EE-551 Final Project: TikTok Data analysis

Shiqi Xie 10478634 [sxie7@stevens.edu](mailto:sxie7@stevens.edu)

My final project is a data analysis project about TikTok users' personality analysis. The project was mainly based on a dataset of TikTok user usage data obtained from the web, washing the data, and creating graphs of overall usage. By exploring the logins of the users, studying their correlations, and performing personality analysis on the users, based on the results of the analysis, labeling all users in the dataset.

The dataset includes the id of the user, the city number of the login, the work id, the author id, the time and date of the watch, the finish broadcast, the likes, the length of the watch and other related data. The project is developed on Jupiter Notebook based on Python 3 as the programming language. The main tools used are pandas, pyecharts, numpy, matplotlib, etc. The whole project is divided into 3 steps: 1st.data inspection and pre-processing, 2nd. audience profiling and 3rd. correlation analysis.

In the data inspection and data pre-processing step, I used the pandas’ package for data cleaning and organizing. I used pandas’ shape function to query the number of data, and also used info(), isna() and other functions to query the data type and missing values respectively, used drop\_duplicates function to remove duplicate elements. The data types modification: user\_city and item\_city and music\_id to int type and then to string type and converting real\_time and date to time type.

In the work of user personality analysis, the total number of users and the total number of cities were obtained by removing duplicate data. Then the bar chart was plotted using the pyecharts package to get the top 20 most frequently used cities. From the data, we can know that the total number of cities is 377, and from the graph, we can see that the top 5 cities are used more than 1000, and the next few cities are concentrated between 800-600. After that, we also use pyecharts to draw a line chart, and we can see that most people use TikTok after 18:00 at night, and there is a decreasing trend after midnight. After that, we drew the pie chart of the number of viewers per week, from the chart, we can see that the number of viewers from the first to the third day of the week is higher, accounting for almost half of the number of viewers in a week. Next, to study the overall video quality on the platform, two new variables, like rate and finish rate, need to be declared. Use count() to calculate the total number of completes and likes respectively, and the like rate and completion rate are the total number of completes and likes divided by the number of video plays respectively. Using pyechart to plot the dashboard image, the completion rate is 40.11% and the like rate is 0.97%. Using numique() to calculate the data of the number of unique values of browse finish podcast and like, respectively, to draw a funnel analysis graph, we can see that the conversion rate of view to like is low.

In the step of correlation analysis, I calculate the number of Page View (PV) and Unique Visitor (UV) in different dates, the x-axis is set to the date, and the y-axis to configured PV and UV respectively, from the graph, in 2021.10.18 into the peak phase of user usage. I created a new table for storing the customer classification, which counts the most viewed periods per user, the user's preferred video length and the user's frequency of use, respectively, based on the uid number.

In general, this project helps platform operators to promote more efficiently by mapping data visualization work and user classification through user login, video playback information and other data. For example, the project can promote videos with better quality during the time when there are more users, promote in popular cities, and promote different types of videos according to different types of customers.

Github: <https://github.com/GqiqiPS/TikTok_Data_Analysis>

YouTube: