# Spam vs Ham Analysis Documentation

Your Name

August 27, 2025

## 1 Introduction

This document presents a Python-based spam vs ham analysis project. The dataset is processed using `pandas`, visualized with `matplotlib` and `seaborn`, and word frequency is shown using `WordCloud`.

The project includes:

- A bar chart showing the distribution of spam and ham messages
- A word cloud for spam messages
- A word cloud for ham messages

## 2 Python Code

The Python script used in this project is shown below:

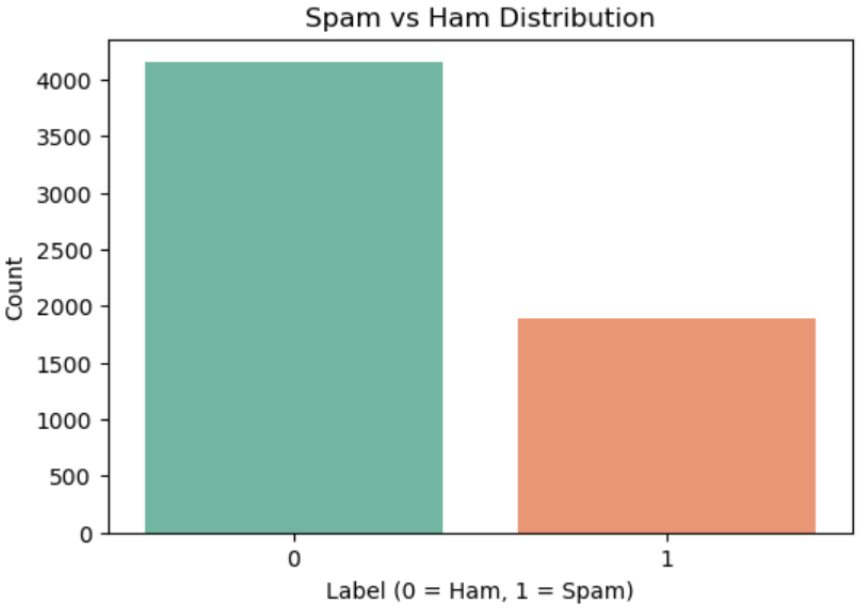
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from wordcloud import WordCloud

# Load your CSV
data = pd.read_csv(r"D:\project\spam_assassin.csv", encoding="latin-1")

# Keep only needed columns
data = data[['Label', 'Body']]
data.columns = ['label', 'message']

# Drop missing values
data = data.dropna(subset=['message'])

# --- 1. Bar chart of spam vs ham ---
plt.figure(figsize=(6,4))
sns.countplot(x='label', data=data, palette="Set2")
plt.title("Spam vs Ham Distribution")
plt.xlabel("Label (0 = Ham, 1 = Spam)")
plt.ylabel("Count")
plt.show()

# --- 2. WordCloud for Spam messages ---
spam_text = " ".join(data[data['label']==1]['message'])
```

```
spam_wc = WordCloud(width=600, height=400,
                    background_color="white").generate(spam_text)

plt.figure(figsize=(8,6))
plt.imshow(spam_wc, interpolation="bilinear")
plt.axis("off")
plt.title("Word Cloud - Spam Messages")
plt.show()

# --- 3. WordCloud for Ham messages ---
ham_text = " ".join(data[data['label']==0]['message'])
ham_wc = WordCloud(width=600, height=400,
                   background_color="white").generate(ham_text)

plt.figure(figsize=(8,6))
plt.imshow(ham_wc, interpolation="bilinear")
plt.axis("off")
plt.title("Word Cloud - Ham Messages")
plt.show()
```

# 3 Results

The following plots were generated:

## 3.1 Spam vs Ham Distribution



### 3.2 Word Cloud - Ham Messages


### 3.3 Word Cloud - Spam Messages


## 4 Conclusion

This analysis shows the proportion of spam vs ham messages and highlights the most frequent words in each category. Such visualization techniques help understand the dataset better and can be used as preprocessing for building a spam detection model.