Appendix

- spam.csv: raw dataset
- cleanedData.csv: cleaned dataset that was analyzed

Our data set is cleaned from a CSV file found from Kaggle, containing both spam and non-spam texts. The analyzed data set has the message text body, type of text, number of words in the text message, and a boolean variable marking if the text is spam or not. It contains 5575 rows and 4 columns.

Figure 1: Data Dictionary

Column	Description	Potential Responses
Category	String detailing whether the message is real or spam	Spam, Real
Message	String of text from the contents of the message	"SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs apply Reply HL 4 info" “Cool, text me when you're ready”
Word_Count	Integer count of space-delimited Strings in the Message	28
Spam	String boolean whether the message is real or spam	0, 1

Figure 2: Distribution of Data

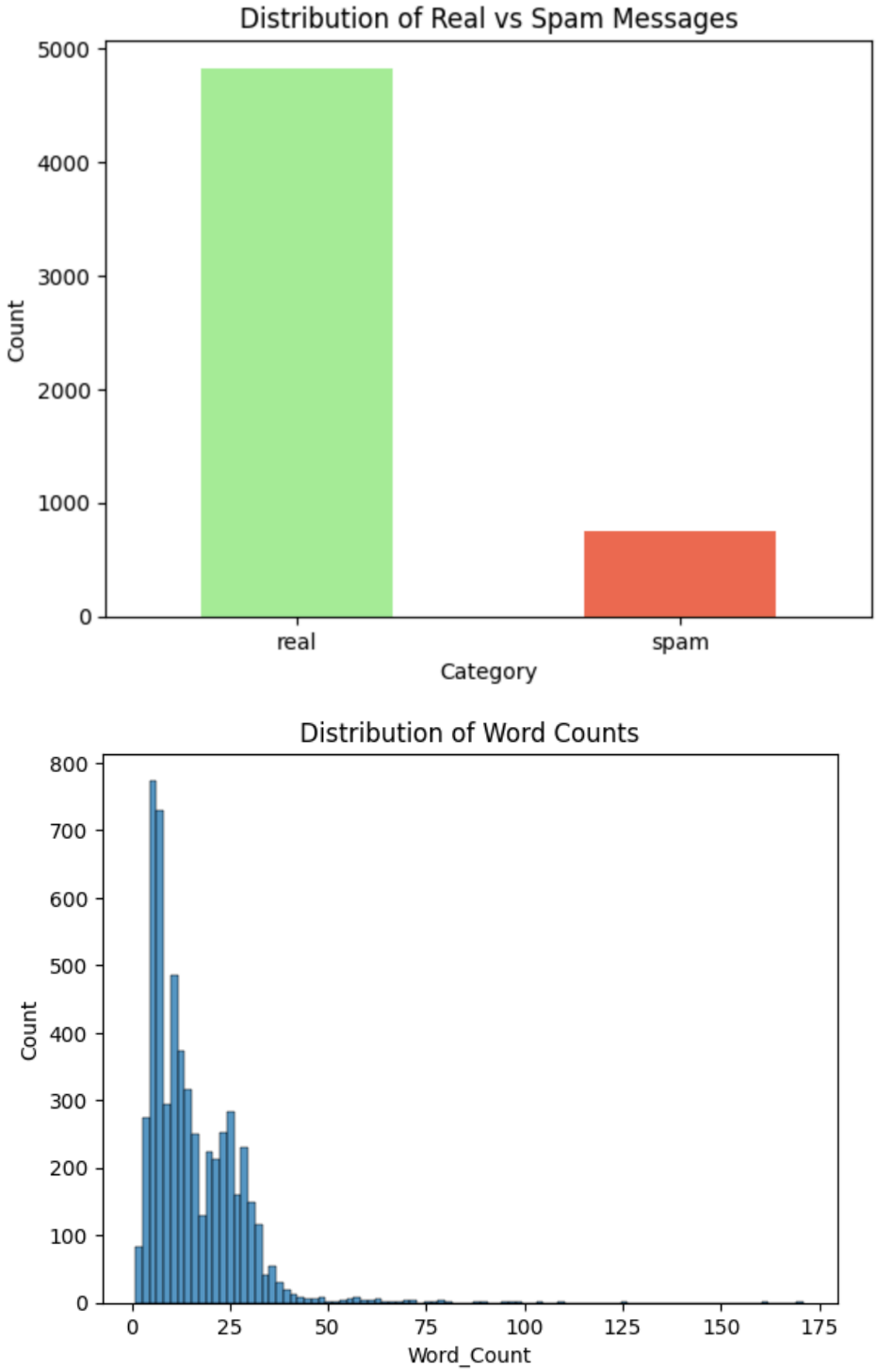

Figure 3: Summary Statistics

Variable	Mean	STD	Min	25%	50%	75%	Max
Word_Count	15.58	11.40	1.00	7.00	12.00	23.00	171.00