# FAQ Bot for a Small Business Website – Dataset Analysis and Visualization Report

**By Gopalakrishnan Kumar, MTech IIT-Bombay,**

**Freelance Data Science Consultant**

LinkedIn: Profile Link : https://www.linkedin.com/in/gopalakrishnankumar-a73301110/

Github:https://www.github.com/Gopalakrishnan-Kumar/

Kaggle URL- https://www.kaggle.com/gopalkk2

# Google Colab URL

- https://colab.research.google.com/drive/1ZqtZ8AaGCJ2Em-_MRwoj9hT66vChoePB?usp=sharing

# Project Overview

- The goal of this project is to analyze and visualize the dataset used to build a FAQ (Frequently Asked Questions) chatbot for a small business website. The dataset consists of questions and their respective answers, with a focus on understanding the content structure, response length, and category-wise distribution. Such insights are valuable for refining chatbot training and improving customer experience.

# Dataset Description

- **Filename:** faq_dataset.csv
- **Fields:**
  - **Question:** The user's inquiry or query.
  - **Answer:** The chatbot's predefined response to the question.

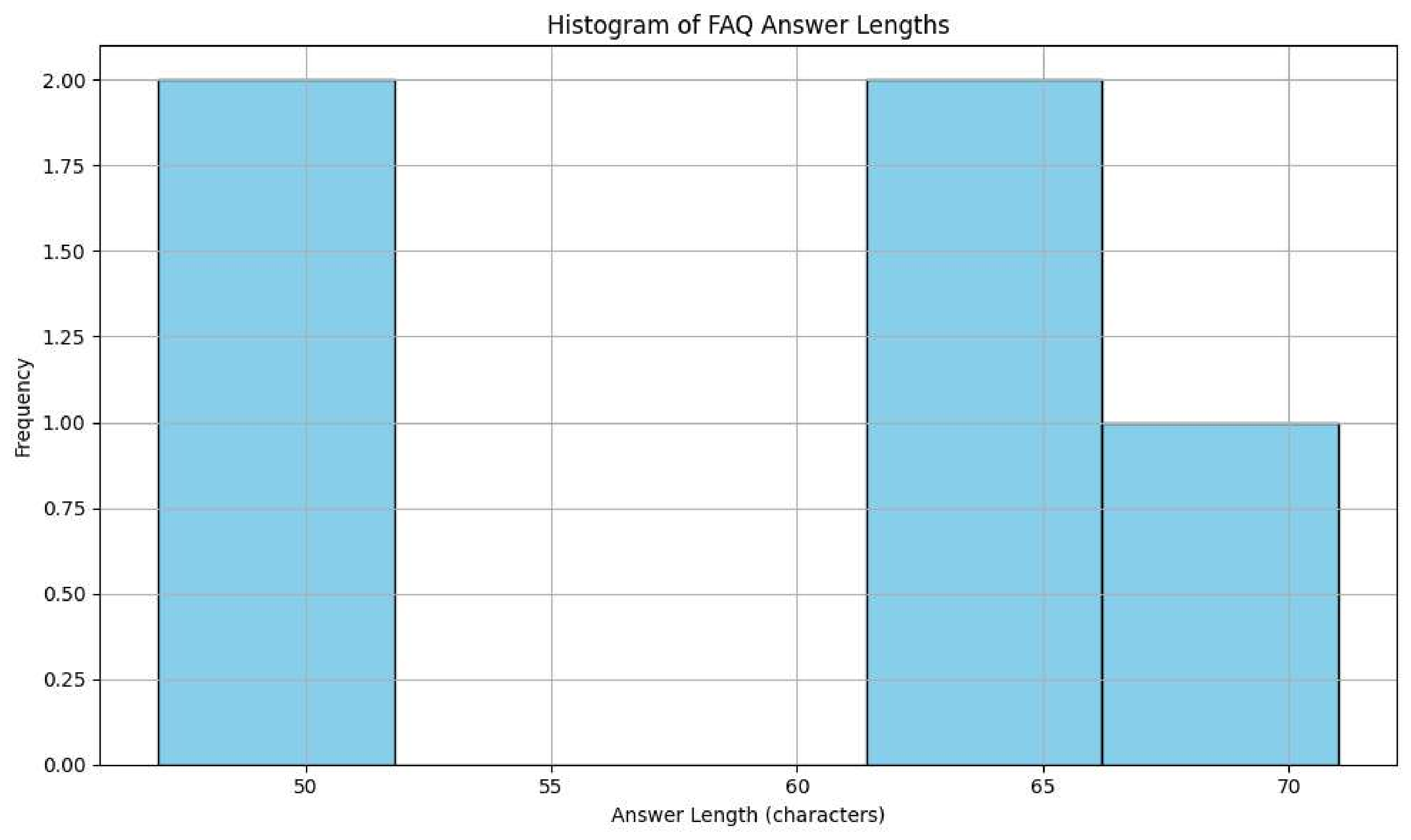
*Optional simulated field added for analysis:*

- **Category:** A label such as 'Billing', 'Technical', or 'General' to group questions (useful for visualization).
- **Answer_Length:** Character length of each answer.
- **Length_Group:** Categorized as 'Short', 'Medium', or 'Long'.

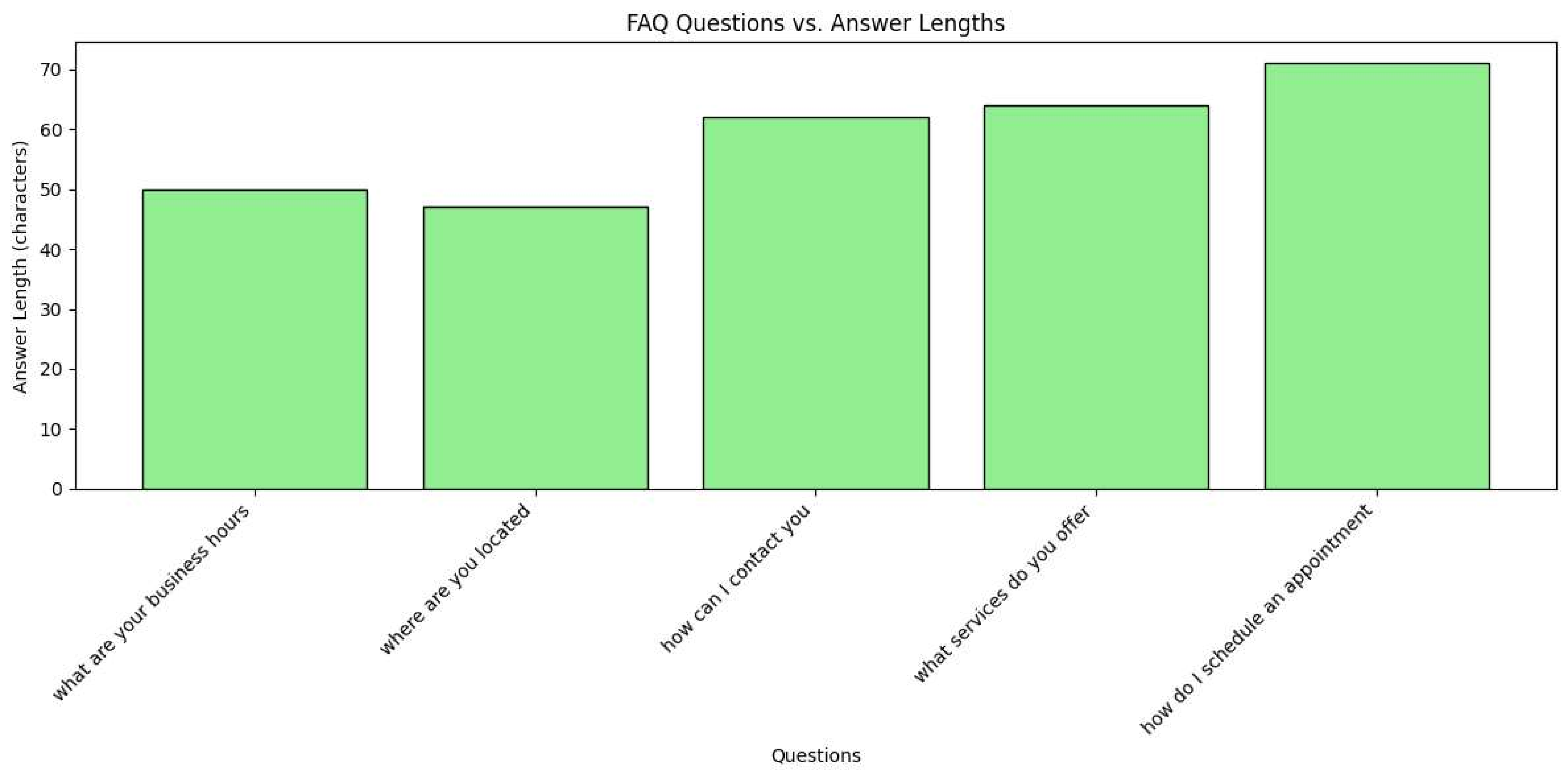
# Data Preprocessing

•Calculated the **length of each answer** using character count.
•Created a **Length Group** classification:
•Short: < 50 characters
•Medium: 50–120 characters
•Long: > 120 characters
•Simulated **Category labels** for grouping answers (in absence of such field in the uploaded dataset).

# Data Visualization



# Data Visualization

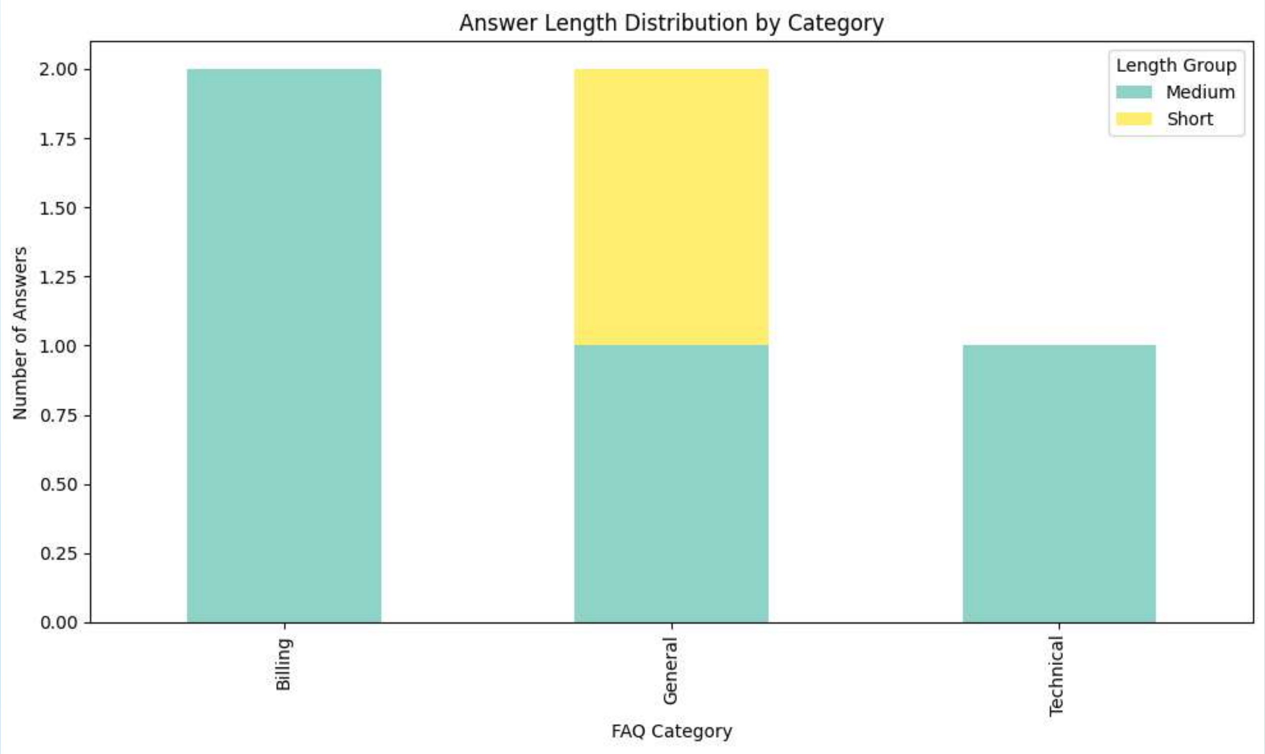


# Key Insights

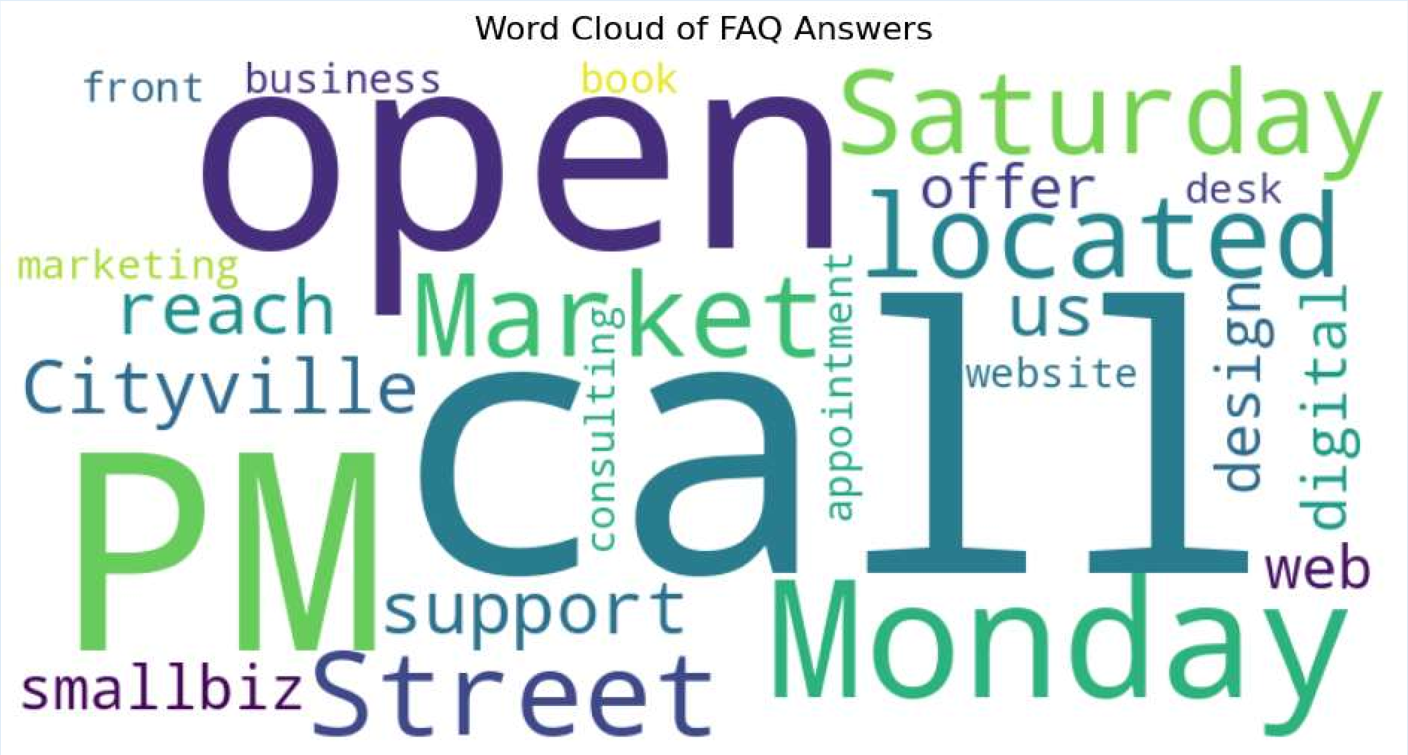
- Plotted a histogram to understand the distribution of answer lengths.
- Most responses were found to be **medium-length**, ensuring conciseness without sacrificing clarity.

- With the simulated Category column, a bar plot was generated.
- Insight: The majority of queries are **Technical**, followed by **Billing** and **General**.

# Data Visualization



# Data Visualization



# Key Insights

•This plot shows how each category contains different proportions of short, medium, and long answers.
•Example: Technical queries often have longer answers; Billing queries are typically shorter.

•A word cloud was generated to visualize frequently occurring words in the answers.
•Common terms: "account", "support", "password", "invoice", and "service".

# Observations

- **Technical questions** tend to be longer and more detailed.
- **Billing questions** are more concise.
- There is an opportunity to **standardize answer lengths** for a more consistent experience.
- Adding **clear categories** to each Q&A pair could improve chatbot performance and facilitate easier training using NLP models.

# Recommendations

- **Tag each Q&A with categories** (e.g., Billing, Technical, General) in the source dataset.
- **Enhance answers** to provide structured guidance for the chatbot
(e.g., step-by-step instructions).
- **Integrate user feedback logs** in the dataset to continuously refine responses.

# Conclusion

The dataset provides a solid foundation for a customer-facing chatbot. With minor improvements like categorization and answer tuning, the FAQ chatbot can become a more intelligent and responsive tool for customer support. Data visualization highlights key patterns that help in understanding content effectiveness and coverage.