DALHOUSIE UNIVERSITY

CSCI 5408 – Data Management, Warehousing, Analytics

Assignment – 4

Prashit Patel - B00896717

Gitlab Repository
https://git.cs.dal.ca/pppatel/csci-5408-f2021-b00896717-prashit_patel.git

Task 1: Business Intelligence reporting using Cognos

Facts and Dimensions

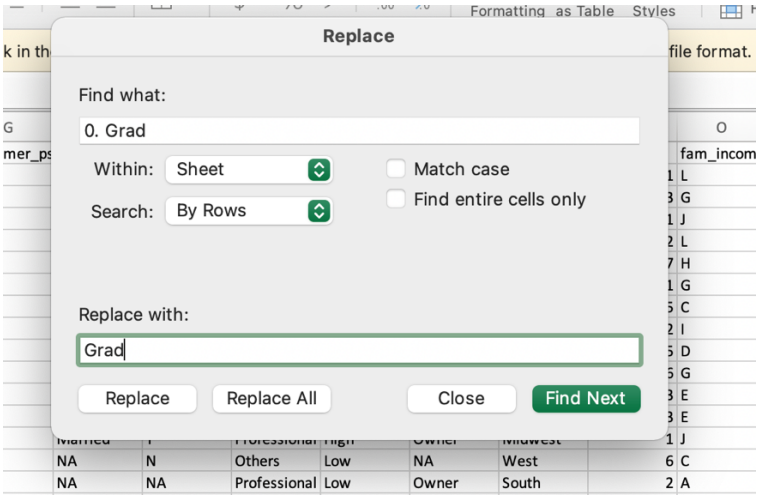
- Sale preferences can be calculated based on different dimensions such as gender, region, age, ownership of the house, occupation, education, product bought or not and online shopping experience. We can gather these sales preferences, study patterns and improve total sales based on the analysis.

Dimensions

1. **Product bought or not**: Y, N - field values describe if the customer bought the product or not. Y specifies that the customer bought the product and N specifies that the customer did not buy the product.
2. **Gender**: Male, Female - field values describe if the customer was a male or a female.
3. **Region**: Northeast, Midwest, West, South, Rest – field values describe the region of the customer.
4. **Age**: Under 25, Under 35, Under 45, Under 55, Under 65, Over 65 – fields describe the age range of the customer.
5. **Ownership of house**: Renter, Owner – describe the ownership of the house of the customer.
6. **Occupation**: Professional, Blue Collar, Sales/Service, Others, Retired, Farm – field values describe the occupation of the customer.
7. **Online shopping experience**: Y, N – field values describe if the customer had any previous experience of online shopping or not. Yes specifies that the customer had a previous online experience while N specifies no previous experience.
8. **Education**: Grad, Bach, College, HS, Less than HS – field values specifies the education of the customer.

Cleaning and Formatting

- Spreadsheet Find and Replace feature was used to clean and format data.



Fields transformed:

- gender – U was converted to NA.

- education – 0. <HS was converted to Less than HS, 1. HS was converted to HS, 2. Some College was converted to College, 3. Bach was converted to Bach, 4. Grad was converted to Grad, blanks were converted to NA.

- age – 1_Unk was converted to NA, 2_<=25 was converted to Under 25, 3_<=35 was converted to Under 35, 4<=45 was converted to Under 45, 5<=55 was converted to Under 55, 6_<=65 was converted to Under 65, 7_>65 was converted to Over 65.

- customer psy – U was converted to NA.

- Marriage – blanks were converted to NA.

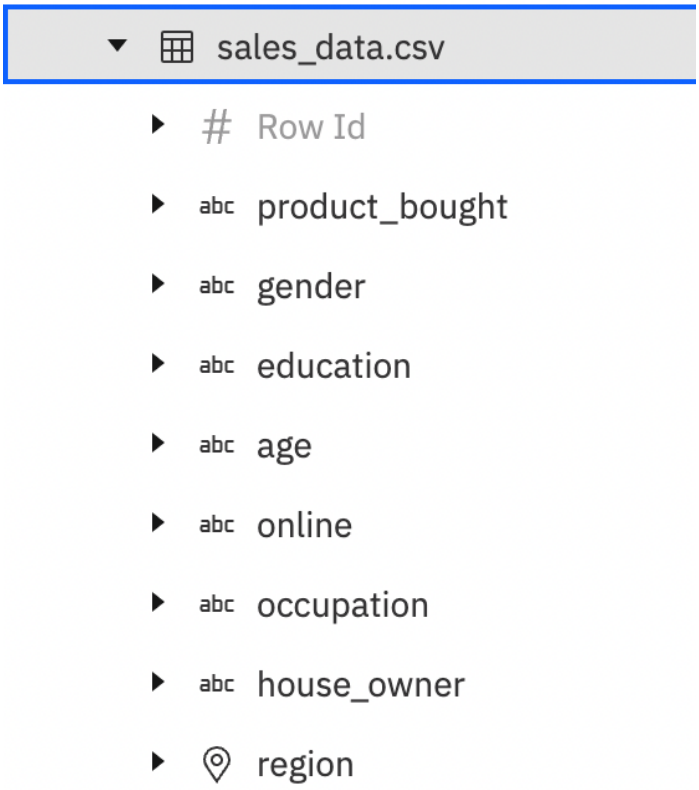
- child – 0 was converted to N, U was converted to NA.

- mortgage – 1Low was converted to Low, 2Med was converted to Med, 3High was converted to High.

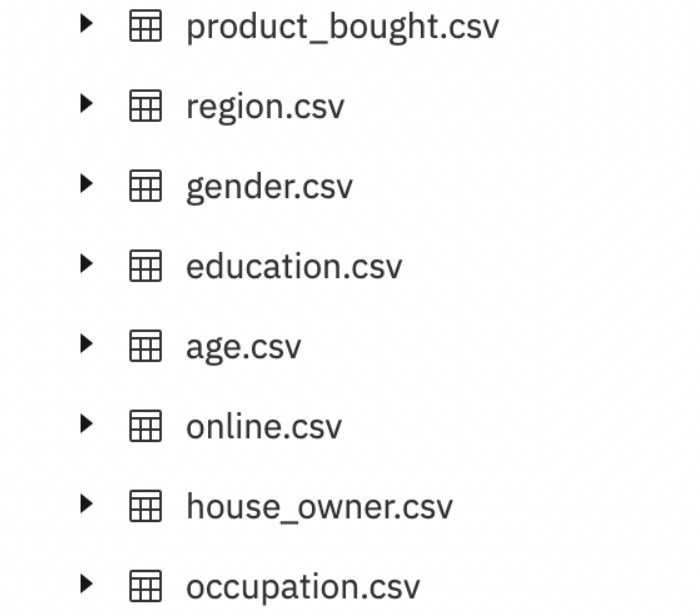
- house_owner - blanks were converted to NA.

- fam_income – U was converted to NA.

Fact Table:

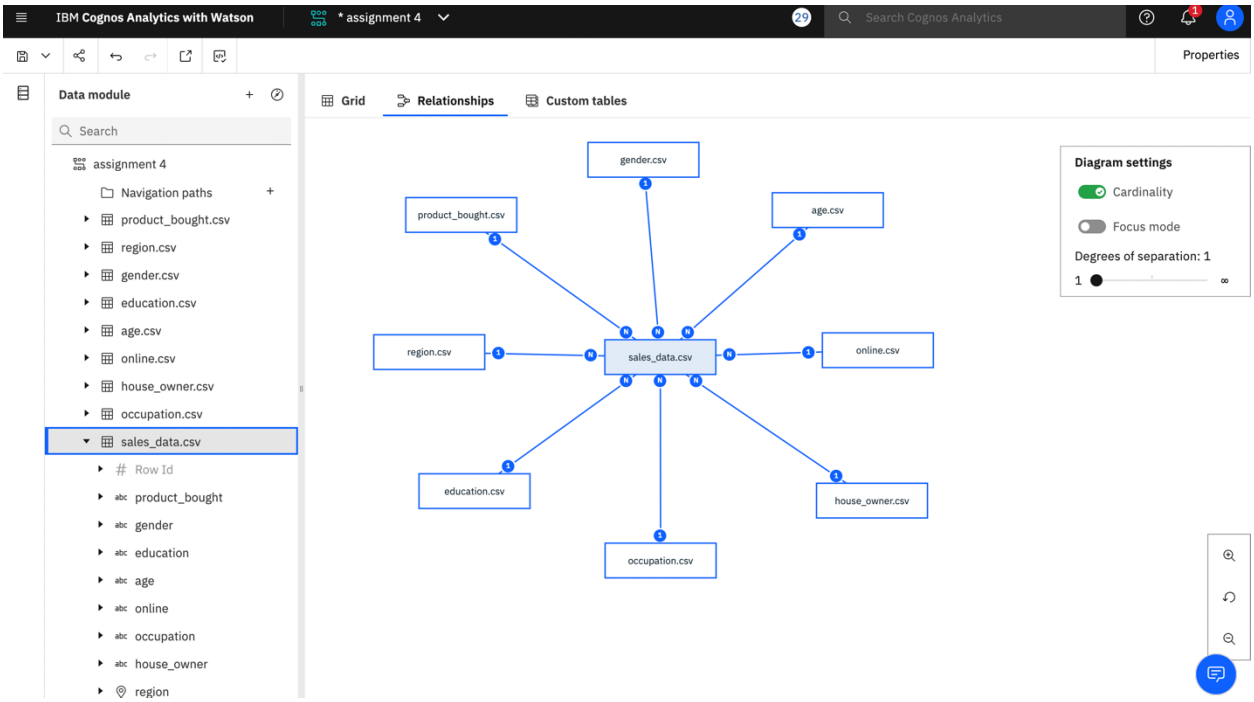


Dimensions Table:



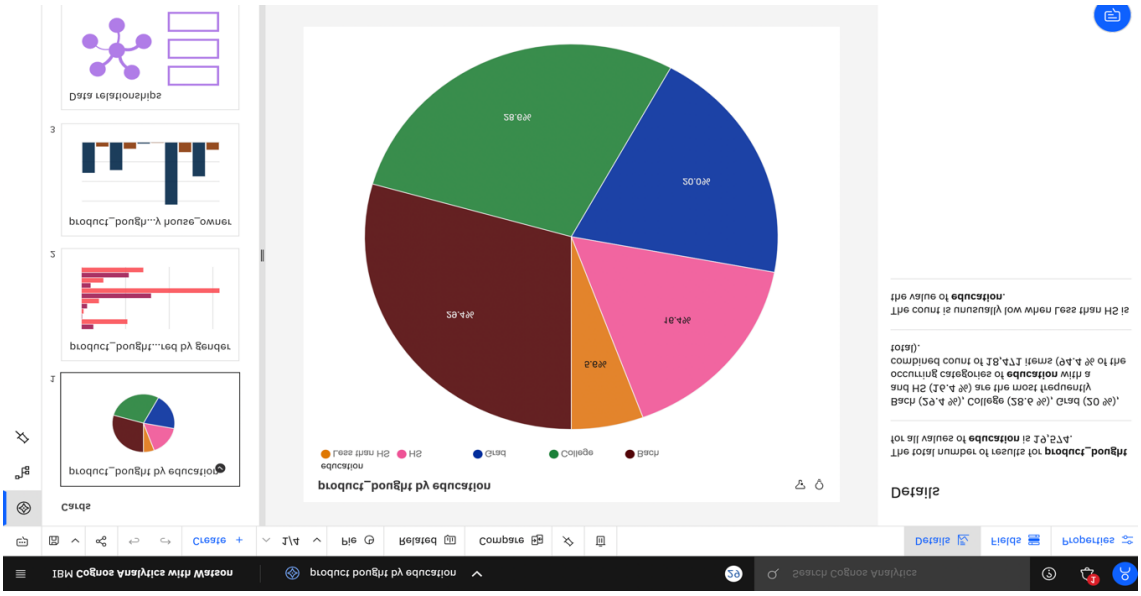
Star Schema

The above facts and dimensions were used to create a star schema. All the CSV files were imported as tables. Then, many(N) to 1 relationship was created between imported facts and tables.

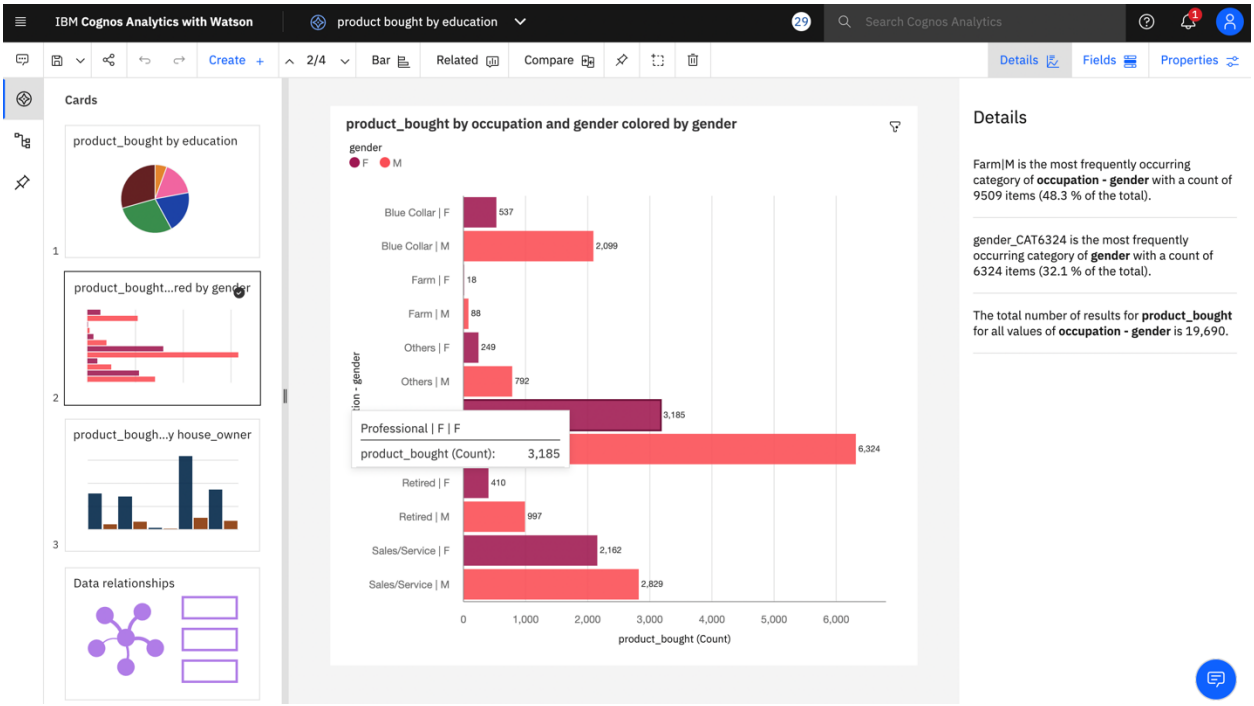


Visual Analysis

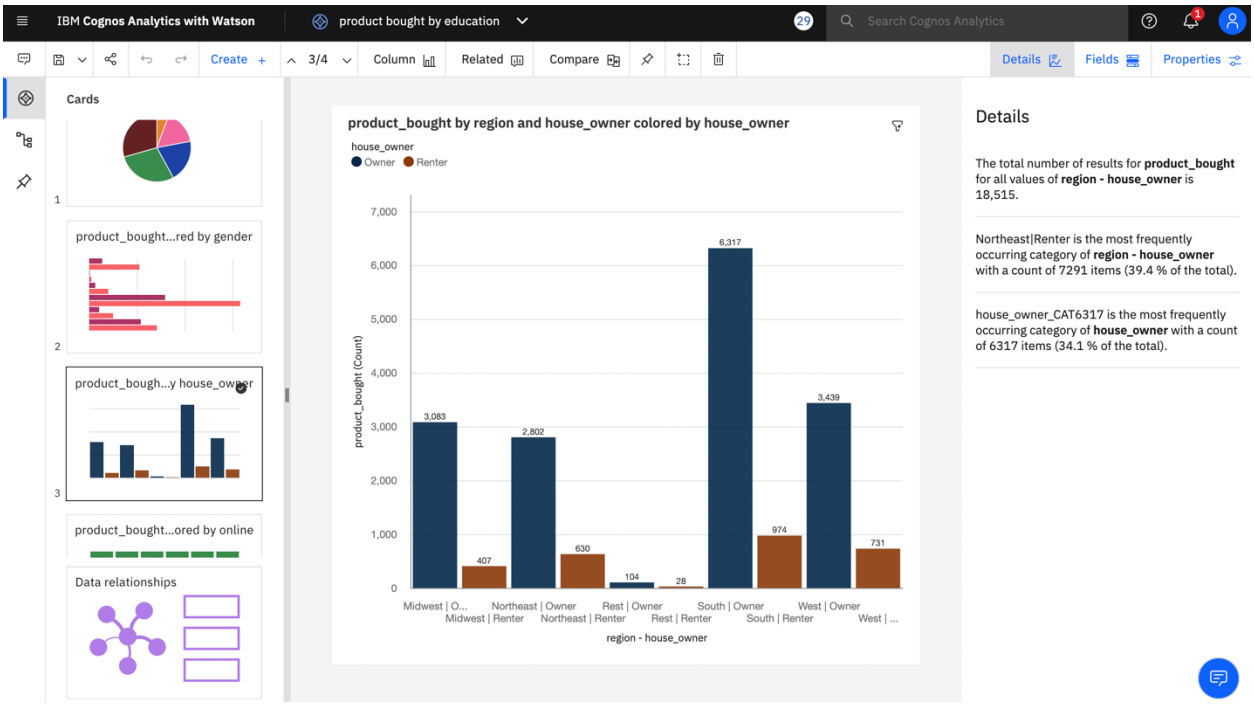
1. product bought by education: Percentages can be obtained for the type of education background of the customers that bought the product.



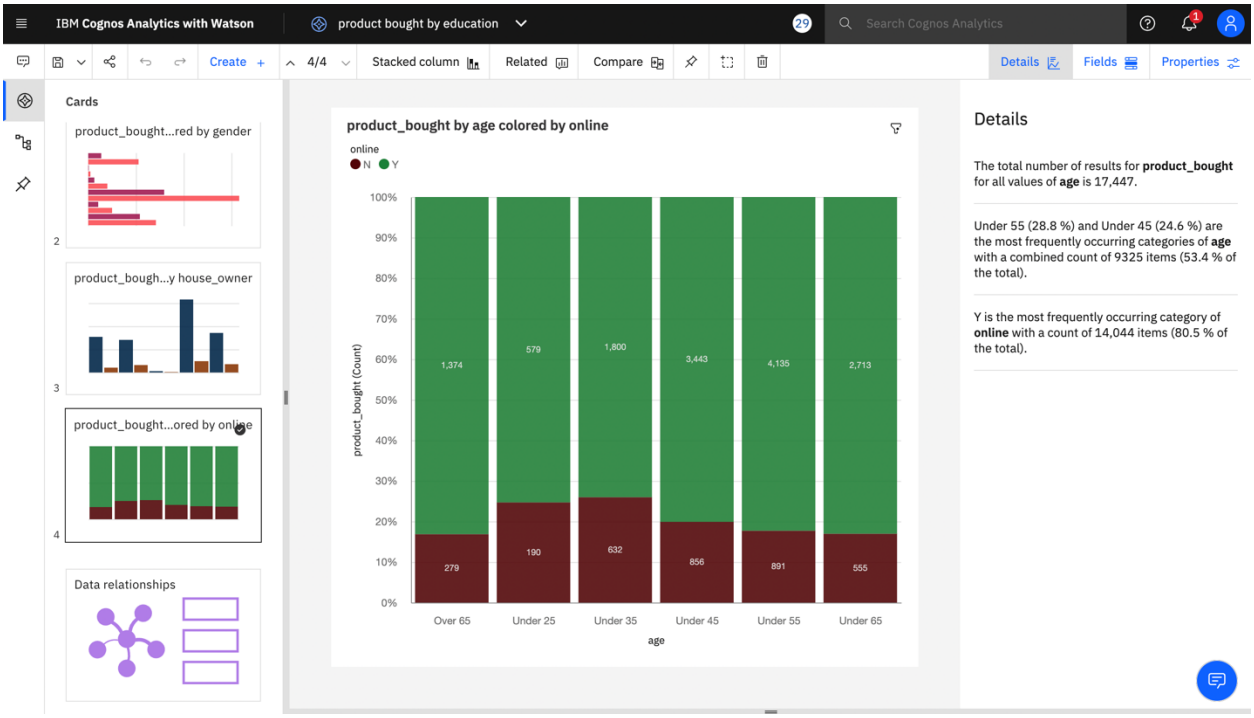
2. product bought by occupation and gender: The total number of customers that bought the product based on the occupation and gender.



3. product bought by region and house_owner: Total number of products bought according to regions and house ownership type of the customer.

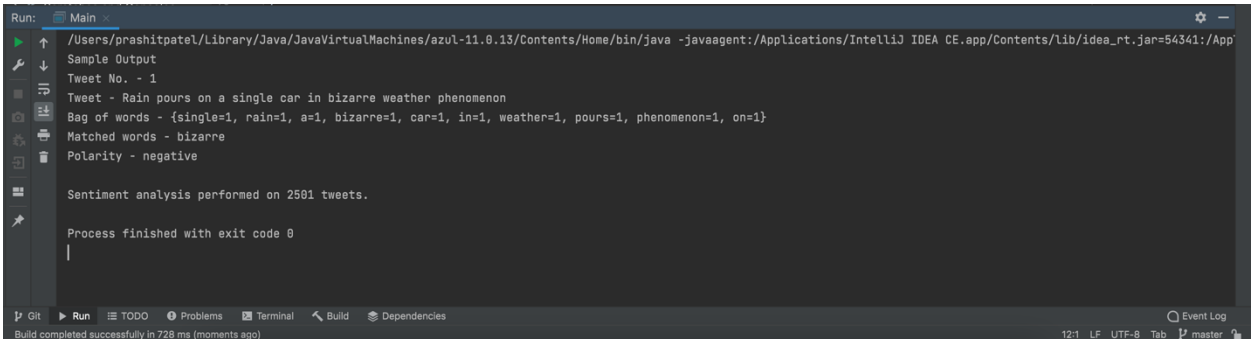


4. product bought by age and online shopping experience: Percentage and the total number of customers that bought the product based on age group and previous online experience amongst them.

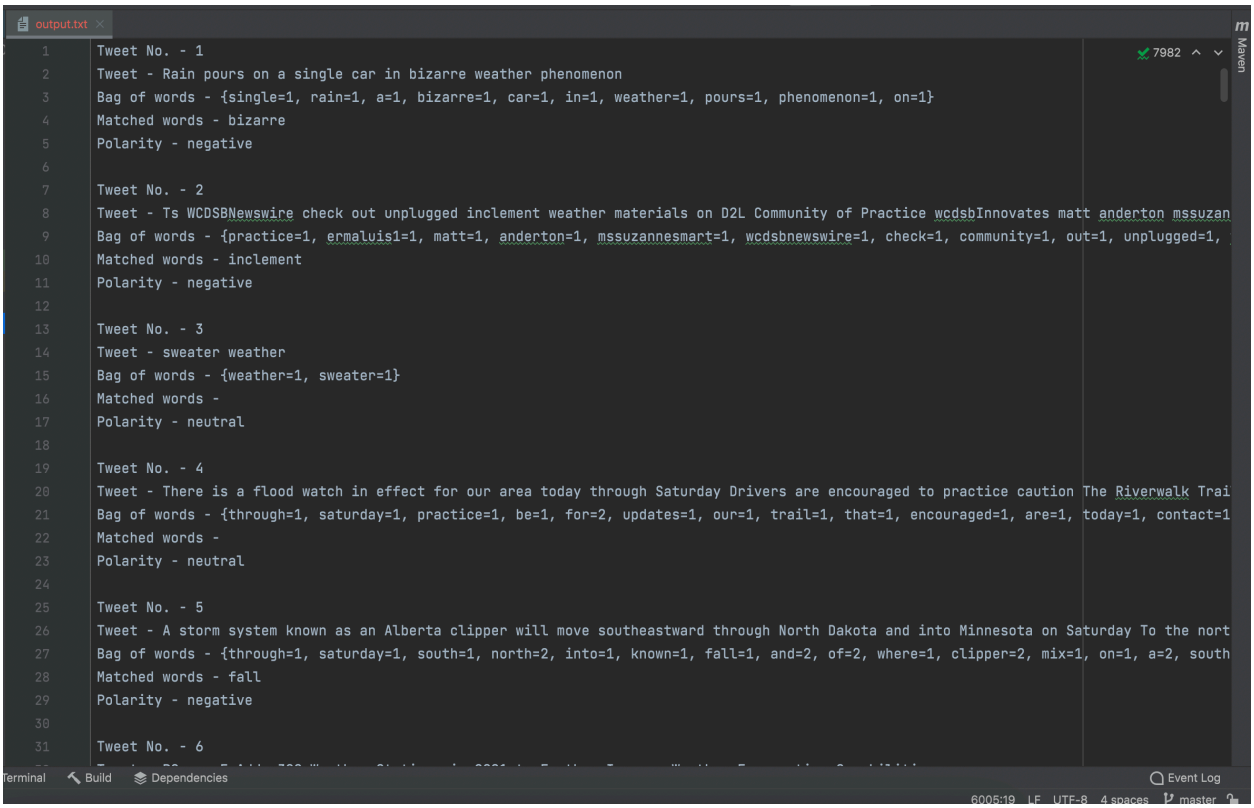


Task 2: Sentiment Analysis

Screenshot of sample output:



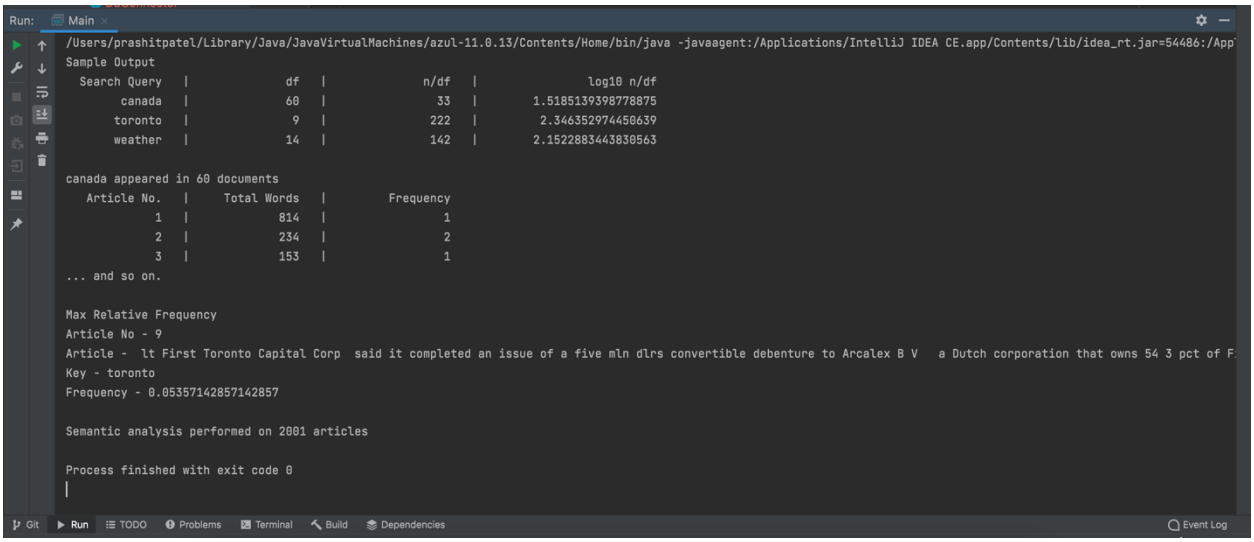
The complete output of the program can be found in the output file in program directory on GitLab.



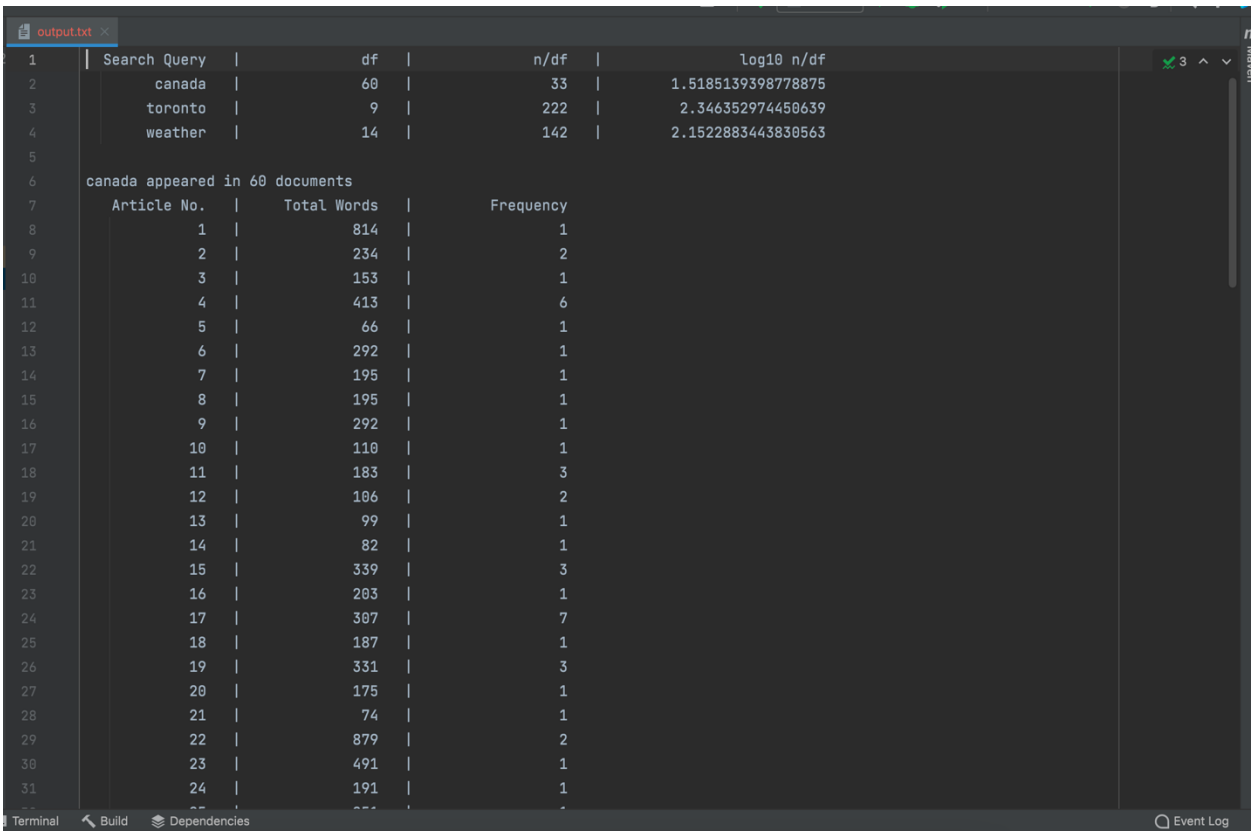
Task 3: Semantic Analysis

- Moncton was found in 0 articles. So, weather keyword is considered in place of Moncton.

Screenshot of sample output:



The complete output of the program can be found in the output file in program directory on GitLab.



References

[1] “Positive words”, Ptrckprry [Online].
Available: https://ptrckprry.com/course/ssd/data/positive-words.txt
[Accessed on 24th November 2021]

[2] “Negative words,” Github [Online].
Available: https://gist.github.com/mkulakowski2/4289441 [Accessed on 24th November 2021]

[3] “Print tables using formatter,” Javatpoint [Online].
Available: https://www.javatpoint.com/how-to-print-table-in-java-using-formatter [Accessed on 24th November 2021]

[4] “Individual Company Sales Data,” Kaggle [Online]
Available: https://www.kaggle.com/mickey1968/individual-company-sales-data [Accessed on 25h November 2021]

[5] “Cognos Analytics”, IBM [Online].
Available: https://www.ibm.com/products/cognos-analytics [Accessed on 25th November 2021]

[6] Minqing Hu and Bing Liu. "Mining and Summarizing Customer Reviews." Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004), Aug 22-25, 2004, Seattle, Washington, USA [Accessed on 24th November 2021]