Assignment - 4
CSCI 5408 - Data Management, Warehousing, Analytics

Gitlab Repository: https://git.cs.dal.ca/adimurthy/5408-assignment-4

1.1 Business Intelligence Reporting using Cognos

1.2 Dataset: https://www.kaggle.com/mickey1968/individual- company-sales-data [1]

1.3 Facts and Dimensions explanation:

1. Measurable Field: The candidates for facts are house value and flag (whether the product was bought or not by the customer). Even though it is recommended to consider facts with numerical values, the flag column is a better choice among the two as it's a quantifiable result compared to the housing value and represents an end result or a successful action item that can be further evaluated with respect to other dimensions.

2. Dimension(s) can be broadly classified as Customer Dimension/details apart from the flag, which would result in a Snow Flake schema, where the customer dimension is further split into more dimensions such as location, housing, and income.

3. On the other hand, a star schema is also a valid usage. Dimensions: Customer (occupation, marital status, has children, age, gender, education, and much more), Location (region), housing (owner of the house, mortgage, the value of the current house, probability of owning a car), and income (family income) are the individual dimensions.

4. To summarize, based on various dimensions, we can now conclude the possible factors that influenced the purchase of the product. However, the dataset lacks details of the product to make a fair assumption.

1.4 Cleaning the dataset:

1.4.1 Common:

1. Capitalize enums, i.e., fields with low cardinality values (uniqueness of data values).
2. Remove numeric prefixes.
3. Rename column names with complete words. Example: fam_income to family_income.

1.4.2 Specific Fields:

1. Education**:** Covert the data into enums as the cardinality is high. GRAD, BACH, SOME_COLLEGE, HS, BELOW_HS.

2. Age: Assuming that <=55 translates to the range 45 to 55, similarly, <=45 represents 35 to 45. Hence, converting the age column to an enum, we have UNKNOWN, 25_35, 35_45, 45_55, 55_65, GREATER_THAN_65. Optionally another approach is to have min_age and max_age as two different columns.

3. Marriage: SINGLE and MARRIED - Capitalized to emphasize the usage of constants.

4. Mortgage: Remove numerical prefixes: LOW, MEDIUM, HIGH.

5. House owner: OWNER and RENTER.

1.4.3 Snippet of cleaned fact and dimension tables:

Table 1: Customer Dimension Table

id	occupation	has_children	marital_status	customer_psyc	ecommerce_exper	age	education	gender
1	Professional	U	MARRIED	A	Y	UNKNOWN	BACHELORS	M
2	Blue Collar	Y	SINGLE	A	N	UNKNOWN	LESS_THAN_HS	M
3	Professional	U	MARRIED	A	N	UNKNOWN	SOME_COLLEGE	M
4	Professional	U		A	N	UNKNOWN	SOME_COLLEGE	U
5	Sales/Service	N		A	Y	UNKNOWN	HS	U
6	Professional	U	SINGLE	A	Y	UNKNOWN	SOME_COLLEGE	M
7	Professional	U	MARRIED	A	N	UNKNOWN	SOME_COLLEGE	F
8	Professional	U	MARRIED	A	Y	UNKNOWN	HS	F
9	Professional	Y		A	Y	UNKNOWN	BACHELORS	F
10	Professional	U		A	N	UNKNOWN		M
11	Professional	N	SINGLE	A	N	UNKNOWN	SOME_COLLEGE	M
12	Sales/Service	U	SINGLE	A	Y	UNKNOWN	SOME_COLLEGE	M
13	Sales/Service	U		A	N	UNKNOWN	LESS_THAN_HS	F
14	Sales/Service	N	MARRIED	A	Y	UNKNOWN	HS	M
15	Sales/Service	N	SINGLE	A	N	UNKNOWN	BACHELORS	M
16	Professional	U		A	Y	UNKNOWN	SOME_COLLEGE	F
17	Blue Collar	U	MARRIED	A	Y	UNKNOWN	HS	M
18	Blue Collar	Y		A	N	UNKNOWN	LESS_THAN_HS	M
19	Sales/Service	U		A	N	UNKNOWN	HS	M
20	Professional	Y		A	N	UNKNOWN	HS	M
21	Sales/Service	U	MARRIED	A	N	UNKNOWN	GRADUATE	M
22	Professional	U	SINGLE	A	Y	UNKNOWN	GRADUATE	M
23	Sales/Service	N	SINGLE	A	Y	UNKNOWN	SOME_COLLEGE	F
24	Professional	N	SINGLE	A	N	UNKNOWN	LESS_THAN_HS	M
25	Sales/Service	N	SINGLE	A	N	UNKNOWN		M

Table 2: Housing and Income Dimensions Table

id	house_owner	mortgage	value	car_probability
1		LOW	2000000	5
2		LOW	0	7
3	RENTER	LOW	0	8
4	OWNER	Med	239560	2
5	RENTER	LOW	0	9
6		LOW	280173	1
7	RENTER	LOW	809434	2
8		LOW	242455	6
9	RENTER	LOW	287119	1
10		LOW	631764	5
11	RENTER	LOW	321634	4
12		LOW	452786	9
13	RENTER	LOW	0	9
14		LOW	0	8
15	RENTER	LOW	0	8
16	OWNER	LOW	159040	1
17	RENTER	LOW	0	5
18		LOW	0	3
19	RENTER	LOW	0	3
20	RENTER	LOW	682390	5
21	RENTER	LOW	483830	6
22	OWNER	LOW	474150	8
23	RENTER	LOW	0	9
24		LOW	2262622	6
25	OWNER	LOW	319086	8

id	family_income
1	C
2	D
3	E
4	E
5	B
6	E
7	H
8	E
9	F
10	I
11	K
12	H
13	D
14	C
15	A
16	E
17	F
18	E
19	D
20	G
21	K
22	C
23	A
24	L
25	F

Table 3: Sales Fact Table and Location Dimension Table

is_product_purchased	id	region_id	income_id	expense_id
Y	1	1	1	1
N	2	3	2	2
N	3	3	3	3
N	4	4	4	4
Y	5	4	5	5
Y	6	4	6	6
N	7	4	7	7
N	8	3	8	8
Y	9	3	9	9
Y	10	4	10	10
Y	11	3	11	11
N	12	4	12	12
N	13	3	13	13
Y	14	3	14	14
Y	15	4	15	15
N	16	4	16	16
N	17	4	17	17
N	18	3	18	18
N	19	1	19	19
Y	20	1	20	20
Y	21	3	21	21
Y	22	4	22	22
Y	23	4	23	23
Y	24	1	24	24
N	25	4	25	25

id	region
1	Northeast
2	Midwest
3	South
4	West
5	Rest

1.5 Fact and Dimension tables (Imported to Cognos) [2]:
Fact and Dimension tables imported in IBM Cognos:

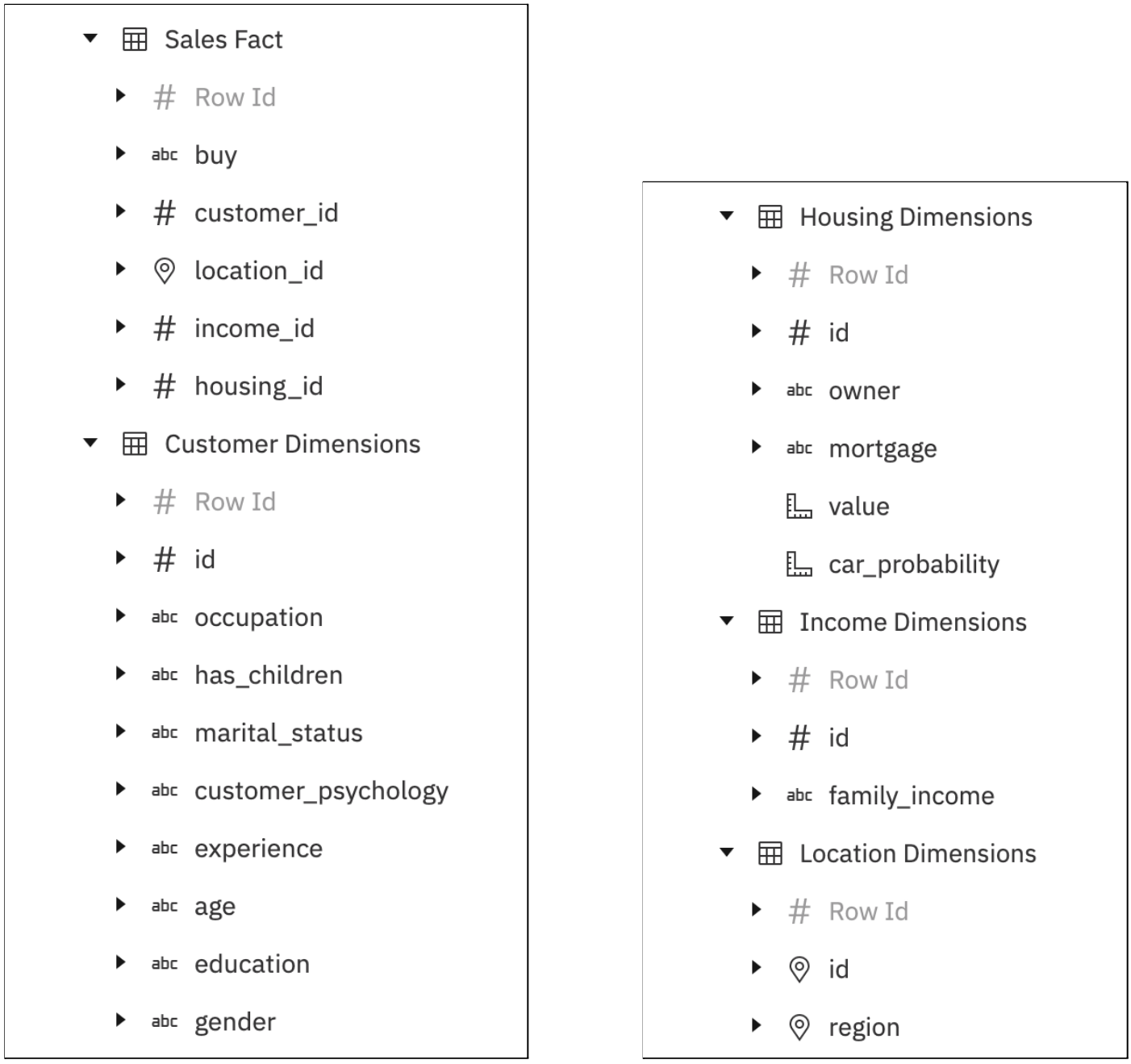

Figure 1: Fact and Dimensions tables for Sales data

1.6 Star Schema (Imported to Cognos) [2]:

As mentioned earlier, using a star schema could be a better option to keep extensibility in mind, where each of the dimensions such as the customer personal information, location, income, and housing details can be further extended and can be further extended to smaller dimensions, thereby reducing the complexity of relations between dimensions.

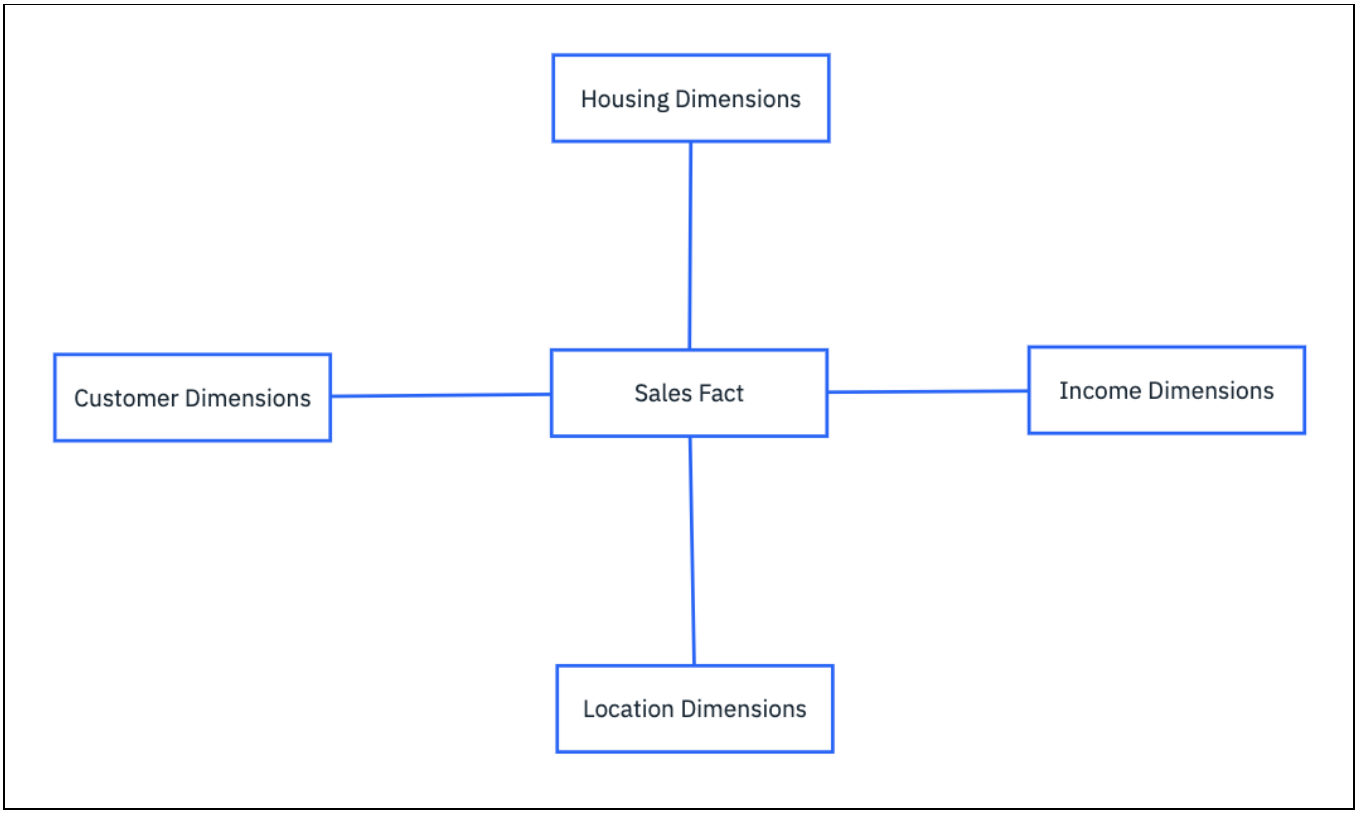

Figure 2: Star Schema for sales data

1.7 Snow flake Schema (Imported to Cognos):

The data set clearly lacks details. Assuming that the number of columns in the dataset will not drastically grow, it is fair to assume that there is only one fact and one dimension table. All fields except the fact represent the customer details. With this approach, the schema would be a snowflake, where the customer dimension is further split into smaller dimensions, i.e., location, housing, and income.

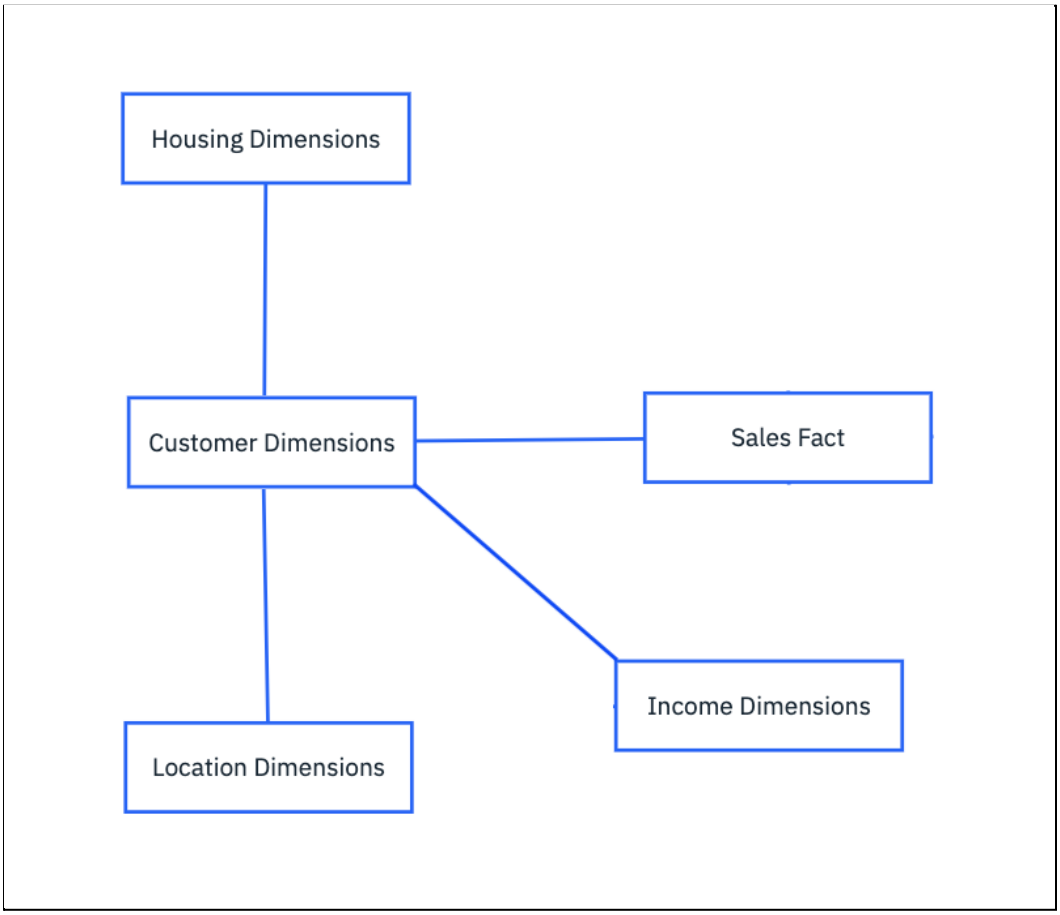

Figure 3: Snow flake Schema for sales data

1.8 Conclusion

To conclude, for this example of the dataset and considering that the dataset still lacks information and has a strong potential to grow. It's is better to use a star schema over a snow flake schema.

1.9 Analysis

Since the fact table here is the deciding factor whether the customer purchased the product or not, every other field across dimensions has a crucial role; even before comparing those fields with the fact, it's essential to validate these fields with others to understand it's volatility and derive conclusions.

In the below example, the housing value, which is the only numeric representation of a field, is first compared with location and mortgage to understand its variance with other fields and finally compared with the flag of whether the customer purchased the product or not.

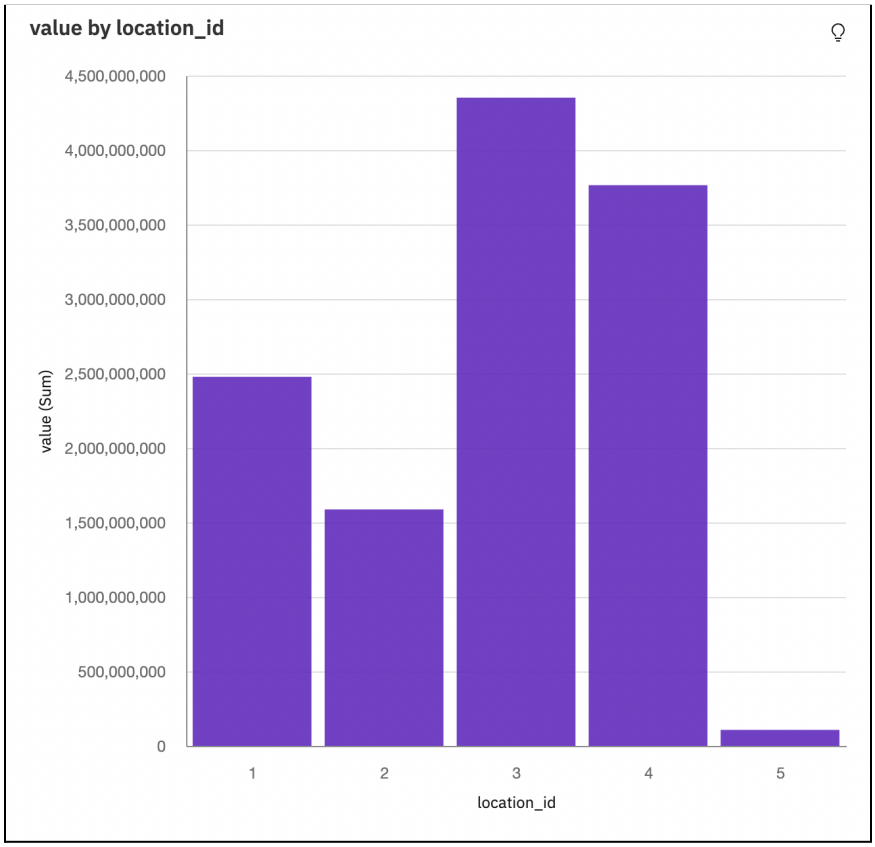

Figure 4: Housing Value by location

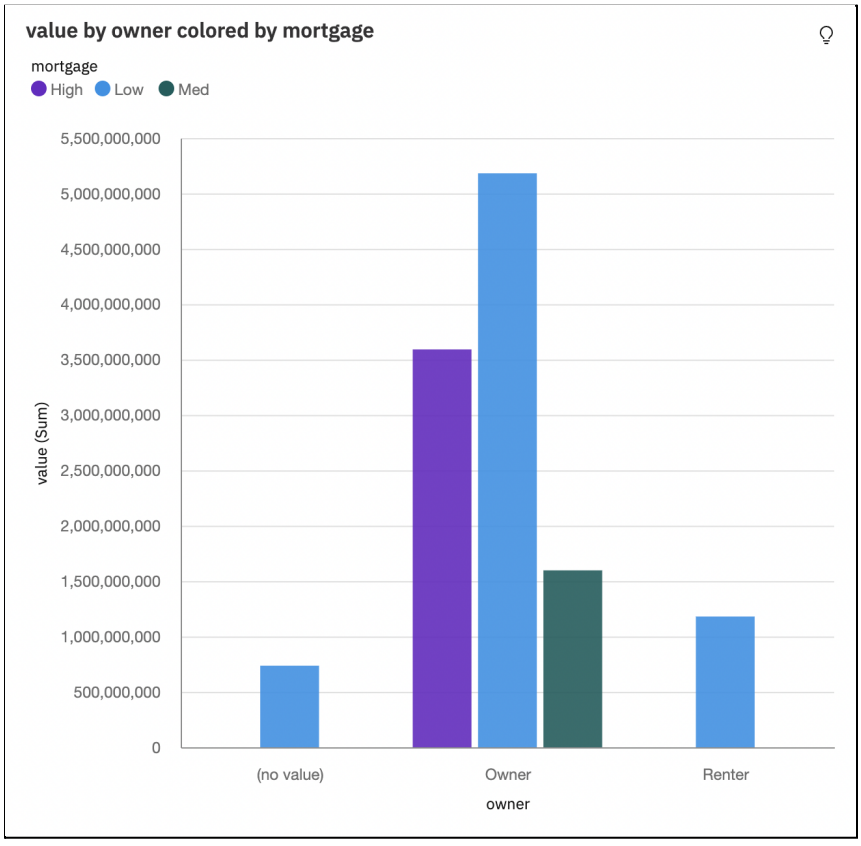

Figure 5: Housing value by Owner for Mortgage

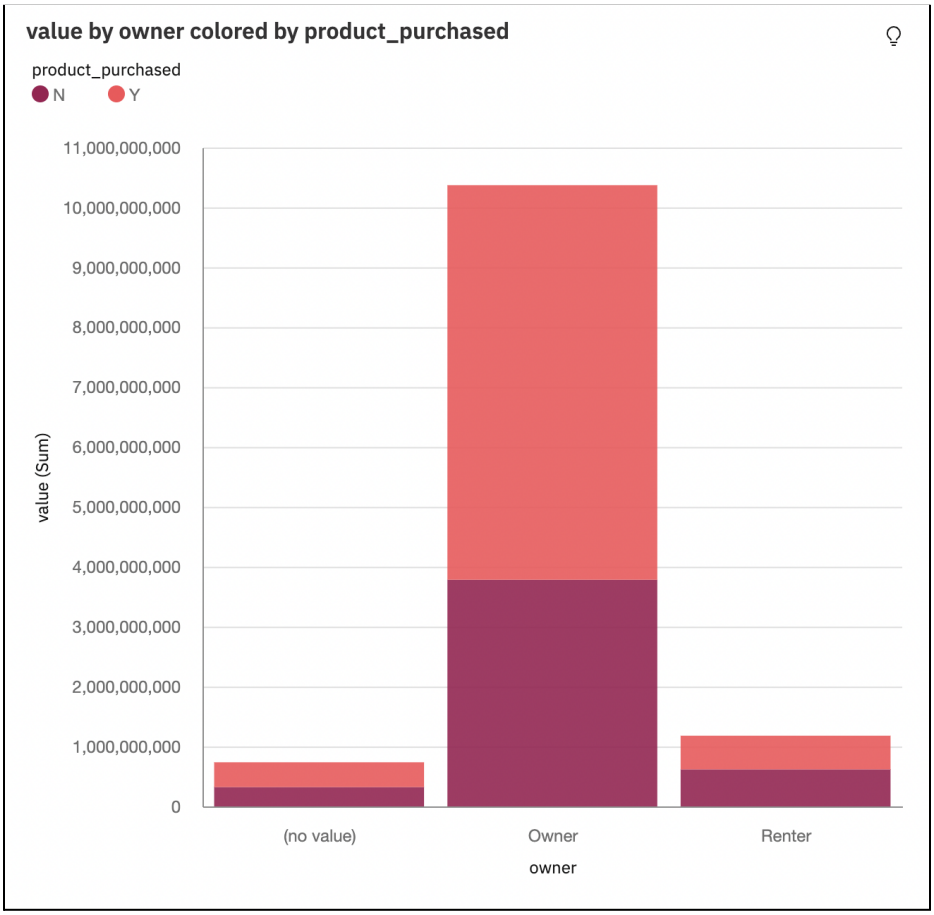

Figure 6: Housing value by Owner for product purchased or not

2.1 Sentiment Analysis: Bag of words

2.2 Steps performed:

1. Gather the tweets from MongoDB (Database: ProcessedDb, Collection: tweets).
2. Get positive and negative word count for every tweet. Using the list of positive and negative words from: https://gist.github.com/mkulakowski2/4289441 [3] and https://gist.github.com/mkulakowski2/4289437 [4]
3. Determine the sentiment polarity (POSITIVE, NEGATIVE, or NEUTRAL) from the number of positive and negative words of each tweet.
4. Output the Tweet, Match words, and Polarity in a tabular format using java.util.Formatter.

2.3 Gitlab Repository:
https://git.cs.dal.ca/adimurthy/5408-assignment-4/-/blob/main/src/main/java/com/example/SentimentAnalysis.java

2.4 Result Snippet (text file included in the zip folder):

Tweet	Match	Polarity
weather data 432 PM 351F Humidity 99 Wind 78 WNW	[]	NEUTRAL
MagisterLatro We get snowstorms sometimes and the	[cold, enough, kill, nice]	NEUTRAL
F1 BrazilGP WEATHER UPDATE INTERLAGOS Moisture flo	[cloudy, breaks, good, warmer]	NEUTRAL
Forecast map for this evening Get Maps and Radars	[]	NEUTRAL
TheLeftistMom Could simply be the weather Wild flu	[wild]	NEGATIVE
NWSAtlanta Are you going to tone the Freeze Watch	[like, freeze]	POSITIVE
RT RyuichiShigaki 1113 1111	[]	NEUTRAL
Perfect red weather today Grey and raining	[perfect]	POSITIVE
RT catbirdandteaco sweater weather huggies are bei	[]	NEUTRAL
Day 12 While Im grateful to live in a place with n	[grateful, perfect, fall, reasonable]	POSITIVE
RT khilanii if the weather keeps this up thanksgiv	[]	NEUTRAL
3050 wild hogs strike again	[hogs, wild, strike]	NEGATIVE
RT FOX55FortWayne Its a chilly evening Temperature	[spotty, chilly]	NEGATIVE
isitokthat the UK media has reported COP26Glasgow	[bad]	NEGATIVE
RT AlviaAlcedo Last days of May and no summer feel	[cold]	NEGATIVE
The GMA woman Amy robot the woman might freeze Arc	[wow, mercy, freeze]	POSITIVE
The weather today in Portland	[]	NEUTRAL
Great weather NYC you never know how it could be h	[great]	POSITIVE
RT RyuichiShigaki 1113 1111	[]	NEUTRAL
RT pieceofcake28 guys heartbreak weather by niall	[]	NEUTRAL
RT MetOfficeSci Our global temperature dataviz by	[rapid]	POSITIVE
RT BBCBreakfast The Uists Islands are on the front	[unpredictable]	NEGATIVE
RT MaryBellavita This cloudy weather has me wantin	[cloudy]	NEGATIVE
RT GuisadoPepe Segn estudios cientficos parece ser	[]	NEUTRAL
WEATHER ALERT DANDRUFF WARNING FOR A JAR OF METAMU	[warning]	NEGATIVE
RT RachelNotley The pandemic isnt over We must get	[vigilant]	POSITIVE
weatherab Sending thoughts and prayers Hope you fe	[better]	POSITIVE
RT mansdnh Reminder The ChooseLoveNH bus tour come	[good]	POSITIVE
I just wanna say bless you to the rainy UK gods fo	[bless]	POSITIVE
i swear everytime i do that sweater weather diet i	[lose]	NEGATIVE
ricandess Bitch this weather worse then these nigg	[bitch, worse]	NEGATIVE
Its shocking I tell ya Shocking what potholes and	[well, shocking, smooth, damaged]	NEGATIVE
emilybdepalma Idk about you but personally the wea	[strong, depression, worse, helping]	NEUTRAL
Such a joy to teach how to build trust as a leader	[trust, wonderful, perfect, joy, leading]	POSITIVE
Theres now a weather advisory for heavy rain amp t	[bad]	NEGATIVE
This 90 degree weather is disrespectful to say the	[disrespectful]	NEGATIVE
I really was not ready for this weather change	[ready]	POSITIVE

Figure 1: Polarity of each tweet.

Note: The tweet text is a sub-string of the first 50 characters for better tabular representation.

3.1 Semantic Analysis

3.2 Steps performed:

1. Gather ~2000 (N) news articles from MongoDB (Database: ReutersDb, Collection: news) [5-7].
2. Search for the number of documents with the words: Canada, Moncton, and Toronto. (Note: keywords and news articles are in lowercase)
3. Calculate Total Documents (N = 2000) / Number of documents term appeared (df) and Log(Base 10) (N/df).
4. Find relative Frequency for each news article for the keyword: Canada, using Total Words (m) / Frequency (f).

3.3 Gitlab Repository:
https://git.cs.dal.ca/adimurthy/5408-assignment-4/-/blob/main/src/main/java/com/example/SemanticAnalysis.java

3.4 Result Snippet (text file included in the zip folder):

N: 2000

Search Query	df	N/df	Log10(N/df)
canada	67	29.850746268656717	1.4749551929631548
toronto	43	46.51162790697674	1.6675615400843946

Figure 2: Number of documents for the search keywords

Article #	Text	Total Words	Frequency
1411	1411	182	3
133	133	296	1
1541	1541	87	1
1926	1926	104	1
137	137	212	1
1673	1673	220	1
268	268	189	3
781	781	438	2
403	403	118	2
531	531	312	8
21	21	809	1
792	792	489	1
282	282	211	1
922	922	313	1
1564	1564	168	1
542	542	185	1
672	672	491	1
417	417	105	1
673	673	196	1
1699	1699	97	1
1319	1319	44	2
553	553	335	4
431	431	84	1
559	559	179	1
1712	1712	220	1
945	945	423	2
59	59	234	2
60	60	168	1
444	444	349	4
700	700	268	1
956	956	171	1
1212	1212	193	1
1857	1857	479	1
834	834	915	1
1986	1986	383	1
1732	1732	145	2
1478	1478	72	2
1862	1862	273	4
1990	1990	12	1

Figure 3: Relative frequency for all news articles for the keyword "Canada"

```
------------------------------------------------------------------
Highest relative frequency: 0.083333 for the news
article: <TITLE>BACHE SECURITIES CANADA BUYS TORONTO EXCHANGE SEAT FOR 301,000 DLRS</TITLE>Blah blah blah.
```

Figure 4: News article with highest relative frequency

4.1 References

[1] Kaggle Company Sales Dataset. Accessed on: Nov 26, 2021. [Online]. Available: https://www.kaggle.com/mickey1968/individual- company-sales-data

[2] IBM Cognos Analytics. Accessed on: Nov 26, 2021. [Online]. Available: https://www.ibm.com/products/cognos-analytics

[3] Negative words. Accessed on: Nov 26, 2021. [Online]. Available: https://gist.github.com/mkulakowski2/4289441

[4] Positive words. Accessed on: Nov 26, 2021. [Online]. Available: https://gist.github.com/mkulakowski2/4289437

[5] MongoDB java client. Accessed on: Nov 6, 2021. [Online]. Available: https://docs.mongodb.com/drivers/java-drivers/

[6] Maven MongoDB java client. Accessed on: Nov 6, 2021. [Online]. Available: https://mvnrepository.com/artifact/org.mongodb/mongo-java-driver

[7] MongoDB Compass. Accessed on: Nov 6, 2021. [Online]. Available: https://www.mongodb.com/products/compass