National College of Ireland

Data Warehousing and Business Intelligence Project

on

Detailed Analysis of Causes and Effects of Crime in Northern Ireland

Sanket Dayama
x18143652

MSc/PGDip Data Analytics – 2019/20

Submitted to: Prof. Sean Heeney

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Sanket Dayama
Student ID:	18143652
Programme:	MSc Data Analytics
Year:	2019/20
Module:	Data Warehousing and Business Intelligence
Lecturer:	Prof. Sean Heeney
Submission Due Date:	8/04/2019
Project Title:	Detailed Analysis of Causes and Effects of Crime in Northern Ireland

I hereby certify that the information contained in this (my submission) is information pertaining to my own individual work that I conducted for this project. All information other than my own contribution is fully and appropriately referenced and listed in the relevant bibliography section. I assert that I have not referred to any work(s) other than those listed. I also include my TurnItIn report with this submission.

ALL materials used must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is an act of plagiarism and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	April 12, 2019

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Table 1: Mark sheet – do not edit

Criteria	Mark Awarded	Comment(s)
Objectives	of 5	
Related Work	of 10	
Data	of 25	
ETL	of 20	
Application	of 30	
Video	of 10	
Presentation	of 10	
Total	of 100	

Project Check List

This section capture the core requirements that the project entails represented as a check list for convenience.

- ☒ Used LaTeX template
- ☐ Three Business Requirements listed in introduction
- ☐ At least one structured data source
- ☐ At least one unstructured data source
- ☐ At least three sources of data
- ☐ Described all sources of data
- ☐ All sources of data are less than one year old, i.e. released after 17/09/2017
- ☐ Inserted and discussed star schema
- ☐ Completed logical data map
- ☐ Discussed the high level ETL strategy
- ☐ Provided 3 BI queries
- ☐ Detailed the sources of data used in each query
- ☐ Discussed the implications of results in each query
- ☐ Reviewed at least 5-10 appropriate papers on topic of your DWBI project

Detailed Analysis of Causes and Effects of Crime in Northern Ireland

Sanket Dayama
18143652

April 12, 2019

Abstract

Crime is a type of unwanted reaction or behaviour on any human by harming the person with an intent or without intent by considering various factors of the offender and making different types of damages and leaving harmful effects on the society. There has been a fluctuation of crime rates based on different factors which gave a rise in crime in the society. The government need to seek some solutions to the problems which are being neglected or not taken into consideration for controlling the crime rate and reduce it gradually. The main objective is appropriate investments need to be sanctioned and applied in the right areas in order to gain a control over the crime and grow awareness in people as to give an importance of significance of the crime if not reduced. The concerns mentioned can be solved and the solutions can be drawn by the apparatus I have developed in Data Warehouse and Business Intelligence with the data collected of Northern Ireland consisting of the number of offences recorded by the police per year , the education area was monitored within Northern Ireland, the number of unemployment rate in Northern Ireland, the net migration from northern Ireland and the type of crime performed with respect to the immediate custody of the offender found at or nearby the crime place and sent to prison.

1 Introduction

In todays world the crime developing in the society has many factors associated which are overlooked by the government which can result in a problem to control the crime. There are some factors which contribute to the rate of crime such as the unemployment rate which is a major concern where if people are jobless and the effects of being unemployed as it says empty mind is a devils workshop. The education is another important factor need to be considered because the crime and the illiteracy rate go hand in hand. The uneducated people are higher at a higher risk to commit a crime as to less awareness of the effects of it after causing the harm. The key insight for the government is to get the number of offences reported to police and the offenders sent to immediate custody while the police were patrolling or were at the spot instantly after the crime was reported and the effect of it on migration of the people. These are the factors which have been included. The main aim of the project is to compare the number of offences and the factors leading to them and the need of investment associated with it that the government should put in, in order to control the crime in Northern Ireland. A video of the project was created

to show the working model and the youtube link is given in the appendix 8.1.
There are four structured datasets collected from four different sources and one unstructured dataset of immediate custody of the offender in prison which was a PDF, it was extracted using R , cleaned , and then the data was pushed in the database to analyse the insights. Section 7.

The motivation for building this data warehouse is to make government aware about the factors which can scale up the crime and and the effect of it on migration of the people because of the crime happening in the country. In order to avoid the consequences the government should plan investments in some areas highlighted and recruiting of police officers should be scaled up with some technological equipment. From the research paper,Bennett & Ouazad (2016) suggests that the crime rate substantially increases with the increase in the unemployment rate. It also suggest that low education leads to the higher risk of the crime to be performed by the people. The major layoffs also can affect the crime rate. The article from the Irish times Gaiman (2014) tells us about the children when not educated properly and consisting of the fear of underachievement did not do well because of improper guidance and it increases their risk to execute any offence. A paper sit Christina Clark (2008) gives the idea of the literacy present in the prison of the prisoners which was a relatable finding for taking a factor education. It is also observed that people migrate because of increasing crime in the specific area because there always resides a risk of being a victim to a crime. A research in Foote (2015), paper also discusses about the migration of the people to a different country with an increasing crime rate. An estimate was carried out from the publisher of the migration using the regression techniques by relating the other areas such as demand for labour and crime offences. A report from the police services in Northern Ireland (PSNI) indicates that the problem of the ratio to number of offences occurred and the offender associated with crime was arrested and taken into immediate custody by the police. This gives us an idea of the quick response from the police services whenever an offence was occured and reported. These factors led me to create a data ware house in which a multidimensional online analytical process will be developed to fulfil my 3 requirements,

1. What is the relation between the number of offences and the number of unemployment rate yearly?
2. What is the relation between the number of offences and the number of people enrolling into colleges with different levels of education like First degree, other undergraduate and Post graduate?
3. What is the relation between the number of offences and the number of offenders sent to immediate custody and the number of net migration rate of the people from Northern Ireland?

2 Data Sources

For implementing this data ware house project I have used in total of five datasets from five different sources in which four are structured and one is unstructured.

Source	Type	Brief Summary
Statista	Structured	It consists the data of number of offences per year which is the backbone of the project.
Office for National Statistics	Structured	It provides the data for the unemployment rate a factor to be considered.
data.gov.uk	Structured	It provides the data for the net migration of the people from Northern Ireland
Department for the Economy	Structured	It provides the data for the number of enrolments in the colleges.
Department of Justice	Unstructured	A PDF report on offenders in custody for police reported crimes.

Table 2: Summary of sources of data used in the project

2.1 Source 1: Statista

1. Released Date : January 2018

The structured data was obtained from the link `https://www.statista.com/statistics/915961/number-of-crimes-in-northern-ireland/` . This dataset consists of the total number of crime offences which were recorded by the police service in Northern Ireland from 2002 to 2018. There were two sheets in the data the overview and the Data sheet, with the help of R code only the Data sheet was chosen. This dataset was important for all BI queries as to relate with the number of offences happened with the other mentioned factors. The number of offences will be my main fact or measure to be compared with every factor causing and effecting the crime. The columns present in database and used for analysis were,

1. Years
2. Number of Offences

2.2 Source 2: Office for National Statistics

1. Released Date : 19 Feb 2019

This Structured data was obtained from the link `https://www.ons.gov.uk/employmentandlabourmarkpeoplenotinwork/unemployment/timeseries/zsfa`. This dataset consists of the years and the unemployment rate in thousands from 1993 to 2018. It also consisted the quarterly data of every year which was removed using R programming because the quarterly data was not useful as was creating redundancy while relating with other datasets and when calculated the value was approximately the same as the per year value. This dataset consists for two columns the year and the number of unemployment rate in thousands. This dataset was used in the first BI query to get the relation between number of offences and unemployment rate. The columns present in database and used for analysis were,

1. Years
2. Unemployment.rate.in.thousands

2.3 Source 3: data.gov.uk

1. File added on the open data website : 28 June 2018

This Structured data was obtained from the link `https://data.gov.uk/dataset/6e4c17da-ed97-49ce-northern-ireland-net-migration` . This dataset consists of the net migration rate of the people from Northern Ireland. It had columns like Years and Net Migration which were used for the analysis and the columns which were not in sync had to be dropped such as GeoName , GeoCode with the use of R programming as the data was not useful for the analysis and was not synchronizing in order get the irredundant data with other sources. This data was used in the third BI. The columns present in database and used for analysis were,

1. Years
2. NetMigration

2.4 Source 4: Department for the Economy

1. Date published : 28 February 2019

This Structured data was obtained from the link `https://www.economy-ni.gov.uk/publications/enrolments-uk-higher-education-institutions-northern-ireland-analysis-20` dataset consists of the enrolments of the students in UK higher education institutions. This data consisted of Full time, part time and the total number of students in different categories such as First Degree, Other Undergraduate, Post graduate and Total. There was a column which was for Great Britain which needed to be dropped and only Northern Ireland was taken and the other counts such as only part time and only full time were also dropped using R programming and the total of all such parameters(First Degree, Other Undergraduate, Post Graduate) were taken for the analysis as the sum was the same. This dataset was used in the second BI query in relation with the number of offences. The columns present in database and used for analysis were,

1. Years
2. First Degree
3. Other Undergraduate
4. Post Graduate
5. Total

2.5 Source 5: Department of Justice (Unstructured)

1. Date published : 26 September 2018

This unstructured dataset was obtained from the link `https://www.justice-ni.gov.uk/publications/r-s-bulletin-262018-northern-ireland-prison-population-2017-18`. This dataset is a report from department of justice Northern Ireland which consists of the information of the immediate custody count of the offenders as per the crime type and the years. This dataset gives us the information of the count of the offenders taken into custody immediately when the crime was reported to police or a crime was found when the patrolling was done by the police. This dataset required a lot of cleaning but first data was extracted using R programming and then cleaning was done by using libraries such as tabulizer, rjava, tidyverse and pdf tools. The functions used to format some complex data were separate, subset and matrix and some unwanted data was removed. A regular expression was used to recognize a pattern of the tab separated values. A permission to use this dataset was asked to the respective authority and it was given. The screenshot is attached in the appendix figure [10]. The columns present in the database and used for analysis were,

1. Years
2. Crime Type
3. Count of Immediate Custody

3 Related Work

In today's world the Crime is a major concern for all the people in the society and the rate at which the crime grows has some factors which are related to the daily life of the people which should be considered by the Government of that country. As I am developing a data ware house and doing an analysis based on Northern Ireland, the government should consider some factors where the crime rate can be controlled by investing in various programs and schemes for the people so that people are aware of the consequences. The proper analysis done will lead to the right investment from the government in right place so that the solutions can be implemented to the root cause of the problems and solve them. As mentioned in the abstract and introduction the factors taken such as unemployment, education, custody count and migration of people should be focused.

From the research paper Witt (1999) it is shown that there is a correlation of rise in crime rate with the rise of unemployment rate.It also indicates that mainly the crime types in property crimes category such as burglary, theft and vehicle offences were measured to be rising and it was understood that there was a need of recruitment of police officials.A dynamic panel data model was created to estimate the relation. Another paper Raphael (2001) also suggested the same results mentioned in the introduction[1]. It described the rise in unemployment by performing regression on the data which had dependent variable

as crime and independent variable as time and unemployment. This motivated me to take the dataset for the unemployment as to analyse the impact of the jobless people on crime . This influenced me to create my first BI query to relate the number of offences and the unemployment rate.

According to the paper Christina Clark (2008) , the prisoners in the United Kingdom are the highest in Europe. The literacy of these prisoners is surveyed in this paper and has the ratio of 70 percent of the prisoners or offenders have problems in basic literacy and 25 percent were have a bit of reading skills. From the research paper Groot & Brink (2007) discusses on the effect of education on crime and tells us about the decrease in the crime ratio of burglary, robbery and other property crimes as the people were educated and knew the difference of things which are not correct and could be an offence. This led me to take the dataset for the education enrolments for studies which gave me an idea as to how many students are enrolling for the different levels of programs such as first degree, other under graduate and post graduate as mentioned in the data source 4(2.4). This gave me the idea of literacy rate related to offenders which is less literacy will lead to more crime. This influenced me to design my second BI query to relate the number of enrolments in colleges to the number of offences.

A Journal paper also states that the net migration of the people from and to cities was on high because of crime. Wooldredge (1986). It also states that the migration out of the country had strongly linked with crime with the multiple regression analysis performed by the publishers. Another researcher, from his paper DUGAN (1999) suggests that the household takes a decision to move out of the area as there is a risk associated of the crime getting repeated on them. An analysis was done by taking the data from nationally representative survey which conveyed positive results for percieved risk of crime on the victim which forces them to leave the area. This was particulary in the domain of property crimes. This gave me an idea to take the dataset of the net migration rate of Northern Ireland which can be related with the number of offences and the immediate custody count of the offenders. The report of the department of justice for Northern Ireland consisted the data for the count of immediate custody with respect to the crime types and the years. It consists of the statistics as to how many offenders were sent to immediate custody the rate which was increased or decreased with respect to any type of crime. This led me to have my third BI query which would be the number of offences vs the number of offenders sent to immediate custody and the migration of people due to the crime.

4 Data Model

The literature survey conducted by me gave me a direction as to which factors should be considered to create the business requirements so that it gives an idea to the authorities to identify the problems to tackle the crime in the above mentioned areas.

In order to build a data warehouse, there are two approaches to build it they are the Kimballs bottom-up approach and Inmons top-down approach.Both approaches have their own perspectives and advantages according to the different requirements. The

Inmons top-down approach created by Bill Inmon consists of the facts or measures stored first and the dimensional data marts are derived later from the measures and the data warehouse is created. The Kimballs bottom-up approach is to create the datamarts first from the data obtained from the sources and then create the dimensions accordingly which are related to the facts or measures. The fact table is created by relating the dimensions and the facts present in the data to get the business requirements and run the queries. In this project Kimballs bottom-up approach is used as the sources were discovered and the data marts were formed for each source. A study [A Survey on Data Warehouse Approaches for Higher Education Institution Panacea Makele , Srinath Doss 2018] reveals that High level of performance provided by the Kimballs approach to the data warehouse and it also focuses on ease of end users accessibility. It also gives an incremental development framework.

The dimensions were created from the data sources which needed to be joined on the measures depending on them. There are two dimensions of this data warehouse built which is the year dimension and crime type dimension.
Justification for two dimensions : The data which I got from my sources mostly consisted of more measures than the columns supporting to it. These dimensions satisfied the need of all the measures. In order to remove the redundancy a new dataset was added and the one causing problems was removed, this limited to have two dimensions. The BI queries were also satisfied with the current model and design.
These dimensions were created from the data sources by joining them on the common values or attributes so as to get the desired output. The years column which was present in every dataset was joined with every year column present in the data from different sources and the dimension for years was created with having the common years present in the dataset. While joining the tables it took the common years existing from all the tables and the unstructured data had less years compared to all which led to reduction of data. The crime types were taken as the second dimension as they represent a unique values from which the different facts can be measured. In this way two dimensions were created and the dimensions are discussed below,
YearDimension : This dimension was derived from all the five datasources listed in the data sources section 2.It is the most important dimension in the project to measure the data per year and to get the insights accordingly. It consists of two columns the YearID and the Years. The YearID is a primary key created to identify the unique values and match the facts on the basis of it. The Years column consists of the common years present in all datasets.The table was created in SSMS and the data was inserted from the SSIS by creating a script of inserting and inner join in SQL.
CrimeTypeDimension : This dimension was created consisting unique crime types present in the unstructured dataset in data source 2.5. The importance of this dimension is to have the different crime types listed and a fact is to be measured on the basis of these crime types. These crime types will be an important factor to measure the number of offenders in custody. This dimension has two columns CrimeTypeID and CrimeType. The CrimeTypeID is a primary key consists of unique ids assigned to each CrimeType and the CrimeType column consists of the types listed in it. The table was created in SSMS and the data was inserted from SSIS creating a script of inserting the data in SQL.
FactTable : The fact table created consists of the measures which are to be related with the dimensions using joins and make the BI query. Fact table consists of the soultions to the problems listed by the analysis of the different measures present in it. Below men-

tioned are the columns in my fact table,
YearID : The primary key of the YearDimension table.
CrimeTypeID : The primary key of the CrimeTypeDimension table.
NumberOfOffences : It consists of the count of number of offences per year.
Unemployment.rate.in.Thousands : It consists of the count of number of unemployment rate per year.
FirstDegree : The count of number of students enrolled for first degree in colleges.
OtherUndergraduate : The count of number of students enrolled for other undergraduate degree in colleges.
Postgraduate : The count of number of students enrolled for post graduate degrees in colleges.
Total : The total number of count of students enrolled in all the levels of education in colleges.
Details : The levels of education can help the data analyst to drill down the data against the total number of education enrolments where he can identify the trend more efficiently.
CountOfImmediateCustody : The count of offenders taken into custody after police recorded crime and while patrolling.
NetMigration : The count of people migrating from Northern Ireland.
With the help of Kimballs approach of data modelling and above mentioned dimensions and facts, a star schema is formed shown in the figure below. The star schema was drawn using a tool which is a website called lucidchart.

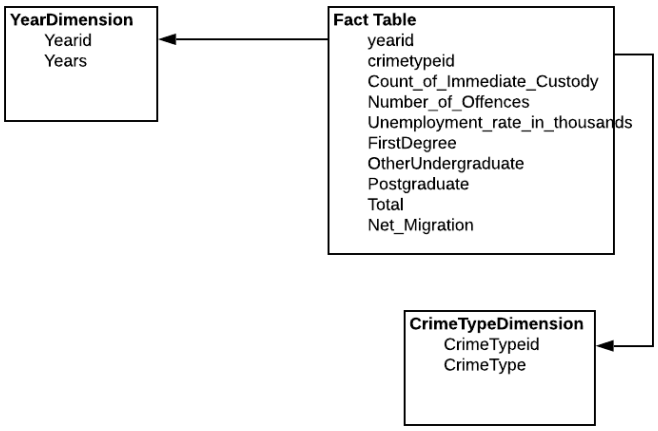

Figure 1: Star Schema for Analysis of Crime in Northern Ireland

5 Logical Data Map

Below is the Logical data map table containing the information of all the facts and dimensions and their transformations.

Table 3: Logical Data Map describing all transformations, sources and destinations for all components of the data model illustrated in Figure 1

Source	Column	Destination	Column	Type	Transformation
Statista, Office for National Statistics, Department for the Economy,data.gov.uk, Department for Justice	Years	YearDimension	Years	Dimension	taken as next year for 2015/16 to 2016 as '/' was needed to be removed and changed datatype to numeric.
Statista	Numberof Offences	FactTable	Numberof Offences	Fact	Taken as it is given in the source in the cleaned format.
Office for National Statistics	Unemployment rate in thousands	Facttable	Unemployment. rate. in. thousands	Fact	removed the attached quarterly data column as the average was same.
Department for the Economy	First degree	FactTable	FirstDegree	Fact	removed the unwanted rows of labels and columns other than Northern Ireland

Continued on next page

Table 3 – *Continued from previous page*

Source	**Column**	**Destination**	**Column**	**Type**	**Transformation**
Department for the Economy	Other Undergraduate	FactTable	Other Undergraduate	Fact	removed the unwanted rows of labels and columns other than Northern Ireland
Department for the Economy	Postgraduate	FactTable	Postgraduate	Fact	removed the unwanted rows of labels and columns other than Northern Ireland
Department for the Economy	Total	FactTable	Total	Fact	removed the unwanted rows of labels and columns other than Northern Ireland
data.gov.uk	NetMigration	FactTable	NetMigration	Fact	taken as it is from the source.
Department of Justice	CrimeType	CrimeType Dimension	CrimeType	Dimension	Extracted the data from the PDF and cleaned the NAs by na.omit() funtion.
Department of Justice	Count ofImmediate Custody	FactTable	Countof Immediate Custody	Fact	Extracted the data from the PDF and cleaned the NAs by na.omit() funtion and splitted the single column with spaces and added the values with vector.

6 ETL Process

The ETL process is the backbone for any project in data warehousing. The three stages such as extract, transform and load all have their own importance and significance related to the different stages. The extraction stage consists of the data to be extracted from different sources.The transformation stage is to transform or clean the extracted data obtained from the different sources. This transformed data should be free from any redundancy and any garbage values.The last stage is loading, here the transformed data is loaded into the database. Surajit Chaudhuri (1997). In this project the Extraction and Transformation process was automated by integrating the R scripts into SSIS and the loading was also automated by integrating with SSMS.

6.1 Extraction

In this project I have extracted the data using R scripts from 5 different sources. The data was then extracted with the help of R programming and was written into an csv file after the cleaning. A function read.csv(url("link")) from the library "readxl" was used for extracting the data from the source, here the data was extracted into a data frame and then was stored in csv file. My four datasets consisted of an excel and the fifth was a PDF where different libraries and functions were used as per the need of requirement in cleaning.The excel data was extracted by using by using the readxl library and the readexcel() function was used to get the data into the dataframe and later perform operations on it. For the unstructured data, as it was a PDF different libraries such as pdftools, tabulizer and rJava were used. For libraries like tabulizer and rJava there was a need to install java and set the environment variable for them. A latest version of java was installed and then the path of the java environment needed to be entered in the R code to use the libraries. In order to extract the data from the page 20 of the PDF a for loop was created to iterate the pages and a function extracttables() was used to get the table data from the PDF.

6.2 Transformation

Firstly the data was scanned by me according to the requirements BI queries where I got to know the data useful for the project and the unwanted data which could create some irregularities and redundancy. The unwanted data was removed from the datasets by using the index (for example dataframe[-1:-13,]) of the data which is visible in R console.The unwanted columns were removed by using Null string to remove the data stored in the data frame. The columns were renamed because the R was giving the column a temporary name when the data was pulled into the dataframe, colnames() function was used to solve this issue. The unstructured data PDF had columns which

had tab separated values which needed to be split using the separate() command with giving the token as tab. These values were separated into different columns and added again into a single column by adding the values by making as.numeric() and storing the data into a dataframe. The subset() function was used to gather all the columns which were stored in a vector and then with the help of select command the data was stored in the data frame. The data was in the different format, so matrix() function was used and the data was transposed into the required row-column format. This data was stored into a data frame and then was written in to a csv file. The cleaning process of unstructured data was a tedious task and had some challenges while cleaning it which consumed a lot of time. The greatest challenge was to identify the pattern wether it was tab separated or space separated which was identified in a software called editplus and then cleaned using R programming. The above operations were performed in R programming using RStudio and the code is available in the appendix.

6.3 Load

After the transformation phase of the data, it needs to be loaded into the database and the staging needs to be completed. SSMS(SQL Server Management Studio) and SSIS(Sql Server Integration Services) were used in order to prepare the staging area and load the raw data into the database. A dataflow task was taken into the SSIS and was named after a dataset. The dataflow task consisted of a flat file source where the file to be imported was chosen and was mapped to an OLEDB destination. The columns were mapped and the connection was created for each dataset so that, on that connection the data would be loaded into the table. Here the table is created and the data is loaded into SSMS through SSIS. This step was done five times as i have five datasets which needed to be loaded into the database. As the raw tables are loaded, we go to the process of creating the dimensions where the dimension tables were created in SSMS and an execute sql task was chosen in the SSIS and the join query of data insertion was written in it so that the common data from all the sources for the dimensions was loaded into the dimension tables in SSMS through SSIS. The next phase was to create a fact table which is the most important phase of the project. I used lookup transformation to create and populate the fact table a feature available in SSIS. An OLEDB source was selected in which a table was selected and then a lookup transformation was selected which was mapped to the OLEDB source on the common columns which were present in both the tables. Later the look ups were selected for each table and were mapped with each other on common columns and the measures or facts were collected into OLEDB destination where the fact table was created and the data was inserted in to the SSMS through SSIS. The truncate command was used to delete the data from the fact table when the whole process was executed again so that there is no re-insertion of the data into the fact table. After populating the fact table the star schema was prepared by selecting a SSAS project and loading all the facts and dimensions in it. Now at the final stage the MOLAP cube was created which was named the Crimecube and the existing fact and dimension tables were selected to populate it. The deployment of the OLAP cube was also automated by creating an analysis services processing task into the SSIS and then the connection of the cube from SSAS was added to make the process fully automated.

7 Application

After the ETL process and cube deployment we can now address the business queries which are created in tableau. The business requirements mentioned in introduction that is in section [1] are designed with the related work done in section [3]. The results of the evaluation of the requirements are listed below.

7.1 BI Query 1: What is the relation between the number of offences and the unemployment rate yearly?

For this query, the contributing sources of data are: source(2.1) which consisted of the number of offences in Northern Ireland and source(2.2) which consisted the unemployment rate in Northern Ireland.

The general idea as illustrated in Figure 2 is the relationship between the Number of offences and the unemployment rate. This figure shows that the number of offences decreases with the decrease in the number of unemployment per year. In the year 2016 Northern Ireland experienced highest number of offences that is crime as the unemployment rate was the highest as compared to the years 2017 and 2018. The Number of offences decreased in the year 2017 compared to 2016 as the Unemployment rate was decreased. The offences were on a rise again in 2018 but less than 2016.

Figure 2: Results for BI Query 1

7.2 BI Query 2: What is the relation between the number of offences and the number of people enrolling into colleges with different levels of education like First degree, other undergraduate and Post graduate?

For this query, the contributing sources of data are: source(2.1) which consisted of the number of offences in Northern Ireland and source(2.4) which consisted the number of students getting enrolled for the education in Northern Ireland.

The general idea as illustrated in Figure 3 is relationship between the number of offences and students enrolling for education at different levels. We can observe that the number of enrolments were less in the year 2018 hence the number of offences were increased comparing to the year 2017. Now if we check the details, it also tells us that the number of students enrolling for different levels of education such as first degree, other under graduate and post graduate differed with less rate in 2018 hence the offences increased. In the year 2016 we can see the increase in the rate of offences with increase in the enrolments for education, but it can be concluded that the general idea is the number offences increase with less education enrolments.

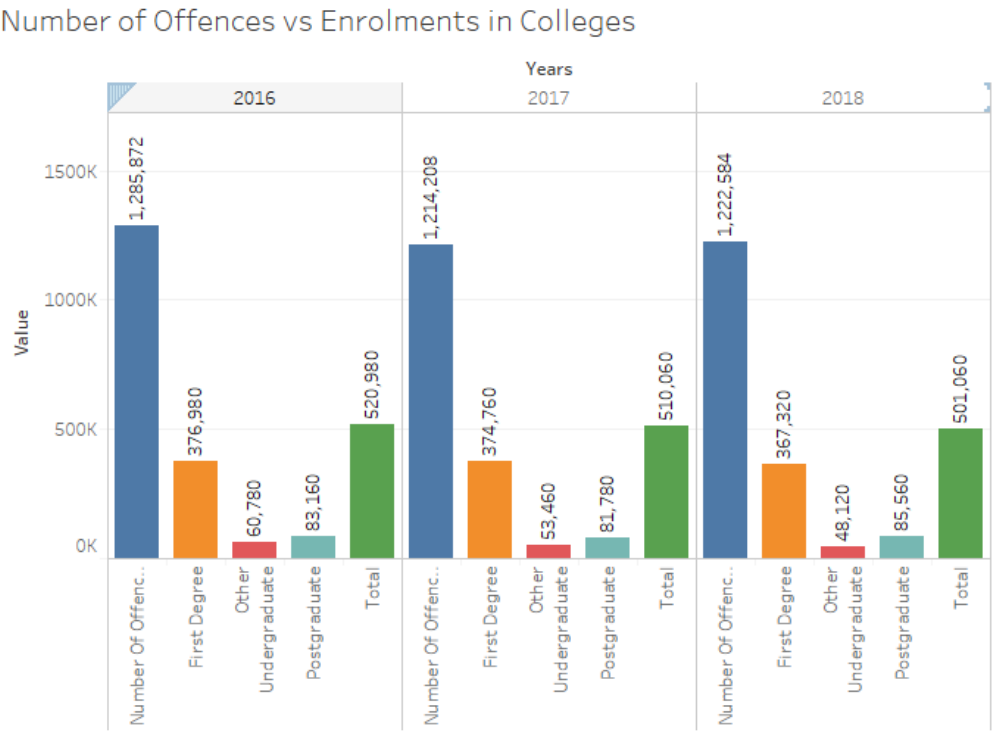

Figure 3: Results for BI Query 2

7.3 BI Query 3: What is the relation between the number of offences and the number of offenders sent to immediate custody and effect of number of net migration rate of the people from Northern Ireland?

For this query, the contributing sources of data are: source(2.1) which consisted of the number of offences in Northern Ireland, source(2.3) which consisted the migration of people from Northern Ireland and source(2.5) which consisted the number of offenders sent to immediate custody after getting caught by the police.

The general idea can be observed as illustrated in Figure 4 can be observed that compared to the number of offences recorded the count of immediate custody is so less that it is merely visible and with that the migration of people due to crime.The offences are divided into different crime types and have the custody count of each of the offence. It can be observed that for the crime type burglary in 2016 the number of offences where the highest that is 2.14 million compared to the count of immediate custody was just 68 which led to migration of the people from Northern Ireland which was 2916. In the year 2017, the offences were decreased to 2.02 million and the immediate custody count was 65 which is almost the same which was in 2016. It is a good sign where the offences were decreased and the custody rate was not much differed, hence less people migrated from the country which was 2350 in 2017. In the year 2018 the number of offences increased to 2.03 million where as the immediate custody count was decreased again to 63 which shows the greater risk of control over the offence. Hence we can see that the migration increased to 2500 in 2018 as compared to 2017.

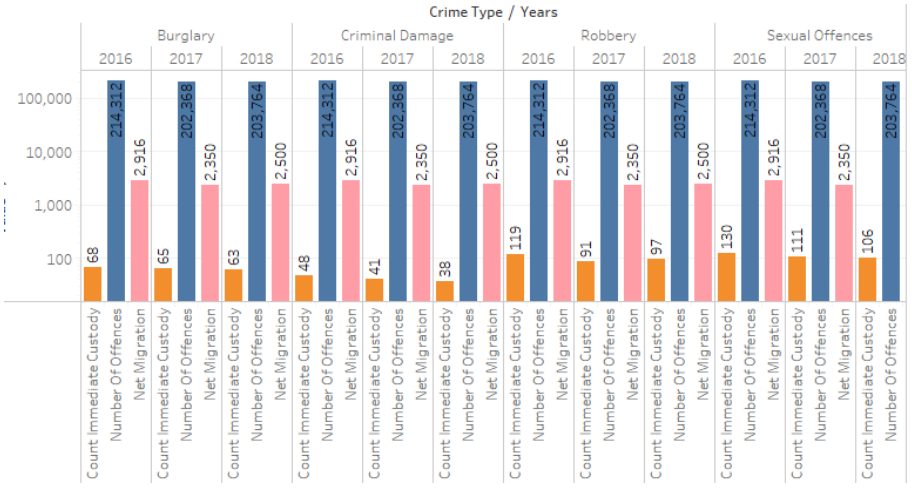

Figure 4: Results for BI Query 3

7.4 Discussion

As we now have the results for our business requirements from our business queries, the trends can be observed and studied for making decisions. This report gives a brief idea on analysis of three years as to what steps should the government should take in order to decrease the number of crime. In the first business query it can be seen that the number of offences decrease when the unemployment rate decreases as studied from Witt (1999) in related work[3] . But also there is a significant decrease of unemployment rate in the year 2018 where the offences are increased but less than 2016. It can be assumed that social media offences might be the factor for them. For the second business query it can be seen that the number of offences decrease when the number of enrollments increase in colleges.The number of offences were high in 2016 and the enrolments were high too which shows a limitation. In the year 2017 the number of offences decreased because the enrolments were increased, whereas in 2018 we can see a rise in number of offences as the enrolments in the colleges were decreased.Here the literacy rate is the key factor measured as mentioned from the paper Groot & Brink (2007) where it is analysed that the behaviour and the intellect of the people is developed of what will be the significance of such unwanted behaviours. The level of education also matters where, how many people were enrolled in first degree, other undergraduate and post graduate, which defines the level of literacy and can be a major factor to decrease the crime. As we can see in the year 2016 we see that the education rate is high and the offences are high too the cyber crimes can cause the increase in crime as these offenders are considered literate hence there is a need of educating the ethics from the educational institutions to guide the students in the right direction Gaiman (2014). For the third business query which involves the factor of the migration on the basis of number of offences and the count of immediate custody of the offenders by their crime type it can be seen as where the number of offences are high and the custody count is low, we can see the migration of people is high as mentioned in the related work in section [3] . We can understand the areas where there is low security and less offenders are being arrested the area to be lived is risky and eventually people will migrate to a safer area. Hence the result of my third query supports the paper Wooldredge (1986) where migration is boosted at the places of crime.

8 Conclusion and Future Work

We have seen the trend or the idea of increase and decrease of number of offences because of the factors affecting it. These factors also imply the effect of it that is migration from the country. I tried to answer a question by building this data warehouse is what is the relation of the factors like unemployment, education and immediate count of custody on the crime (number of offences) and its effect which is migration ? The result from data warehouse built by me was to combine and co-relate the different domains mentioned in the question and their data so that it will help the government and police forces to invest the money in the right areas such as pushing to create jobs by funding start ups and more recruitments in police to help reduce the crime. The investments from the education

institutions can also help by giving scholarships to low poverty people collaborating with the government. An ideal investment into technological investments can also help monitor the offences which will reduce the crime. The limitations of the project are such as the enrolments for the colleges for education could be more detailed into the ones doing part-time and full-time courses. It could give a better idea of which category is contributing more to crime. Another limitation is which gender is more affected to the number of offences happened which will give a detail of any gender specific crimes. This project gives us the idea of crime in Northern Ireland for property crimes but it can be detailed more involving some future work by getting into all the domains like cyber crimes and location wise distribution. Another thing I would suggest for the future work is to measure the justice given per offence for the offender and also different data sets which are based on gender and race will modify the schema and dimensions can be added, which will give a detailed analysis of crime and the time spent in prison which can be done having a team and this analysis will be help the government to improve the judicial system.

References

Bennett, P. & Ouazad, A. (2016), 'The relationship between job displacement and crime 29 october 2016'.

Christina Clark, G. D. (2008), 'Literacy changes lives'.

Doss, P. M. . S. (2018), 'A survey on data warehouse approaches for higher education institution'.
URL: *http://www.ijirase.com/assets/paper/issue*$_1$1/*volume*$_1$/*Issue* − 11 − 223 − 227.*pdf*

DUGAN, L. (1999), 'The effect of criminal victimization on a households moving decision criminology'.

Foote, A. (2015), 'Decomposing the effect of crime on population changes'.

Gaiman, N. (2014), 'https://www.irishtimes.com/news/politics/literacy-is-key-to-keeping-children-from-life-of-disadvantage-and-crime-1.1669346'.

Groot, W. & Brink, H. M. V. D. (2007), 'The effects of education on crime'.

Raphael, S., . W. R. (2001), 'Identifying the effect of unemployment on crime'.

Sinha, A. S. . A. (n.d.), 'A comparison of data warehousing methodologies'.

Surajit Chaudhuri, U. D. (1997), 'An overview of data warehousing and olap technology'.

Witt, R. (1999), 'Crime and economic activity. a panel data approach. british journal of criminology'.

Wooldredge, R. J. S. J. D. (1986), 'Evidence that high crime rate encourage migration away from central cities'.

Appendix

8.1 video of working project :

https://www.youtube.com/watch?v=SWbf2STQITM

R code

```
###Unstructured data###
getwd()
setwd("C:\\data\\DA2019\\DWBI\\datasets\\Crime2\\Crimedatatemp")
library(pdftools)
library(tidyverse)
library(tabulizer)
#prisondata <- pdf_text("C:\\data\\DA2019\\DWBI\\datasets\\Crime2\\Northern
prisondata <- c(data_frame())
for (i in 20){
  out <- as.data.frame(extract_tables("C:\\data\\DA2019\\DWBI\\datasets\\Cri
  prisondata[[i]] <- out
}
prisonout = out[-1:-3,]
prisonout
prisonout1<-separate(out, X2, into = c("V9", "V10"), sep = "␣")
prisonout2<-separate(out1, X3, into = c("V11", "V12"), sep = "␣")
prisonout3<-separate(out2, X4, into = c("V13", "V14"), sep = "␣")
prisonout3

prisonout3$year20152016custody <- as.numeric(prisonout3$V9) + as.numeric(pri
prisonout3$year20162017custody <- as.numeric(prisonout3$V11) + as.numeric(ou
prisonout3$year20172018custody <- as.numeric(prisonout3$V13) + as.numeric(pr
out3
out3 = subset(prisonout3,select=-c(V9,V10,V11,V12,V13,V14))
out3

data<-t(as.matrix(prisonout3[,2:4]))
#the needed columns are cbinded
prisonoutresout3<-cbind(out3[rep(1:nrow(out3),each=3),1],  #this repeats the
           Year <- c("2015-16","2016-17","2017-18"),
           #Year= as.vector(data[,]),#taking the unique values from the data
           prisonout3_VAlue = as.vector(data[1:3, ]))

prisonoutresout3 = prisonoutresout3[-97:-102,]
prisonoutresout3 = prisonoutresout3[-43:-51,]

write.csv(prisonoutresout3, "C:\\data\\DA2019\\DWBI\\datasets\\Crime2\\Crime
```

```
prisondata <- data.frame(read.csv(file="C:/data/DA2019/DWBI/datasets/Crime2/
prisondata = prisondata[-76:-87,]
prisondata = prisondata[-64:-72,]
prisondata = prisondata[-55:-57,]
prisondata = prisondata[-49:-51,]
prisondata = prisondata[-34:-42,]
prisondata = prisondata[-22:-30,]
prisondata = prisondata[-13:-15,]
prisondata = prisondata[-7:-9,]
#prisondata$X <- NULL
#prisondata$X.1 <- ""
write.csv(prisondata, "C:\\data\\DA2019\\DWBI\\datasets\\Crime2\\Crimedatate


##Unemployment data##
library(readxl)
library(plyr)
library(tibble)
library(dplyr)
library(sqldf)
X <- read.csv(url("https://www.ons.gov.uk/generator?format=csv&uri=/employme
CrimeUnemp1 <- X #saving into a dataframe
CrimeUnemp1
CrimeUnemp1 = CrimeUnemp1[-1:-17,]
CrimeUnemp1 = CrimeUnemp1[-17:-461,]
##CrimeUnemp1 = CrimeUnemp1[-35:-41,]
##CrimeUnemp1 = CrimeUnemp1[-7]
write.csv(CrimeUnemp1, "C:\\data\\DA2019\\DWBI\\datasets\\Crime2\\Crimedatat

################statista data##########
library(readxl)
library(plyr)
library(tibble)
library(dplyr)
library(sqldf)
Statista <- data.frame(read_excel("C:/data/DA2019/DWBI/datasets/Crime2/Crime
Number_of_police_recorded_crimes_in_Northern_Ireland_2002_2018 <- Statista[-
Number_of_offences <- Statista[-1:-2, 2]
Statista <- data.frame(Number_of_police_recorded_crimes_in_Northern_Ireland_
Statista
write.csv(Statista, "C:\\data\\DA2019\\DWBI\\datasets\\Crime2\\Crimedatatemp
##education data####
library(readxl)
library(plyr)
library(tibble)
library(dplyr)
library(sqldf)
edudata <- data.frame(read_excel("C:/data/DA2019/DWBI/datasets/Crime2/Crimed
#removing the unwanted columns
str(edudata)
edudata$..18 <- NULL
```

```
edudata$..3 <- NULL
edudata$..4 <- NULL
edudata$..5 <- NULL
edudata$..7 <- NULL
edudata$..8 <- NULL
edudata$..9 <- NULL
edudata$..11 <- NULL
edudata$..12 <- NULL
edudata$..13 <- NULL
edudata$..15 <- NULL
edudata$..16 <- NULL
edudata$..17 <- NULL
edudata$..18 <- NULL
edudata$..19 <- NULL
edudata = edudata[-38:-43,]
edudata = edudata[-3:-26,]
edudata
edudata = edudata[-1,]
edudata = edudata[-2,]
#renaming the columns
colnames(edudata)[which(names(edudata) == "Table.1..Northern.Ireland.domicil
colnames(edudata)[which(names(edudata) == "..2")] <- "FirstDegree"
colnames(edudata)[which(names(edudata) == "..6")] <- "OtherUndergraduate"
colnames(edudata)[which(names(edudata) == "..10")] <- "Postgraduate"
colnames(edudata)[which(names(edudata) == "..14")] <- "Total"
edudata = edudata[-1,]
write.csv(edudata, "C:\\data\\DA2019\\DWBI\\datasets\\Crime2\\Crimedatatemp\
##Migration excel###
library(readxl)
library(plyr)
library(tibble)
library(dplyr)
library(sqldf)
Migration <- data.frame(read_excel("C:/Users/MOLAP/Desktop/Crime2/Crimedatat
Migration$Geo_Name <- NULL
Migration$Geo_Code <- NULL
Migration$Gender <- NULL
Migration$Age <- NULL
Migration$Mid_Year_to_Mid_Year <-NULL
Migration$Net_Migration <- NULL
Migration = Migration[-18:-1456,]
Migration$MigrationID <- 301:317
str(Migration)
Migration$Years <- 2002:2018
write.csv(Migration, "C:\\Users\\MOLAP\\Desktop\\Crime2\\Crimedatatemp\\Migr
```

8.2 Screenshots of data sources.

Below are the data sources used in this project mentioned in section 2.

1. Statista data source

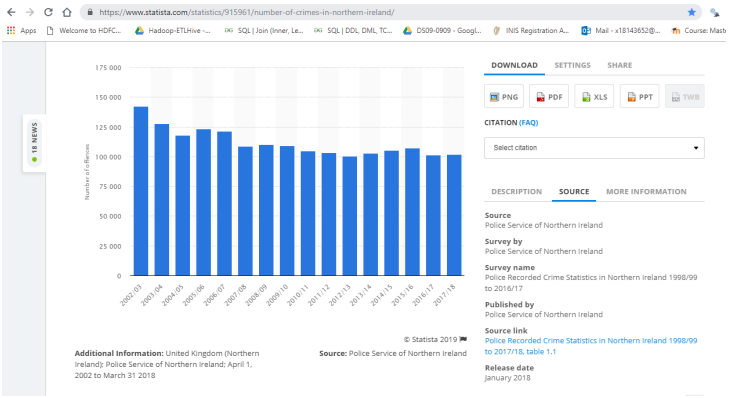

Figure 5: Statista data

2. Office for National Statistics

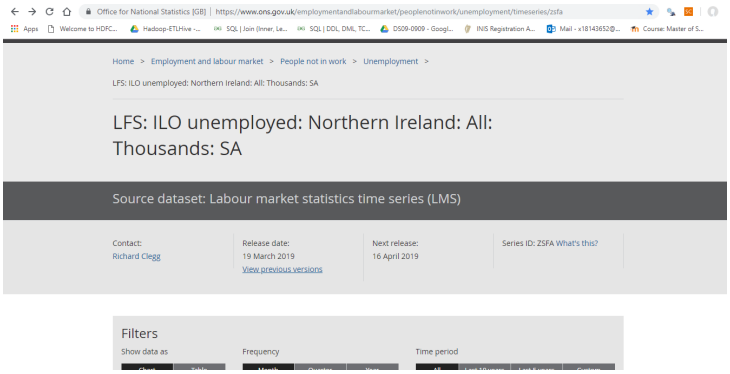

Figure 6: Office for National Statistics : unemployment data

3. data.gov.uk
4. Department of Economy
5. Department of Justice
6. Permission to use the Department of justice data

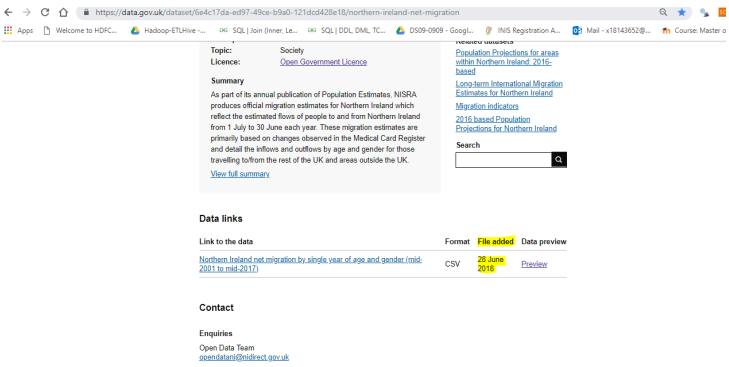

Figure 7: data.gov.uk : Migration data

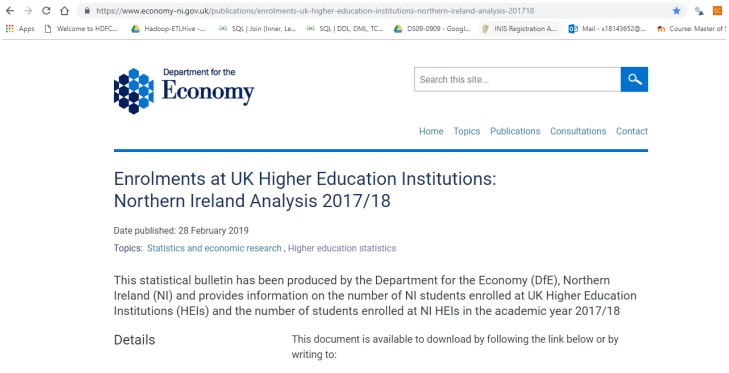

Figure 8: Department of Economy : Education data

Figure 9: Prison Custody data

Figure 10: Permission to use data