**7350 Final Report**

Members’ name&sid:

NI Haokun 22455744

LU Tianqing 22413731

ZHENG Jiale 22443126

HUANG Xinjing 22427201

1. **Research Background**
   1. **Reasons for Topic Selection**

With computer science and technology development, programmers must continuously improve their programming skills to meet the needs of enterprises and the market. The Programming Question Bank website was created to provide a more systematic and comprehensive programming training platform for programmers. For those preparing for employment and interviews, analyzing the data in the question bank and studying the information on question types, difficulty, and user distribution can help them better understand the current technology trends in the industry, so that they can learn and improve their professional skill level in a more targeted manner.

* 1. **Research significance**

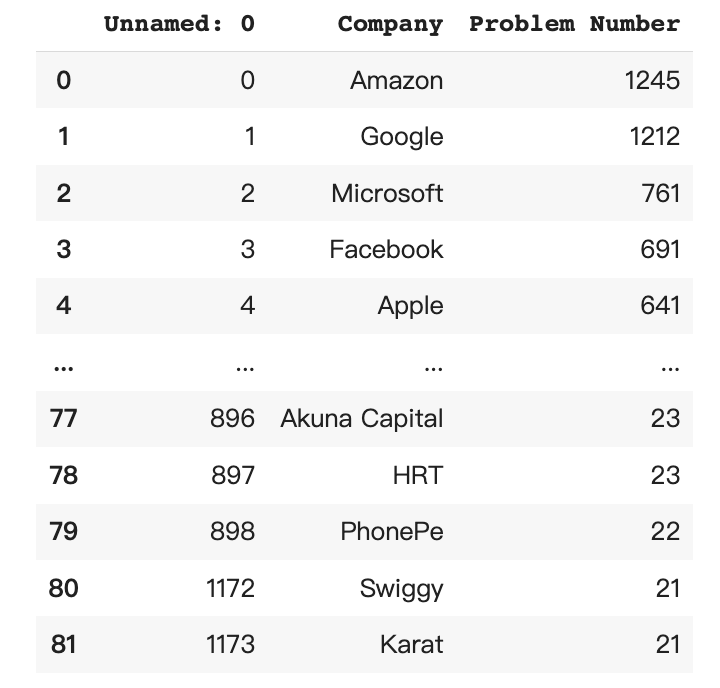
1. Improve interview success rate: By analyzing information such as the type and difficulty of questions and tags on the programming question bank website, you can effectively understand the knowledge and skill requirements that are of more concern in the industry, thus helping to better prepare for interviews and improve interview success rates.
2. Understand global programmers' technical level and learning trends: By analyzing information such as the distribution of global users and the skills and languages that the top users prefer to master, you can better understand the global programmers' technical level and learning trends, and provide a reference for your career planning.
3. Predicting future exam difficulty: Through machine learning methods, future exam top difficulties can be predicted to help candidates prepare for exams in a more focused manner. Help interviewees allocate their time and energy more effectively.
4. **Data source**
   1. **Website link**
5. Chinese version：[https://leetcode.cn/](https://leetcode.cn/company/)
6. United State version：<https://leetcode.com/>
   1. **Introduction**

This website is an online programming practice platform for programmers. On this site, users can solve various algorithm and data structure problems, and participate in global programming competitions. In addition, the website also provides many practical tools and resources, such as topic classification, a code editor, an online discussion community, etc., to help users learn and improve their programming skills.

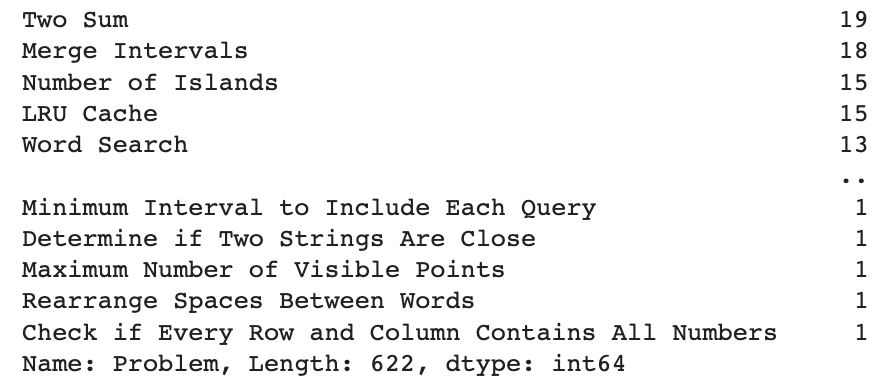
* 1. **Data set content**

1. companies\_df.csv
   * Company
   * Problem Number
2. global\_ranking\_df.csv
   * Rank
   * User Name
   * Contests Attended
   * Country
3. problems\_info.csv
   * Title
   * Topic Label
   * Accetpance
   * Difficulty
   * Frequency
   * Accepted
   * Submissions
   * Premium
4. ranking\_user\_Info\_df.csv
   * User Name
   * Languages
   * Advanced skill
   * Intermediate skill
   * Fundamental skill
   * 技能
5. Each\_Company\_Problem folder
   * Eachl companies problem
6. **Research Questions**
   1. **What questions do these companies prefer for companies with a total number of questions greater than 20?**

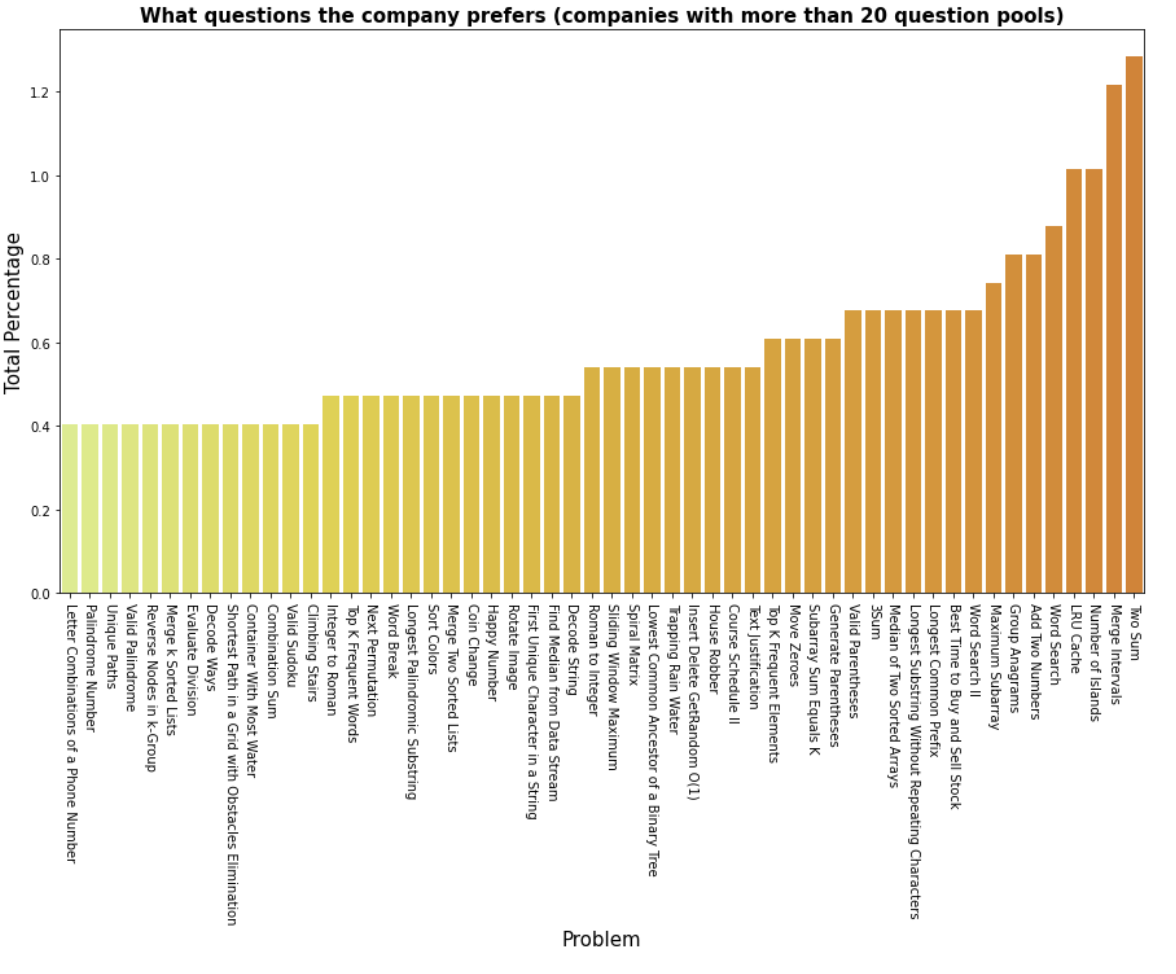
Of the 273 companies, a total of 23 companies have a total number of questions greater than 20.



Use the value\_counts() function to count the duplicate values of each question in the question bank of these 82 companies.



The average value of the number of occurrences of all problems is 0.16% of the total, and the maximum value of the number of occurrences of problems is 1.2% of the total, which are not large values. In order to analyze these data effectively, we only take the data with the value of the number of problems above 0.4% for visualization.



As can be seen from the graph, companies with a total number of questions greater than 20 favored these questions: Two Sum, Merge Intervals, Number of Islands, LRU Cache, Word Search, Add Two Numbers, Group Anagrams

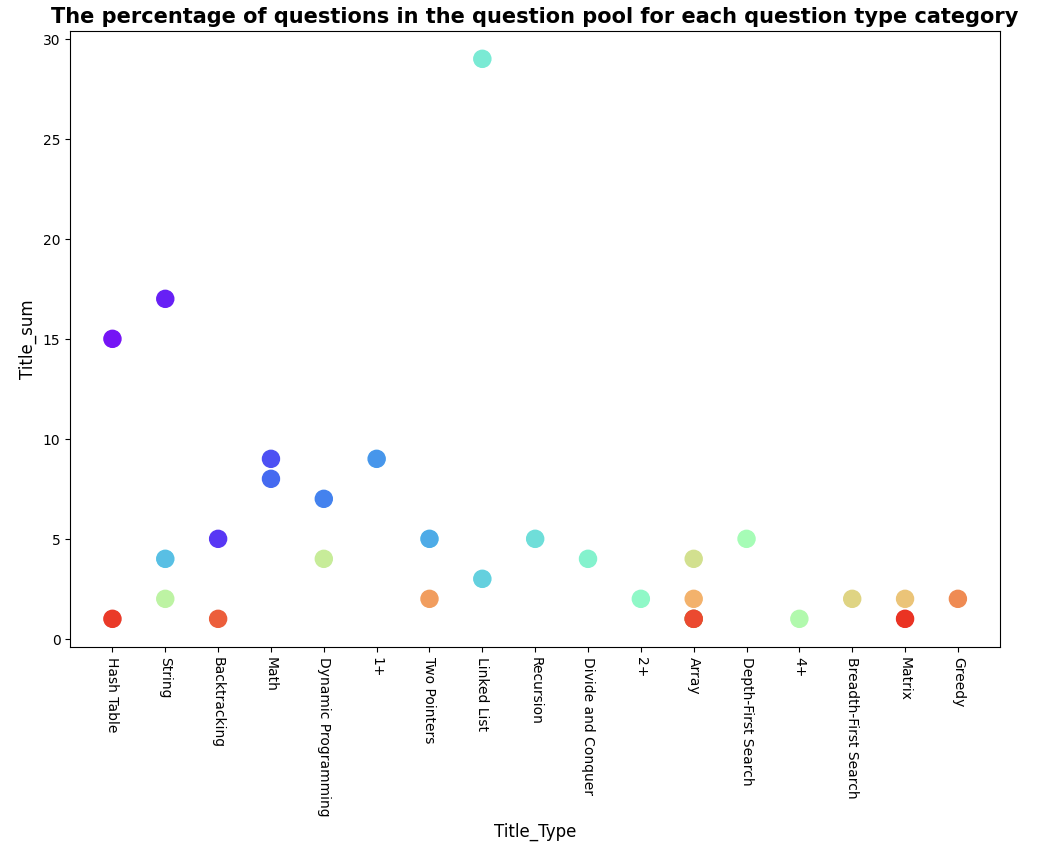
* 1. **The percentage of questions in the question bank for each question category**

First, only the column with the title of the data frame is selected, and the deviated data are processed by the regular method and then merged into the original data frame.

Since the question categories are stored in arrays, we then use a for loop to separate each word; then we use the Counter() function to count the number of occurrences of each category of words and store them in the list.

Finally, the number of question types is visualized.

As can be seen from the graph, the Linked List question type has the highest number of occurrences, followed by the String question type, closely followed by Hash Table, Math, 1+ question type, and the remaining 22 question types have a count of 7 or less.

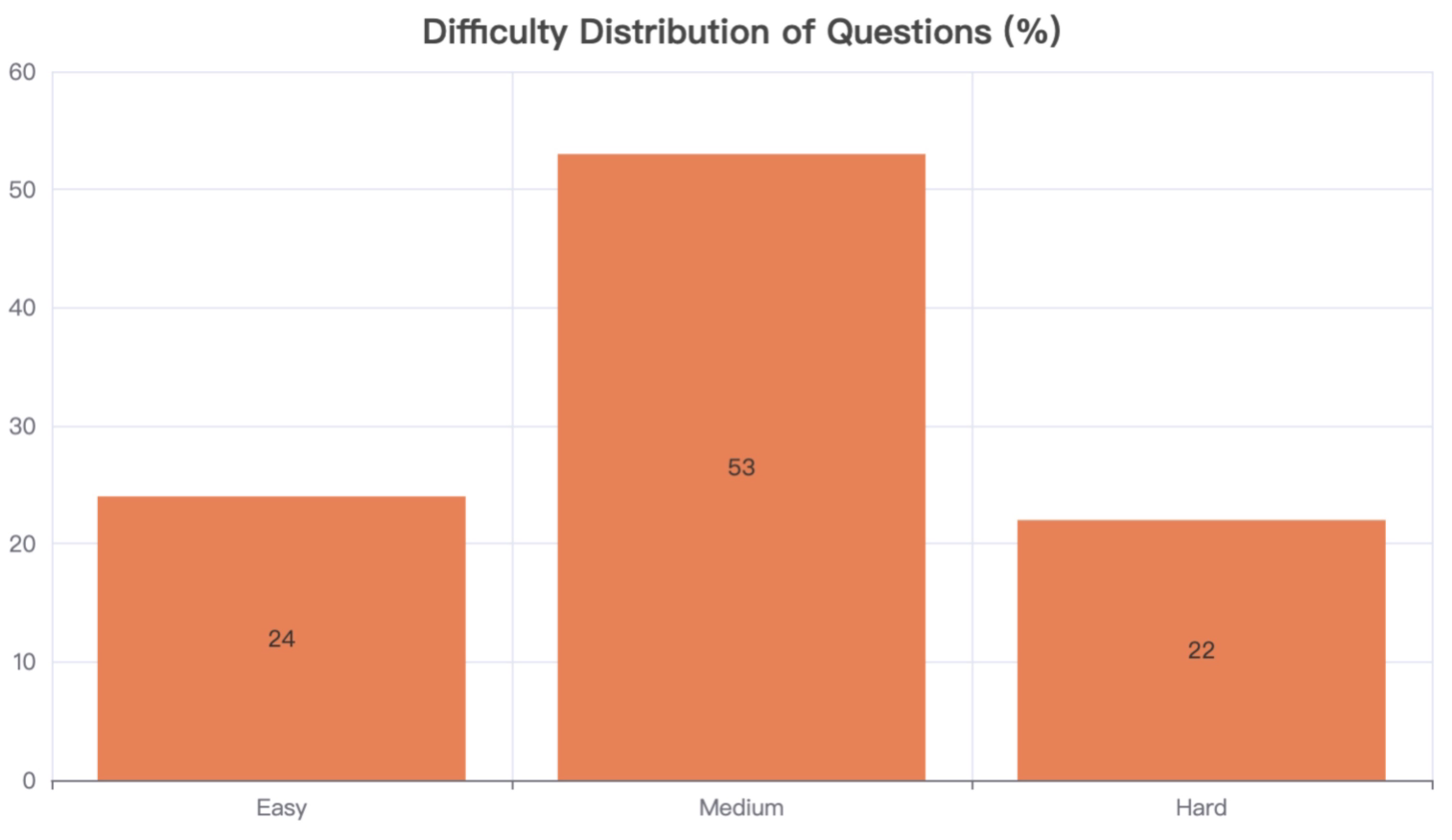


* 1. **What’s the difficult distribution of questions from question bank**

Knowing the difficult distribution of the questions is profitable for students to practice targeted exam questions. Therefore, we add a part exploring the difficult distribution of questions from the question bank.

Firstly, we use the function of value\_counts() to require the number of difficulty distributions, then calculate the percentage of difficulty distribution and interactively visualize the results by the pie charts package.

The plot shows median level questions account for over 50%, while the proportion of easy and hard level questions is about 20%. That means students can concentrate on the median level’s questions when preparation time is limited.



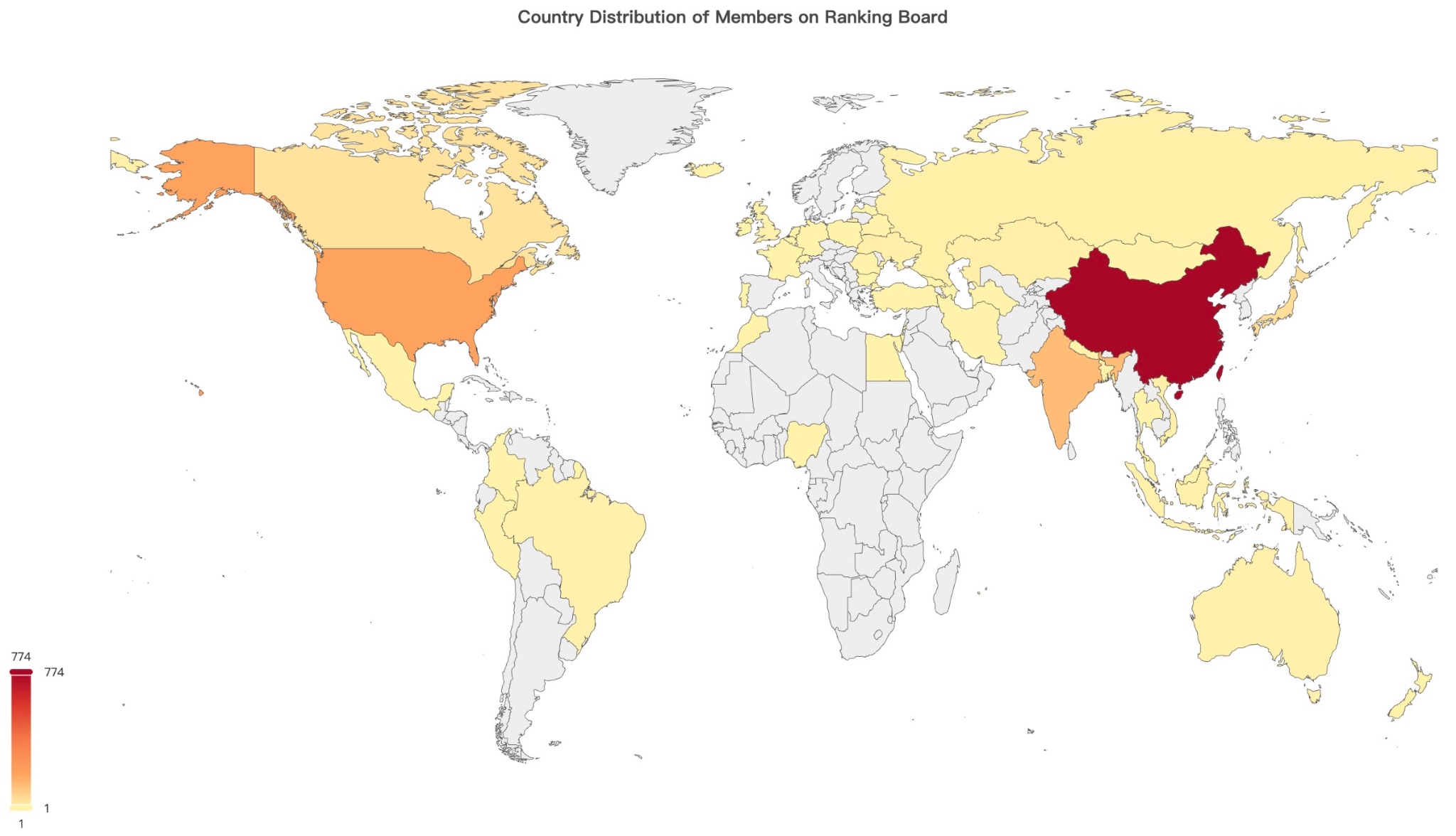
* 1. **What’s the country distribution of members on the ranking board**

LeetCode raises contests weekly and publishes online ranks. We collect 1725 ranking data and explore the country distribution of members on the ranking board.

In the preprocessing stage, we delete the rows with no value in the country column and replace Chinese values with English values.

Then in the visualization part, we visualize the global distribution by a map from the pie charts package.

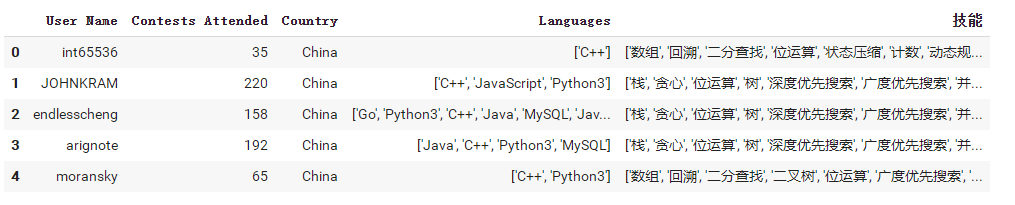
This map reflects most members are from China, then from The United States and India, which means Chinese programmers perform well in LeetCode's races.

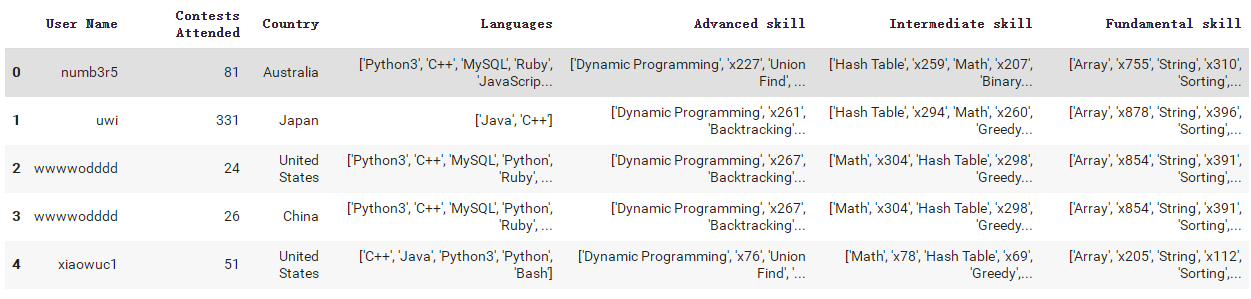


* 1. **What language(s) and skill(s) do worldwide users prefer?**

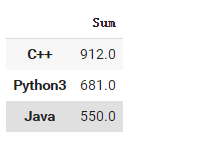
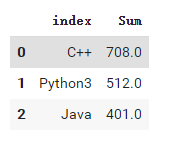
LeetCode has worldwide users, each user is capable of using more than one programming language, such as C++, C, Python. What’s more, each one is equipped with varied skills, like String, sorting. We want to know what are the most common and preferred programming languages and skills among world users. It can offer help to readers who want to prepare for IT related interviews.

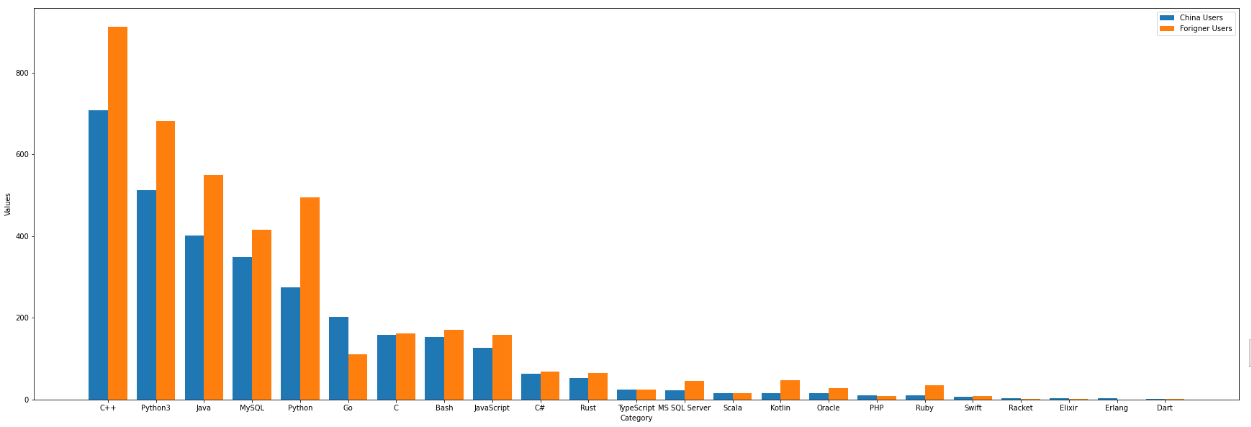
As for the data of Chinese users and foreign users are a bit different (*the below screenshots*), we separate worldwide users into two parts, namely China users and foreigner users.





By calculating, we get the number of each language frequency. We can see the differences in the 3 most popular languages between China users (*left*) and foreigner users (*right*).

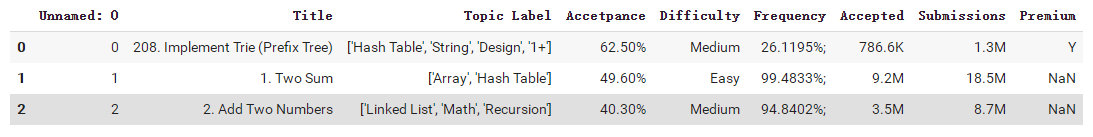


The double-column bar chart shows the worldwide language distribution between China users and foreigner users. Although there are differences in the number of users of each language, C++, Python3, Java, MySQL, Python are the 5 most popular programming languages, while Dart has the least demand. 

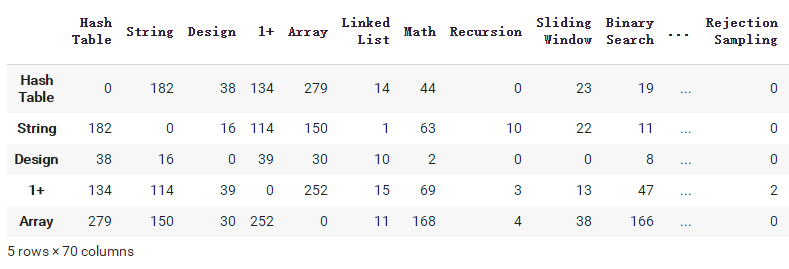
What’s more, the number of foreigner users’ application of the 5 most popular languages are higher than China users.

* 1. **What’s the correlation among subsets in topic labels?**

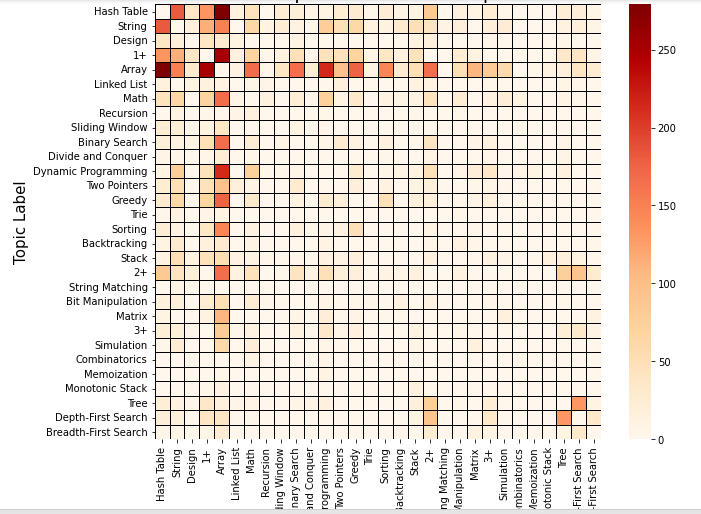
As the data shows, each problem topic (the Title column) has several subsets of topic labels (the Topic Label column ), by knowing the correlation among those subsets in topic labels we can have a clear idea of which subset is more related to which subset.



We use a matrix to calculate the correlation in order to see how many times each one relates to each other.



Then we use the hot-map to visualize the correlation distribution. The darker the color the more correlated between two subsets. As we can see, ‘Hash Table’ and ‘Array’ have the highest correlation; ‘Array’ and ‘1+’ have the second highest correlation. Other highly correlated matches include ‘Array’ and ‘Dynamic Programming’, ‘Array’ and ‘Math’, ‘Array’ and ‘Binary Search’ , and etc..

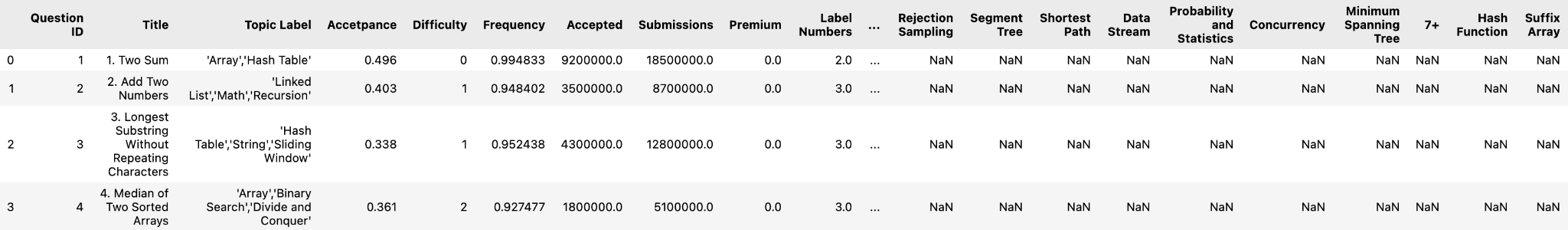


From the matrix, what cannot be missed is the significance of Array, as the most obvious two orthogonal straight lines show. It gives a hint that Array is the most ‘flexible’ and ‘malleable’ one to be connected with other parts.

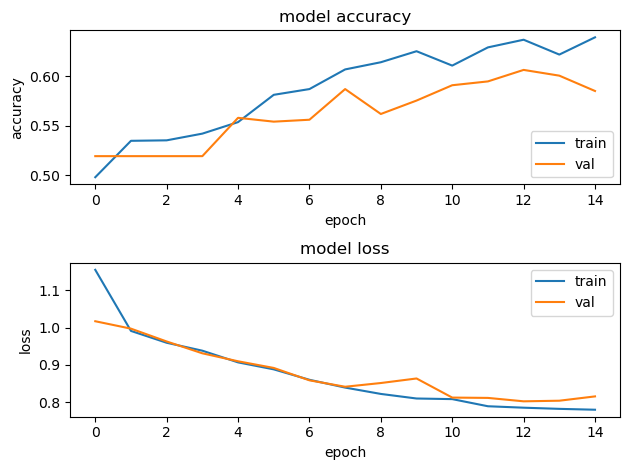
As a result, for anyone who is learning or preparing for an interview, by using this map, (s)he can know what needs to be learned and reviewed together more effectively.

* 1. **Using Machine Learning to Evaluate the Difficulty of Exam Questions**

We preprocessed the problem dataset: splitting all the labels in the 'topic label' column and found a total of 70 unique non-repeated labels. Then, we inserted these labels into the dataframe which is shown below figure and calculated the labels involved in each question for subsequent machine-learning modeling. We utilized the TensorFlow package to build a neural network model. For the input layer, it is used by 76 features from the problem dataframe, excluding the 'Difficulty' column. The 'Difficulty' column was used as the output. Additionally, 20% of the data was used for testing, while 80% of the data was used for training.



The following graph shows the model accuracy and model loss for the training and test sets during the modeling process. The blue line represents the training set, while the orange line means the test set. Based on the results, neither the test set nor the training set's accuracy surpassed 70%. From the perspective of model evaluation, this model is not considered very good. We attempted several methods, such as adding L1 and L2 regularization, reducing the number of neural network layers, and adjusting the number of training iterations. During this debugging process, the best result achieved was: the training set's accuracy exceeded 70%, while the test set's accuracy remained between 55% and 64%.



1. **Conclusion**
2. Companies with a total number of questions greater than 20 favored these questions: Two Sum, Merge Intervals, Number of Islands, LRU Cache, Word Search, Add Two Numbers, and Group Anagrams.
3. Most popular question types: Linked List, String, Hash Table, Math, 1+
4. Median-level questions account for over 50% of the question bank.
5. Most members of the ranking board are from China.
6. The 5 most popular languages among worldwide users: C++, Python3, Java, MySQL, Python
7. The 5 most correlated subsets are: ‘Hash Table’ and ‘Array’, ‘Array’ and ‘1+’, ‘Array’ and ‘Dynamic Programming’, ‘Array’ and ‘Math’, ‘Array’ and ‘Binary Search’. From which we can tell ‘Array’ is the core part to prepare for an interview.
8. **Workload distribution**