Name: Priyanka Kattishetti

Coding Assessment

1. **My observations about the data are as follows:**

   a. **Data structure:** There are 469308 entries spread across 5 columns in the data. The period of data collection is 8/7/2022–14/7/2022.
   b. **Column information:** ‘keyword_id’ and ‘keyword_rank’ appear to represent identifiers and ranking positions, respectively. ‘date’ represents the date of the observations. ‘searches’ indicates the number of searches for a particular keyword on a specific date. ‘search_engine’ is a categorical variable representing the search engine used.
   c. **Data Summary**
   The ‘search_engine’ column ranges from 0 to 25, representing different search engines.
   d. **Data type:**
   ‘Keyword_id’, ‘keyword_rank’, ‘searches’, ‘search_engine’ have data type as **‘int’** and date has ‘**object’**.
   e. **Descriptive analysis:** I conducted a descriptive analysis, and these statistics facilitated me in comprehending the central tendencies, variability, and distribution of datasets, helping to understand the characteristics of each column.
      - Count: All columns have the same count 469308 indicating no missing value.
      - Standard deviation: This helped me to understand amount of variation in a set of values.
      - Similarly, min and max values were calculated to understand the max and min values for each column.

   I also noticed variations in ‘keyword_rank’ for the same ‘keyword_id’. Specifically, there are instances where identical ‘keyword_id’ exhibit different ranks while sharing the same search engine. This suggests that keyword rankings are not consistent across all occurrences of a particular ‘keyword_id’, indicating potential fluctuations or changes in ranking positions for the same keyword across different time periods.

2. **What You Attempted and Why**

   I began my search for keyword_rank = ‘1’, but soon found out that there were no keywords matching that rank explicitly. Consequently, I did the following.

   I conducted a comprehensive analysis focused on a specified ‘search _engine’. This involves aggregating the number of searches (‘searches’ column) for each ‘keyword_id’ across various dates within the chosen ‘search_engine’. Subsequently, the analysis aimed to pinpoint the specific ‘keyword_id’ that exhibits the maximum total searches within the designated ‘search_engine’. The ‘keyword_id’ with maximum searches is considered to be **‘top-ranked (rank ‘1’)’** for individual ‘search_engine’. By undertaking this targeted examination, I gained a valuable insight into the performance and popularity of individual ‘keyword_id’ within a particular ‘search_engine’ over time. The cumulative search count for each keyword is evaluated, enabling the identification of ‘keyword_id’ that resonate most with users of the chosen ‘search_engine’.

   I opted for this method due to the fluctuating nature of ‘keyword_rank’, where the same ‘keyword_id’ has different ‘keyword_rank’ distributed across different ‘date’. The ‘keyword_rank’ is assigned in a random order, making it **impractical** to use them directly for the analysis to identify instances where rank equals

1. However, upon closer examination, I observed that specific ‘keyword_id’ were consistently recurring only within certain ‘search_engine’. With the help of this observation, I made the decision to proceed with my analysis, focusing on the 'keyword_id', 'date', 'searches', and 'search_engine' variables to determine the top-ranked ‘keyword_id’ (rank=1) for each individual ‘search_engine’. This tailored approach allowed me to navigate the variability in ‘keyword_rank’ and pinpoint specific keyword performance within distinct ‘search_engine’ contexts.

3. Visualization:

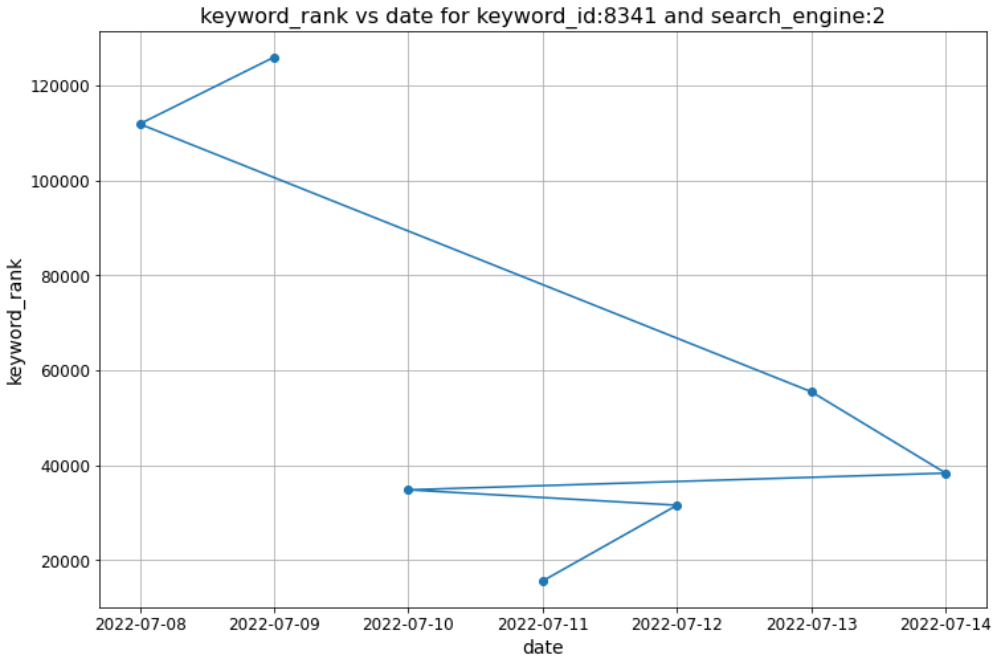

Fig 1: Keyword_rank vs date for keyword_id: 8341 and search_engine:2

The Fig:1 illustrates how ‘keyword_rank’ varies with ‘date’ for a specific ‘keyword_id’ and ‘search_engine’. Here in the above plot, I am considering ‘keyword_id: 8341’ and ‘search_engine: 2’ for illustration. Recognizing that keyword rankings are influenced by the number of searches, it became evident that to conduct a meaningful analysis on the number of searches for a top-ranked keyword (rank ‘1’) in a particular engine, I needed to aggregate the searches across all dates for that specific **‘keyword_id’** and **‘search_engine’**.

4. Model and Evaluation

For the model, I used the ‘for’ loop that iterates over the unique values in the 'search_engine' column of the DataFrame. For each iteration, it creates a subset of the DataFrame (engine_data) containing only the rows corresponding to the current ‘search_engine’. Then, using the groupby function, it calculates the sum of searches for each unique ‘keyword_id’ within that particular ‘search_engine’.

The variable keyword_sum holds the aggregated sum of searches for each ‘keyword_id’ in the current search engine. The code then uses the idxmax() function to find the keyword_id with the maximum sum of searches and retrieves the corresponding maximum search count.

Overall, this code efficiently performs a per-search-engine analysis, calculating the sum of searches for each ‘keyword_id’ and identifying the ‘keyword_id’ with the highest total searches within each specified ‘search_engine’.

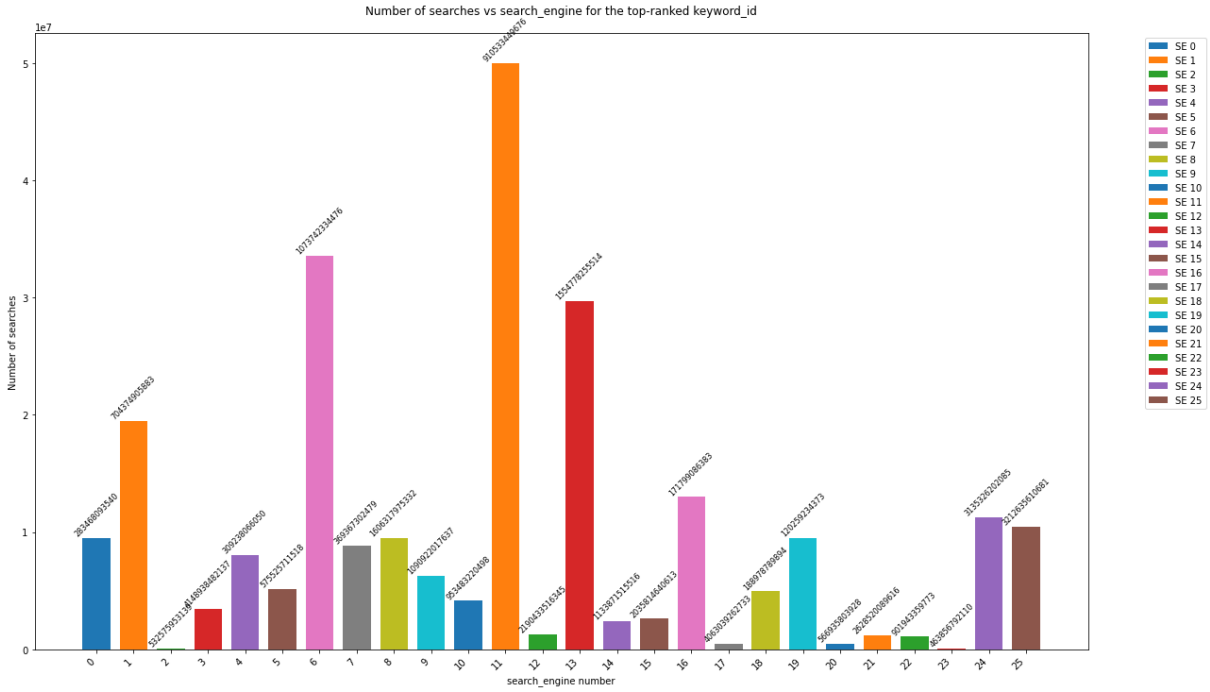

Fig 2: Number of searches for the top-ranked keyword_id across different search_engine.

Evaluation

Figure 2 illustrates the cumulative searches for the top-ranked ‘keyword_id’ corresponding to each ‘search_engine’. In the Fig 2, above each rectangle the ‘keyword_id’ is displayed which is the id of the keyword which has the most searches. The y-axis represents the sum of searches, while the x-axis denotes the ‘search_engine’ number.

a. Notably, each ‘search_engine’ is linked to unique ‘keyword_id’, as evident from the Fig 2.

b. Among the search _engines, ‘search_engine’: **‘11’** stands out with the highest search count, attributed to ‘keyword_id’: **‘910533449676’**, whereas search_engine: **‘2’** reports the lowest searches.
c. Search engines **11**, **6**, **13**, and **1** exhibit slightly higher search counts compared to their counterparts, while Search engines **2**, **23**, **17**, and **20** register lower search counts.

5. Conclusion:

The following conclusions were made.

a. Search Engine Performance Variation:

The analysis reveals notable variations in the cumulative searches for the top-ranked ‘keyword_id’ across ‘search_engines’. Search engines exhibit diverse levels of performance, as evidenced by the range of search counts represented on the y-axis.

b. Distinct Keyword_id’s per Search Engine:

Each ‘search_engine’ is associated with unique ‘keyword_id’, emphasizing the specificity of ‘search_engine’ dynamics. This implies that certain keywords may be more prominent or relevant within specific ‘search_engines’ compared to others.

c. Performance Disparity Among Search Engines:

Search_engine **‘11’** emerges as the standout performer with the highest search count, linked to ‘keyword_id’ 910533449676. Conversely, ‘search_engine’ **‘2’** reports the lowest searches, highlighting a significant performance disparity among different ‘search_engines’.

d. Variation in Search Counts:

Certain ‘search _engine’, namely **11**, **6**, **13**, and **1**, exhibit slightly higher search counts, indicating potentially higher user interest or search activity for keywords associated with these engines. Conversely, ‘search_engine’ **2**, **23**, **17**, and **20** register lower search volumes, suggesting comparatively lower popularity or relevance of keywords in these engines.

In conclusion, the findings emphasize the importance of understanding the specific dynamics of each search engine, tailoring keyword strategies accordingly, and recognizing the diverse performance of ‘keyword_id’ across different search engines. These insights can inform more effective search engine optimization (SEO) strategies and enhance the overall performance of keywords in online searches.