Akash Bharti
B00837970

# CSCI 5408
# Data Management, Warehousing and Analytics
# Assignment 1

## A) Feasibility Analysis

1. Yes, I have found enough datasets that can be used to identify the key performance indicators to improve business, education, lifestyle, and safety of the Halifax region.

2. I found the following datasets for identifying the KPI.

### 2.1 Safety:

1. Name: Crime
   URL:
   http://catalogue-hrm.opendata.arcgis.com/datasets/f6921c5b12e64d17b5cd173cafb23677_0
   Reason: this dataset provides the crimes reported in the last 7 days along with their accurate location.

2. Name: Annual Numbers and Rates of Accident Fatalities in NS by Zone of Residence
   URL:
   https://data.novascotia.ca/Health-and-Wellness/Annual-Numbers-and-Rates-of-Accident-Fatalities-in/nbw4-zth5
   Reason: this dataset will help us identify the accidents that have happened in the past, fatality rate and frequency of the accidents.

3. Name: Department of Justice: Justice Centres/Courts
   URL:
   https://data.novascotia.ca/Crime-and-Justice/Department-of-Justice-Justice-Centres-Courts/xdjw-yy9m
   Reason: this dataset lists the location of Justice centres and courts in the area.

### 2.2 Education:

1. Name: Graduation Rates
   URL:
   https://data.novascotia.ca/Education-Primary-to-Grade-12/Graduation-Rates/fqau-nfyv
   Reason: This dataset provides insight into the graduation rates of high schools.

2. Name: Nova Scotia Public School Contact Information
   URL:
   https://data.novascotia.ca/Education-Primary-to-Grade-12/Nova-Scotia-Public-School-C

ontact-Information/iyap-ttn5
Reason: This dataset provides information about public schools and the area in which they are located.

3. Name: Early Childhood Education Training Institutions
   URL:
   https://data.novascotia.ca/Education-Early-Childhood/Early-Childhood-Education-Training-Institutions/5zyd-q83j
   Reason: This dataset provides a list of institutions that offer early childhood education training.

4. Name: Recognized Private Schools Granting the Nova Scotia High School Leaving Certificate
   URL:
   https://data.novascotia.ca/Education-Primary-to-Grade-12/Recognized-Private-Schools-Granting-the-Nova-Scoti/mjkg-93xg
   Reason: This dataset provides a list of private schools that offer high school leaving certificates.

**2.3 Lifestyle**

1. Name: Nova Scotia Government Pay Scales
   URL:
   https://data.novascotia.ca/Employment-and-Labour/Nova-Scotia-Government-Pay-Scales/hn6q-5dmm
   Reason: This dataset provides pay scales for government jobs in the province.

2. Name: Bus Routes
   URL:
   http://catalogue-hrm.opendata.arcgis.com/datasets/e3b2bfdd61154176822c00602504c950_0/data
   Reason: This dataset provides information about all the bus routes in Halifax.

3. Name: HRM Tax Rates
   URL:
   http://catalogue-hrm.opendata.arcgis.com/datasets/9dd46591618948efb0efb52f9f54efd1_0
   Reason: This dataset provides tax rates for various ares rates and taxes within HRM.

4. Name: Outdoor Recreation Areas
   URL:
   http://catalogue-hrm.opendata.arcgis.com/datasets/a52efe6252044ba482ff35fac4590e34_0

Reason: This dataset provides information about the various outdoor activities available for the people in Halifax.

5. Name: HRM Parks
   URL: http://catalogue-hrm.opendata.arcgis.com/datasets/3df29a3d088a42d890f11d027ea1c0be_0
   Reason: This dataset provides information about various parks in HRM.

6. Name: Hospitals
   URL: https://data.novascotia.ca/Health-and-Wellness/Hospitals/tmfr-3h8a
   Reason: This dataset provides information of all the hospitals in the province.

7. Name: Boat Facilities
   URL: http://catalogue-hrm.opendata.arcgis.com/datasets/47fe59959754485c9025e7249444c94f_0
   Reason: This dataset provides information about the various boat facilities available in HRM which makes it an important recreational factor in improving the lifestyle

**Business:**

1. Name: Tourism Nova Scotia Visitation
   URL: https://data.novascotia.ca/Business-and-Industry/Tourism-Nova-Scotia-Visitation/n783-4gmh
   Reason: This dataset provides information about the tourism industry of Nova Scotia. Not sure

2. Name: Nova Scotia Business Inc. Export Development and Investment Attraction Activity
   URL: https://data.novascotia.ca/Business-and-Industry/Nova-Scotia-Business-Inc-Export-Development-and-In/6aac-8xtn
   Reason: This dataset provides information about the various services provided by Nova Scotia Business Inc. to help businesses across the province growth.

3. Name: Innovacorp Incubation Resident Companies 2018-2019
   URL: https://data.novascotia.ca/Business-and-Industry/Innovacorp-Incubation-Resident-Companies-2018-2019/qhzg-5qkx
   Reason: This dataset provides information about Innovacorp's incubation companies.

4. Name: Innovacorp Venture Capital Investments 2018-2019
   URL:
   https://data.novascotia.ca/Business-and-Industry/Innovacorp-Venture-Capital-Investments-2018-2019/qbpb-7r4c
   Reason: This dataset provides information about Venture capital investment in startups in the province.

5. Name: [ARCHIVED] Business Establishments 2010-2011
   URL:
   https://data.novascotia.ca/Business-and-Industry/-ARCHIVED-Business-Establishments-2010-2011/wa8g-ji9a
   Reason: This dataset reports the number of business establishments in the province.

6. Name: Nova Scotia Business Inc. Film and Television Production Incentive Fund Activity
   URL:
   https://data.novascotia.ca/Business-and-Industry/Nova-Scotia-Business-Inc-Film-and-Television-Produ/upw3-yx9z
   Reason: This dataset provides information about film and television industry in the province.

7. Name: Capital Projects
   URL:
   http://catalogue-hrm.opendata.arcgis.com/datasets/3d468db830e3430b8e4340015e11517e_0
   Reason: This dataset provides information about HRM capital projects for each construction season.

8. Name:  Awarded Public Tenders
   URL:
   https://data.novascotia.ca/Procurement-and-Contracts/Awarded-Public-Tenders/m6ps-8j6u
   Reason: This dataset contains information about government and public sector tenders awarded to different vendors.

3. I searched for the required dataset on the websites[1] [2] provided and then downloaded the dataset which I thought would be appropriate for identifying the KPI. I then used **Microsoft Excel** to remove the unwanted column from each datasets. I used **python**[3] script to format the data in which I wanted.

Akash Bharti
B00837970

4. Entities:

**Strong Entities**

1. **Crime** - taken from CRIME dataset and used only 4 fields.

2. **Justice Centres** - taken from “Department of Justice: Justice Centres/Courts” dataset and used only 6 fields.

3. **Parks-** taken from “HRM Parks” dataset and used only 6 fields.

4. **Outdoor Recreation Areas** - taken from “Outdoor Recreation Areas” dataset and used only 3 fields.

5. **Tourist Visitors**- taken from “Tourism Nova Scotia Visitation” dataset and used only 5 fields.

6. **Transit-** taken from “Bus Routes” dataset and used only 4 fields.

7. **Accidents-** taken from “Annual Numbers and Rates of Accident Fatalities in NS by Zone of Residence” dataset and used only 5 fields.

8. **Hospitals-** taken from “Hospitals” dataset and used only 4 fields.

9. **Venture Capital Investments**- taken from “Innovacorp Venture Capital Investments 2018-2019” dataset and used only 5 fields.

10. **Incubation Companies** - taken from “Innovacorp Incubation Resident Companies 2018-2019” dataset and used only 5 fields.

11. **Entertainment Productions** - taken from “ Nova Scotia Business Inc. Film and Television Production Incentive Fund Activity” dataset and used only 6 fields.

12. **Business Establishments** - taken from “ [ARCHIVED] Business Establishments 2010-2011” dataset and used only 6 fields.

13. **Business Activities** - taken from “Nova Scotia Business Inc. Export Development and Investment Attraction Activity” dataset and used only 5 fields.

14. **Capital Projects** - taken from “Capital Projects” dataset and used only 3 fields.

15. **Government Pay Scales** - taken from “Nova Scotia Government Pay Scales” dataset and used only 6 fields.

16. **Tax Rates** - taken from “HRM Tax Rates” dataset and used only 7 fields.

17. **Public Schools** - taken from “Nova Scotia Public School Enrolment by Board and School” dataset and used only 5 fields.

18. **Private Schools** - taken from “Recognized Private Schools Granting the Nova Scotia High School Leaving Certificate” dataset and used only 3 fields.

19. **Childhood Training Institutions** - taken from “Early Childhood Education Training Institutions” dataset and used only 4 fields.

20. **Boat Facility** - taken from “Boat Facilities” dataset and used only 5 fields

**Weak Entities**

21. **Public Tenders** - taken from “Awarded Public Tenders” dataset and used only 5 fields.
22. **Graduates** - taken from “Graduation Rates” dataset and used only 5 fields.

**B. Data Modelling**

The rough sketch and extended ERD images are included in the zip file. I user draw.io [4] to make the extended ERD. After making the ERD, I was not able to find any design issues that were to be solved. So, that is why I have only attached 2 ERD images in the zip file.

Akash Bharti
B00837970

## C. DDL (Data definition language) & DML (data manipulation language):

**Question 1.** Which business organization or type of business organization has the highest employees?

Ans. Construction business has the most employees in Halifax.

```
select * from businessestablishments where GeogName = "Halifax County" and DataGroups != "All Industries" order by `Total employees` desc;
```

Result Grid | Filter Rows: | Edit: | Export/Import: | Wrap Cell Content:

| ID | Company ID | GeogName | DataGroups | Year | Total employees |
|---|---|---|---|---|---|
| 318 | cnt1209 | Halifax County | Construction | 2010 | 2749 |
| 319 | cnt1209 | Halifax County | Construction | 2011 | 2748 |
| 539 | cnt1209 | Halifax County | Retail Trade | 2011 | 2328 |
| 538 | cnt1209 | Halifax County | Retail Trade | 2010 | 2308 |
| 464 | cnt1209 | Halifax County | Wholesale Trade | 2010 | 1244 |
| 465 | cnt1209 | Halifax County | Wholesale Trade | 2011 | 1133 |
| 613 | cnt1209 | Halifax County | Transportation and Warehousing | 2010 | 941 |
| 614 | cnt1209 | Halifax County | Transportation and Warehousing | 2011 | 936 |
| 390 | cnt1209 | Halifax County | Manufacturing | 2010 | 553 |
| 391 | cnt1209 | Halifax County | Manufacturing | 2011 | 498 |
| 624 | cnt1209 | Halifax County | Information and Cultural Industr... | 2011 | 481 |
| 95 | cnt1209 | Halifax County | Agriculture, Forestry, Fishing an... | 2011 | 321 |
| 94 | cnt1209 | Halifax County | Agriculture, Forestry, Fishing an... | 2010 | 287 |
| 170 | cnt1209 | Halifax County | Mining, Quarrying, and Oil and ... | 2010 | 89 |
| 171 | cnt1209 | Halifax County | Mining, Quarrying, and Oil and ... | 2011 | 85 |
| 244 | cnt1209 | Halifax County | Utilities | 2010 | 27 |
| 245 | cnt1209 | Halifax County | Utilities | 2011 | 23 |

businessestablishments 57

**Question 2.** Which area in the Halifax region has more schools?

Ans. Both Halifax city and Dartmouth city have the highest number of schools in the Halifax region.

```
select count(ID) as num, location from publicschools where `BOARD NAME` = "Halifax Regional Centre for Education" group by location order by num desc ;
```

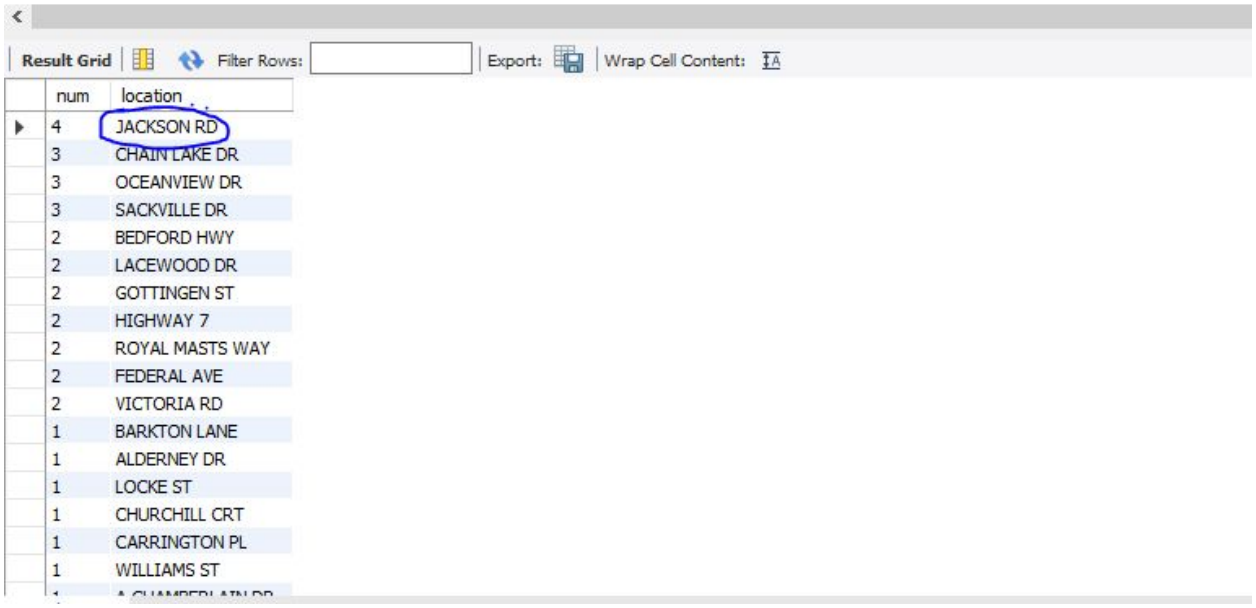
Result Grid | Filter Rows: | Export: | Wrap Cell Content:

| num | location |
|---|---|
| 32 | Dartmouth NS |
| 32 | Halifax NS |
| 10 | Lower Sackville NS |
| 6 | Bedford NS |
| 5 | Eastern Passage NS |
| 3 | Fall River NS |
| 3 | Upper Tantallon NS |
| 3 | Hammonds Plains NS |
| 3 | Beaver Bank NS |
| 3 | Middle Sackville NS |
| 2 | Hatchet Lake NS |
| 2 | Middle Musquodobo... |
| 2 | Cole Harbour NS |
| 2 | Porters Lake NS |
| 2 | Timberlea NS |
| 2 | Herring Cove NS |
| 1 | Hubley NS |

Akash Bharti
B00837970

**Question 3.** Which street in the Halifax region has the most reported crimes.
Ans: Jackson Road.

```
1 •  use mydb;
2
3 •  show tables;
4 •  select count(OBJECTID) as num, location  from crime group by location order by num desc;
```



Result Grid | Filter Rows: | Export: | Wrap Cell Content:

| num | location |
|---|---|
| 4 | JACKSON RD |
| 3 | CHAIN LAKE DR |
| 3 | OCEANVIEW DR |
| 3 | SACKVILLE DR |
| 2 | BEDFORD HWY |
| 2 | LACEWOOD DR |
| 2 | GOTTINGEN ST |
| 2 | HIGHWAY 7 |
| 2 | ROYAL MASTS WAY |
| 2 | FEDERAL AVE |
| 2 | VICTORIA RD |
| 1 | BARKTON LANE |
| 1 | ALDERNEY DR |
| 1 | LOCKE ST |
| 1 | CHURCHILL CRT |
| 1 | CARRINGTON PL |
| 1 | WILLIAMS ST |

**Normalization**

While cleaning the data and removing unwanted columns, I made sure that all the datasets have primary key consisting of a single column which makes sure that my data is in second normal form. All while cleaning the data, I ensured that no transitive dependencies occur within the dataset. However, there were 1 or 2 instances where I had to remove the transitive dependencies and make the table in 3NF. I have included 2 copies of the database - one copy containing tables before normalization and the other one having tables after normalization.

Akash Bharti
B00837970

**References:**

1. [1] “Nova Scotia Government - Open Data Portal | Open Data | Nova Scotia,” *Socrata*, 2018. [Online]. Available: https://data.novascotia.ca/. [Accessed: 9-Feb-2020].

2. [2] “Halifax Regional Municipality,” *Arcgis.com*, 2017. [Online]. Available: http://catalogue-hrm.opendata.arcgis.com/. [Accessed: 9-Feb-2020].

3. [3] Real Python, “Reading and Writing CSV Files in Python,” *Realpython.com*, 16-Jul-2018. [Online]. Available: https://realpython.com/python-csv/. [Accessed: 9-Feb-2020].

4. [4] draw.io - free flowchart maker and diagrams online, “Flowchart Maker & Online Diagram Software,” *Draw.io*, 2020. [Online]. Available: https://www.draw.io/. [Accessed: 9-Feb-2020].