1. **Introduction**

As an unofficial source of documentation, Q&A sites such as Stack Overflow are immensely popular among programmers, and that popularity is only increasing as time passes. As of October 19, 2014, Stack Overflow has over 8.2 million questions and 14 million answers. A mere year ago, Stack Overflow had over 5.5 million questions and about 10.5 million answers. Both of these are nearly 50% increases over a single year. Stack Overflow is constantly growing, and shows no signs of slowing down. However, as with all forms of community-driven content, the quality of the material on Stack Overflow is subject to questioning. With the increasing popularity of Stack Overflow as not only a Q&A forum, but as a form of unofficial documentation to those who come after, low-quality content is a very bad thing to have.

The primary forms of quality control on Stack Overflow are the "voting" and "accepted answer" systems. For any given post, be it question or answer, any Stack Overflow user can vote the post either up or down, and these votes are aggregated into a "score" on the answer. The simplest and most obvious use of this is that posts that score higher are more visible. Users are encouraged to participate and vote through the Reputation system. Therefore, generally speaking, a highly-voted post will be a high-quality, useful post as judged by the Stack Overflow community while a low (or negative) scoring post has been judged to be of little or no value. On the other hand, the accepted answer system is only available to question askers. The asker of a question can choose to mark any answer to his question as the "accepted answer", causing it to appear at or near the top of the answer list for his question, along with the highly-voted posts. Practically, this is very similar to the voting system, though there is the key difference that the only person who determines the accepted answer is the question asker, instead of the whole community.

Our problem focuses specifically on the answers that have been deemed to be high quality. Specifically, we will be examining answers above a certain, as-yet-undetermined score threshold and the set of accepted answers. These answers will be evaluated against a set of metrics to answer the question of "What makes for a good answer on Stack Overflow?". We'll be examining the presence of code snippets, the presence of links, the length of the answer (in words), the answerer's reputation, and the response time for each answer to attempt to answer this question. The accepted answers and the highly-voted answers will, nominally, comprise two different data sets, but both are being measured in the same way. Our primary goal is to find trends in these metrics across the data, so that we can identify aspects of the "good" answers and, perhaps, be able to apply these trends to the general problem of writing better documentation, both in crowdsourced form and otherwise.

1. **Input and Output**

The two source sets of data we will be using are the "accepted anaswers" and "Highly voted answer" against the metrics of "Response Time", "Number of tags", Length of the answer (in words)", "reputaion of the answerer", "links provided as asnwers" and "code snippets"

We would extract this inofrmation by running various queries on the Stack dump provided onto an excel sheet or any tablular format for that matter and plot them againast the above two datasets of "highly voted answers" and "Accepted answers"

The OUTPUT would be in the form of bar charts and graphs to conduct analysis and draw conclusions from the chart statistics.

We can then put forward an analysis and conclusion stating how to analyse the quality of an answer to build a higher reputation and have your answer accepted and highly voted on Stack overflow.

1. **Analysis**
2. **Presence of Code Snippets**
3. **Presence of Links**
4. **Length of the Answer**
5. **Answerer’s Reputation**
6. **Response Time**
7. **Sources of Information**

The information we use in our analysis is derived from two sources – highly voted answers (HVA) and accepted answers (AA).

1. **Highly Voted Answers (HVA)**

Highly voted answers are a subset of all submitted Stack Overflow answers which contains only the highest rated answers as voted by the community. To obtain this subset, we first look at the user-voted rating of each answer and determine a rating threshold derived from the average score. The set of highly voted answers is then defined as all answers that exceed the average score.

1. **Accepted Answers (AA)**

Accepted answers are answers to questions which the asker felt most adequately answered the prompt. Contrary to highly voted answers, accepted answers do not have to meet a certain score threshold and, in extraordinary cases, may even be the least preferred answer as rated by the community.