MIST7770 - Assignment 4,

Group 3: **Collin Ladina, Judd Douglas, Matthew Karnatz, Ryan Cullen, Shaan Patel**

FEEDBACK FROM RIOS 10/19/2024

“ This seems interesting. Please see my comments below.

*** While you can't calculate the mean or standard deviation for your categorical variables, you can still calculate the relative frequency of value. That will help you describe your variable.**

*** There are other ways to deal with missing values that are not excluding them. You still need to analyze how many you have in each variable.**

Please review these analyses before the presentation so we can discuss them during the presentations.

Score

6 / 6 - 100 %

- The presentation should focus on at least these first four elements. Context and Questions, Data and Variables, Descriptive Visualizations, and Model Selection.
- In addition to understanding the features better, you want to use the *Descriptive Visualizations* to show how your features are related to the target variable. This is more of an exploratory analysis.
- Notice that you don’t have to build any model yet. For this submission, you only need to select the appropriate model for your final analysis and describe why you picked that model.

Instructions
The objective of this assignment is for you to work with your team and select a topic of your own choosing for your Final Project of the course. Just for context, part of the Final Project's description is included below.

"..., you prepare and analyze data, create visualizations and appropriate models, and interpret findings in a clear story. The data set may come from any source and any application. Good examples include government record information, financial information, sports statistics, data that can be scrapped from HTML sites, social media data, such as X (Twitter), and openly available datasets from sites like 538 (for politics), Kaggle and UCI ML Repository (for more general data science explorations).

Once you are settled on your question(s) and on which tool(s) to use, the next step for you is to answer your project questions. This step entails building a predictive model, making supporting visualizations, and/or creating a dashboard as necessary. Finally, consider recommending a solution that can be implemented."

Now, see below the expected structure of the current assignment (Research Question and Data). Submit your assignment answering the questions below in a document of maximum 3 pages.

Context and Question

Write a brief description (a paragraph or two) of the context and the dataset (the relevant industries, actors, etc.) In plain language, formulate your problem: what is/are the question(s) you are trying to answer? Why is/are this/these (a) useful question(s) to answer?

The main question this dataset aims to answer is: Can we accurately classify an email as spam or non-spam based on its content and structure? This is a key question for email security because it allows us to make spam filters and prevent unwanted emails from flooding inboxes. It is also the first line of defense against scams and helps keep us productive by removing distracting/worthless emails from our inboxes.

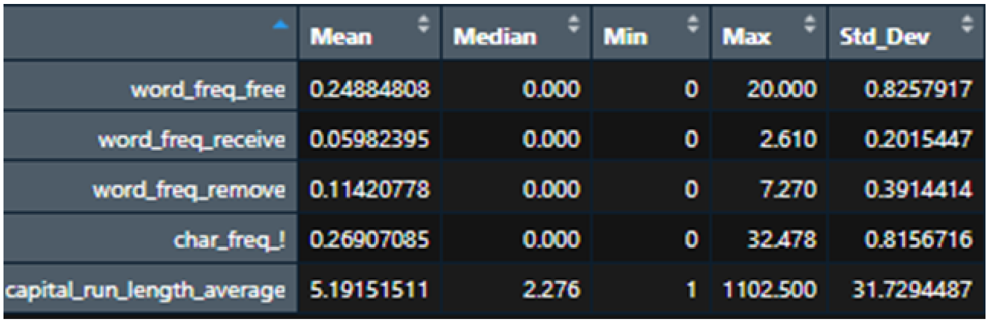
The Spambase dataset, from the University of California Irvine Repository, includes 4,601 instances and 57 features, representing various attributes of email content. Spam emails in the dataset consist of diverse content, such as advertisements, get-rich-quick schemes, chain letters, and pornography, while non-spam emails are primarily personal or work-related. The classification task again is determining whether an email is spam based on these features and can be deleted from a person's inbox if it's malicious or unnecessary.

Data and Variables

Our dataset “Spambase” was obtained from the UC Irvine Machine Learning Repository website (https://archive.ics.uci.edu/dataset/94/spambase). We were able to convert it to an excel file, where it was then imported to RStudio. It is an adequate dataset to address our problem because it has various numeric features that allow us to build a model that can predict whether or not an email is classified as spam or not. It resembles real-world data and usage that is commonplace in contemporary emails that an individual or business may send/receive.

Spambase has 58 total variables, with the vast majority of them being frequencies of specific words, characters, or special symbols used. Other variables include runtimes/lengths of strings, like consecutive or total capital letters. Our target variable is the binary classifier “is_spam”, which categorizes where an email is spam or not (1,0). Due to the various forms a spam email can embody, we plan to utilize most if not all the frequency variables in our analysis. Our dataset does not include any dates or times.

Due to the related nature of the variables, we only included 5 of the frequencies and performed descriptive statistics on it. The results are posted below.



	Mean	Median	Min	Max	Std_Dev
word_freq_free	0.24884808	0.000	0	20.000	0.8257917
word_freq_receive	0.05982395	0.000	0	2.610	0.2015447
word_freq_remove	0.11420778	0.000	0	7.270	0.3914414
char_freq_!	0.26907085	0.000	0	32.478	0.8156716
capital_run_length_average	5.19151511	2.276	1	1102.500	31.7294487

Spambase contains 4,601 observations. And although outliers are very abundant throughout the dataset, we do not think removing these outliers is a good idea. We hypothesize that including these outliers allows us to use the same “blackbox” as the trato cover as many predictions as possible. We may decide to remove some extreme outliers for regularization, if we decide to go down that route.

Review for this data set:
https://archive.ics.uci.edu/dataset/94/spambase

Judd’s dataset work
https://outlookuga-my.sharepoint.com/:x:/g/personal/sjd53041_uga_edu/EceJ-bRrsv1JlImi2rCdP2QB-egoaI2PbsqwnORMUuisJg?rtime=qkMId1Xq3Eg

Data and Variables
Answer the following questions discussing your project. Include paragraphs, not just bullets.

Where and how did you obtain the data from? Our dataset ”Spambase” was obtained from the UC Irvine Machine Learning Repository website. (https://archive.ics.uci.edu/dataset/94/spambase).
Why is this data adequate to address your problem? It is an adequate dataset to address our problem because it has various features that allow us to build a model that can predict whether or not an email is classified as spam or not. It resembles real-world data and usage that is commonplace in contemporary emails that an individual or business may send/receive.
Describe your variables. Your dataset needs at least ten variables. What is the target variable of interest? Spambase has 58 total variables, with the vast majority of them being frequencies of specific words, characters, or special symbols used. Other variables include runtimes/lengths of strings, like consecutive or total capital letters. Our target variable is the binary classifier “is_spam”, which categorizes where an email is spam or not.
What are the predictors/features you plan to include in the analysis?
Due to the various forms a spam email can embody, we plan to utilize most if not all the frequency variables in our analysis.
Is your target variable numeric or a class type?
Do you have any date type of variables? Our dataset does not include any dates or times.
Report the relevant descriptive statistics on each variable (mean/median, standard deviation, etc.). Due to the related nature of the variables, we only included 10 of the frequencies and performed descriptive statistics on it.
You can use a table to report the descriptive statistics and their descriptions combined.
How many observations are in your dataset? Your dataset needs at least 1,000 observations. Our dataset contains 4,601 observations. And although outliers are very abundant throughout the dataset, we do not think removing these outliers is a good idea. We hypothesize that this occurrence is due to _______?___?___?_______

Do you have outliers or missing values? How are you dealing with them?

The dataset was compiled using two sources: spam emails collected from a postmaster and individuals who had reported spam, and non-spam emails collected from personal and work-related correspondence. The dataset is adequate for addressing the problem of email classification because it includes a wide range of features (57 in total) that represent different aspects of email content, such as word frequencies, the presence of certain keywords, and patterns in the email structure. These features are crucial for distinguishing between spam and legitimate emails.

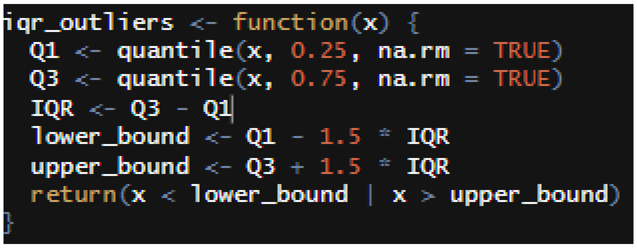
The dataset includes 57 variables, most of which are continuous and represent the frequency of specific words or characters in the email content (e.g., “make,” “address,” “our”). The target variable is binary and indicates whether an email is classified as spam or non-spam. Predictors include the frequency of specific words and characters, as well as the percentage of capital letters in the email body.

The target variable is a class type, as it separates emails into two categories: spam or non-spam. There are no date-type variables in this dataset.

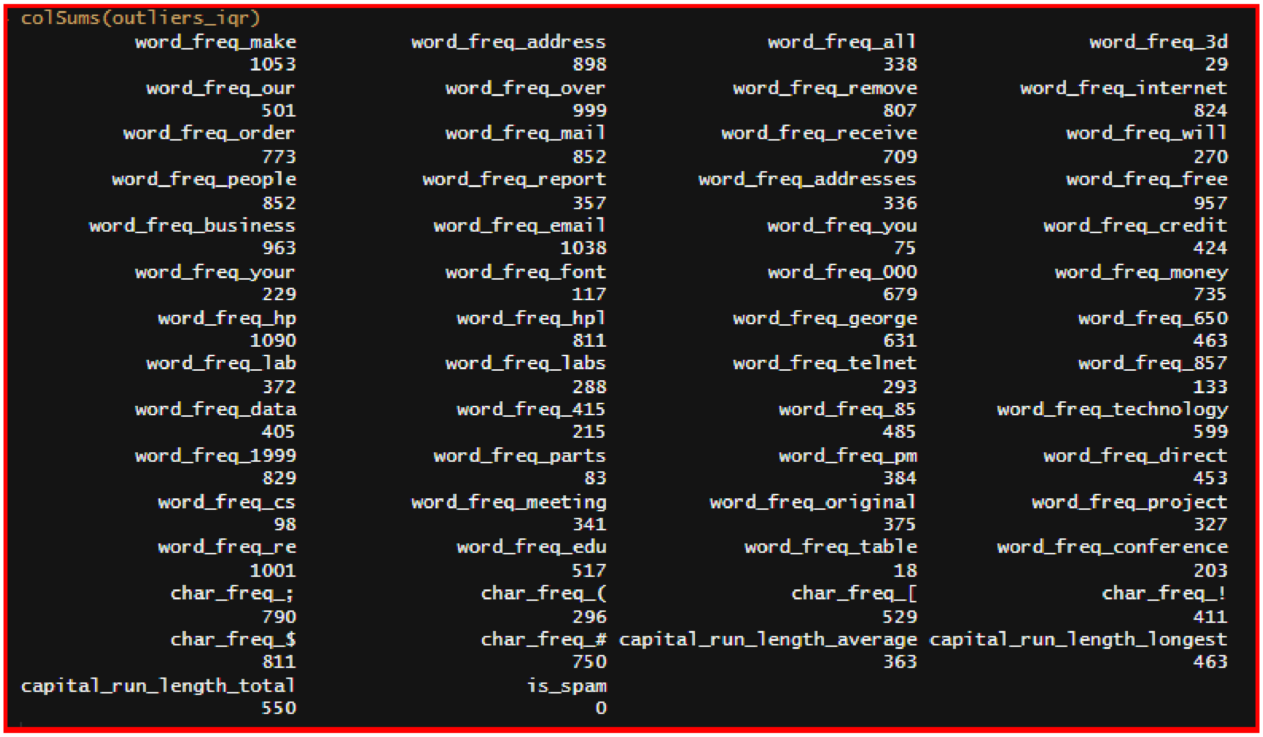
Descriptive Statistics:

Variable	Mean	Median	Standard Deviation
Word Frequency (‘make’)	0.21	0.0	0.67
Word Frequency (‘address’)	0.06	0.0	0.24
Word Frequency (‘our’)	0.31	0.06	0.58
Char Frequency (%) (’!’)	0.06	0.0	0.3
Capital Letters (%)	1.43	0.0	3.26

The dataset contains 4,601 observations, which is more than sufficient for analysis. The dataset has some missing values that will need to be addressed through appropriate data imputation techniques. Outliers may also be present in variables like the percentage of capital letters, which could influence the classification performance. These will be handled using standard statistical techniques, such as z-score normalization or removing extreme outliers.



```
iqr_outliers <- function(x) {
  Q1 <- quantile(x, 0.25, na.rm = TRUE)
  Q3 <- quantile(x, 0.75, na.rm = TRUE)
  IQR <- Q3 - Q1
  lower_bound <- Q1 - 1.5 * IQR
  upper_bound <- Q3 + 1.5 * IQR
  return(x < lower_bound | x > upper_bound)
}
```



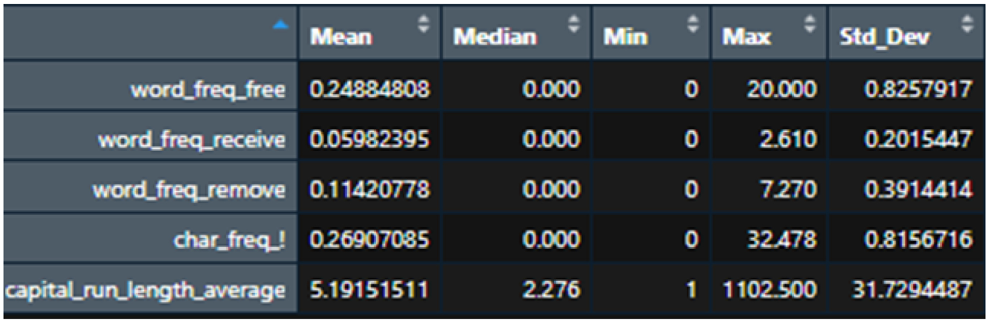
```
> colSums(outliers_iqr)
            word_freq_make          word_freq_address              word_freq_all               word_freq_3d
                      1053                        898                        338                         29
             word_freq_our             word_freq_over           word_freq_remove         word_freq_internet
                       501                        999                        807                        824
           word_freq_order             word_freq_mail          word_freq_receive             word_freq_will
                       773                        852                        709                        270
          word_freq_people           word_freq_report        word_freq_addresses             word_freq_free
                       852                        357                        336                        957
        word_freq_business            word_freq_email              word_freq_you           word_freq_credit
                       963                       1038                         75                        424
            word_freq_your             word_freq_font              word_freq_000            word_freq_money
                       229                        117                        679                        735
              word_freq_hp              word_freq_hpl           word_freq_george              word_freq_650
                      1090                        811                        631                        463
             word_freq_lab             word_freq_labs           word_freq_telnet              word_freq_857
                       372                        288                        293                        133
            word_freq_data              word_freq_415               word_freq_85       word_freq_technology
                       405                        215                        485                        599
            word_freq_1999            word_freq_parts               word_freq_pm           word_freq_direct
                       829                         83                        384                        453
              word_freq_cs          word_freq_meeting         word_freq_original          word_freq_project
                        98                        341                        375                        327
              word_freq_re              word_freq_edu            word_freq_table       word_freq_conference
                      1001                        517                         18                        203
               char_freq_;                char_freq_(                char_freq_[                char_freq_!
                       790                        296                        529                        411
               char_freq_$                char_freq_# capital_run_length_average capital_run_length_longest
                       811                        750                        363                        463
  capital_run_length_total                    is_spam
                       550                          0
```

Data and Variables

Our dataset “Spambase” was obtained from the UC Irvine Machine Learning Repository website (https://archive.ics.uci.edu/dataset/94/spambase). We were able to convert it to an excel file, where it was then imported to RStudio. It is an adequate dataset to address our problem because it has various numeric features that allow us to build a model that can predict whether or not an email is classified as spam or not. It resembles real-world data and usage that is commonplace in contemporary emails that an individual or business may send/receive.

Spambase has 58 total variables, with the vast majority of them being frequencies of specific words, characters, or special symbols used. Other variables include runtimes/lengths of strings, like consecutive or total capital letters. Our target variable is the binary classifier “is_spam”, which categorizes where an email is spam or not (1,0). Due to the various forms a spam email can embody, we plan to utilize most if not all the frequency variables in our analysis. Our dataset does not include any dates or times.

Due to the related nature of the variables, we only included 5 of the frequencies and performed descriptive statistics on it. The results are posted below.



	Mean	Median	Min	Max	Std_Dev
word_freq_free	0.24884808	0.000	0	20.000	0.8257917
word_freq_receive	0.05982395	0.000	0	2.610	0.2015447
word_freq_remove	0.11420778	0.000	0	7.270	0.3914414
char_freq_!	0.26907085	0.000	0	32.478	0.8156716
capital_run_length_average	5.19151511	2.276	1	1102.500	31.7294487

Spambase contains 4,601 observations. And although outliers are very abundant throughout the dataset, we do not think removing these outliers is a good idea. We hypothesize that including these outliers allows us to use the same “blackbox” as the trato cover as many predictions as possible. We may decide to remove some extreme outliers for regularization, if we decide to go down that route.