that contains the year-month, Primary Type, and # rows, but only for the primary types of Theft, Battery, and Narcotics. Started this data set in 2004 and end it in 2013.

**R code:**

#Read the file "Chicago crimes.dat" into Rstudio and name it to "mycrime"

mycrime = read.delim(file.choose(), sep=",", stringsAsFactors=FALSE, na.strings="")

#Change the data type of column "Date" in "mycrime" data set from "character" to "date"

mycrime$Date <- as.Date(mycrime$Date , format="%m/%d/%Y")

#Change the format of "Date" column in "mycrime" table from "MM/DD/YYYY" to "YYYY-MM"

mycrime$Date <- strftime(mycrime$Date, format="%Y-%m")

#Filter the "Primary type" column as to get data related to only "THEFT", "BATTERY" and "NARCOTICS"

#thereby creating a new table "mycrimept" with that corresponding data

mycrimept <- mycrime[(mycrime$Primary.Type == "THEFT" |

mycrime$Primary.Type == "BATTERY" |

mycrime$Primary.Type == "NARCOTICS"),]

#Filter the table "mycrimept" as to get data only from 2004 to 2013

mycrimept <- mycrimept[mycrimept$Year >= 2004 & mycrimept$Year <= 2013,]

#select Date, Primarytype, count(Primarytype) from mycrimept Group by Date, Primarytype

#Place the data obtained after getting count of rows in a new table "mycrimefinal"

mycrimefinal <- aggregate(mycrimept,

by=list(mycrimept$Date, mycrimept$Primary.Type),

FUN=length);

#Select only the first three required columns, "Year month", "Primary type" and "No of events"

#from the table "mycrimefinal"

mycrimefinal <- mycrimefinal[,c(1,2,3)]

#Rename the column names from previous names to "YearMonth","PrimaryType" and "NoOfEvents"

names(mycrimefinal) <- c("YearMonth","PrimaryType","NoOfEvents")

#Order the data set to start from 2004 and end with 2013

mycrimefinal <- mycrimefinal[order(mycrimefinal$Year),]

#Export the table "mycrimefinal" to "mycrimefinal.xlsx" in Downloads folder from PC

library(xlsx)

write.xlsx(mycrimefinal, "/Users/mothiki/Downloads/mydata.xlsx")

**Part 2**

Implemented Regression in excel to predict the number of events during the first 6 months of 2014 for each of the three primary crime types.

I have divided the 360 rows of the table “mycrimefinal” to three independent tables “mycrimetheft”, “mycrimebattery” and “mycrimenarcotics”, each of 120 rows each filtering them with “primary type”.

R code to generate original data of “mycrimetheft”, “mycrimebattery” and “mycrimenarcotics” from Jan 14 to June 14 and export them to "mycrimefinal\_2.xlsx" in Downloads folder from PC is:

**R code:**

#Read the file "Chicago crimes.dat" into Rstudio and name it to "mycrime"

mycrime = read.delim(file.choose(), sep=",", stringsAsFactors=FALSE, na.strings="")

#Change the data type of column "Date" in "mycrime" data set from "character" to "date"

mycrime$Date <- as.Date(mycrime$Date , format="%m/%d/%Y")

#Change the format of "Date" column in "mycrime" table from "MM/DD/YYYY" to "YYYY-MM"

mycrime$Date <- strftime(mycrime$Date, format="%Y-%m")

#Filter the "Primary type" column as to get data related to only "THEFT", "BATTERY" and "NARCOTICS"

#thereby creating a new table "mycrimept" with that corresponding data

mycrimept <- mycrime[(mycrime$Primary.Type == "THEFT" |

mycrime$Primary.Type == "BATTERY" |

mycrime$Primary.Type == "NARCOTICS"),]

#Filter the table "mycrimept" as to get data only from 2014 Jan to 2014 June

mycrimept <- mycrimept[(mycrimept$Date == "2014-01"|

mycrimept$Date == "2014-02"|

mycrimept$Date == "2014-03"|

mycrimept$Date == "2014-04"|

mycrimept$Date == "2014-05"|

mycrimept$Date == "2014-06"),]

#select Date, Primarytype, count(Primarytype) from mycrimept Group by Date, Primarytype

#Place the data obtained after getting count of rows in a new table "mycrimefinal"

mycrimefinal <- aggregate(mycrimept,

by=list(mycrimept$Date, mycrimept$Primary.Type),

FUN=length);

#Select only the first three required columns, "Year month", "Primary type" and "No of events"

#from the table "mycrimefinal"

mycrimefinal <- mycrimefinal[,c(1,2,3)]

#Rename the column names from previous names to "YearMonth","PrimaryType" and "NoOfEvents"

names(mycrimefinal) <- c("YearMonth","PrimaryType","NoOfEvents")

#Order the data set to start from 2014 Jan and end with 2014 June

mycrimefinal <- mycrimefinal[order(mycrimefinal$Year),]

#Export the table "mycrimefinal" to "mycrimefinal\_2.xlsx" in Downloads folder from PC

library(xlsx)

write.xlsx(mycrimefinal, "/Users/mothiki/Downloads/mycrimefinal\_2.xlsx")

* Please find below, the excel sheet attached containing the deseasonalized forecast prediction of data “mycrimetheft”, “mycrimebattery” and “mycrimenarcotics” using regression.
* To the right of each table, we can see Regression graphs, Regression summary output and seasonality calculation table for smoothing.
* To the bottom of the table we can see the required forecast parameter.



Procedure for regression calculation:

* The moving average and centered moving averages(CMA) were calculated. With the help of CMA values, seasonality was calculated using the below formula
* Seasonality\*Irregularity= Number of Events/CMA
* This seasonality\*Irregularity parameter was used to calculate seasonality using a separate table. With the help of seasonality, deseasonalized parameters are obtained by dividing the No of events parameter with seasonality.
* Regression is made using deseasonalized number of events as y and time as x.
* Trend is calculated using the obtained slope and intercept.
* Forecast component is obtained by multiplying trend component with seasonality component.
* All the obtained parameters were available in the excel.

Below are the comparisons with Predicted values and Original data for Regression:

**BATTERY:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Year Month*** | ***Primary Type*** | ***Original No Of Events*** | ***Predicted Events*** | ***Difference (Predicted-Original)*** |
| 2014-01 | BATTERY | 3329 | 3696.357937 | 367.357937 |
| 2014-02 | BATTERY | 3173 | 3382.840733 | 209.840733 |
| 2014-03 | BATTERY | 3973 | 4367.526877 | 394.5268771 |
| 2014-04 | BATTERY | 4184 | 4377.480325 | 193.480325 |
| 2014-05 | BATTERY | 4801 | 4968.341486 | 167.3414859 |
| 2014-06 | BATTERY | 4816 | 4924.419999 | 108.4199992 |

**NARCOTICS:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Year Month*** | ***Primary Type*** | ***Original No Of Events*** | ***Predicted Events*** | ***Difference (Predicted-Original)*** |
| 2014-01 | NARCOTICS | 2223 | 2733.301907 | 510.3019074 |
| 2014-02 | NARCOTICS | 2354 | 2616.935345 | 262.9353452 |
| 2014-03 | NARCOTICS | 2634 | 2880.79789 | 246.7978896 |
| 2014-04 | NARCOTICS | 2597 | 2609.949272 | 12.94927226 |
| 2014-05 | NARCOTICS | 2606 | 2618.707262 | 12.70726239 |
| 2014-06 | NARCOTICS | 2424 | 2476.613041 | 52.61304084 |

**THEFT:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Year Month*** | ***Primary Type*** | ***Original No Of Events*** | ***Predicted Events*** | ***Difference (Predicted-Original)*** |
| 2014-01 | THEFT | 4412 | 5319.04387 | 907.0438703 |
| 2014-02 | THEFT | 4012 | 4374.304298 | 362.3042981 |
| 2014-03 | THEFT | 4498 | 5372.908773 | 874.9087735 |
| 2014-04 | THEFT | 4815 | 5551.754101 | 736.7541006 |
| 2014-05 | THEFT | 5329 | 5997.069694 | 668.0696936 |
| 2014-06 | THEFT | 5807 | 6188.315913 | 381.3159131 |

* The predicted values for “theft” are more diverted from the original values. For all the Primary types, predicted values are coming closer to original values with increase in time from Jan to June.

**Part 3**

Implemented HoltWinters prediction with monthly seasonality to predict the first 6 months of 2014 for each of the three primary crime types.

The Holt winters prediction is made using R studio. Please find below the code used. We have considered system provided values alpha= 0.212, beta= 0.005 and gamma= 0.534 as we have found the forecast to be very effective.

* We tried with other values of alpha, beta and gamma between 0.1 to 0.4 to avoid overfitting but, the system provided values look closer to the original values compared to other sets.

**R code:**

#Read the file "Chicago crimes.dat" into Rstudio and name it to "mycrime"

mycrime = read.delim(file.choose(), sep=",", stringsAsFactors=FALSE, na.strings="")

#Change the data type of column "Date" in "mycrime" data set from "character" to "date"

mycrime$Date <- as.Date(mycrime$Date , format="%m/%d/%Y")

#Change the format of "Date" column in "mycrim" table from "MM/DD/YYYY" to "YYYY-MM"

mycrime$Date <- strftime(mycrime$Date, format="%Y-%m")

#Filter the "Primary type" column as to get data related to only "THEFT", "BATTERY" and "NARCOTICS"

#thereby creating a new table "mycrimept" with that corresponding data

mycrimept <- mycrime[(mycrime$Primary.Type == "THEFT" |

mycrime$Primary.Type == "BATTERY" |

mycrime$Primary.Type == "NARCOTICS"),]

#Filter the table "mycrimept" as to get data only from 2004 to 2013

mycrimept <- mycrimept[mycrimept$Year >= 2004 & mycrimept$Year <= 2013,]

#select Date, Primarytype, count(Primarytype) from mycrimept Group by Date, Primarytype

#Place the data obtained after getting count of rows in a new table "mycrimefinal"

mycrimefinal <- aggregate(mycrimept,

by=list(mycrimept$Date, mycrimept$Primary.Type),

FUN=length);

#Select only the first three required columns, "Year month", "Primary type" and "No of events"

#from the table "mycrimefinal"

mycrimefinal <- mycrimefinal[,c(1,2,3)]

#Rename the column names from previous names to "YearMonth","PrimaryType" and "NoOfEvents"

names(mycrimefinal) <- c("YearMonth","PrimaryType","NoOfEvents")

#Order the data set to start from 2004 and end with 2013

mycrimefinal <- mycrimefinal[order(mycrimefinal$Year),]

#Divide the data set "mycrimefinal" to independent datasets, "mycrimetheft", "mycrimebattery" and "mycrimenarcotics"

mycrimetheft <- mycrimefinal[mycrimefinal$PrimaryType == "THEFT", ]

mycrimebattery <- mycrimefinal[mycrimefinal$PrimaryType == "BATTERY", ]

mycrimenarcotics <- mycrimefinal[mycrimefinal$PrimaryType == "NARCOTICS", ]

**Prediction for THEFT:**

The below code is used to predict and plot the THEFT data set thereby exporting the forecasted dataset to “mycrimefinal\_t.xlsx”, which is used for comparison.

#Create a new vector for "mycrimetheft" table, using the time series, to be used as x-axis

mycrimetheft\_ts <- ts(mycrimetheft$NoOfEvents, start=2004, frequency=12)

#Apply Holt winter prediction

hw <- HoltWinters(mycrimetheft\_ts)

#Forecast 2014 Jan to 2014 June using "predict"

forecast <- predict(hw, n.ahead = 6)

plot(hw, forecast)

#Export the table "forecast" to "mycrimefinal\_t.xlsx" in Downloads folder from PC

library(xlsx)

write.xlsx(forecast\_t, "/Users/mothiki/Downloads/mycrimefinal\_t.xlsx")

Below graph is the obtained plot for Holt winter prediction.



Below is the comparison table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Year Month** | **Primary Type** | **Original  No Of Events** | **Predicted  Events** | **Difference (Predicted - Original)** |
| 2014-01 | THEFT | 4412 | 5127.401753 | 715.4017527 |
| 2014-02 | THEFT | 4012 | 4199.112201 | 187.1122015 |
| 2014-03 | THEFT | 4498 | 5136.569762 | 638.5697623 |
| 2014-04 | THEFT | 4815 | 5392.606918 | 577.606918 |
| 2014-05 | THEFT | 5329 | 5926.493482 | 597.4934819 |
| 2014-06 | THEFT | 5807 | 6187.941185 | 380.9411852 |

**Prediction for NARCOTICS:**

The below code is used to predict and plot the NARCOTICS data set thereby exporting the forecasted dataset to “mycrimefinal\_n.xlsx”, which is used for comparison.

#Create a new vector for "mycrimenarcotics" table, using the time series, to be used as x-axis

mycrimenarcotics\_ts <- ts(mycrimenarcotics$NoOfEvents, start=2004, frequency=12)

#Apply Holt winter prediction

hw <- HoltWinters(mycrimenarcotics\_ts)

#Forecast 2014 Jan to 2014 June using "predict"

forecast\_n <- predict(hw, n.ahead = 6)

plot(hw, forecast\_n)

#Export the table "forecast" to "mycrimefinal\_n.xlsx" in Downloads folder from PC

library(xlsx)

write.xlsx(forecast\_n, "/Users/mothiki/Downloads/mycrimefinal\_n.xlsx")

Below graph is the obtained plot for Holt winter prediction.



Below is the comparison table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Year Month** | **Primary Type** | **Original  No Of Events** | **Predicted  Events** | **Difference (Predicted - Original)** |
| 2014-01 | NARCOTICS | 2223 | 2780.041849 | 557.0418487 |
| 2014-02 | NARCOTICS | 2354 | 2658.273724 | 304.2737243 |
| 2014-03 | NARCOTICS | 2634 | 2995.049695 | 361.0496954 |
| 2014-04 | NARCOTICS | 2597 | 2581.728116 | -15.27188387 |
| 2014-05 | NARCOTICS | 2606 | 2635.5293 | 29.5292997 |
| 2014-06 | NARCOTICS | 2424 | 2462.419065 | 38.41906525 |

**Prediction for BATTERY:**

The below code is used to predict and plot the BATTERY data set thereby exporting the forecasted dataset to “mycrimefinal\_b.xlsx”, which is used for comparison.

#Create a new vector for "mycrimebattery" table, using the time series, to be used as x-axis

mycrimebattery\_ts <- ts(mycrimebattery$NoOfEvents, start=2004, frequency=12)

#Apply Holt winter prediction

hw <- HoltWinters(mycrimebattery\_ts)

#Forecast 2014 Jan to 2014 June using "predict"

forecast\_b <- predict(hw, n.ahead = 6)

plot(hw, forecast\_b)

#Export the table "forecast" to "mycrimefinal\_b.xlsx" in Downloads folder from PC

library(xlsx)

write.xlsx(forecast\_b, "/Users/mothiki/Downloads/mycrimefinal\_b.xlsx")

Below graph is the obtained plot for Holt winter prediction.



Below is the comparison table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Year Month** | **Primary Type** | **Original  No Of Events** | **Predicted  Events** | **Difference (Predicted - Original)** |
| 2014-01 | BATTERY | 3329 | 3548.472773 | 219.4727728 |
| 2014-02 | BATTERY | 3173 | 3114.752365 | -58.24763452 |
| 2014-03 | BATTERY | 3973 | 4281.972441 | 308.9724409 |
| 2014-04 | BATTERY | 4184 | 4271.176653 | 87.17665268 |
| 2014-05 | BATTERY | 4801 | 5055.057218 | 254.0572182 |
| 2014-06 | BATTERY | 4816 | 5174.119727 | 358.1197274 |

* Even for the holt winter prediction, I observed that the predicted values for “theft” are more diverted from the original values. For all the Primary types, predicted values are coming closer to original values with increase in time from Jan to June.

**Parts 4**

Compared both the Holtwinters and Regression forecasts for all the three Primary types from Jan 2014 to June 2014.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Year Month** | **Primary Type** | **Original  No Of Events** | **Holt Winter** | **Regression** |
| 2014-01 | BATTERY | 3329 | 3548.472773 | 3696.357937 |
| 2014-02 | BATTERY | 3173 | 3114.752365 | 3382.840733 |
| 2014-03 | BATTERY | 3973 | 4281.972441 | 4367.526877 |
| 2014-04 | BATTERY | 4184 | 4271.176653 | 4377.480325 |
| 2014-05 | BATTERY | 4801 | 5055.057218 | 4968.341486 |
| 2014-06 | BATTERY | 4816 | 5174.119727 | 4924.419999 |
| 2014-01 | NARCOTICS | 2223 | 2780.041849 | 2733.301907 |
| 2014-02 | NARCOTICS | 2354 | 2658.273724 | 2616.935345 |
| 2014-03 | NARCOTICS | 2634 | 2995.049695 | 2880.79789 |
| 2014-04 | NARCOTICS | 2597 | 2581.728116 | 2609.949272 |
| 2014-05 | NARCOTICS | 2606 | 2635.5293 | 2618.707262 |
| 2014-06 | NARCOTICS | 2424 | 2462.419065 | 2476.613041 |
| 2014-01 | THEFT | 4412 | 5127.401753 | 5319.04387 |
| 2014-02 | THEFT | 4012 | 4199.112201 | 4374.304298 |
| 2014-03 | THEFT | 4498 | 5136.569762 | 5372.908773 |
| 2014-04 | THEFT | 4815 | 5392.606918 | 5551.754101 |
| 2014-05 | THEFT | 5329 | 5926.493482 | 5997.069694 |
| 2014-06 | THEFT | 5807 | 6187.941185 | 6188.315913 |

Please find below, the graphs for the above data which tells us the exact differences between both the predictions.

**BATTERY:**

**NARCOTICS:**

**THEFT:**

**Conclusion:** From the above details, we can see that both Holt winter and Regression predictions are similar. But, Holt winter prediction is closer to the Original number of events compared to the regression.

Both the graphs are moving closer to original data with increase in x-coordinate.

My definition of accuracy was depending on the “Number of Events” grouped by Year-Month and Primary Type.