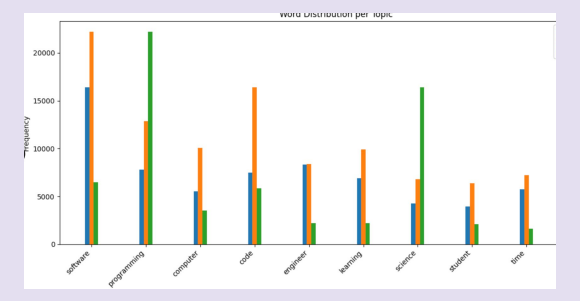
Need to know about Your Users? User Segmentation is the way

Trishala Jayesh Ahalpara, Chen Lin, Daria Popova, Divya Nalam, Xiaoyi Wang

[Geographic Distribution of StackOverflow Users: Top 10 Countries](#)

[Top 10 Highest Scored Questions on Stack Overflow](#)

[Top 10 keyword of About Me Section and it's distribution using LDA](#)



Background

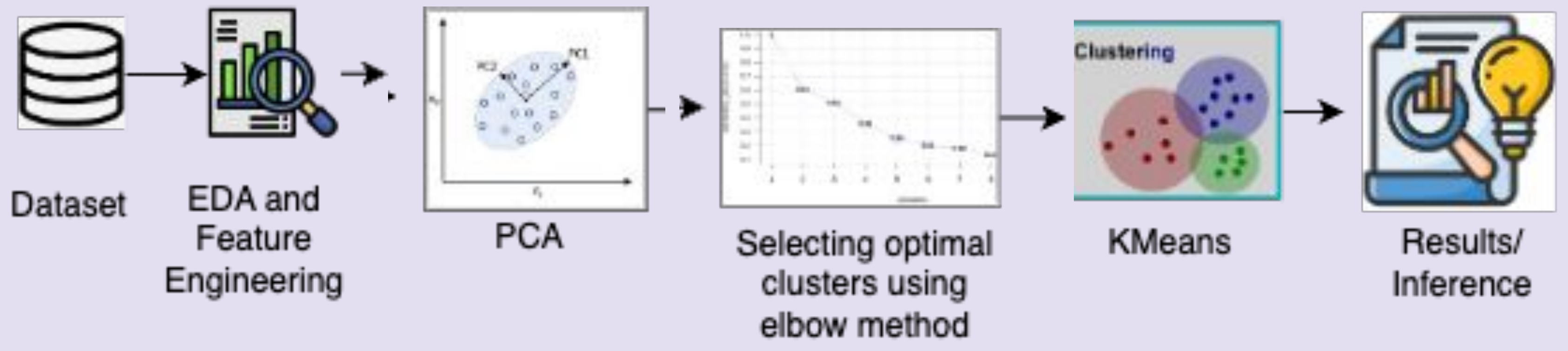
User segmentation is important because it enables businesses to target specific groups of customers, allocate resources efficiently, and deliver a personalized customer experience.

Data Processing and Analysis

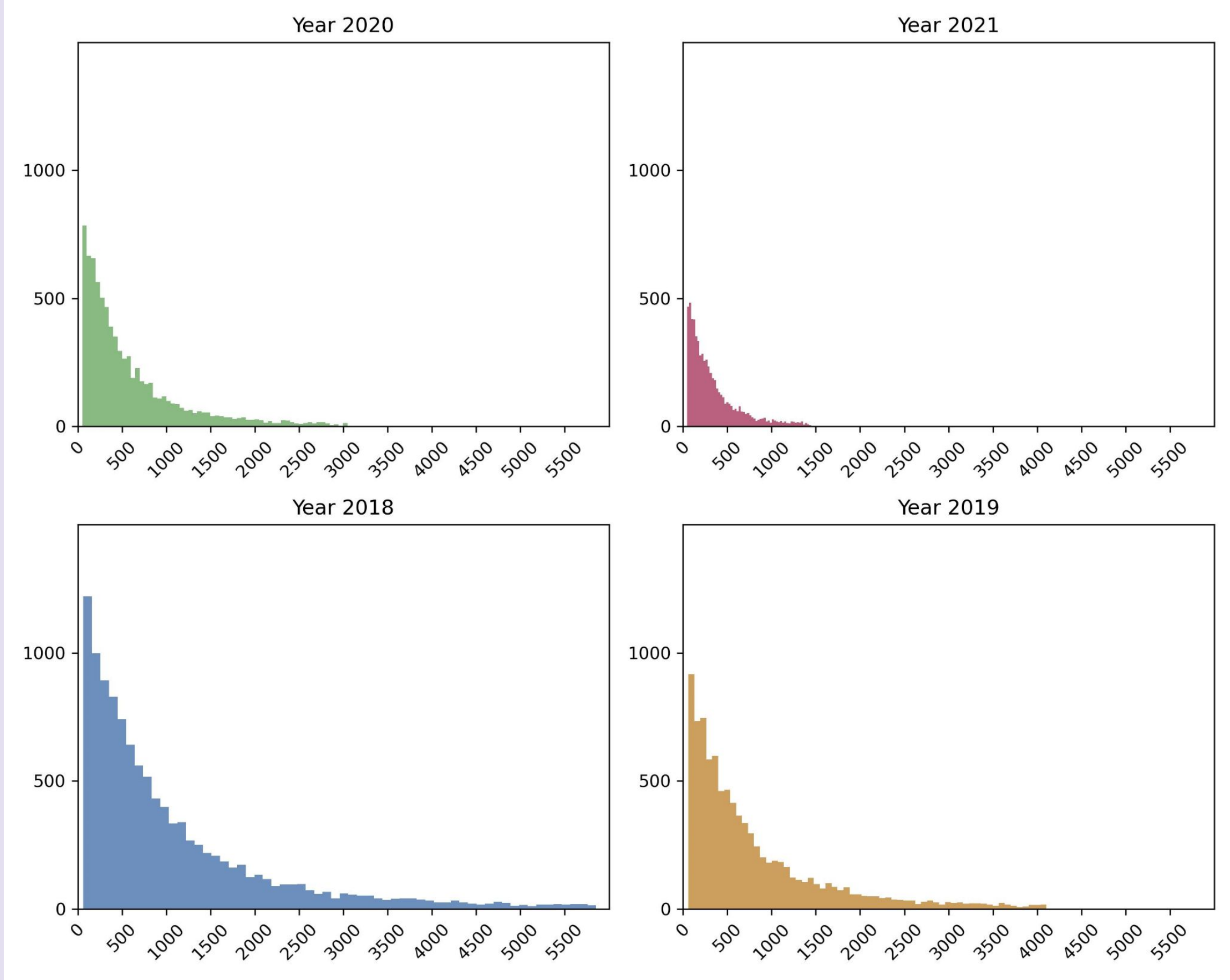
The Stack Overflow BigQuery table is a large dataset that contains information from the Stack Overflow question and answer platform.

We have processed and merged tables using big query to create our primary dataset with user information from 2017-2021

Model



Dataset → EDA and Feature Engineering → PCA → Selecting optimal clusters using elbow method → KMeans → Results/ Inference



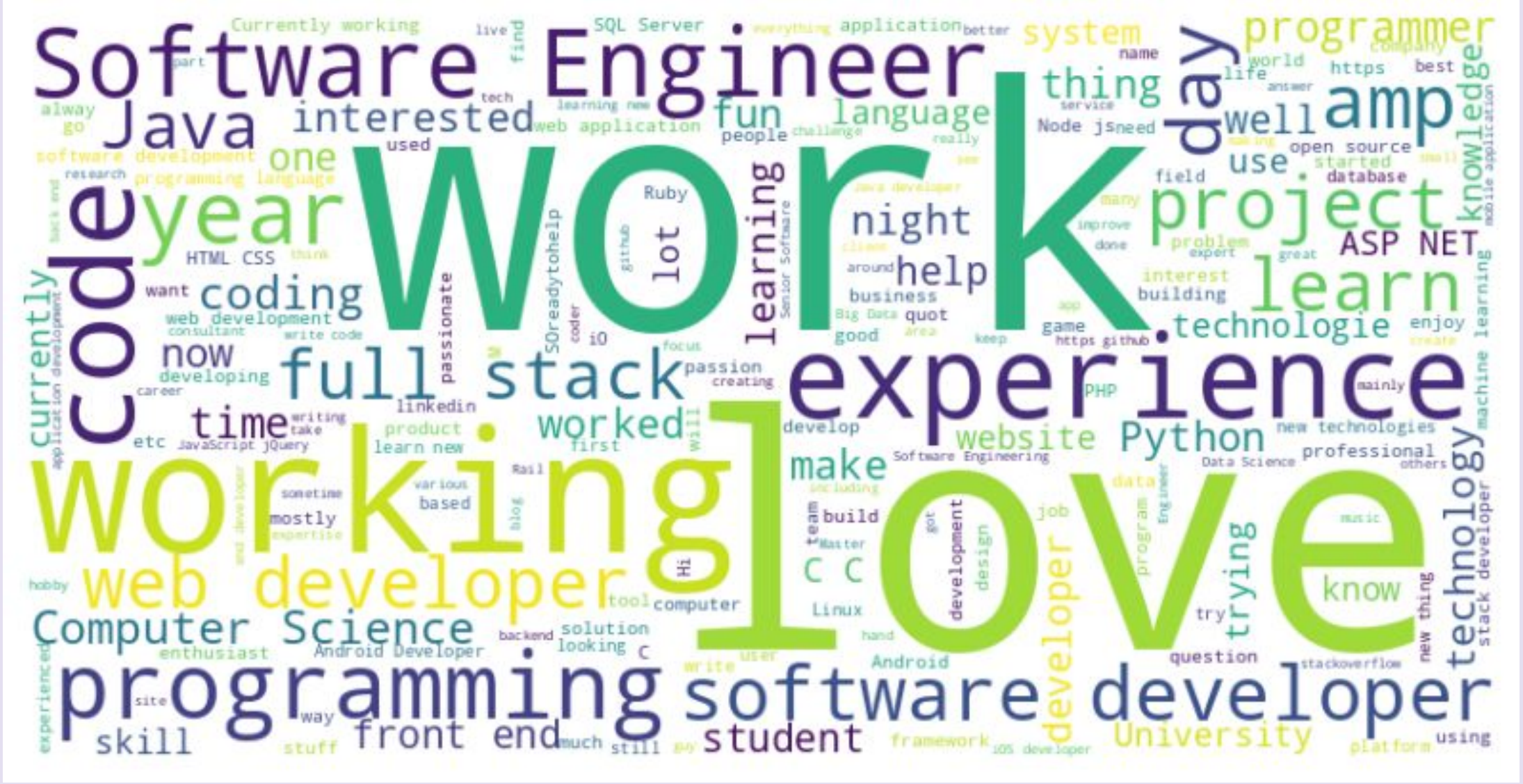
Average Number of Views Received per Post Declining since 2018.

Pre-Pandemic Median
695 view per Post.

Latest Year 2021
245 views per Post.

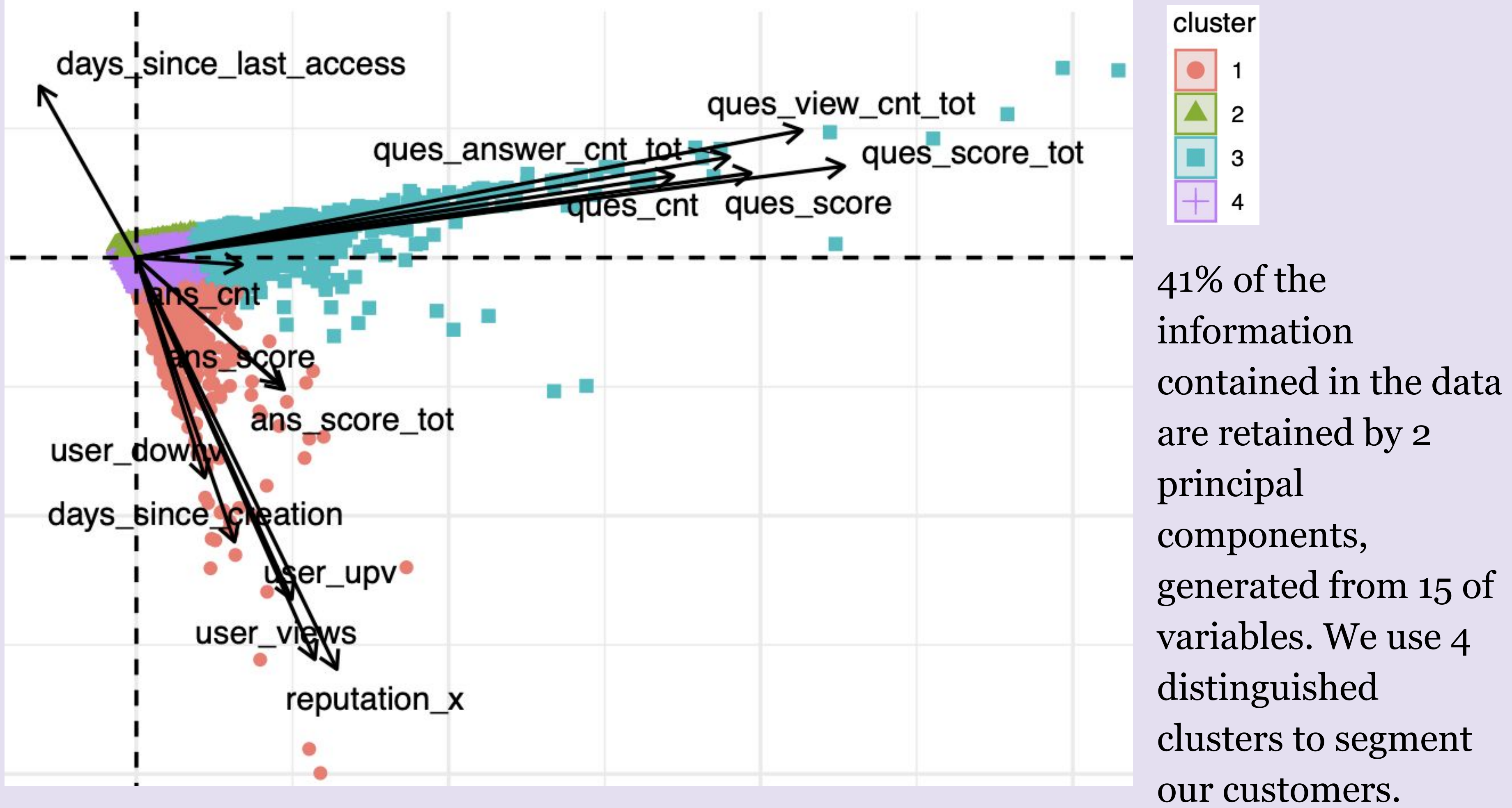
WordCloud

Top keywords in the 'about me' section of our Stackoverflow users. Most of the users **love coding, Software Engineer, and developers.**



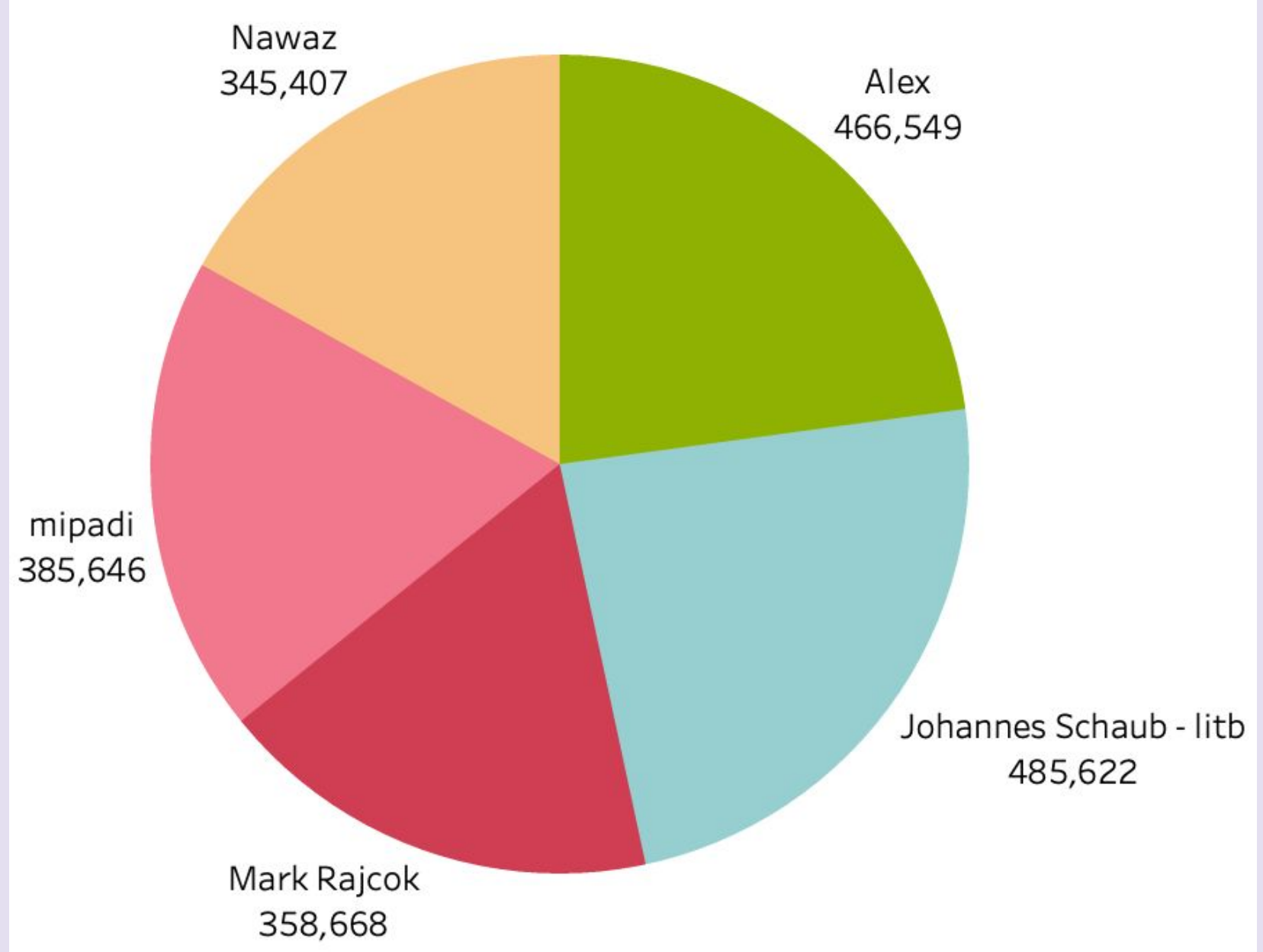
Such wordcloud can help companies with **marketing campaigns** to understand **user demographic** and what kind of **advertisements to target for.**

Biplot of Clusters and Variables

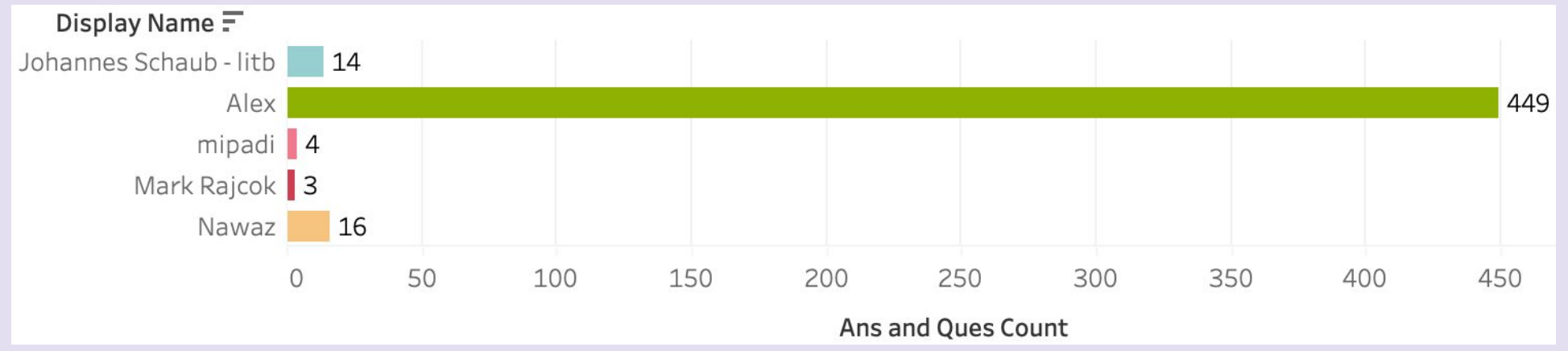


41% of the information contained in the data are retained by 2 principal components, generated from 15 of variables. We use 4 distinguished clusters to segment our customers.

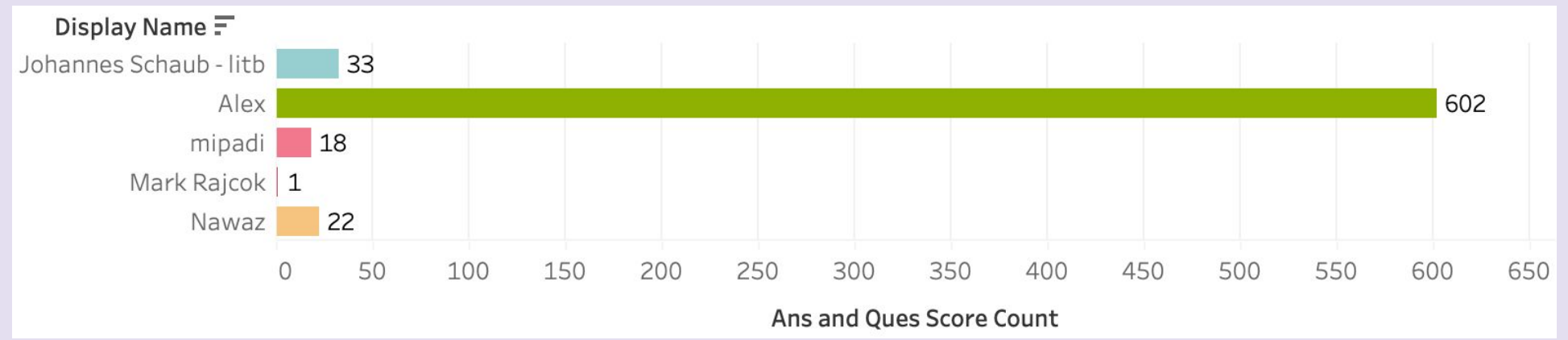
Top 5 Reputed Users



Answers and Questions Count

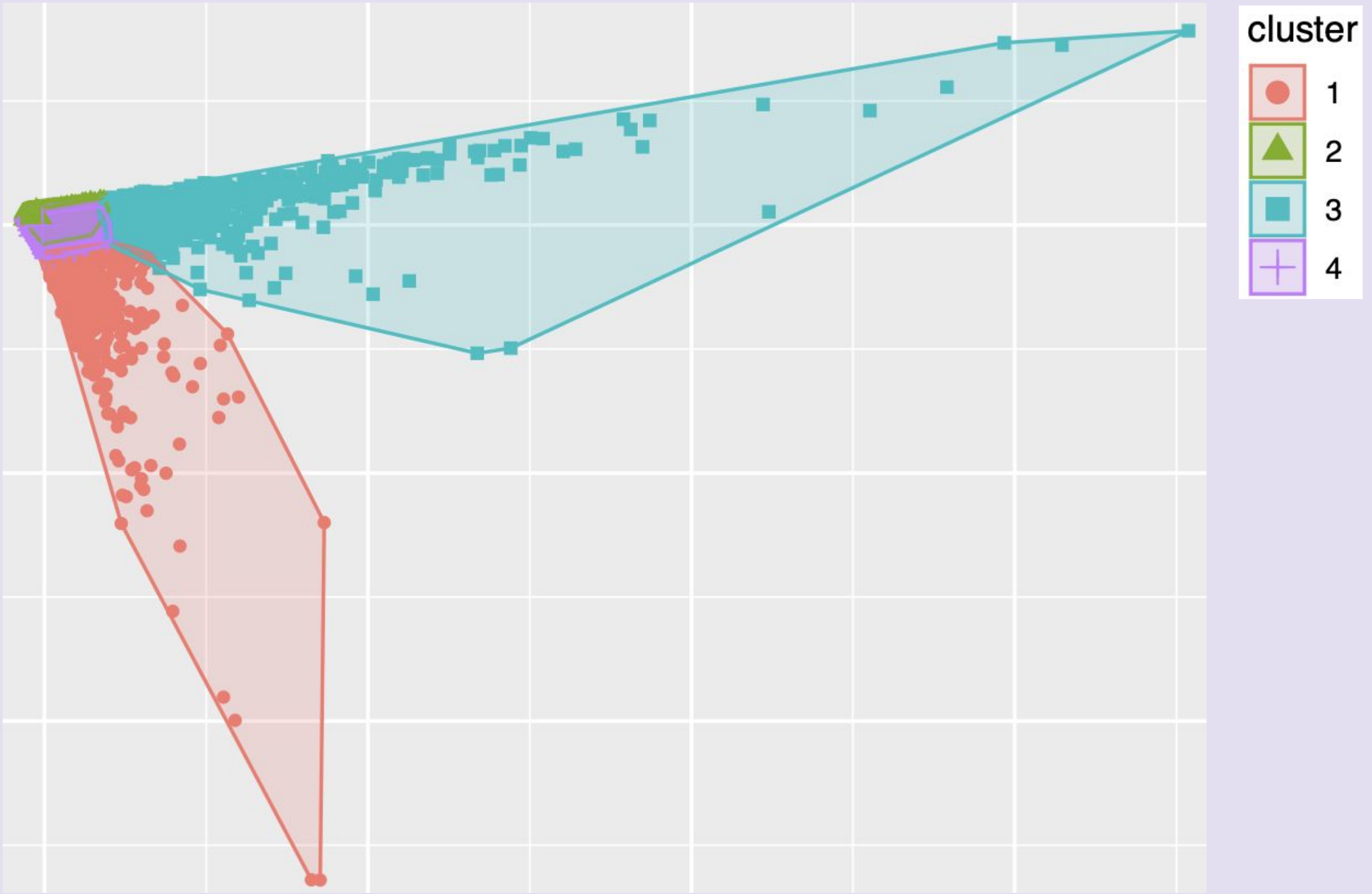


Answers and Questions Score Count



During the years 2017 to 2021, it is evident that the top 5 reputed users, excluding Alex, exhibited comparatively lower levels of activity and made fewer contributions.

Clustering Results



[Link to PDF Report of Clustering Results](#)

Based on Elbow method and Silhouette method, we selected 4 as the optimal number of clusters and 25.4% of variance explained by the clustering.

Used PCA to create components that can help us to perform user segmentation of stack overflow data.

Cluster Group	Size	Description of Cluster
Reputable contributors	**1422**	Reputable, active, long-time dedicated users; great contribution into building the community with high quality contents; willing to offer constructive feedback
Inactive users	**10781**	Long time since last activity; low participation; they stopped making contribution
Curious learners	**1399**	Most active in raising questions; eager to attempt answering (not necessarily providing perfect answers that garner top scores)
Community builders	**64501**	The majority of community users; no specific preference in answering or asking questions