**Phase II : Decision Making**

**Section 1: Used Visualization Tools**

We will use Tableau to visualize the given data for several reasons:

1. Tableau offers the ability to create interactive dashboards that allow users to explore datasets in real-time.

2. Tableau provides a wide range of visualization options, making it effective in representing different data attributes such as categorical variables and numerical metrics.

3. The user-friendly interface of Tableau simplifies the process of creating visualizations, making it accessible to users with varying levels of technical expertise.

4. Tableau seamlessly integrates with external data sources, enabling the visualization of real-time or constantly updated data.

5. Tableau is capable of handling large datasets efficiently, ensuring consistent performance regardless of the dataset size.

6. With its publishing and sharing features, Tableau makes it easy for users to share and collaborate on visualizations.

**Section 2: Explanation of required data pre-processing:**

In our group project, we are focusing on visualizing the LeetCode dataset to gain insights into problem difficulty, acceptance rates, and engagement metrics. To ensure effective data preprocessing, we emphasize the importance of a collaborative approach. Here is a detailed outline of our data processing plan:

Data Cleaning

1. Identification of Missing Values: We will collectively review and determine the best approach to handling missing data in key columns like difficulty, acceptance, likes, dislikes, and discussion\_count.

2. Imputation: In cases of missing numerical values like acceptance, we will reach a consensus on whether to impute using the median or mean based on the data distribution.

3. Removal: If the percentage of missing critical data is minimal and imputation could distort the data, we may opt to remove those rows.

4. Handling Outliers: Addressing outliers in numerical columns is essential.

Data Transformation

1. Normalization and Binning: We will normalize metrics like likes and dislikes as necessary and convert continuous data such as acceptance rate into categorical bins for specific visualizations.

2. Collaborative Binning Strategy: Categories like 'Low', 'Medium', 'High' acceptance will be established through discussion to simplify analysis and visualization.

Data Validation

1. Ensuring Consistency and Accuracy: Our team will guarantee the consistent application of transformations and cleaning steps throughout the dataset. We will validate the data to ensure it adheres to the expected formats and falls within the specified value ranges.

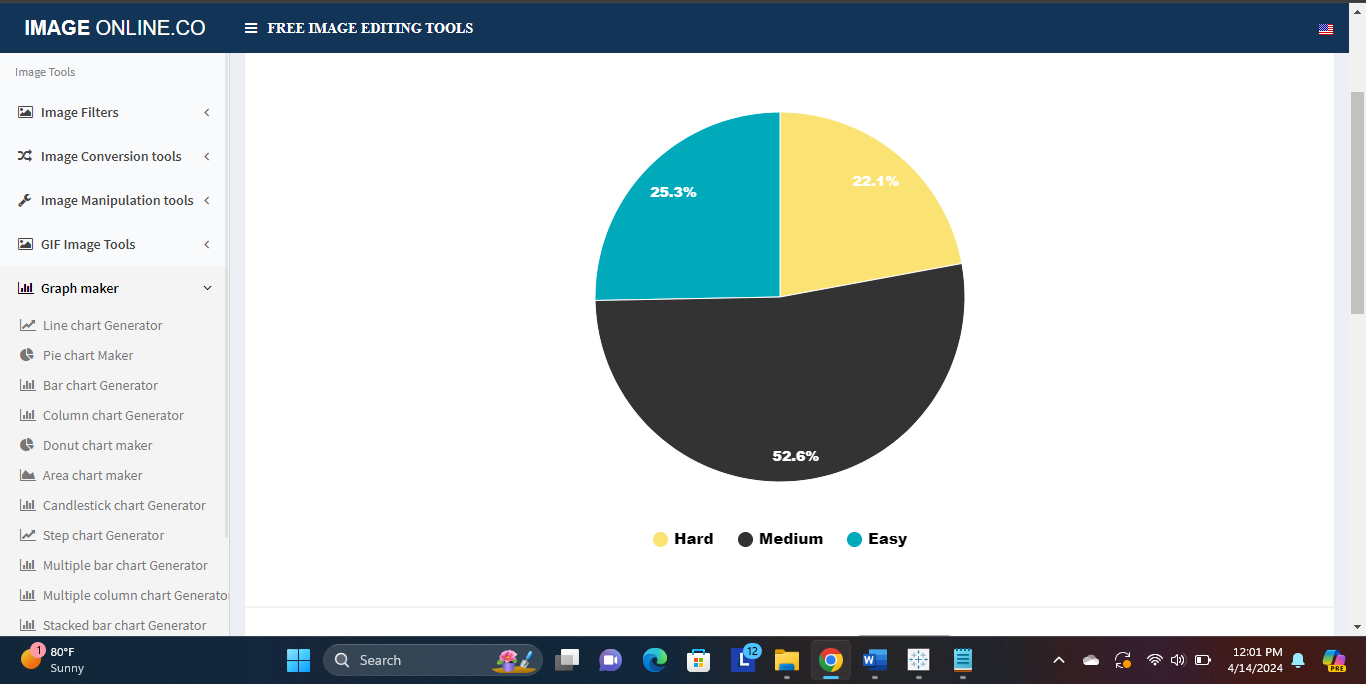
2. Previewing and Testing Visualizations: Prior to finalizing the dashboard, we will conduct a thorough examination of the data using test visualizations. This will help us identify any unforeseen issues or errors in data processing.

**Section 3: List of Final Set of Questions**

1. What is the distribution of problem difficulty levels among the LeetCode problems available in the dataset?
2. How does the acceptance rate vary across different problem difficulty levels?
3. What are the most popular topic tags among the LeetCode problems, and how do they correlate with problem difficulty?
4. Which LeetCode problems have the highest acceptance rates?
5. How does the engagement level (likes, dislikes, discussion counts) vary across different problem difficulty levels?
6. What are the most frequently discussed problems, and what insights can we gain from these discussions?
7. What are the most challenging problem categories based on submission frequency and acceptance rates?
8. What are the top 20 problem tags for recruitment and educational needs?
9. What is the problem difficulty to engagement metrics correlation?
10. What is the likes to dislikes ratio by problem difficulty?
11. What is the relationship between problem difficulty and number of submissions and acceptance rate?
12. What is the cumulative distribution of problem acceptance?

**Section 4: Dashboard Plot Drafts**

**1. Distribution of Problem Difficulty Levels**



Plot: Pie Chart

What it shows: The distribution of problem difficulties within the dataset is depicted in the pie chart, which showcases segments representing Easy, Medium, and Hard problems. The size of each segment corresponds to the proportion of problems classified under their respective difficulty category.

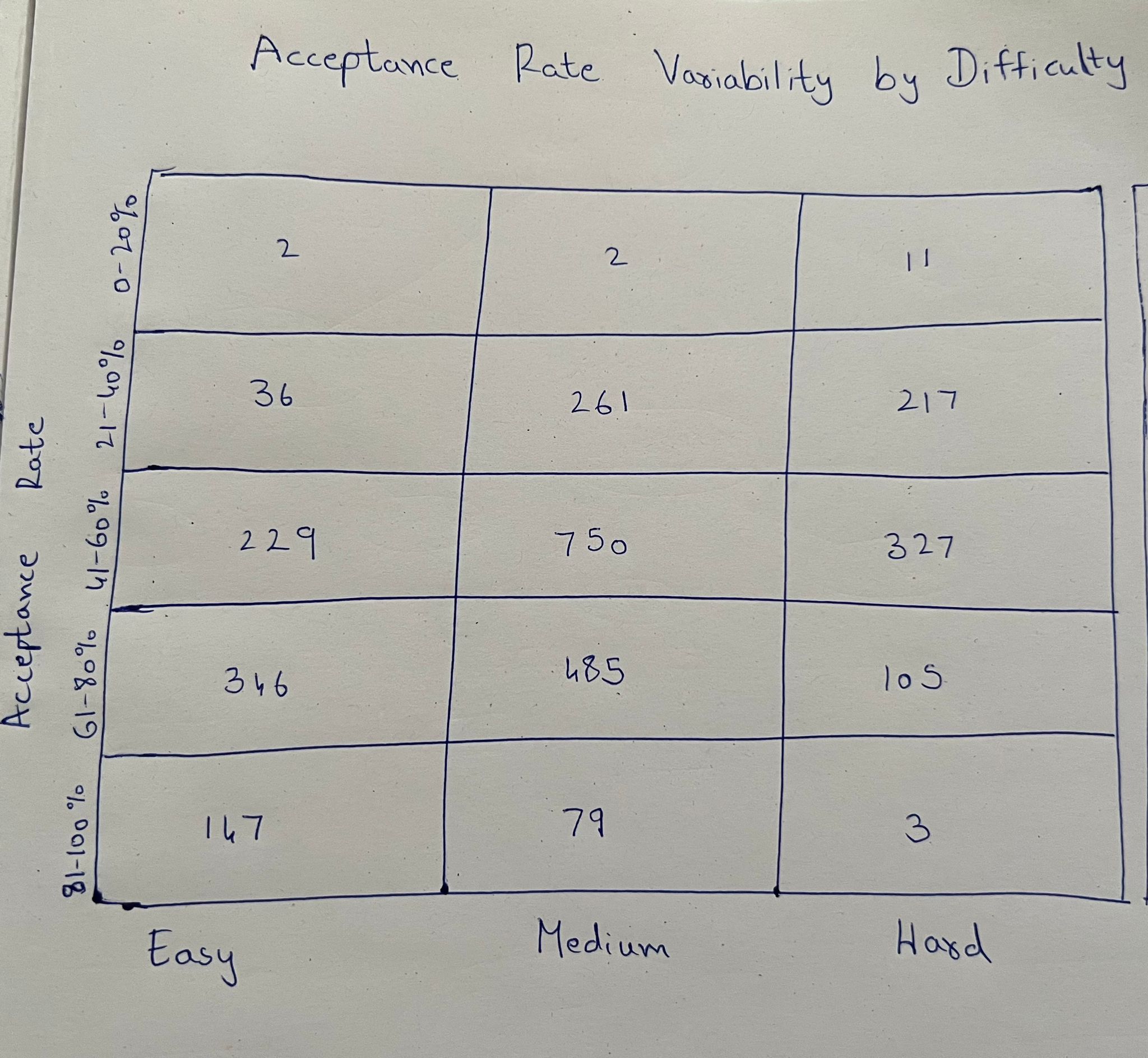
Relation to the question: This visualization effectively tackles the inquiry regarding the distribution of problem difficulties by visually dissecting the complete range of problems into distinct categories of Easy, Medium, and Hard.

Pre-attentive Attributes:

Color: Utilizes distinct shades for each difficulty level

Size: The angular size of each pie slice visually encodes the proportion of each difficulty level.

**2. Acceptance Rate Variability by Difficulty**

****

Plot: Heat Map

What it shows: The heatmap illustrates the distribution of problems based on acceptance rate bins and difficulty levels.

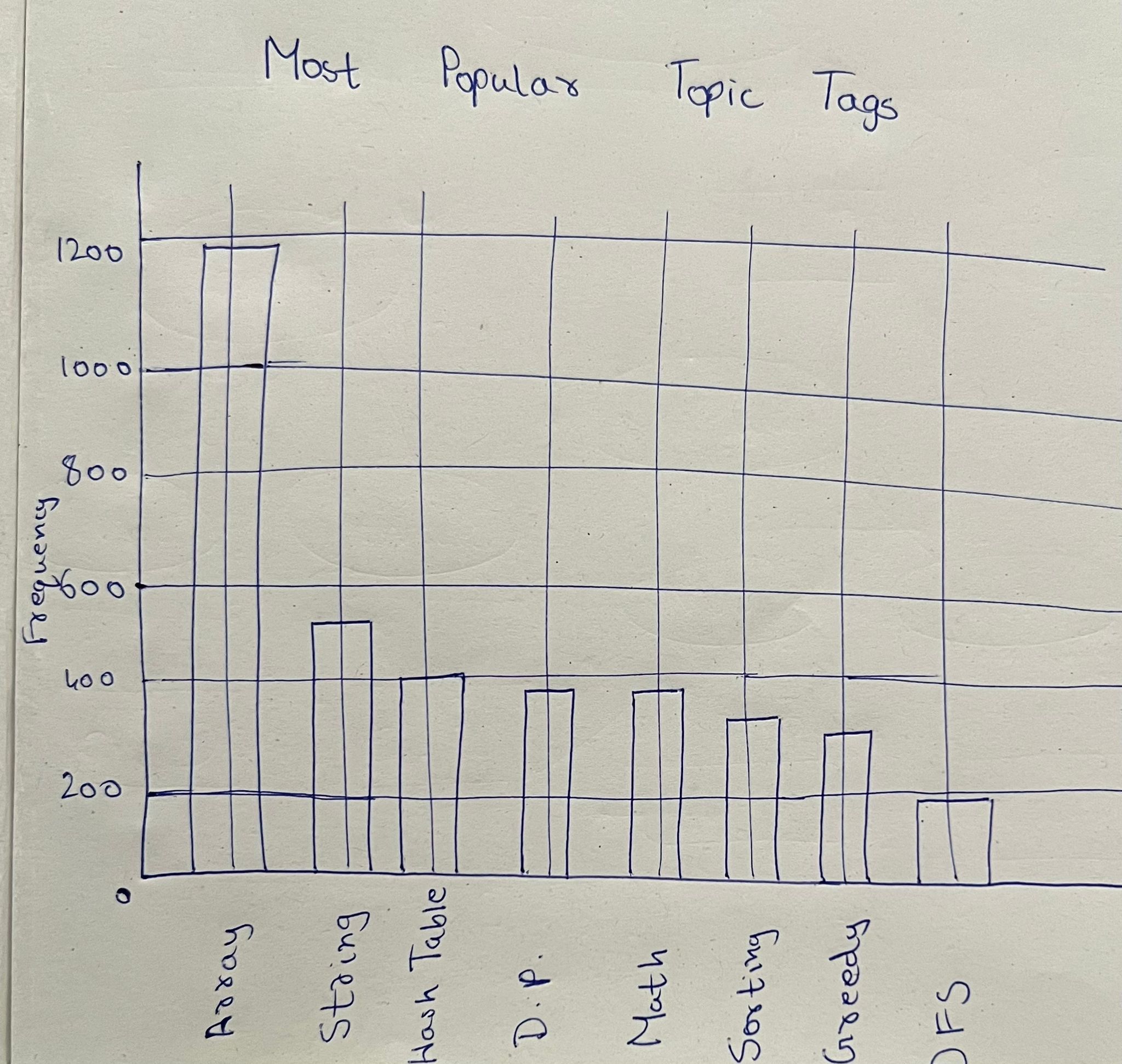
Relation to the question: This visualization effectively addresses the question by displaying the distribution of acceptance rates across different levels of problem difficulty. It offers a straightforward depiction of where problems are concentrated in terms of acceptance, aiding in the identification of which difficulty levels may pose more or less of a challenge depending on the number of problems falling into lower or higher acceptance rate bins.

Pre-attentive Attributes:

Color Intensity: Darker shades represent greater quantities of problems within particular acceptance rate categories.

Position: Arranged in a systematic manner according to acceptance rate bins and difficulty levels.

3. **Most Popular Topic Tags with Difficulty**

****

Plot: Combination of Bar Chart and Box Plot

What it shows:

Bar Chart: The chart exhibits the occurrence frequency of each topic tag in the dataset. Each bar represents a specific topic tag, and its height indicates the number of problems associated with that tag.

Box Plot: In addition to the bars, a box plot is displayed, providing a comprehensive view of the distribution of acceptance rates for problems within each tag. It includes the median, quartiles, and potential outliers, offering insights into the difficulty level of the problems associated with each tag.

Relation to the question: This combined visualization effectively addresses the question by not only showcasing the popularity of topic tags (as depicted in the bar chart) but also illustrating the correlation between those topics and problem difficulty (as demonstrated in the box plot).

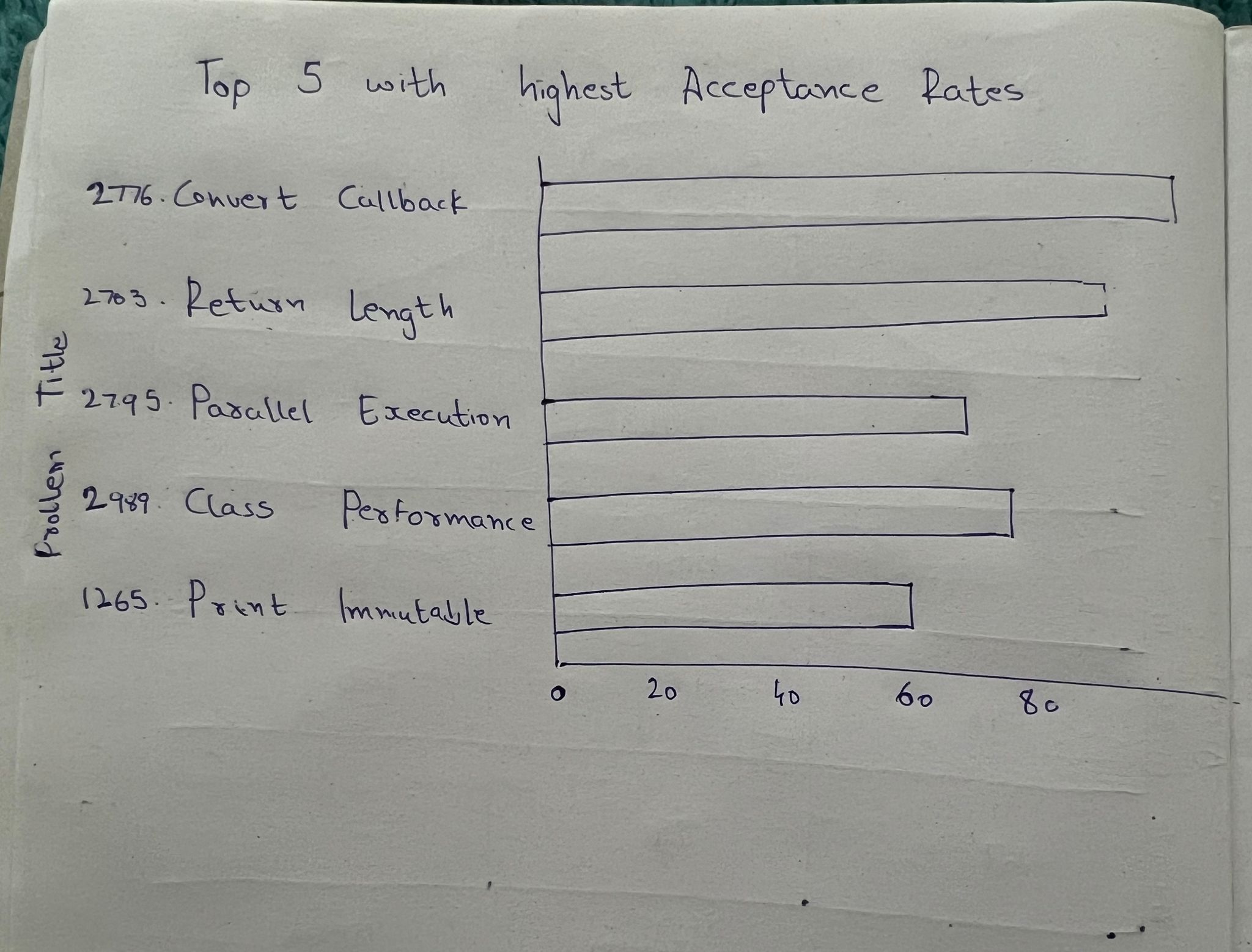
Pre-attentive Attributes:

Size: The size of each bar in the bar chart indicates the popularity of each tag, facilitating easy comparison between tags.

Color: Different colors can be utilized to differentiate between various topic tags in the bar chart, enhancing visual clarity.

Position: By placing bars and box plots closely together, it enables a comprehensive examination of the relationship between frequency and difficulty.

**4. Problems with the Highest Acceptance Rates**

****

Plot: Horizontal Bar Chart

What it shows:

The problems with the highest acceptance rates in the dataset are illustrated in the horizontal bar chart. Each bar corresponds to a specific problem, with the length of the bar reflecting the percentage of acceptance rate.

Relation to the question:

This chart provides a direct response by visually presenting the problems with the highest acceptance rates, offering a clear indication of the problems that are solved successfully most often.

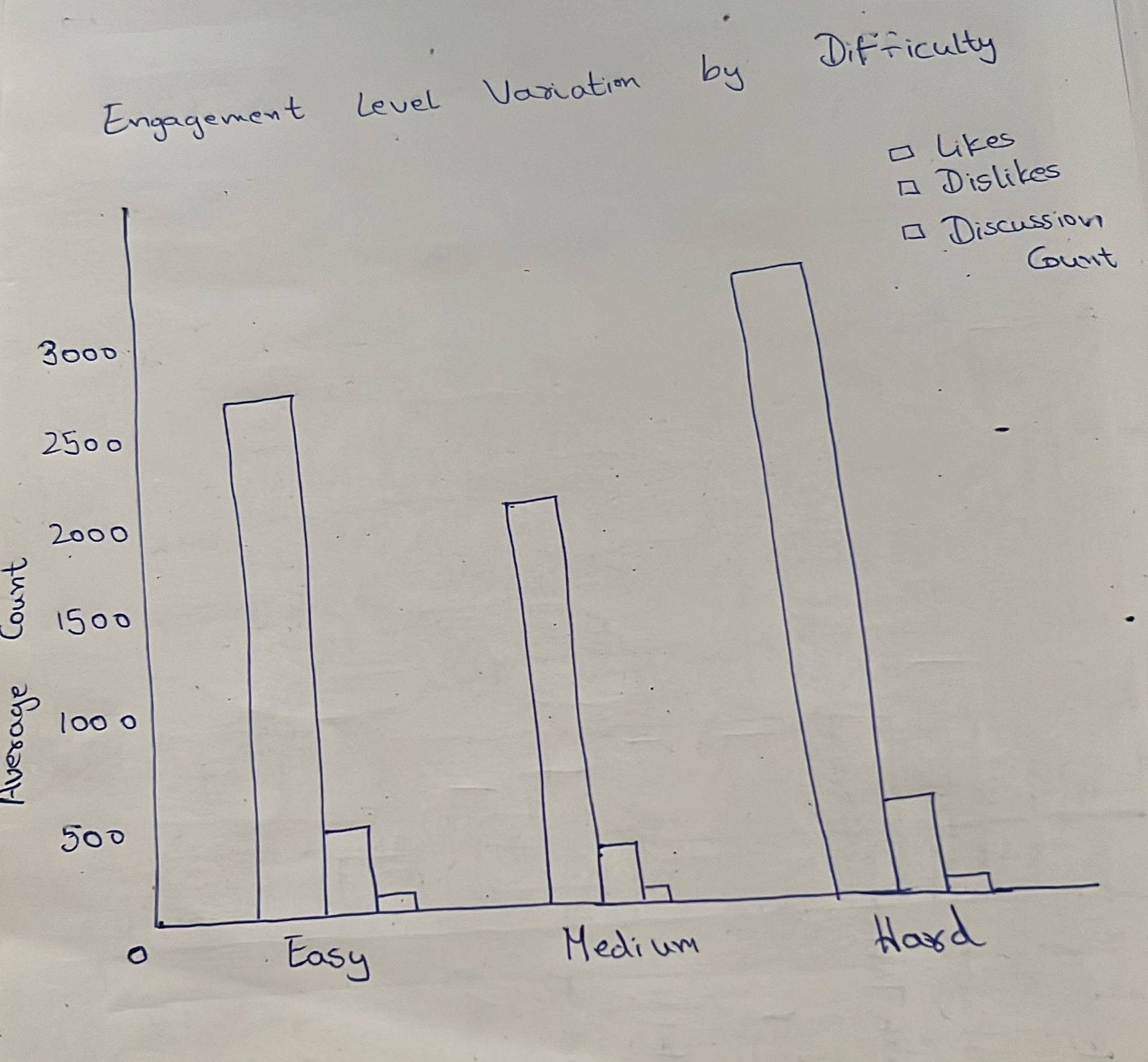
Pre-attentive Attributes and Colors:

Length: The length of each bar is proportional to the acceptance rate.

Color: Different color is used for each bar

Position: Bars are aligned horizontally, which can be easier to read and compare than vertical bars.

**5. Engagement Level Variation by Difficulty**

****

Plot: Grouped Bar Chart

What it shows:

The bar chart presents the mean values of likes, dislikes, and discussion counts for each level of problem difficulty (Easy, Medium, Hard).

Relation to the question:

The chart provides a clear answer to the question by illustrating the variations in engagement metrics based on different levels of problem difficulties. Stakeholders can easily identify which difficulty levels generate higher levels of discussions, likes, or dislikes.

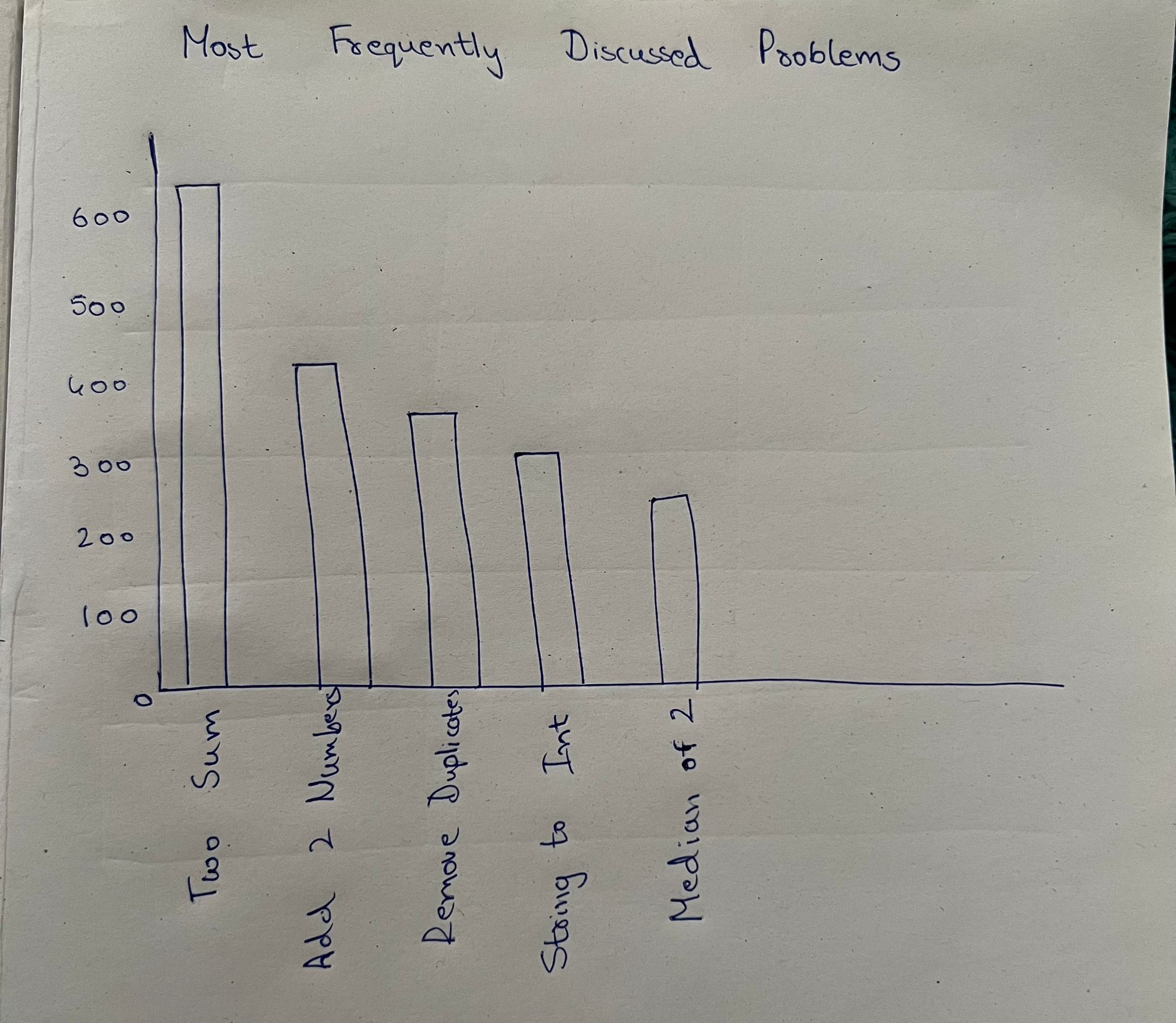
Pre-attentive Attributes:

Length: The length of each bar within the groups represents the average count of each metric.

Color: We can use 3 colors each for each engagement level

Position: The bars are grouped by difficulty level.

**6. Most Frequently Discussed Problems**

****

Plot: Bar Chart

What it shows:

The bar graph illustrates the hierarchy of LeetCode problems according to the number of discussions they have generated. Each bar corresponds to a particular problem.

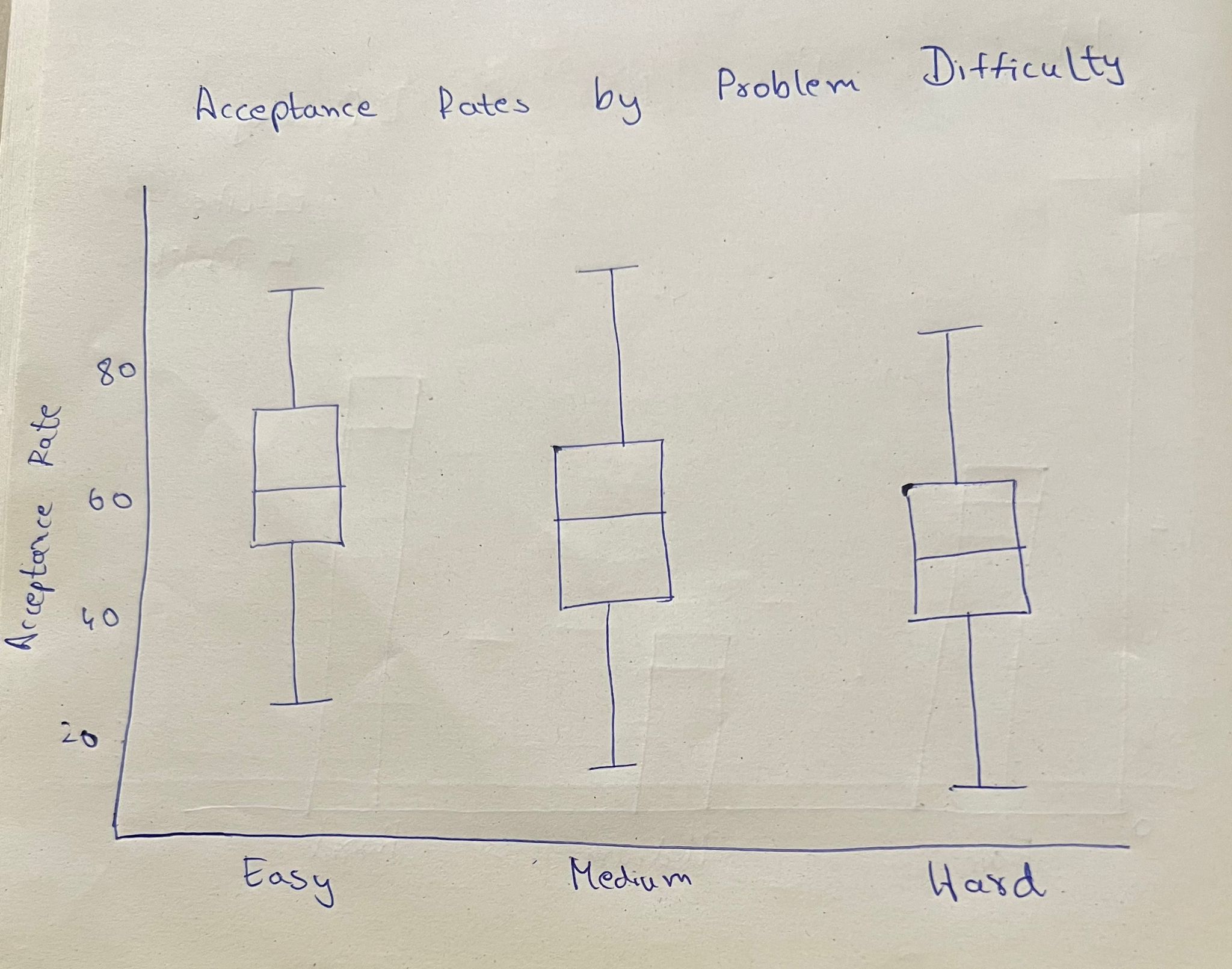
Relation to the question:

This visual representation effectively addresses the query by categorizing and ordering problems based on the volume of discussions they provoke.

Pre-attentive Attributes:

Length: The primary pre-attentive attribute used in this chart is the length of the bars. Longer bars indicate a higher number of discussions.

**7. Most Challenging Problem Categories**

****

Plot: Box plot

What it shows:

The Box and Whisker Plot illustrates the distribution of acceptance rates for LeetCode problems, which are classified into different difficulty levels (Easy, Medium, Hard). Each box represents the interquartile range (IQR) of acceptance rates within its respective difficulty category, showcasing the middle 50% of the data. The horizontal line inside the box represents the median acceptance rate.

Relation to the question:

This visualization is crucial in comprehending the level of difficulty associated with the problems in each category, as it provides insights into their acceptance rates.

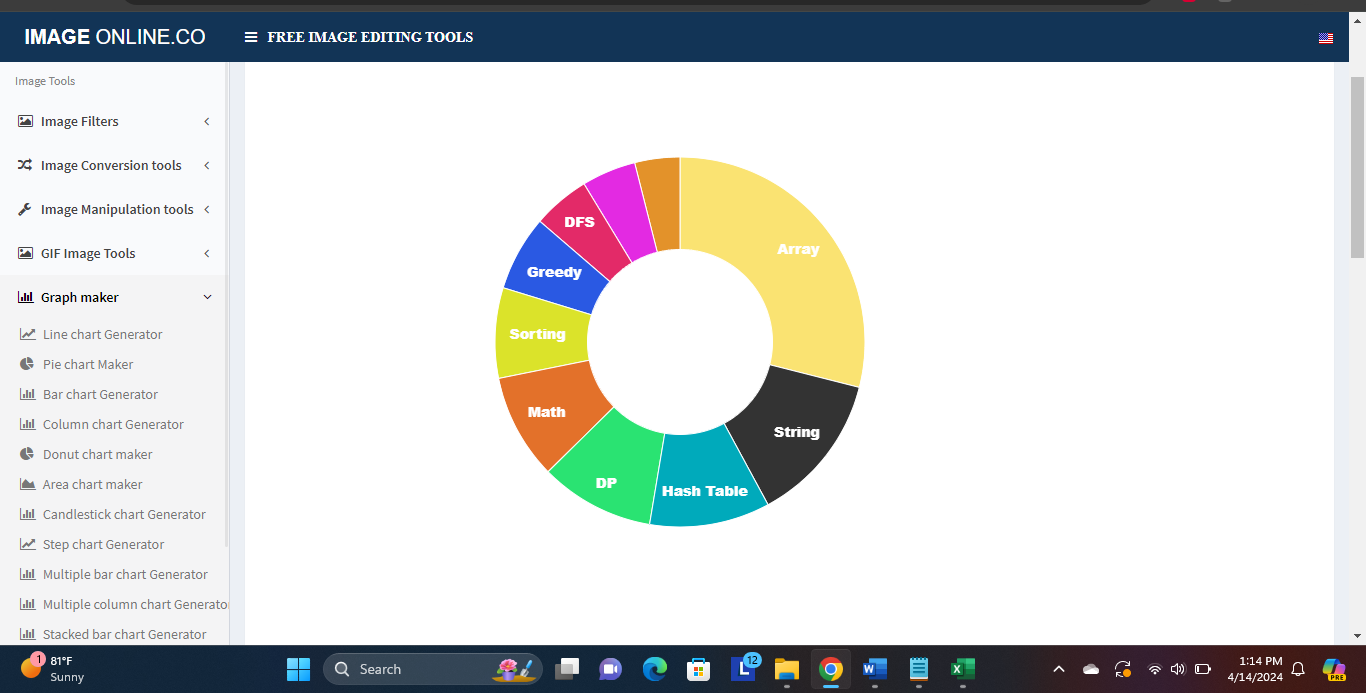
Pre-attentive Attributes:

Length: The range of the box and whiskers provides visual cues about the variability of the acceptance rates.

Color: The boxes are colored with one color for clear visibility, with a different color formedians to emphasize the middle value.

Position: Each box is aligned with its corresponding difficulty category on the x-axis, making it easy to compare across categories.

**8. Top 20 Problem Tags for Recruitment and Educational Needs**

****

Plot: Donut Chart

What it shows: The donut chart visualizes the top 20 problem tags as segments of a donut, with each segment's size proportional to the frequency of that tag in the dataset.

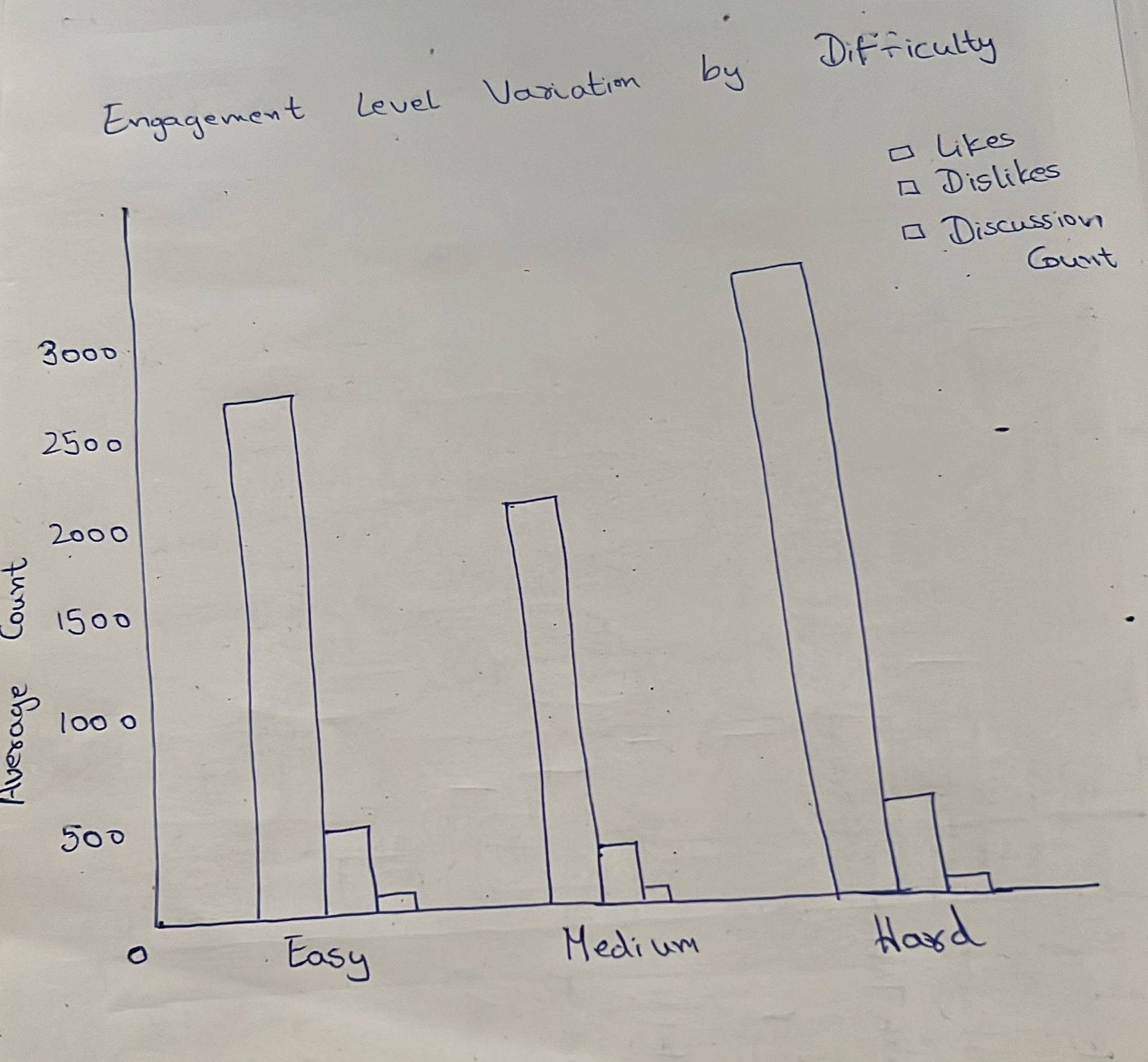
Relation to the question: The visualization directly addresses the question by showing which problem tags are most prevalent.

Pre-attentive Attributes:

Size: The angular extent of each segment shows the tag's frequency relative to others.

Color: Distinct colors for each segment help to differentiate the tags visually.

**9. Problem Difficulty to Engagement Metrics Correlation**

****

Plot: Grouped bar chart

What it shows: The Grouped Bar Chart displays the average values of likes, dislikes, and discussion counts for problems grouped by their difficulty levels (Easy, Medium, Hard).

Relation to the question: This visualization directly addresses the question by showing the correlation between problem difficulty and user engagement metrics.

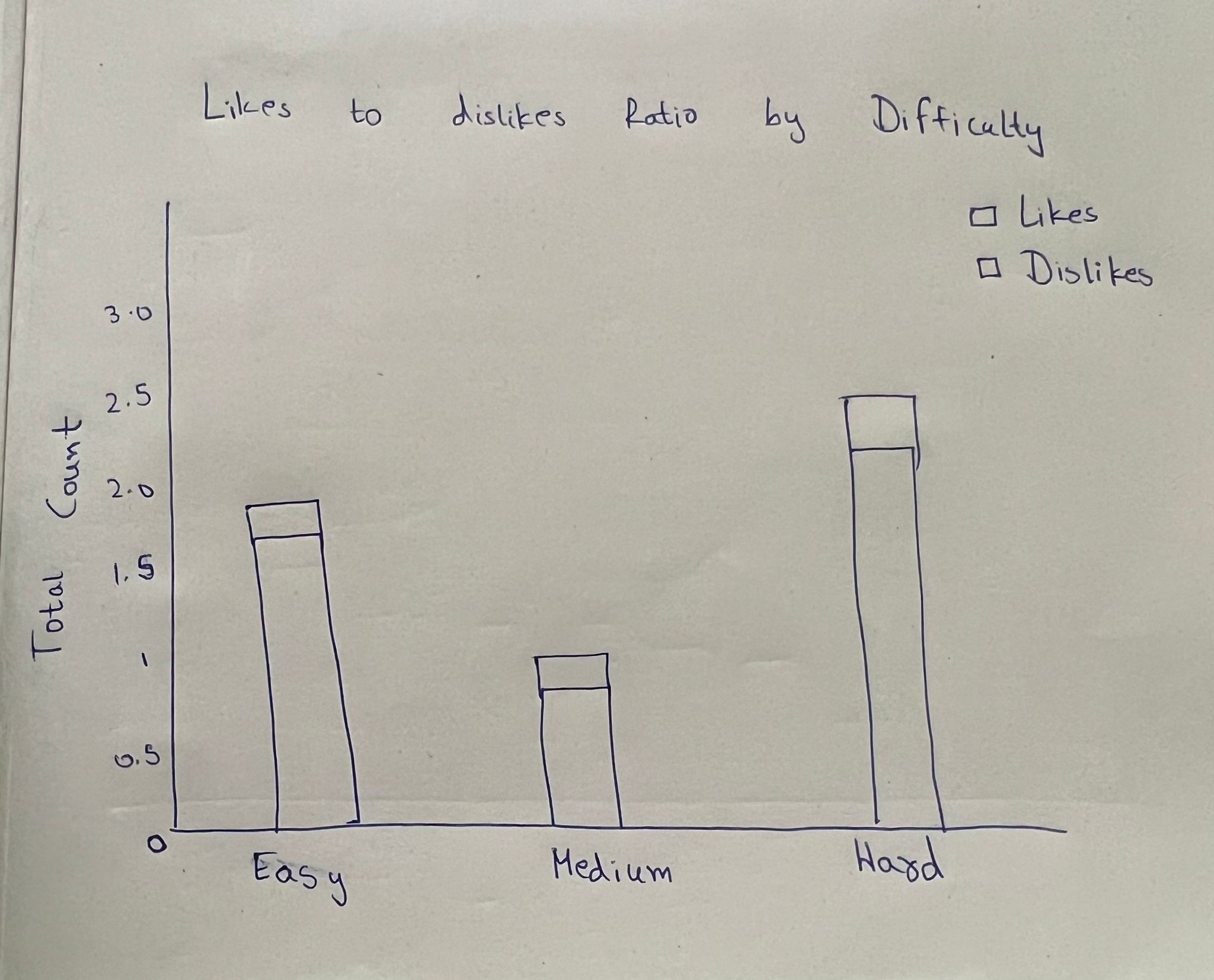
Pre-attentive Attributes:

Position: The grouped bars are positioned according to the problem difficulty on the x-axis.

Height: The height of each bar indicates the average count of each metric.

Color: Distinct colors are used for each metric.

**10. Likes to Dislikes Ratio by Problem Difficulty**

****

Plot: Stacked Bar chart

What it shows: The Stacked Bar Chart displays the total counts of likes and dislikes for each problem difficulty level stacked within each bar.

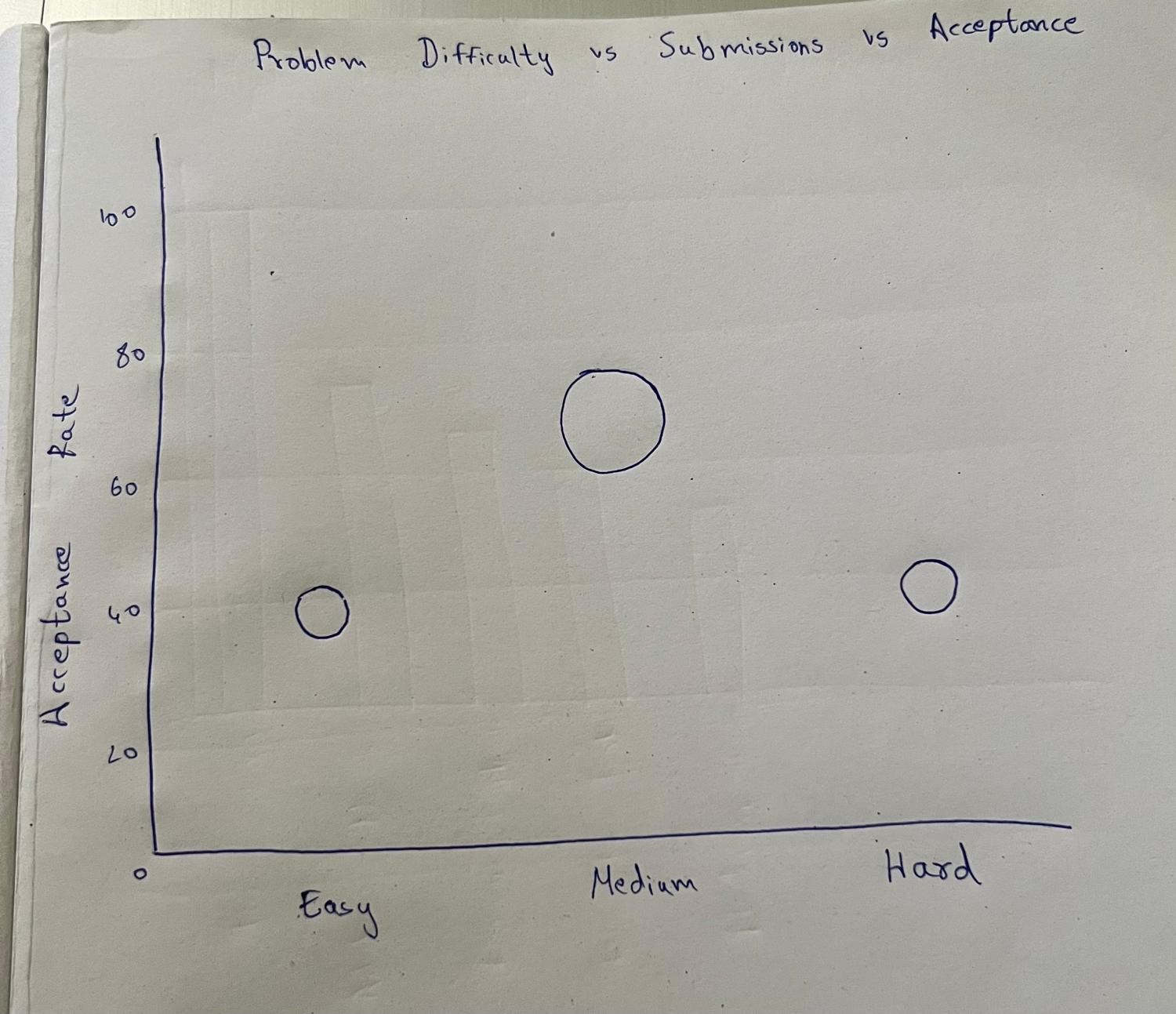
Relation to the question: This visualization directly addresses the question by showing the likes to dislikes ratio visually through the proportions of the stacked segments.

Pre-attentive Attributes:

Height and Proportion: The size of each segment within the bars communicates the absolute number and proportion of likes and dislikes.

Color: We use two different colors for likes and dislikes.

**11. Relationship Between Problem Difficulty, Number of Submissions, and Acceptance Rate**

****

Plot: Bubble Chart

What it shows:

The bubble chart is utilized for mapping problem complexity against approval rates, with individual bubbles symbolizing distinct problems. The horizontal axis denotes the complexity level, the vertical axis showcases the approval rate, and the magnitude of each bubble corresponds to the number of submissions.

Relation to the question:

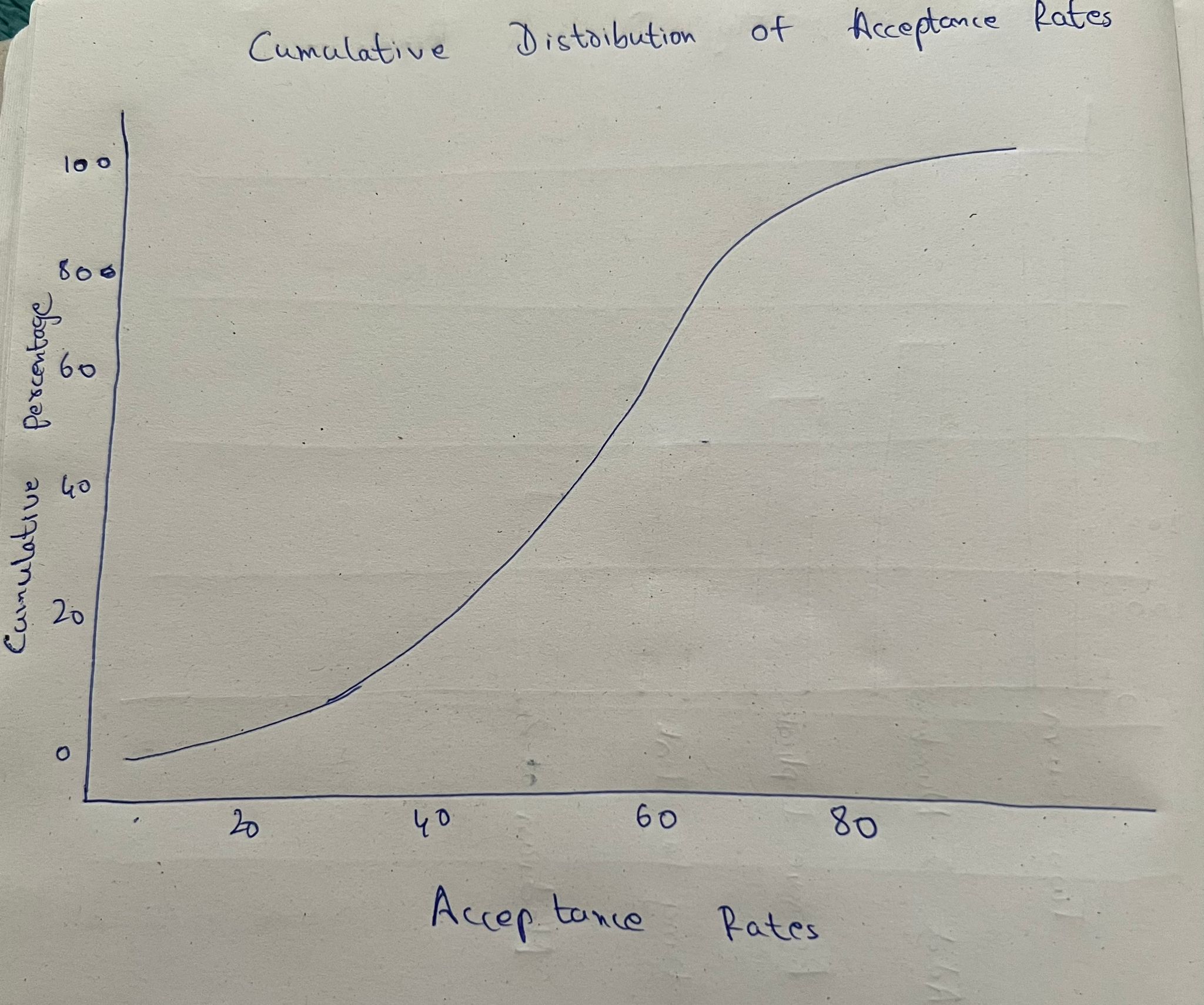
The bubble chart effectively tackles the question by demonstrating the relationship between problem complexity and achievement rates, all the while taking into account the problems' popularity.

Pre-attentive Attributes:

Position: The placement of bubbles along the x and y axes immediately informs about the difficulty and acceptance rate, respectively.

Size: The size of each bubble represents the number of submissions, with larger bubbles indicating a higher count of submissions.

**12. Cumulative Distribution of Problem Acceptance**



Plot: Line Graph

What it shows:

The line graph displays the Cumulative Distribution Function (CDF) of problem acceptance rates within the dataset.

Relation to the question:

The line graph provided offers a clear depiction of how problem difficulty correlates with acceptance rates, aiding stakeholders in grasping the distribution of problems across different difficulty levels.

Pre-attentive Attributes:

Position: The horizontal position of each point along the x-axis represents the acceptance rate, while the vertical position along the y-axis shows the cumulative percentage.

Slope: The slope of the line provides visual cues about the rate of change in acceptance rates across problems.

**Section 5 :**

1. **Difficulty Level Filter:**

**Purpose**: To sort information in different visualizations based on the level of difficulty of the problem (Easy, Medium, Hard).

**Connected plots**: This filter will impact all visualizations that include problem difficulty as a factor, including:

Distribution of Problem Difficulty Levels

Variability of Acceptance Rate based on Difficulty

Variation of Engagement Level based on Difficulty

Problems with the Highest Acceptance Rates

Likes to Dislikes Ratio based on Problem Difficulty

**Value Range**: The available options will be ['Easy', 'Medium', 'Hard'], directly retrieved from the 'difficulty' attribute in the data.

1. **Acceptance Rate Slider**

**Purpose**: Modifying visualizations according to various acceptance rates enables users to analyze the impact of different levels of problem acceptance on metrics.

**Connected Plots:** This slider will impact charts illustrating acceptance rates, including:

Acceptance Rate Variability by Difficulty

Problems with the Highest Acceptance Rates

Most Challenging Problem Categories

Relationship Between Problem Difficulty, Number of Submissions, and Acceptance Rate

Cumulative Distribution of Problem Acceptance

**Value Range**: 0% to 100%, allowing users to set minimum and maximum acceptance rate thresholds.

1. **Data Refresh Button**

**Purpose**: To reload or refresh the dashboard data, useful for keeping the visualizations up-to-date with the latest dataset changes.

**Connected Plots**: All plots on the dashboard.

**Value Range**: This control doesn’t have a value range but acts as an action trigger for data refresh.

**Section 6 :**

Link to the dataset -<https://www.kaggle.com/datasets/jaydeepagravat94583/leetcode>

Link to Mural - <https://app.mural.co/t/team605123/m/team605123/1712279236151/f7b58b853b8797e47d29e720117ffea18007cc61?sender=u989641d48ac1c05fb15d3600>