Quiz 2 (KsqlDB)

DADS6005

Data Streaming and Real-Time Analytics

Authors
Korawee Peerasantikul 6420422007
Kruawun Jankaew 6420422016
Pimchayanan Kusontramas 6420422018
Sorawit Sinlapanurak 6420422020
Napat Phongvichian 6420422022

May 5, 2023

Contents

1 Introduction

This quiz is part of the DADS6005 (Data Streaming and Real-Time Analytics) course at National Institute of Development Administration (NIDA).

1.1 Objectives

The main objective of this quiz is to design and implement "data-streaming and real-time analytics" system in addition to real-time data cleansing and analytics.

2 Datasets

We use a dataset of Food choices College students' food and cooking preferences from kaggle provided by Assistant Prof. Ekarat Rattagan.

3 Methodology

1. Use Python to write the data source line-by-line from the CSV to SQL database (DB) (source part). The writing duration (d) must be random and between 0 and 2 seconds.
2. Connect the SQL DB to a Kafka cluster via topic1.
3. Clean the data in topic1 using ksqlDB and then submit it to topic2.
4. Analyze the data in topic2 using ksqlDB and then submit it to topic3.
5. Analyze the data in topic2 using ksqlDB and then submit it to topic3.
6. Analyze the data in topic2 using ksqlDB and then submit it to topic3.
7. Visualize the real-time data from topic3 using tools such as Plotly, etc.
8. Conduct analytic based on your questions.
9. Set up and answer three questions:
   (a) easy, e.g., simple statistics
   (b) medium, e.g., join, windowed, etc.
   (c) hard, e.g., analytical insight

4 System Design

4.1 Diagram of System Design

Figure 1 shows the workflow and system design.

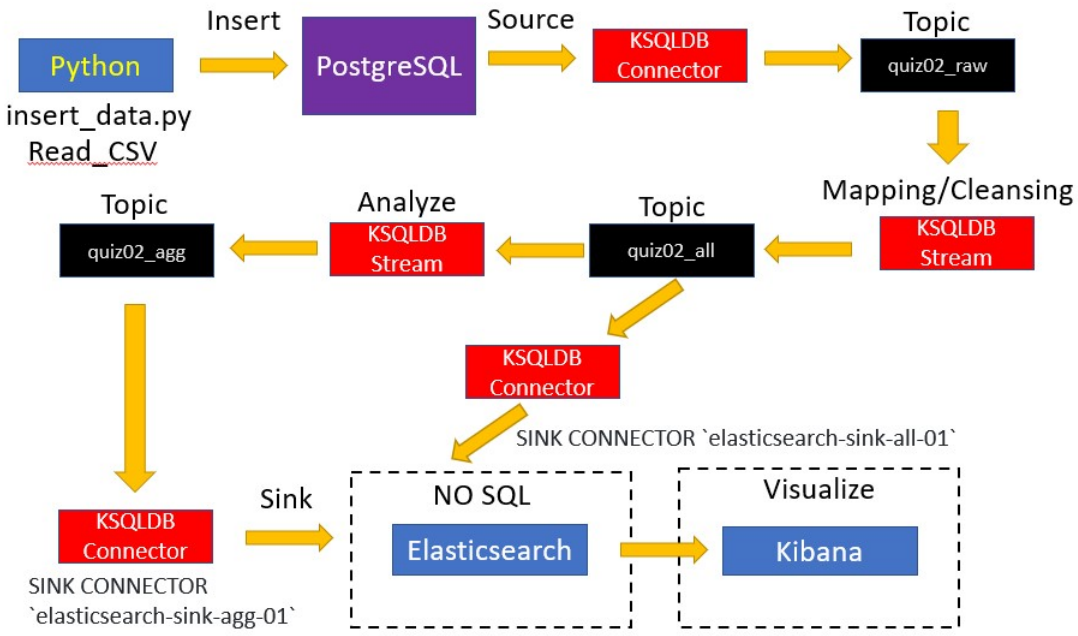

Figure 1: System design diagram

5 Questions

5.1 Easy Question

For easy questions we perform counting tasks relating to demographic distribution of the respondents.

5.1.1 Easy Question 1: Gender Distribution

The first question is the gender distribution.

Majority of college students responded to this survey is female, 60%, while male account for 40% (Figure 2).

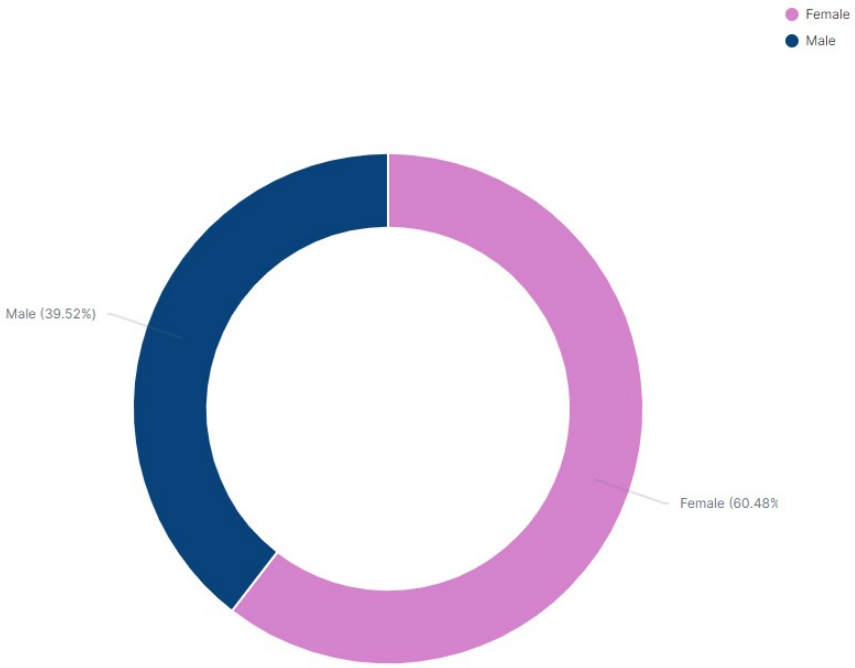

Figure 2: Gender Distribution

5.1.2 Easy Question 2: Employment

The second question is the employment status of college students in this survey.

Majority of college students responded to this survey work part-time (52%) and about 47% do not work (Figure 3).

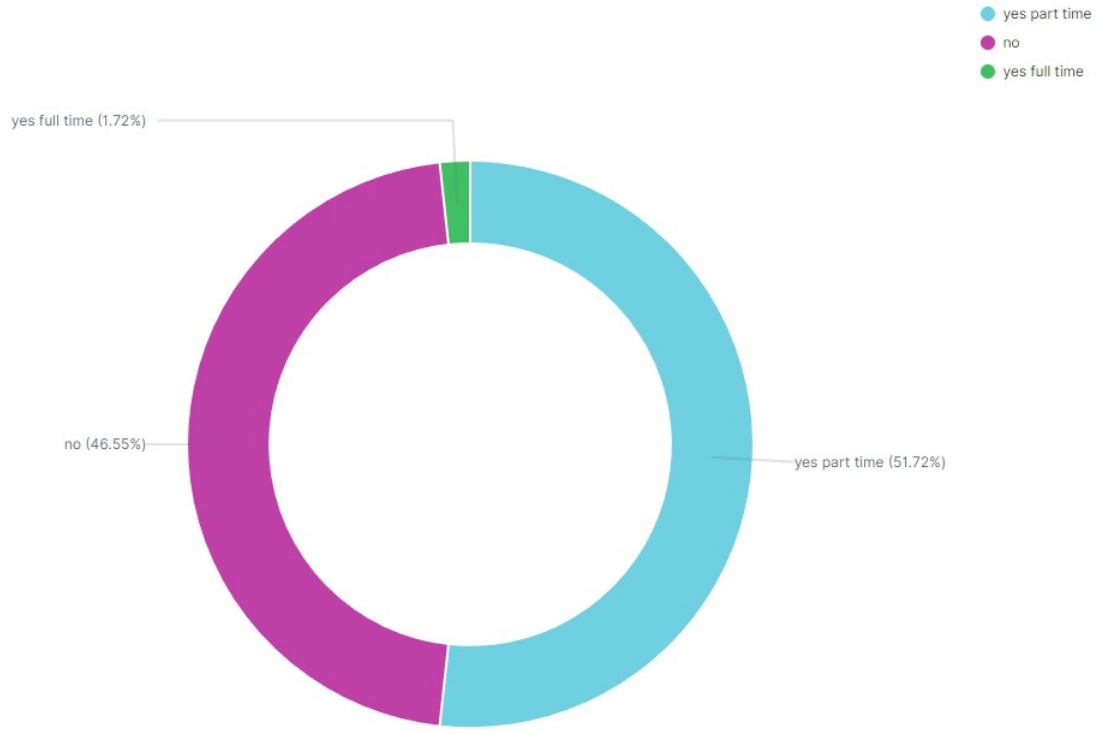

Figure 3: Employment

5.1.3 Easy Question 3: College Grade Distribution

The third question is the college grade of the respondents.

Thirty percent of respondents are freshman and about 30% are from second year (sophomore, 26%) whereas the third and fourth year (junior and senior, respectively) accounted for about 22% each (Figure 4).

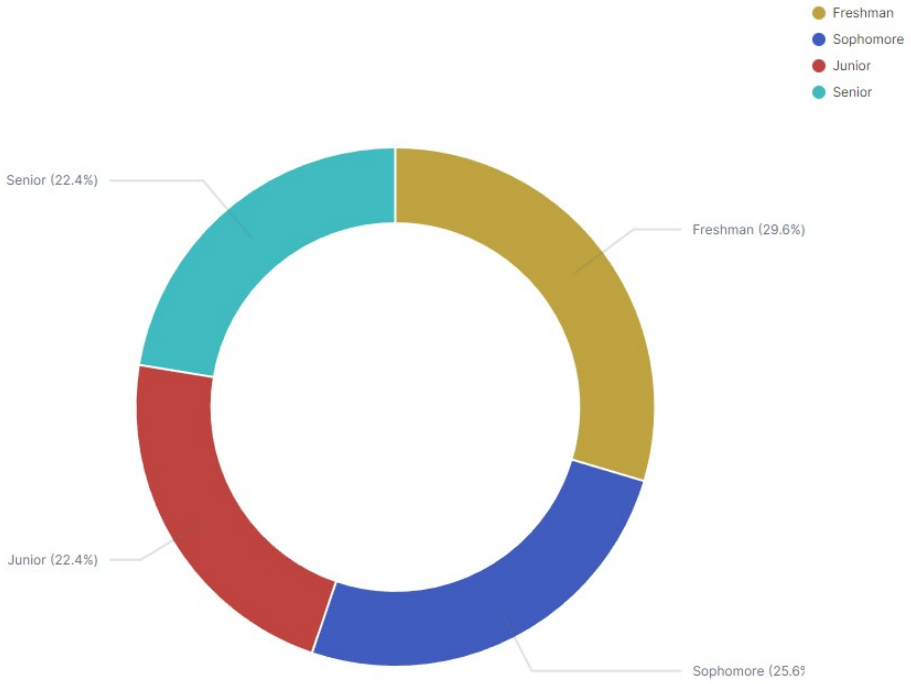

Figure 4: College Grade Distribution

5.1.4 Easy Question 4: Income

The last question is the students income.

The income distribution of college students in this survey surprised us as it appeared that the majority of students (about 33%) responded that they have annual income higher than 100,000$ (Figure 5). The second highest income group is those with income between 70,001 and 100,000$.

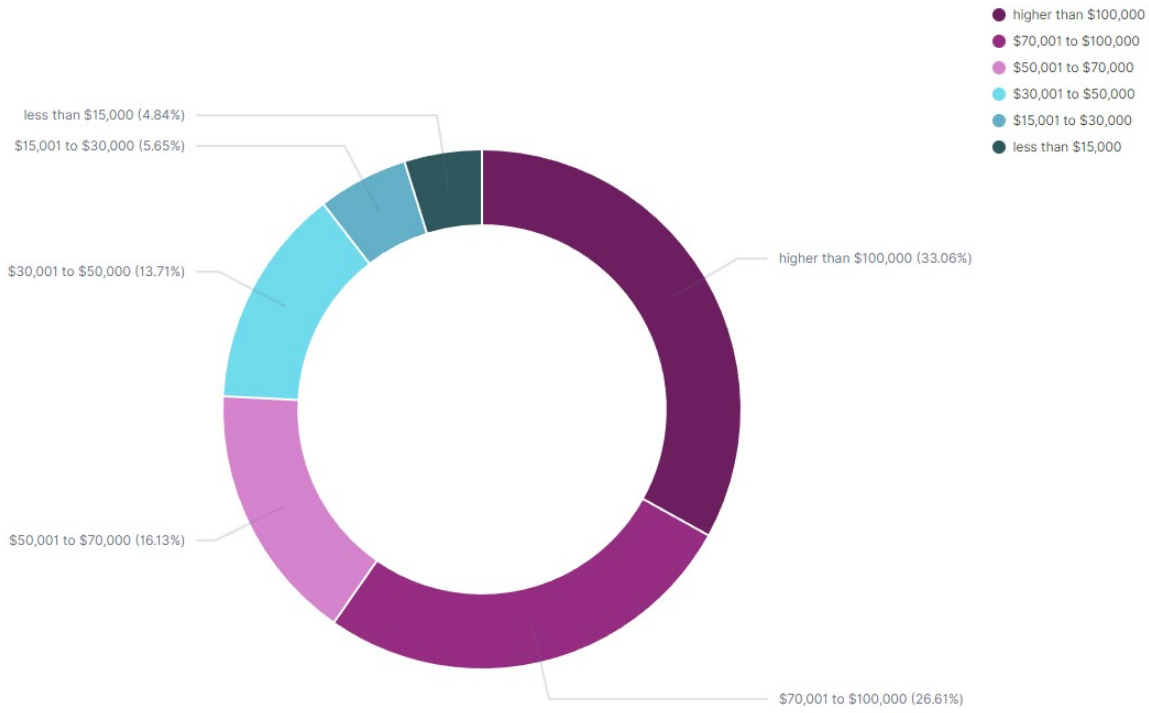

Figure 5: Income

5.2 Medium Question

For medium question we performed groupedby operation.

5.2.1 Medium Question 1: Calories Perception and Exercise

The first question is "Is there a correlation between different perspective on the importance of calories per day (calories_day) with frequency of exercise per week (exercise) ?"

The result is shown in Figure 6 in which we can see that people who think that carolies consumption per day is important do exercise more regularly than those who think that calories per day is not at all important.

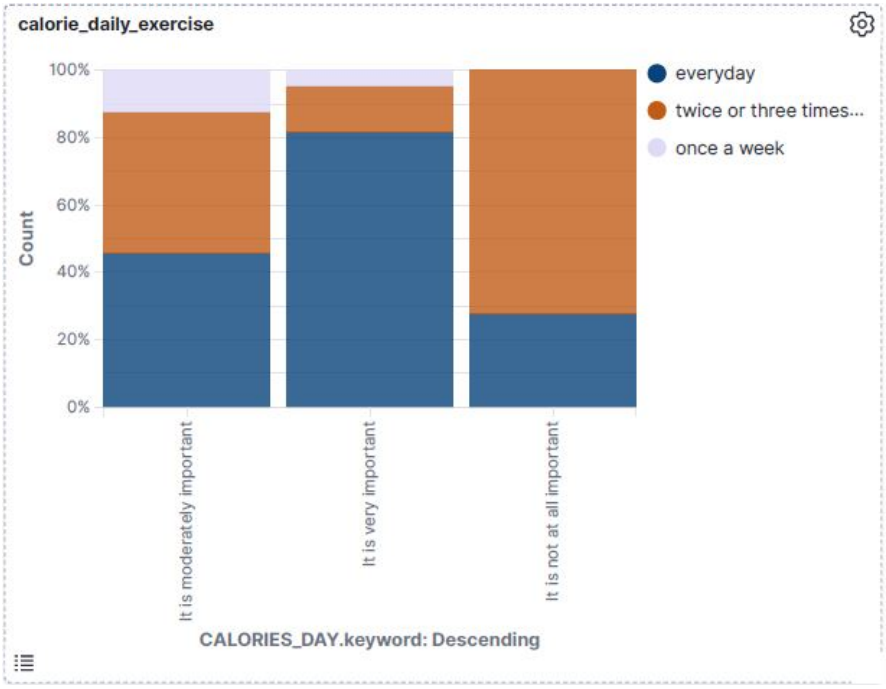

Figure 6: Perspective on importance of calories per day and weekly exercise frequency.

5.2.2 Medium Question 2: Calories Perception and Nutritional Values Check

The second question is "Is there a correlation between different perspective on the importance of calories per day (calories_day) with the frequency of checking nutritional values (nutritional_check) ?"

Figure 7 revealed similar relationship we see in the previous question in that people who think that carolies consumption per day is important frequently check the nutritional values of the products than those who think the calories per day is not important.

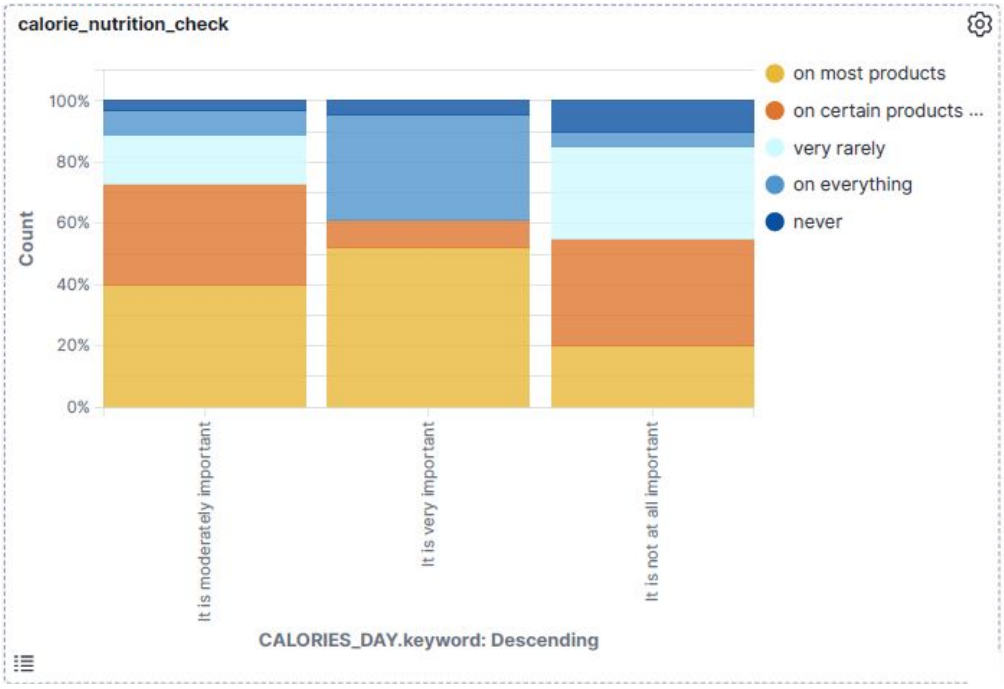

Figure 7: Perspective on importance of calories per day and weekly exercise frequency

5.3 Difficult Question

For hard question we attempt to extract analytical insights of the data.

5.3.1 Difficult Question: Gender and Self Perception of Weight vs. Exercise Frequency

We would like to know if between different genders the self perception of weight (self_perception_weight) will correlate with the frequency of weekly exercise (exercise).

Sunburst in Figure 8 reveals that high proportion of male respondent have a perception that they are physically fit while majority of female think that their weights are just right followed by the perception of feeling slightly overweight. The interesting insight we discover in this figure is that respondents who perceived that they are overweight (both male and female) tend to exercise the least (once a week).

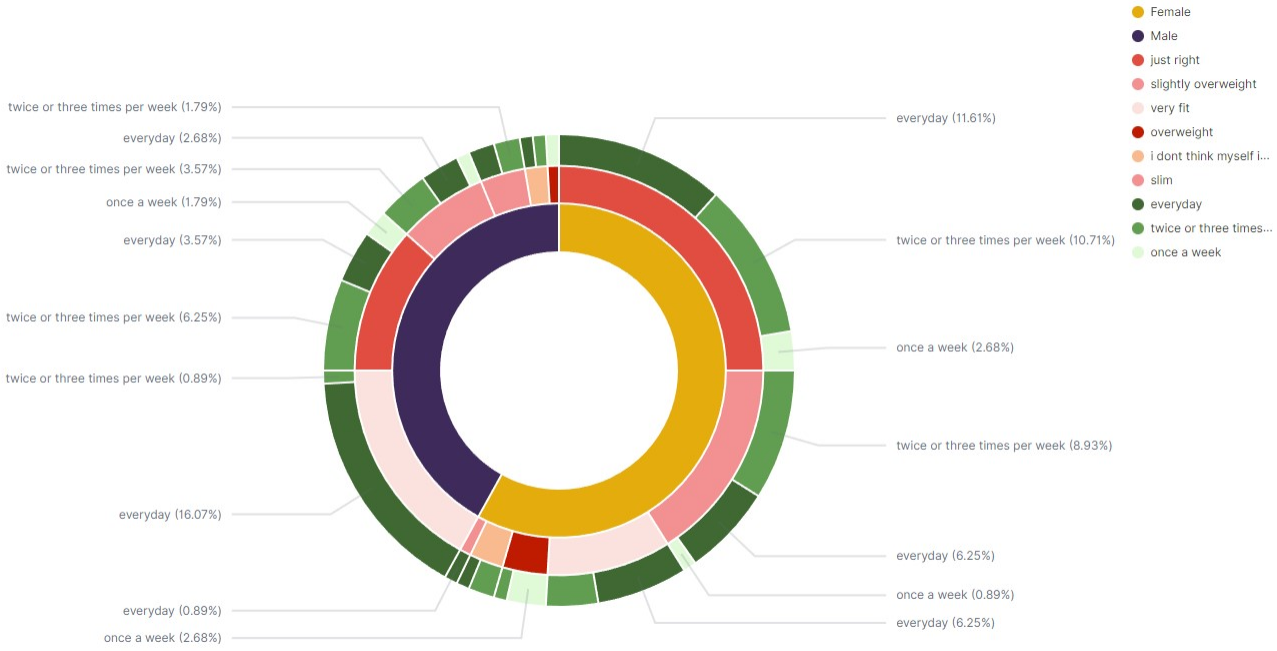

Figure 8: Gender and self perception of weight vs. Exercise frequency.

References

www.kaggle.com/datasets/borapajo/food-choices?select=food_coded.csv