Text-Mining of User-Generated Queries on Menstrual Pain

Visualizing a collection of Dysmenorrhea related posts by Age, Geography, Sentiment and Chronological trends

TEAM MEMBERS

- Ashok Reddy Singam [asingam@iu.edu]
- Bill Screen [wscreen@iu.edu]
- Ha-Lan Nguyen [nguyenhl@iu.edu]
- Sunanda Unni [suunni@iu.edu]

INTRODUCTION

What is Dysmenorrhea and why should you care?
Dysmenorrhea [dis-men-uh-ree-uh] is the medical term for menstrual cramps. It is the leading cause of recurrent short-term school absence in adolescent girls and a common symptom for women of reproductive age [1]. In the workplace, a BBC survey indicated that 50% of female workers experienced 'period pain' that affected performance on the job [2]. Many women who experience dysmenorrhea seek solutions outside of their doctors' office and turn to Internet websites for answers. The objective of this project is to discover key insights for women experiencing dysmenorrhea by performing text-mining on a large dataset of dysmenorrhea related questions posted on a popular question-and-answer website. The scope of analysis will involve examining selected *questions* by: age, geography, sentiment and chronological trends.

METHODS

[Description of how we process and analyze the dataset.]

The ChaCha dataset provided by Dr. Chen consisted of two excel files.

- Survey questions asked by women between the age of 13 – 50, which contained 5,07,327 records.
- Survey questions asked by men of all age groups, which contained 1,13,888 records

Two methods were used for extraction of data for analyzing.

The first method used SQL database and SQL language to extract the data required for the targeted visualization. The data extracted was - Top words used by each age group, Number of queries done as per Season.

The second method used Sci2 tool [1] for extracting tokens from the questions. The default set of stop words provided by Sci2 tool was used during tokenization. Additional cleanup was conducted using regular expression and bare eye inspection of normalized data. Sci2 was used for Burst analysis, Temporal Analysis.

Further visualization could be accomplished using Tableau.

RESULTS

Dysmenorrhea and age

[Visualization + Statistics]

In today's modern society, the average woman's reproductive years are between ages 12 and 51, which means that the **average woman will spend nearly ten years of her life menstruating and experience about 450 periods over her lifetime.**[11] Given the journey from the 'first period' to menopause is a personal adventure, reaching menopause is somewhat analogous to learning to drive a car - challenging in the beginning but less difficult over time. An example of this concept can be seen in the data by observing the high number of questions posted shortly after the 'first period' and the low number of questions leading to perimenopause and menopause. Comparable models to this concept can be seen in Psychology as the 'Power law of practice' and the Learning Curve. Therefore, the data suggests that the more menstrual cramps one experiences, the less questions one posts.

Why it is important: With this visualization we can identify most active ages who are posting questions. We can see that teenage had the maximum number of questions. This is a good insight. It tells that teenagers are new to such problems hence the more questions. Over a time period they may be gaining knowledge and we observe less questions in the higher age group. (Ref – Fig 1).

Top 5 ages that asked most questions are 15, 16, 17, 14 and 19. They were all teenagers. All girls started having their first menstruation during their teenage age. That can explain why they had a lot of questions and preferred to use this platform to ask to avoid embarrassment.

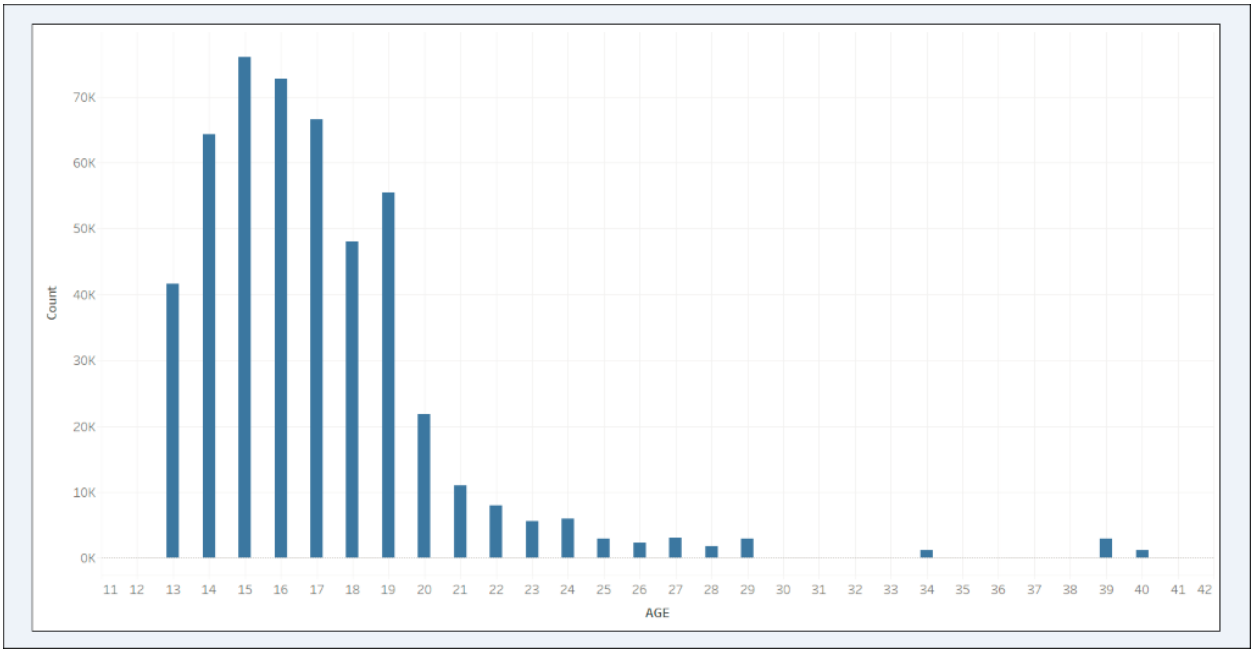

Fig 1. Distribution of count of questions at each age

This visualization identifies what are prominent words used in each age group. This is important as it helps to identify whether the problems are same or different across different age groups.

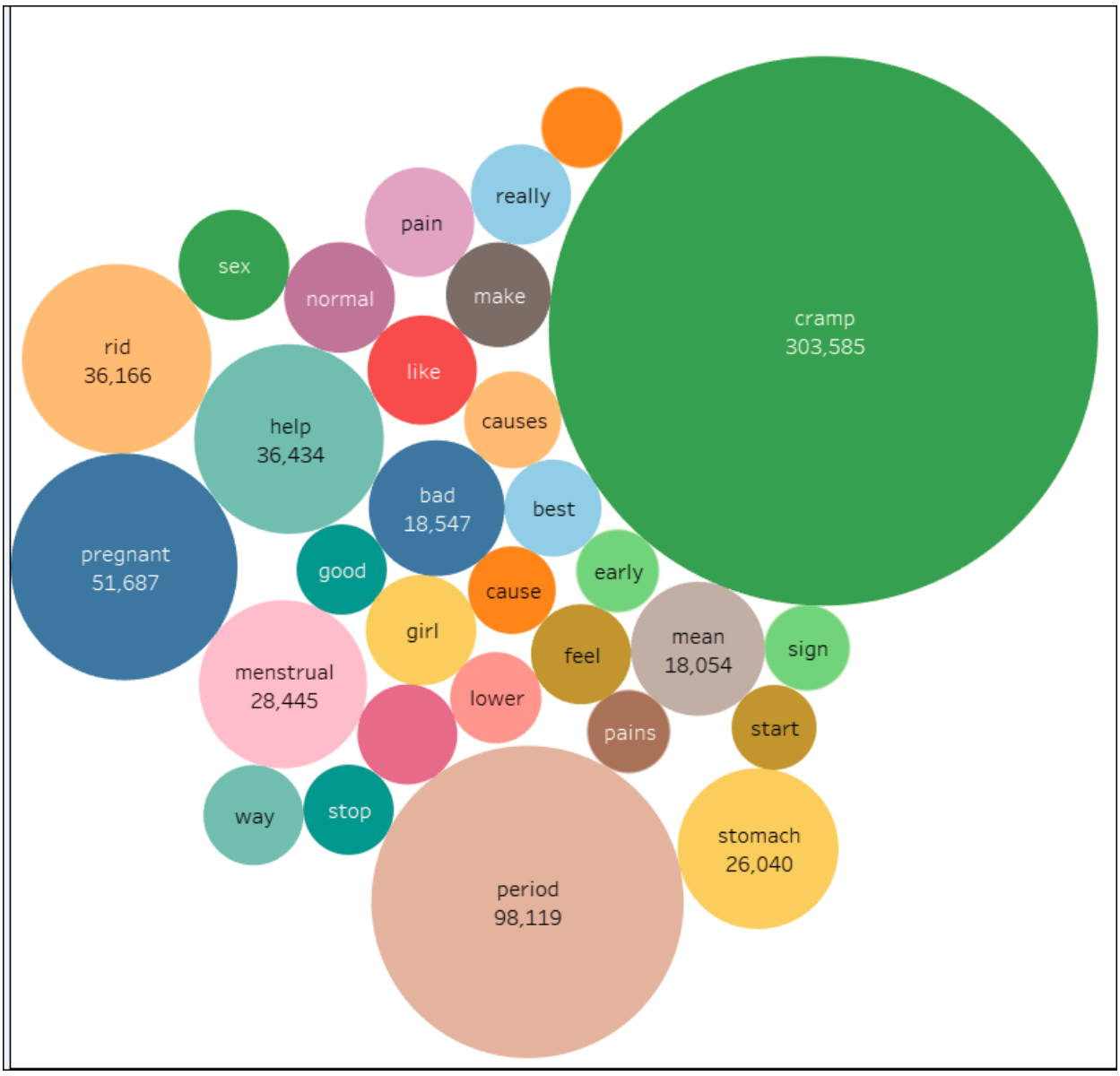

Fig 2. Distribution of the overall Burst words

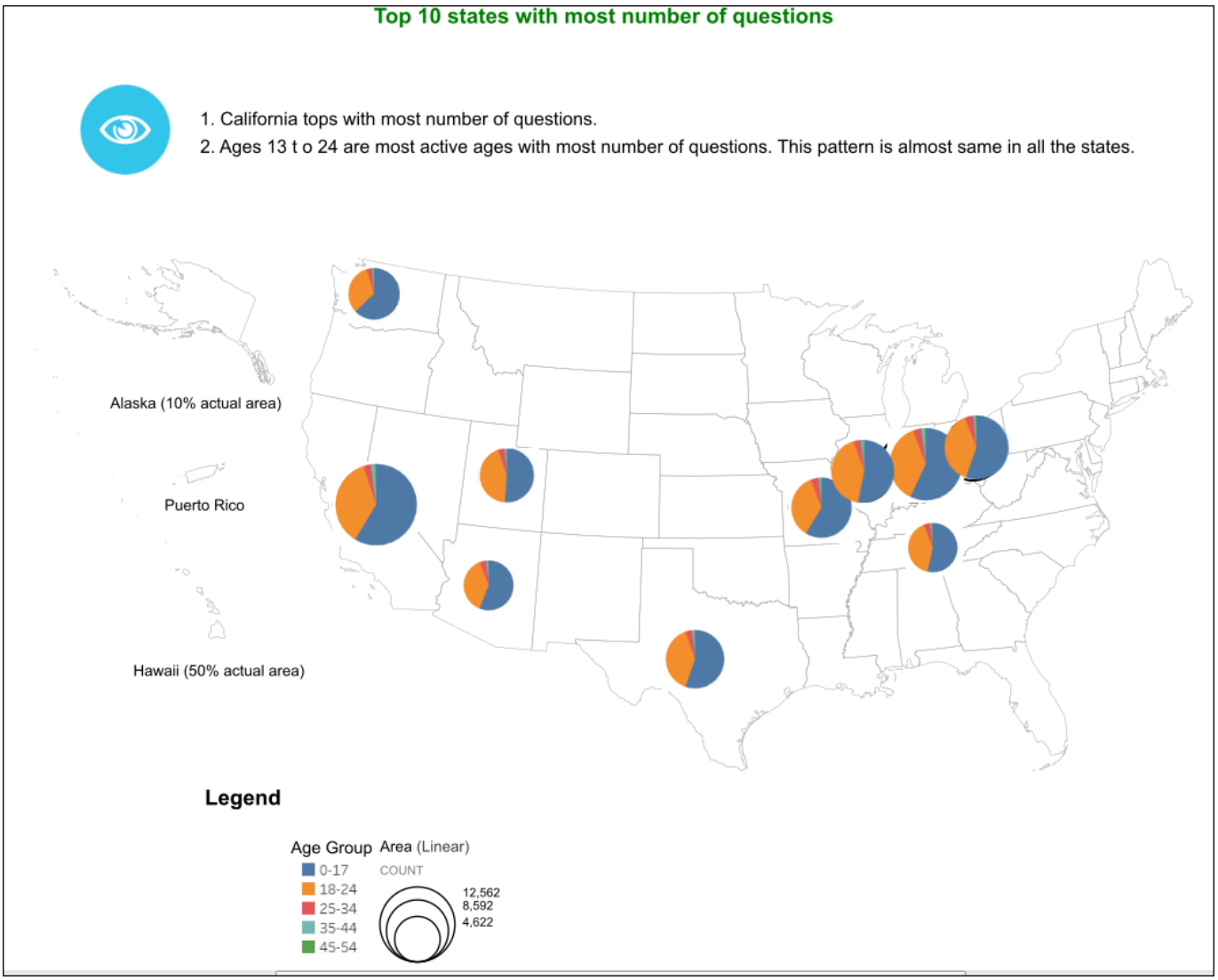

Fig 3. Distribution of queries as per Age and State

Geographic areas and related posts

[Visualization + Statistics]

Why it is important: With the visualization (Ref- Fig 4) we can identify zip codes from where maximum number of questions coming. This may help in understanding the socio-economic conditions of these zip codes have impact on the raise of number of questions. We could also identify that Top 5 states that had the maximum number of questions are California, Indiana, Ohio, Missouri, and Texas. California is the most populous state in the US, Indiana is where ChaCha Inc. was founded, Ohio and Missouri are in Midwest area. All these can explain why ChaCha was popular among people from these states (Ref- Fig 5).

Data issues: There are 373,201 records without zip-code information. We may not get accurate visualization. The Men-Women distribution of questions would have made better conclusion if we could get equal query dataset from both genders.

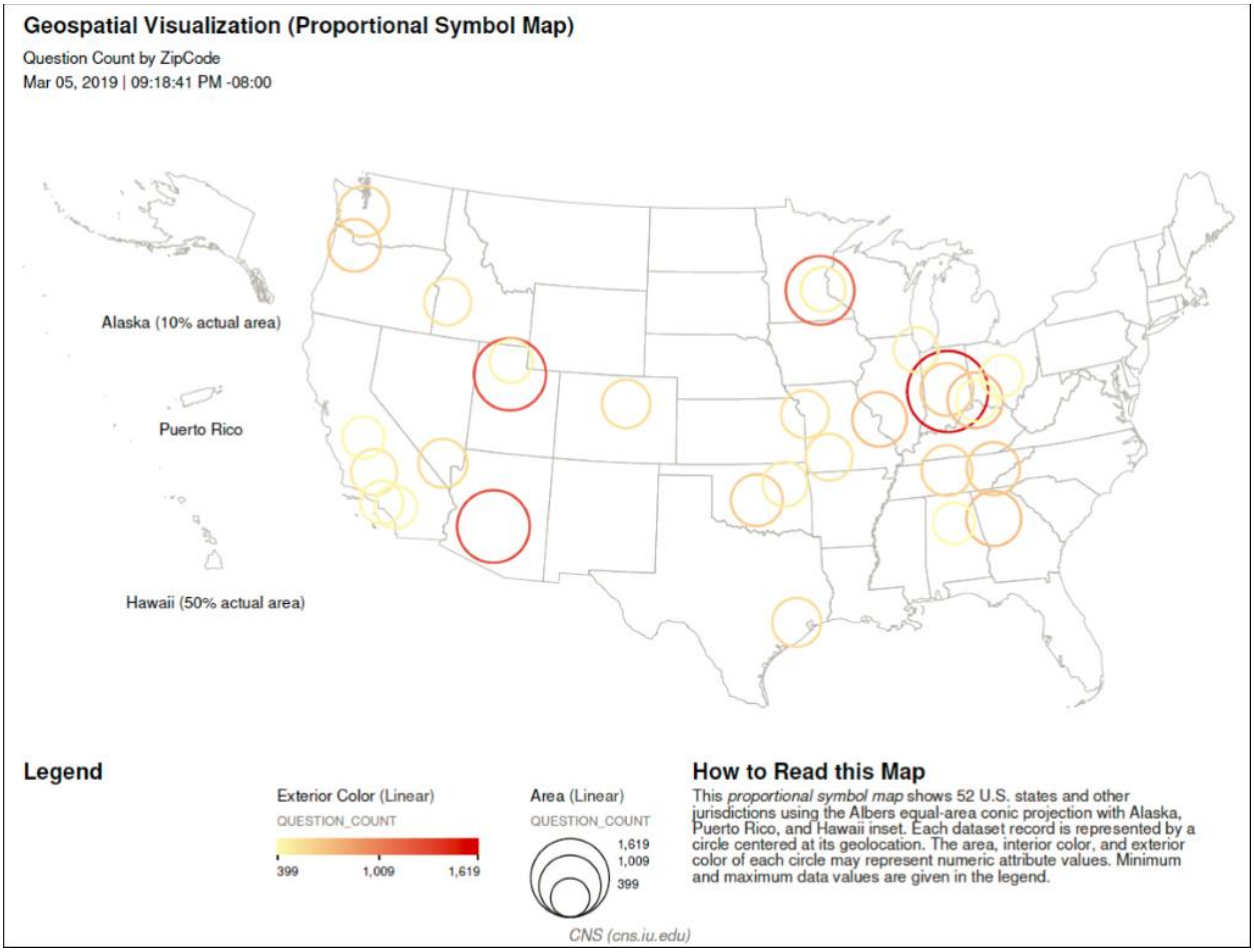

Fig 4. Geospatial Visualization of distribution of queries according to Zip code

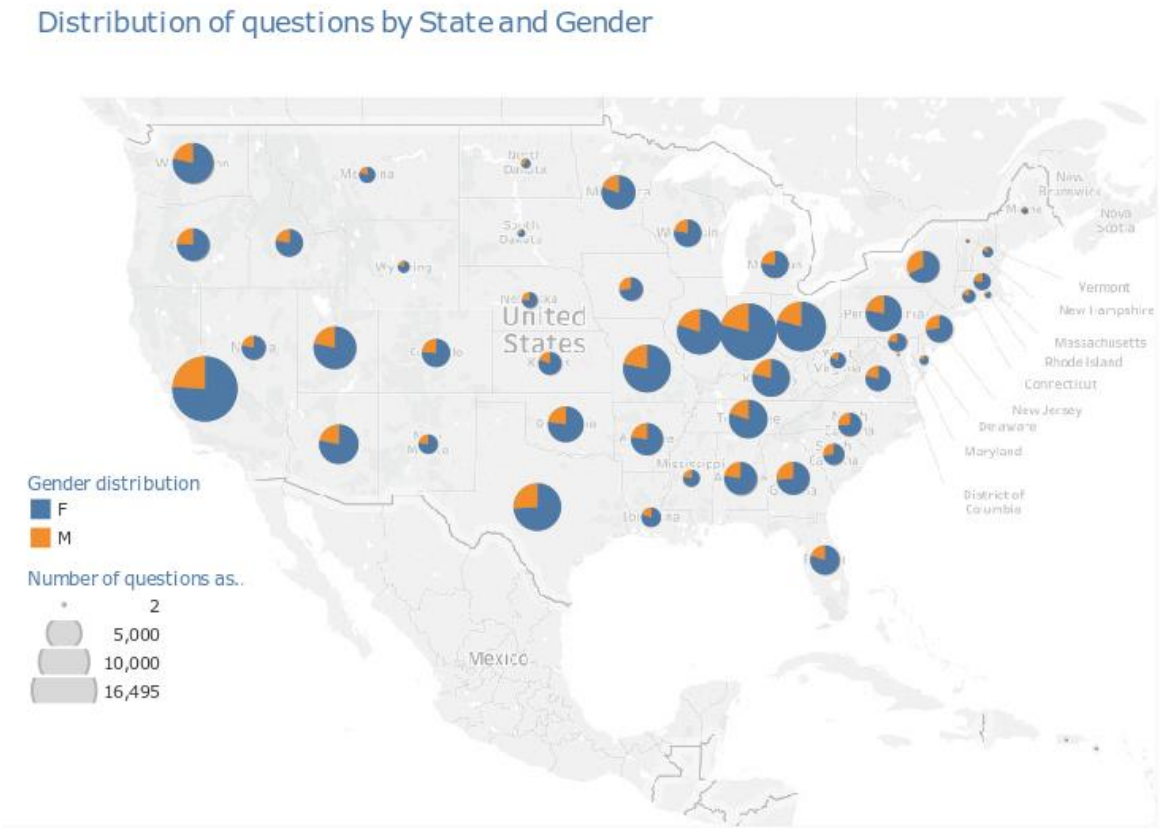

Fig 5. Geospatial visualization of queries as per Sex

Sentiment Analysis of randomly selected dysmenorrhea posts

If women who have more experience with dysmenorrhea post fewer questions related to menstrual cramps, is there a measurable variance in the verbal tone, feeling, opinion, or sentiment in their posts? To analyze this question, we used Sentiment Analysis (aka Opinion Mining) - a discipline within Natural Language Processing (NLP) that uses algorithms to identify and extract 'opinions' (subjective or objective expressions) within text to quantify sentiment. To calculate the sentiment scores for the questions posted in the Dysmenorrhea dataset, we used the Microsoft Cognitive Sentiment Analysis API service. It provides clients with the ability to send a body of text and have its sentiment scored between 0.0 (negative) and 1.0 (positive), with 0.5 as a neutral sentiment. According to Microsoft ' Sentiment score is generated using classification techniques. The input features of the classifier include n-grams, features generated from part-of-speech tags, and word embeddings. The following table illustrates selected questions from the data set and their corresponding score and polarity.

Question	Sentiment Score	Sentiment Polarity
What's a great natural remedy for cramps?	0.9238	Positive
Is it normal to have mild cramping when you are 13 weeks pregnant?	0.5	Neutral
I am having really painful abdominal cramps and I don't	0.0286	Negative

know why or how to get rid of them. Help!!!		

Given that the discomfort level of dysmenorrhea is subjective, and often varies based on number of experiences, we measured the sentiment score of 250 random questions from each age group to determine if there is an appreciable variation in the average sentiment scores between age groups. Based on analysis of mean sentiment scores (Low:0.1267, High: 0.1735) across all age group, the was no substantial variance in sentiment, therefore on average, all age groups showed a strong negative sentiment when posting dysmenorrhea related questions.

Chronological trends in dysmenorrhea posts

[Visualization + Statistics]

Why it is important: This visualization (Ref- Fig 6) identifies the season with maximum number of questions. Summer is top among other seasons. This is a good insight. This may because teenagers are not attending school and have more spare time to clarify their problems. This can also be due to the increased outdoor activities during summer and issues which cause hindrances during the active period.

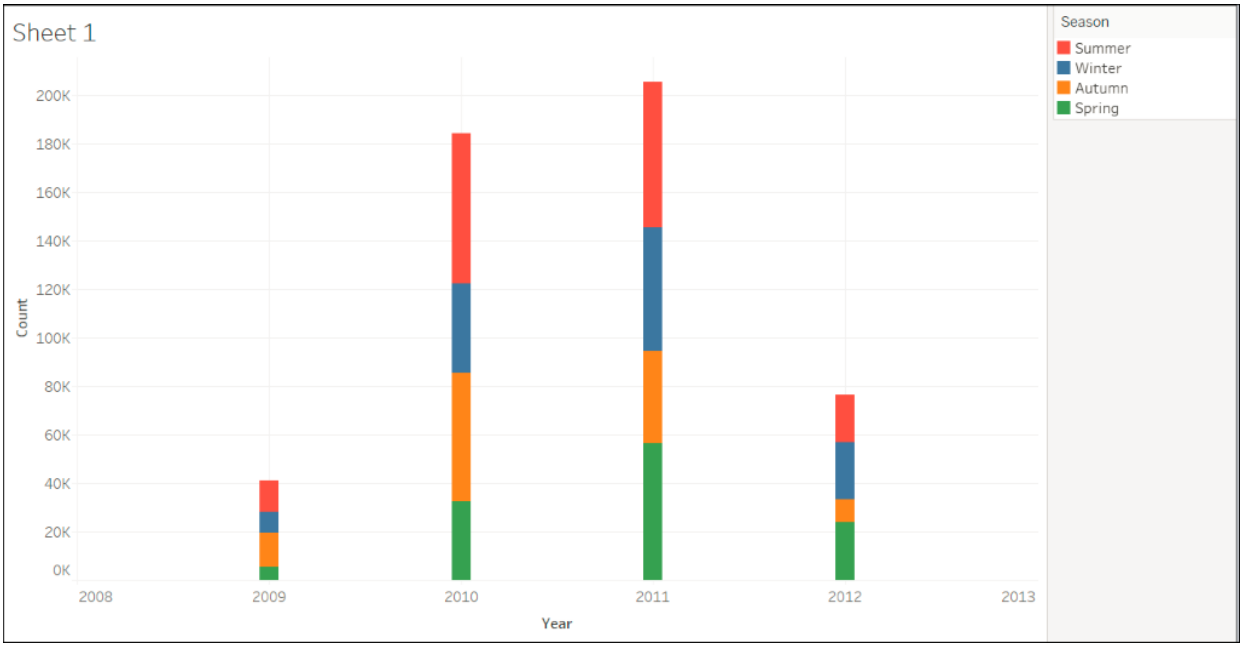

Fig 6. Distribution of queries as per season.

DISCUSSION

[What are our inferences from the above analysis? What are the challenges and what questions are unanswered? What can be done more in the future?]

Most questions were asked in 2010 and 2011 (184,312 and 205,417, respectively). Fewest questions were asked in 2009. It may be explained by the popularity of ChaCha website. The period of 2010 - 2011 was the most prosperous time of ChaCha. In 2010, ChaCha Inc. was recognized as one of the "Hottest Companies in the Midwest" by Lead411.

CONCLUSION

[…]

ACKNOWLEDGEMENTS

[…]

REFERENCES

1. Chen, C. X., Groves, D., Miller, W. R., & Carpenter, J. S. (2018). Big Data and Dysmenorrhea: What Questions Do Women and Men Ask About Menstrual Pain?. *Journal of Women's Health*, *27*(10), 1233-1241.
2. https://www.independent.co.uk/news/uk/home-news/period-pains-women-productivity-work-bbc-radio-5-live-survey-a7324536.html
3.