SixSigma study on Improving Spam Blocking and Classification in EMAIL-SMS Messaging Traffic

Sr.Mallesham Yamulla
Indian Statistical Institute-Hyderabad

2019-01-22

Contents

1.About Organization and team

AdaptiveMobile is the world leader in mobile network security protecting over 2.1 billion subscribers worldwide. With deep expertise and a unique focus on network security, Adaptive Mobile's Threat Intelligence Unit and Network Protection Platform are trusted by the world's largest service providers to secure their critical communications infrastructure and services against current and mobile threats and financial exploitation. Adaptive Mobile's sophisticated, revenue-generating security-as-a-service portfolio empowers consumers and enterprises alike to take greater control of their own security.

AdaptiveMobile was founded in 2004 and boasts some of the world's largest mobile operators as customers and the leading security and telecom equipment vendors as partners. The company is headquartered in Dublin with offices in the North America, Europe, South Africa, Middle East and Asia Pacific.

2.Introduction to SixSigma methodolgies

Six Sigma is a project-based approach for improving effectiveness and efficiency. It is a disciplined, customer-focused, data-driven approach for improving the performance of processes, products, or services.

Now,the term Six Sigma can be used to refer to a philosophy, a performance metric, or a methodology.

As a philosophy, Six Sigma strives for perfection in achieving effectiveness and efficiency in meeting customer and business requirements.Six Sigma is proactive and prevention-based instead of reactive and detection-based.

As a performance metric,Six Sigma refers to a level of quality that is near perfection. It strives for a defect level that is no more than 3.4 parts per million. Sigma is a letter in the Greek alphabet used to represent standard deviation, a measure of variation. A Six Sigma process is very consistent, with very little variation, and therefore has a very small standard deviation.So small that the distance from the mean, or the average, to the nearest specification limit is equivalent to six standard deviations, or six sigma's. As a result, only 3.4 parts per million or less are out of specification.

As a methodology, Six Sigma refers to DMAIC, a methodology for improvement named after its five phases of define, measure, analyse, improve, and control. Using this prescriptive approach, a team focuses on improving what's important to customers, and uses data analysis to diagnose and improve the performance of process, products, or services.

3.DMAIC: Define

3.1. Introduction, Objective and Goals

We are into Threat Intelligence Unit where we do fight against Spammers who keeps on sending the different kinds of Spams to the mobile subscribers, and these spam attacks can be getting through

- EMAIL TO SMS and EMAIL TO MMS
- SMS
- Grey Routes/SIM Banks
- SS7 Signaling/ Diameters

We have already built up the two products called NPP (Network Protection Platform)and SIGIL(Signalling Intelligence) to control the attacks being happened on the above mentioned messaging systems

In this Project, I'm going to work on EMAIL and SMS messaging traffics only, first I would like to give a small introduction on how the messaging gets routed from Spammers to Subscribers, how our team TIU can block the suspicious, unsolicited , bulk messages out from the networks using the different tools

The mobile subscribers go on receiving the variety of messages (Personal, Business, Etc Etc..), from the known-unknown persons-companies., here It's little difficult for them to figure out a message to be Legit or Spam

In this case, they are supposed to pass these suspicious/doubtful/unwanted messages to our system called 7726 that gets these complaints registered for further investigation, these 7726 complaints would also be sent in to our internal tool called Security Centre where we can classify the messages to their respective categories

We have the following 8 different types of Spam Categories in our systems.

- LEGIT-CLEAN Messages
- SPAM- SCAM
- SPAM-LOAN
- SPAM-PHISHING
- SPAM-ADULT
- SPAM-MARKETING
- SPAM-GAMBLING
- SPAM-MALWARE

Our team is given an access to roll though the 7726 logs, carry out an investigation using the listed principles, intelligence and take a necessary action if a message is found to be a spam and these can be of

- SMS
- EMAIL-TO-SMS/EMAIL-MMS

We generally bank upon the following given filtering methods that are part of our NPP product to block out the spams,

- Fingerprinting
- URL Blacklist
- Content Matching
- IP Blacklists
- Domain Blacklists
- Reputation Filters
- Regular Expressions
- And many others

The messaging classification can only be done via Fingerprinting Filter in Security Centre that connects to NPP.

Fingerprinting Filter:

Once messages are sent into Security Centre an analyst from Team has to look through them, and apply a fingerprint to a message by moving it to the SPAM or LEGIT Category so that it gets assigned a 13 digit unique hash code, stored in SC and looks for its variants in the traffic that comes in.

If a fingerprinted messages match with any other messages taking an account of its configured similarity(about 80%) its cluster size would go on increasing and here cluster size means the total No.of similar messages.

If a fingerprint is of SPAM category all of these messages would be blocked out and on the other hand legit fingerprinted messages would be getting allowed.

Challenge:

Our Teams works in three different time zones (US-EUROPE-INDIA) 24-7, here weekly, the messaging traffic gets stored in our systems about 3-5B messages and 10-15mil of them are of spam messages. And Per day about 10K messages which have come from 7726 system would be manually classified in Security Centre by Analysts.

In this continuous manual messaging marking process, we have to get through the below mentioned issues,

- Misclassification of Messages
- Higher False Positive and False Negatives
- Time Taking process in classification

- Spam leakages

As the messaging complaints sends in faster from subscriber our team should proactively mark and classify the messages one by one hourly basis, in his/her manual classification there could be a change to get misclassified messages, as explained below,

Example:

- SPAM Phishing category to SPAM Loan
- SPAM Loan Category to SPAM Scam
- LEGIT Message to any one of SPAM category
- Any of Spam Message to Legit Category

And the first step Misclassification leads to improve the rates of False Positive and False Negative rate which can't be acceptable by Businesses.

False Positive Message - It's a case when a LEGIT message is found to classified as SPAM message

False Negative Message - It's a case when a spam message is found to classified as LEGIT message

In the second place, Analyst has to spend a more time to make up his mind to classify and take an action on messages when he/she comes across any new spam campaigns day to day, here there might be spam leakages if they are not actioned on time.

Here our business goal would be that the current email spam blocking rates should be improved about 20% from around 10% on average while meeting the above specified requierement. In addition to it, the following business questions would also be required to look through

1. How are the spam traffic volumes/behaviours over week/days/hours?.
2. are the spam patterns getting changed frequently ?
3. is there any effect of manual messaging classification done by fingerprint filter on controlling spams ?
4. estimating fingerprint filter active blocking time ?
5. any delays in classfiying a message once its on Security Centre?

3.2. Project charter

Project Charter		
Title	**EMAIL and SMS Spam Messaging Classification**	
General Project Information		
Project Sponsor/Champion	AdaptiveMobile Security - Sr.Mallesham Yamulla	**Team Members**
Team Leader	Sr.Cary Anderson	Sr.Johanson Alsina
Process/Area of Study	Threat Intelligence - Cyber/Messaging Security	
Process Owner/Dept Head	Sr.Stuart McBride	
Describe Business Case, Goals, Objectives, and Deliverables of this Project		
Business Case	In a process of manual messaging classification in our Security Centre being done by Analysts, legitimate messages would get blocked out as they are wrongly classified as SPAM and other hand Spam messages would get through as they are erroneously moved to CLEAN category, here both of these scenarios would never be acceptable by Telco Operators. We are looking through these cases to figure out what has led to cause the false positives and false negatives. We are spending many hours in a day for manual classification , and we would like to get this time down as lesser minutes as posible while bettering blocking rates.	
Problem Statement	Improve Spam Blocking rate in messaging traffic of EMAIL to SMS/MMS	
Goals / Objectives	**1.** Improving accuracy of messaging classification - SMS & EMAIL 2. Identify messaging patterns and trends 3. Estimate an effect of manual messaging classification on controlling spams being carried out by Fingerprint Filter algorithm 4.Make out an impact,pros and cons in case of this messaging classification system gets automated using Text Analytics such as WORD2VEC Algorithms	
Expected Benefits	1. Minimizing False positive and False negative rates in Spam Classiication so as to avoid over blocking/lower blocking of messaging- SMS & EMAIL 2. Spam leakages get controlled	
Define the Project Scope and Schedule		
Within Scope	**1.** List out the standard guidelines and give in KTs for doing classification properly across team	
Outside of Scope	**1.**Review the classified messages in analysis mode before they are moved to Blocking mode	
Describe Project Constraints (Risks, Difficulties, Challenges etc..)		
Constraints	**1.**Manual Classification carried out by different analysts in different time zones **2.** Higher messaging traffic volumes send in peak hours **3.** Improper catridges updates and Synchronization issues in Database **4.** Bulk marking/uploads of messages at one go in Security Center	
Project Milestones		
Define	Expected Start Date : 20-AUG-2018	Expected End date : 31-AUG-2018
Measure	Expected Start Date : 01-SEP-2018	Expected End date : 30-SEP-2018
Analyze	Expected Start Date : 01-OCT-2018	Expected End date : 30-NOV-2018
Improve	Expected Start Date : 01-DEC-2018	Expected End date :14-DEC-2018
Control	Expected Start Date : 15-DEC-2018	Expected End date : 31-DEC-2018

Figure 1: Project Charter

3.3. Voice of Customers:

A. Classify messages correctly

1. Proper guidelines to be followed in classification of messages
2. Experienced analysts
3. Security Centre (SC) GUI appearance / flexibility / smarter features for manual classification

B. Make sure over blocking of legitimate message/lower blocking of spam messages wouldn't get occurred

1. Carry out daily checks on the classified messages to figure out false positive and false negatives
2. Monitor Cartridge updates

C. Get Spams blocked out as faster as possible and Improve spam blocking rates

1. Look for spams in logs continuously
2. Do regular query searches on Database.
3. Do bulk uploads of better spam candidatures.

4. Get fingerprint fiter algorithm worked more efficient
5. Do experimenting with text analytics

3.4. Process Map

A picture is worth a thousand words, considering this stament here a process map is a diagram that provides a visual representation of the process flow, or sequence of activities, or steps that take place in a process from start to finish.,the flow can go from top to bottom or left to right.

A process map enables us to see and understand the process.once a current process has been mapped, the team will also know what's not happening or what's different from what should be happening.

In our process map, It starts from messaging being sent out from spammers towards telcom network through which users receive messages,here it could be personal or business related messages. Users generally want to get a benefit from legitimate messages which give them a useful information regarding a service or a conversational message.

Spammers take an advantage of user's need and go on pumping in unwanted/suspicious/bulk messages to them for doing a fraud, User's woudn't be aware of these messages or found them to be a harmful for them, So here they would have a **system 7726** where these messages can be reported for furhter investigation.

The reported messages to 7726 system will be pushed to a system called **Security Center** in which they are classified, blocked out and contolled from not being sent on to users if they are really spams.

The SC (Security Center) is continously being looked through by analysts in 24/7 hours, SC mechanism works like in this way: a message gets matched with another message considering the configured similarity percentage in fingerprint filter, and they goes on getting clusterd as long as its similarity matches.

Analysts would often go for a method to upload messages that are collected from another system called TSM where the entire messaging traffic gets stored into Security Center to controll spams effectively.

Once these are lsited out in SC Analyst would start classifying them manually banking upon his intelligence/techniques,the classified spam messages are being tripped and blocked out by a fingerprint filter, here spam traffic get handled and not sent to users and in another case the classified legitimate messages are being allowed towards users.

There could be a chance to happen a misclassification on messages that reached to SC, Analyst intelligence would go wrong sometimes and it leads to increase rates of false positives and false negatives, spam leakages as well,as explained here, a legit message is wrongly classified to spam, hence these messages wouldn't be delivered to users, on the other hand a

spam message is categorized as legit,in this case spams would be flown to users and they get effected.

The process gets stopped with blocking out spams and allowing legit messages towards Users.

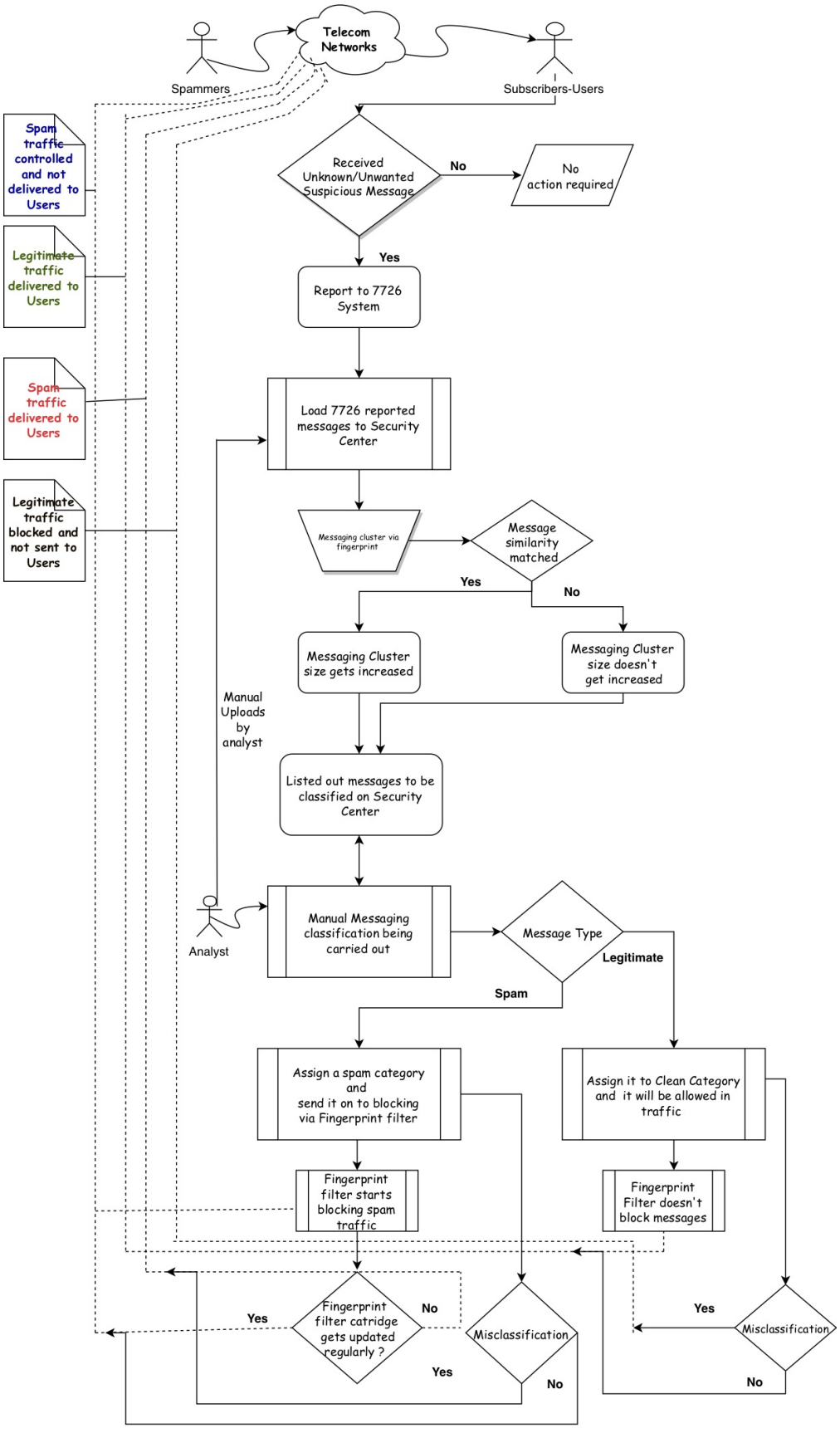


Figure 2: Process Map

4. DMAIC: Measure

4.1. What type of tools/libraries have been made use of? and its setup

I have decided to get this project done using R since it's an open source tool whose community has been growing up in Statistical analysis

- tidyverse(EDA)
- data.table (EDA faster process)
- tidytext/quanteda (Text Analytics)
- gggplot2/Plotly(Data Visualizations)
- sixSigma/qic/Quality Tools - Control charts
- rmarkdown(Reporting)

4.2. Glance at Data

Two expert analysts of our team are dedicated to look through a file that gets generated with the last day messages classified by all the analysts from 3 time zones and reclassify the messages if needed, here they would have to spend at least half of the day (4hrs) for this task itself. On a daily basis the false positive and false negative checks have to be verified so that there wouldn't be an over blocking/lesser blocking of messages.

A single file would consist of about 5-10K classified messages from Security Centre, for this project I have gathered these files from February 2018 to October 2018,I have merged them as monthly wise data in 3 different CSV and overall there are about 350K Observations with 7 Variables.

- **Processing Date:** Day on which a message was classified ?
- **Fingerprint ID :** A unique hash code for a message
- **SPAM_CAMPAIGN:** Identifies which type of category a message belongs to?
- **STATUS:** Indicates which status has given to a message? Here are three status codes (SFS- SPAM+FINGERPRINT, S-SPAM ONLY, CS-CLEAN)
- **CLUSTER_SIZE:** How many time a message gets clustered with its similar messages
- **USERNAME:** Who has marked a message ?
- **CONTENT:** Message description it might contain URLs, CTA's, so on and so forth.

I have been able to add up few more variables to the dataset, and finally it has got around 15variables such as,

1. From Processing Date these three variables were created DAY,HOUR,WEEK
2. Cluster Size were bucketized based on its size like
    - (between(CLUSTER_SIZE,1,2)) ~ "CL_LOW_1",
    - (between(CLUSTER_SIZE,3,4)) ~ "CL_LOW_2",

- (between(CLUSTER_SIZE,5,7)) ~ “CL_MID_1”,
- (between(CLUSTER_SIZE,8,10)) ~ “CL_MID_2”,
- (between(CLUSTER_SIZE,11,50)) ~ “CL_HIGH1”,
- (between(CLUSTER_SIZE,51,100)) ~ “CL_HIGH2”,
- Greater than 101 ~ Others

3. USERNAMES have been converted to 3 Regions (USA,EUROPE,INDIA)
4. CONTENT : Message Type whether it’s an email one or SMS one, URL domains, CTA info were extracted.
5. Message Time related variables:
    - first_message_received (the date on which a message enters into Security Center)
    - processing_date (the date on which a messages is classified into the respective categories)
    - last_message_received (the date on which a message gets clustered to its variant)
    - last_update_date (the date on which a message gets modified such changed its status/category)
    - fp_ttl_days its derived as last_message_received - first_message_received
    - fp_process_days its derived as processing_date-first_message_received
    - fp_last_days : as last_update_date-processing_date
6. classification_time(class_time): here 2 classes have been created based on timestamps such as processing_date, last_message_received

here a class: “antes” is assigned for messaging traffic that had been received before it processed by analyst in Security center, and a class: “despues” is assigned for messaging traffic that had been received after it processed by analyst in security center.

4.2.1 Glance at messaging traffic

processing_date	fingerprint_id	spam_campaign	status	cluster_size	user
2018-05-07 19:52:00	2463a76a-c08d-486d-8670-4d166017	Spam_Loan	SFS	36	USA
2018-04-30 17:49:00	6fedb805-3cc6-480e-a74c-82129d37	Spam_Adult	SFS	2	USA
2018-08-01 21:31:00	59bba62f-f005-4f74-9870-dd67dfc5	Legit_Message	CS	13	USA
2018-03-07 14:56:00	8f179ca9-cd3d-4db4-b664-a77d7d7b	Spam_Scam	SFS	8	EUR
2018-07-09 18:58:00	ec26ea98-94ab-4840-be7c-53e44f45	Spam_Loan	SFS	1	USA
2018-07-13 00:50:00	d54a2e28-0ad5-4d1b-910b-4f845187	Spam_Phishing	SFS	1	USA
2018-09-24 20:45:00	cd8273fe-bfee-4832-bb88-4b834e56	Spam_Scam	SFS	8	USA
2018-06-05 15:18:00	38767739-6bf3-427c-b19c-4a241dbb	Spam_Adult	SFS	2	EUR
2018-07-31 21:25:00	e644d5a8-90d0-471a-9396-04414139	Spam_Phishing	SFS	1	USA
2018-04-06 19:35:00	e9b68395-f103-4949-bd1c-d5c354d9	Spam_Scam	SFS	8	USA

4.2.2 Glance at messaging traffic

content_txt	spam_campaign	status
ld aaor while you re not wearing bngW o e pw	Spam_Scam	SFS
We are looking for Financial Agents up to per month view more here	Spam_Scam	SFS
We have a new job opportunity in your area work for Google	Spam_Scam	SFS
Crucial Declaration Mid	Legit_Message	CS
cz lqm goodbye hug I had grabbed	Spam_Adult	SFS
Your Facebook account requires urgent action follow now	Spam_Phishing	SFS
Howdy Our system issued a new notification for your account Your unique ID	Spam_Phishing	SFS
Hi Please read before rd of July	Spam_Phishing	SFS
WELLS FARG We need to verify your identity Log in login id ssl bnk org tel	Spam_Phishing	SFS
rv rmkxa nights My job isn t just what	Spam_Scam	SFS

4.2.3 Glance at messaging traffic classified in SC

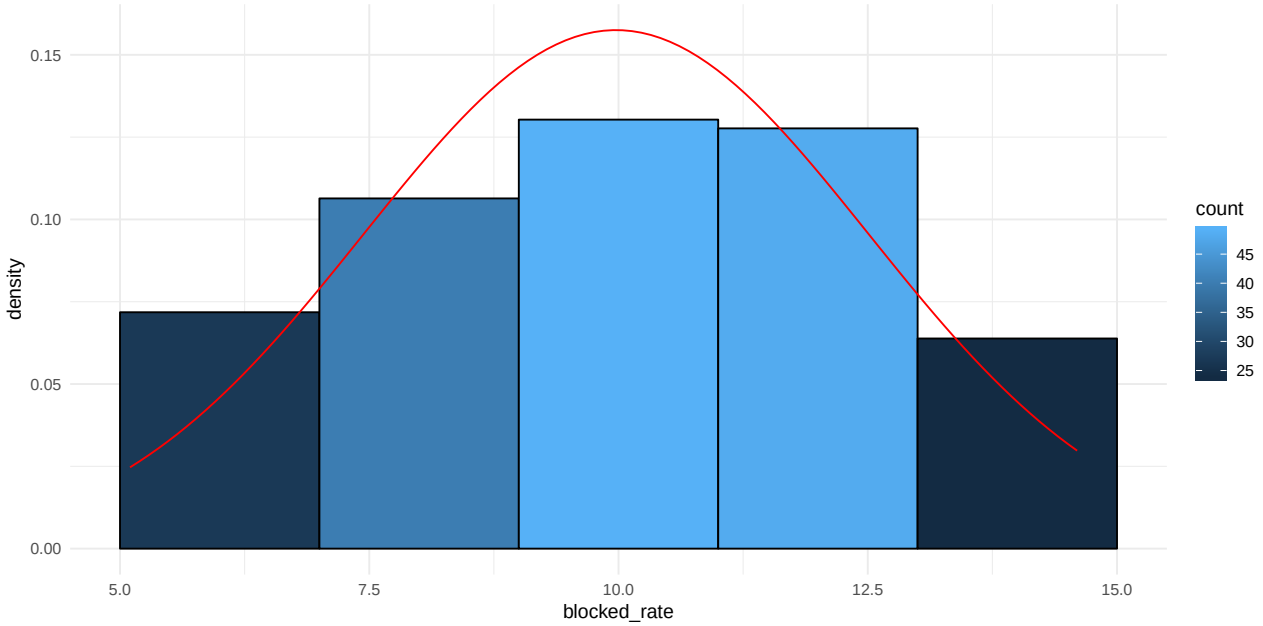
fp_ttl_days	fp_process_days	fp_last_days	class_time	st_change_dias
3.7194444 days	0.6687500 days	0.0000000 days	antes	same-day
0.6805556 days	0.0048611 days	0.0000000 days	antes	same-day
20.7229167 days	0.4500000 days	0.0000000 days	antes	same-day
0.0000000 days	0.0618056 days	0.2500000 days	despues	same-day
0.0000000 days	0.0548611 days	0.0000000 days	despues	same-day
9.7590278 days	1.4729167 days	0.0000000 days	antes	same-day
0.0000000 days	0.0152778 days	0.0000000 days	despues	same-day
0.0000000 days	0.3826389 days	0.0000000 days	despues	same-day
0.0000000 days	0.0569444 days	0.0000000 days	despues	same-day
0.0000000 days	0.6847222 days	0.2069444 days	despues	same-day

4.2.4 Glance at spam blocking rates data

day	blocked_rate
2018-07-07	24.20
2018-09-29	10.43
2018-03-16	6.90
2018-04-22	11.00
2018-08-23	9.40
2018-07-17	17.50
2018-02-23	14.10
2018-04-06	5.50
2018-10-24	15.66
2018-07-25	13.80

4.3. Process Stability Analysis

Histogram

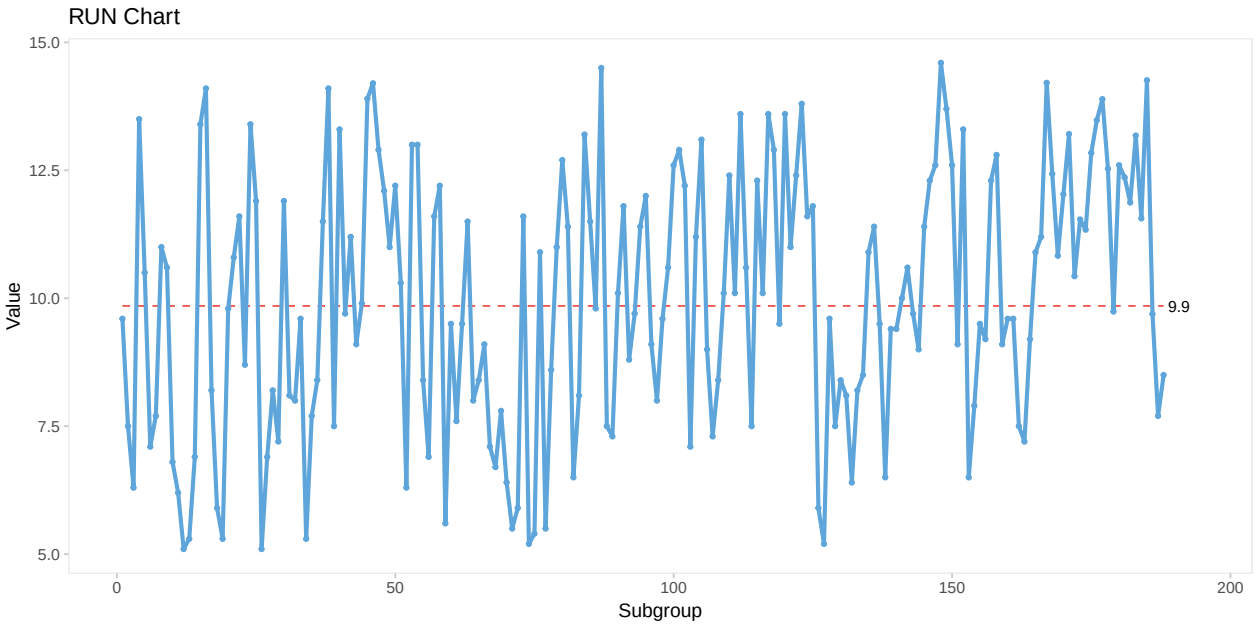



```
#Anderson-Darling Normality test
nortest::ad.test(spam_traffic_brate_vector)
```

```
##
##  Anderson-Darling normality test
##
## data:  spam_traffic_brate_vector
## A = 1.4212, p-value = 0.001098
```

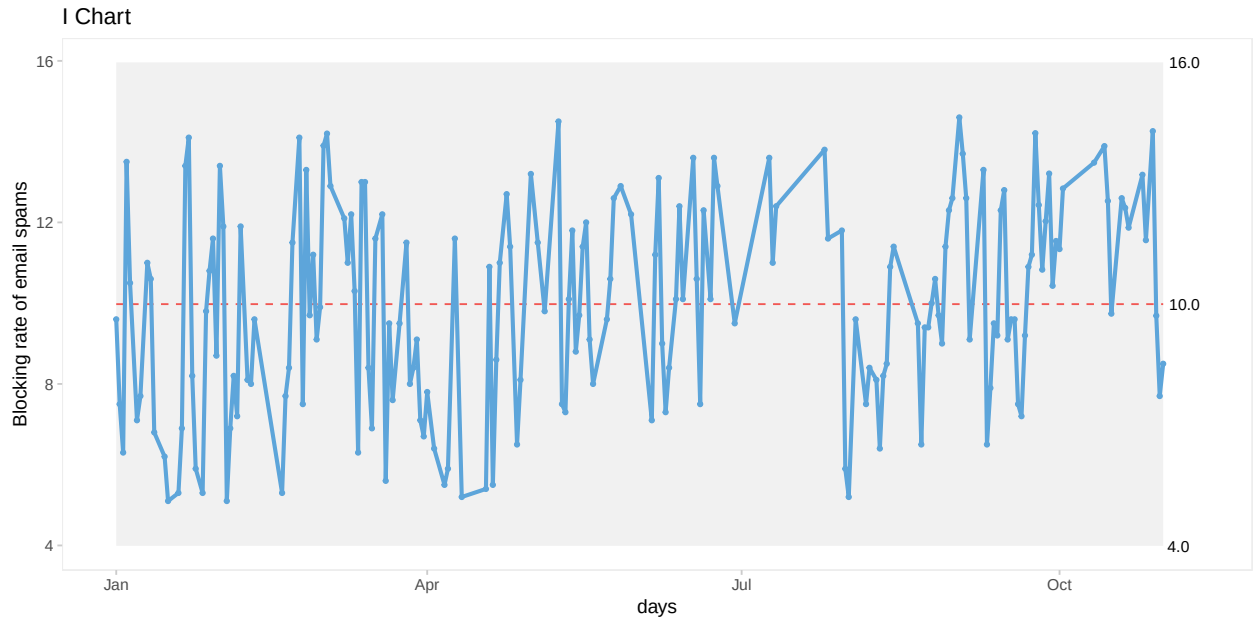
Here it's found that the spam blocking rates recorded are not following normal distibution and these can also be investigated on Run Chart below.

Run chart



Run charts has drawn the spam blocking rates for the days and they have correctly spread in random behavior which is an indication of statistical stability of spam classification process and this can also be predictable, and here the assignable causes have been studied and kept a side.

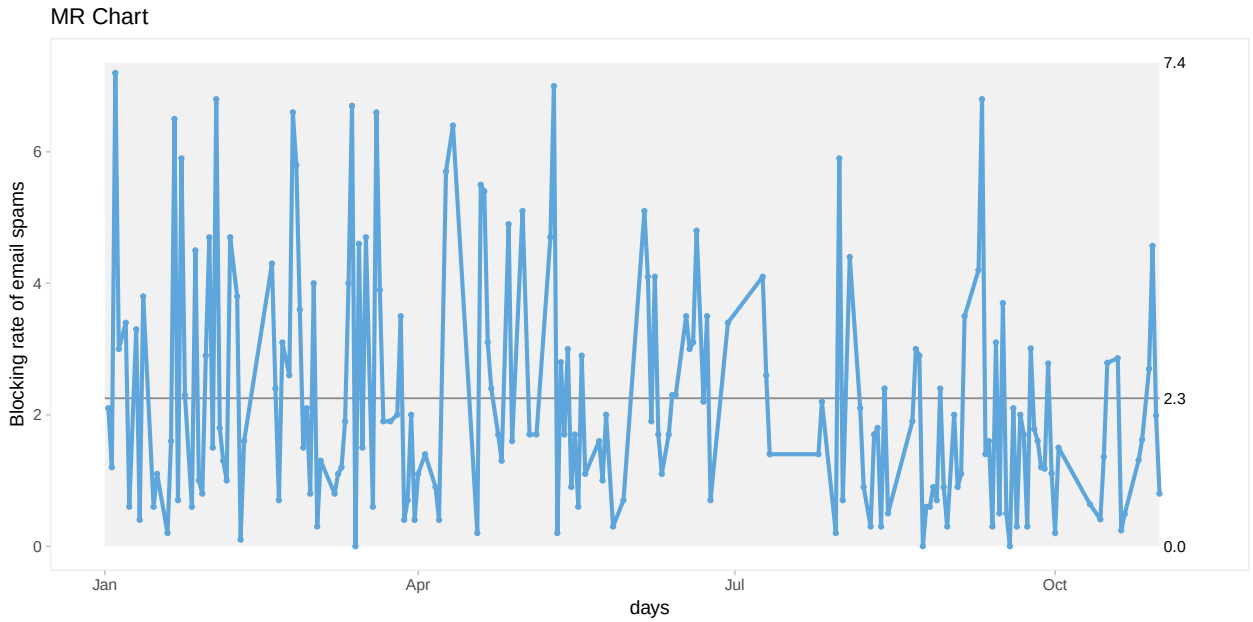
Process Control:I Chart



Here I-chart plots spam blocking rates of each day for the last 10 months and helps us to understand a process "of classifying messaging that improves spam blocking rates" mean is in stable/control state or not, The center line is an estimate of the process average, and it's found to be **10**, and The control limits **(UCL:16.0,LCL:4.0)** on the I chart which are set at a distance of 3 standard deviations above and below the center line, show the amount of

variation that is expected in the individual sample values.Here all the process observations have stranded in between the control limits, no observation has gone beyond the control limits, hence the said process is proved to be a stable and control.

Process Control:MR Chart



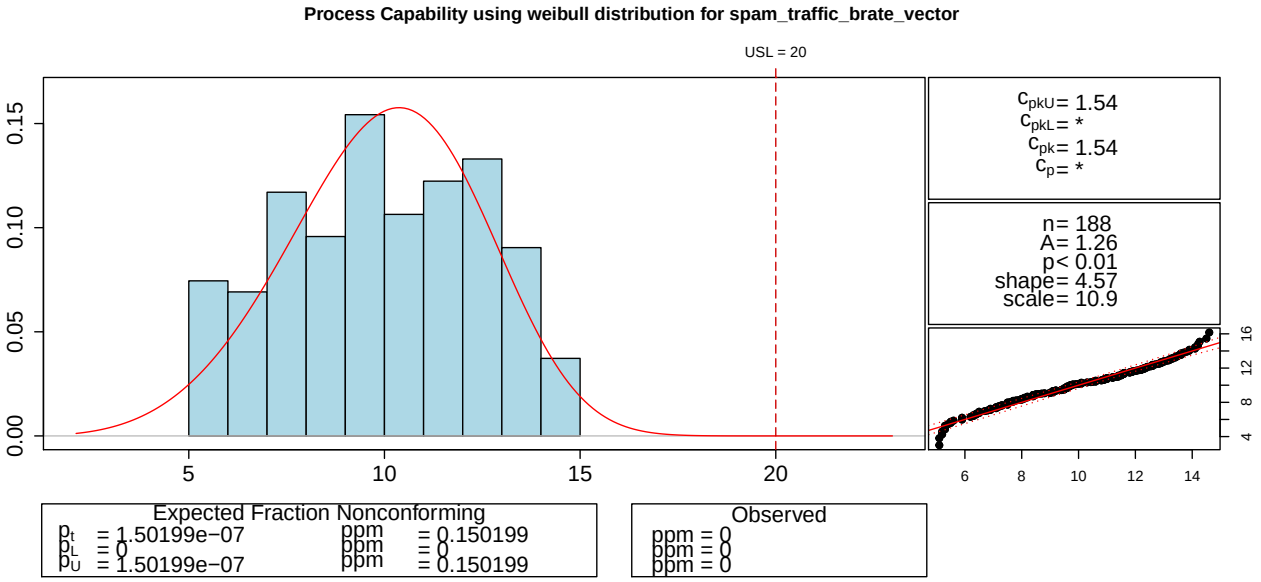
Here MR-Chart plots the moving ranges of spam blocking rates,The center line is the average of all moving ranges i.e **2.3**.The control limits on the moving range chart **(UCL:7.4,LCL:0.0)**, which are set at a distance of 3 standard deviations above and below the center line, show the amount of variation that is expected in the moving ranges of the standardized data..

4.6. Process Capability Analysis - One sided specification

```
# Test for weibull distribution
set.seed(200)
goft::weibull_test(spam_traffic_brate_vector)
```

```
##
##  Test for the Weibull distribution
##
## data:  spam_traffic_brate_vector
## p-value = 0.466
```

Since the available data on spam blocking rates found to be non-normal, a weibull distibution test has been formed on it and it's confirmed that it follows a weibull distrbution




```
##
##  Anderson Darling Test for weibull distribution
##
## data:  spam_traffic_brate_vector
## A = 1.2614, shape = 4.572, scale = 10.949, p-value <= 0.01
## alternative hypothesis: true distribution is not equal to weibull
```

Here a process capability analysis is summarized in indices, these indices show the ability of the process to meet the specified requierements, they can be monitored and reported over time to show how a system is changing, these indices are,

- **Potential capability** C_p: A simple and straight forward indicator of the process performance, it is the ratio of the allowable variation to thr actual natural variation.
- **Achieved capability** C_{pk}: It measures how close a process is to the target and simultaneously how consistent the process is around the average performance.

As our business specification is one sided USL:(10+10)=20 the Achieved capability C_{pk} is only considered to be confirming this process capable of meeting the business specification and it's caluclated at **1.54** that indicates that the process is capable of improving spam blocking rates.

4.4. FMEA (Failure Mode and Effects Analysis)

It is an engineering technique used to define, identify and eliminate known and /or potential failures, problems, errors and so on from the system, design, process, and or service before they reach the customer.

Any FMEA conducted properly and appropriately will provide the practitioner with useful information that can reduce the risk (work) load in the system, design, process and service.

Used to analyze services before they reach the customer. It focuses on failure modes (tasks, errors, mistakes) caused by system or process deficiencies.

Process step 1: Load messages from 7726 system to Security Center for classification

Potential Failure Modes, effects and causes:

1. There is a chance of getting Late report of suspocious/unwanted messages to 7726 system from Users due to this classification can't be done on time and spam can get leaked towards users, this failure can get occured where there is no proper connectivity in between SC, and 7726 systems
2. API Calls gets down between 7726 and SC when there is a huge flow of traffic in between them, when it happens no message will be appeared on SC, because of it no classification can be done and spam gets leaked towards users. here cause could be connectivity issues in SC and 7726 systems.
3. Security Center Storage Disk gets filled up when there is no memory space in SC, since then no message would get into Security center, it leads to no message gets classified and spam gets leaked towards users, here cause could be insufficient storge system built up in SC.

Process Step 2: Message Clustering

Potential Failure Modes, effects and causes:

Different variants/patterns of messages keep on getting sent out from spammers, in this case analyst has to spend more time to understand a message and get it classified to it's respective category,if he come acrooss 100's of variatns it would definetly be a herculean task to do classification, when there is no on time action taken on a spam message it would get leaked and lower blocking can also happen, here cause could be an Ineffective Fingerprint filter algorithm that we have been using in spam filtering.

Process step 3: Manual Messaging classification by analysts

Potential Failure Modes, effects and causes:

- Misclassification
  - Over blocking/lower blocking caused by Inadequate training/Knowledge transfer
- Delayed in classification

  - Spam leakages caused by Inadequate training/Knowledge transfer
- Delayed in processing classified messages
  - Spam leakages caused by Ineffective Fingerprint filter algorithm
- Security Center crashes down
  - Spam leakages caused by Improper System design/Architecture
- Inflexible Security Center UI/UX visibility/options
  - Time consuming for manual classification caused by Improper System design/Architecture
- No proper monitoring of logs during busy shift hours
  - Spam leakages caused by Shortage of analysts in team
- Deprioritize the classification task
  - Spam Leakages caused by Inadequate training/Knowledge transfer
- Left messages unclassified for longer time in Security Center
  - spam Leakages caused by shortage of analysts in team

Process step 4: Manual uploads of messages from TSM system to SC

Potential Failure Modes, effects and causes:

Analyst would have to look for spams on a system TSM where all the messaging traffic gets stored, sometimes users might not be complaining about unknown/unwanted messages to 7726, that's why the non reported messages are not appeared on to SC. in this case analyst would have an option to get the messages from TSM and upload it into SC via .csv files. there is a chance of misclassification of messages when a manual upload is carried out that leads to messaging traffic overblocking/lower blocking, here a cause could be Ineffective Fingerprint filter algorithm.

Process Step 5: Regular Interval Fingerprint filter catridge updates

Potential Failure Modes, effects and causes:

Once messages are classified as SPAM, they go on getting blocked out via fingerprinting filter, here fingerprinting filter gets updated in an time intervals i.e for every 3-5mins.In case of any delay occured in this update time there would be spam leakages as fingerprinting filter stop blocking spams. here cause could be Ineffective Fingerprint filter algorithm

Process Step 6: Review of Classified messages

Potential Failure Modes, effects and causes:

On everyday, a .csv file gets exported automatically with the messages classified on earlier day to figure out if there are any false positives or false negatives caused. this file can be review by an expert analyst without fail, here there is a duplication of work for the same task, and

Process/ Function	Potential Failure Mode	Potential Effects of Failure	SEV	Potential Cause(s) Mechanism(s) of Failure	OCC	Current Process Controls	DET	RPN
Load messages from 7726 system to Security Center for classification	Late report of suspocious/unwanted messages to 7726 system from Users	No classification, Spam leakage	9	Connectivity issues	1		2	18
	API Calls gets down between 7726 and SC	No classification, Spam leakage	9	Connectivity issues	1		2	18
	Security Center Storage Disk full	No classification, Spam leakage	9	Insufficeint Storage	1		1	9
Message Clustering	different variants/patterns of messages	time consuming for manual classification, lower blocking	9	Ineffective Fingerprint filter algorithm	9		9	729
							2	0
Manual Messaging classification by analysts	misclassification	over blocking / lower blocking	9	Inadequate training/Knowledge transfer	9		3	243
	Delayed in classification	Spam leakages	8	Inadequate training/Knowledge transfer	8		9	576
	Delayed in processing classified messages	Spam Leakages	9	Ineffective Fingerprint filter algorithm	8		9	648
	Security Center crashes down	Spam leakages	6	Improper System design/Architecture	3		3	54
	Inflexible Security Center UI/UX visibility/options	time consuming for manual classification	2	Improper System design/Architecture	2		2	8
	No proper monitoring of logs during busy shift hours	Spam leakages	6	Shortage of analysts in team	5		5	150
	Deprioritize the classification task	Spam Leakages	7	Inadequate training/Knowledge transfer	5		5	175
	left messages unclassified for longer time in Security Center	spam Leakages	8	shortage of analysts in team	5		5	200
Manual uploads of messages from TSM system to SC	misclassification	over blocking	9	Ineffective Fingerprint filter algorithm	6		3	162
	Security Center crashes down	Spam leakages	6	Ineffective Fingerprint filter algorithm	5		3	90
Regular Interval Fingerprint filter catridge updates	Delayed in catridges updates	Spam leakages	7	Ineffective Fingerprint filter algorithm	6		5	210
	Shorter time-to-live of catridge in blocking	lower blocking/spam leakages	9	Ineffective Fingerprint filter algorithm	7		8	504
Review of Classified messages	misclassification	Over blocking, time consuming	7	Inadequate training/Knowledge transfer	6		5	210
	Duplication of work	time consuming	5	Inadequate training/Knowledge transfer	5		5	125

Figure 3: FMEA

even an expert analyst might misclassify messages manually if he has messed up with other taks or doen't have an idea about it. here cause could be Inadequate training/Knowledge transfer of analyst

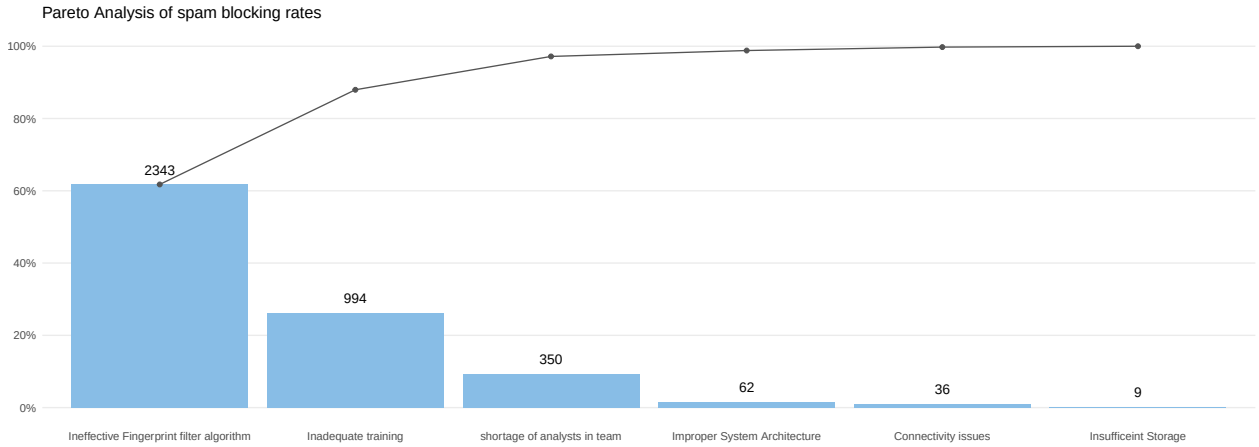
4.5. Pareto Analysis

Pareto analysis is a statistical technique that is used in decision making for the selection of the limited number of tasks that produce the most significant overall effect. It uses the concept based on identifying the top 20% of causes that need to be addressed in order to resolve 80% of the problems.

The potential causes for having recorded lower/improper spam blocking rates are identified as

1. *Insufficient storage*
2. *Connectivity issues*
3. *Improper system architectures*
4. *Shortage of analysts*
5. *Inadquate training*
6. *Inefficeint fingerprint filter algorithm*

These causes can be drawn on pareto chart to investigate which one of them have occured more to make spam blocking rates down or let spams allowed.



From the above pareto diagram it is insighted that about **&95% ofspam leakages/improper blocking has happend because of having ineffective fingerprint filter algorithm, inadquate training to analyst and shortage of analysts in team**.

5. DMAIC : Analysis

Part 1. Exploratory Data Analysis on SMS-EMAIL Messaging Traffic Datasets

Exploratory Data Analysis (EDA) is used on the one hand to answer questions, test business assumptions, generate hypotheses for further analysis. On the other hand, we can also use it to prepare the data for modelling. The thing that these two probably have in common is a good knowledge of our data to either get the answers that we need or to develop an intuition for interpreting the results of future modelling.

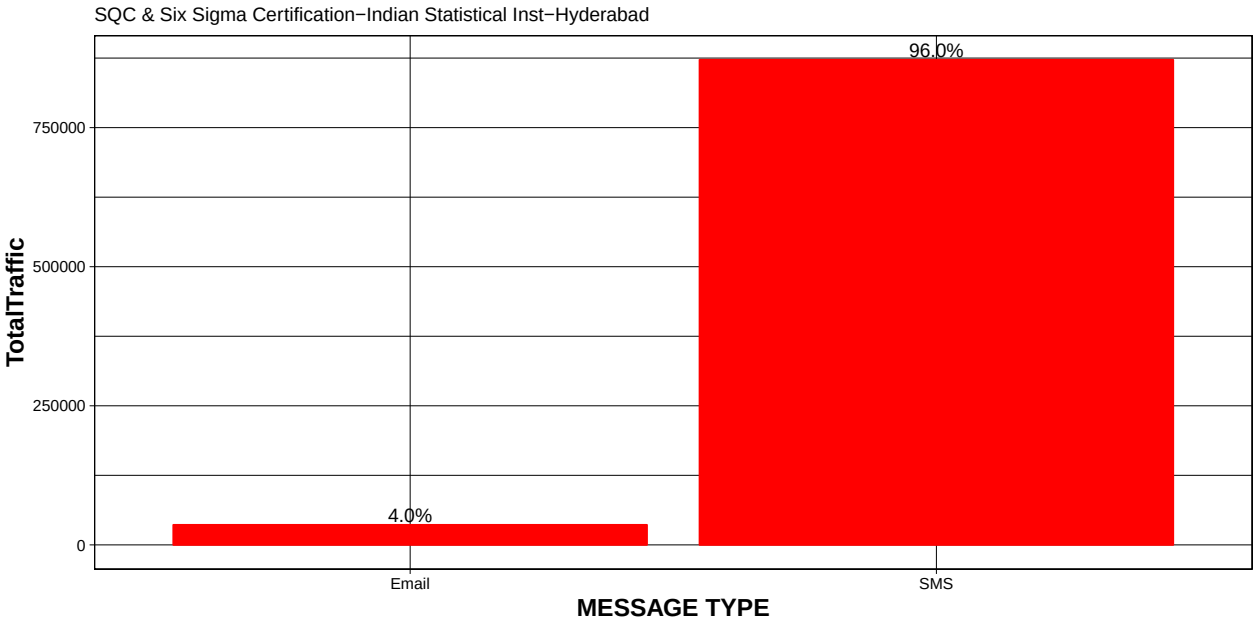
There are a lot of ways to reach these goals: we can get a basic description of the data, visualise it, identify patterns in it, identify challenges of using the data, etc.

One of the things that we will often see when we're reading about EDA is Data profiling. Data profiling is concerned with summarising our dataset through descriptive statistics. We want to use a variety of measurements to better understand our dataset.

The goal of data profiling is to have a solid understanding of our data so we can afterwards start querying and visualising our data in various ways. However, this doesn't mean that we don't have to iterate: exactly because data profiling is concerned with summarising our dataset, it is frequently used to assess the data quality. Depending on the result of the data profiling, we might decide to correct, discard or handle our data differently.

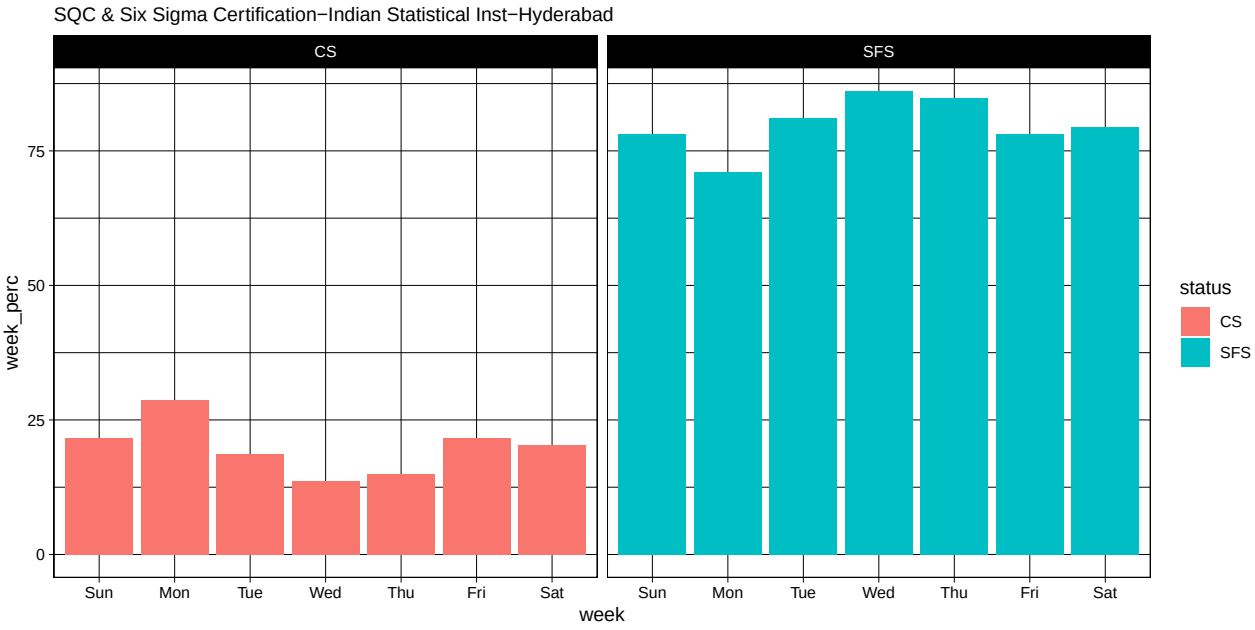
I have made use of R packages called **tidyverse, and data.table** to get EDA carried out as spelled out in the below code chunks. Here I have framed up about 20 EDA questions which help us to undestand how have been the potential causes,followed by hypothesis testing to determine if they are true causes of the problem.

EDA1: How are the messaging volumes?



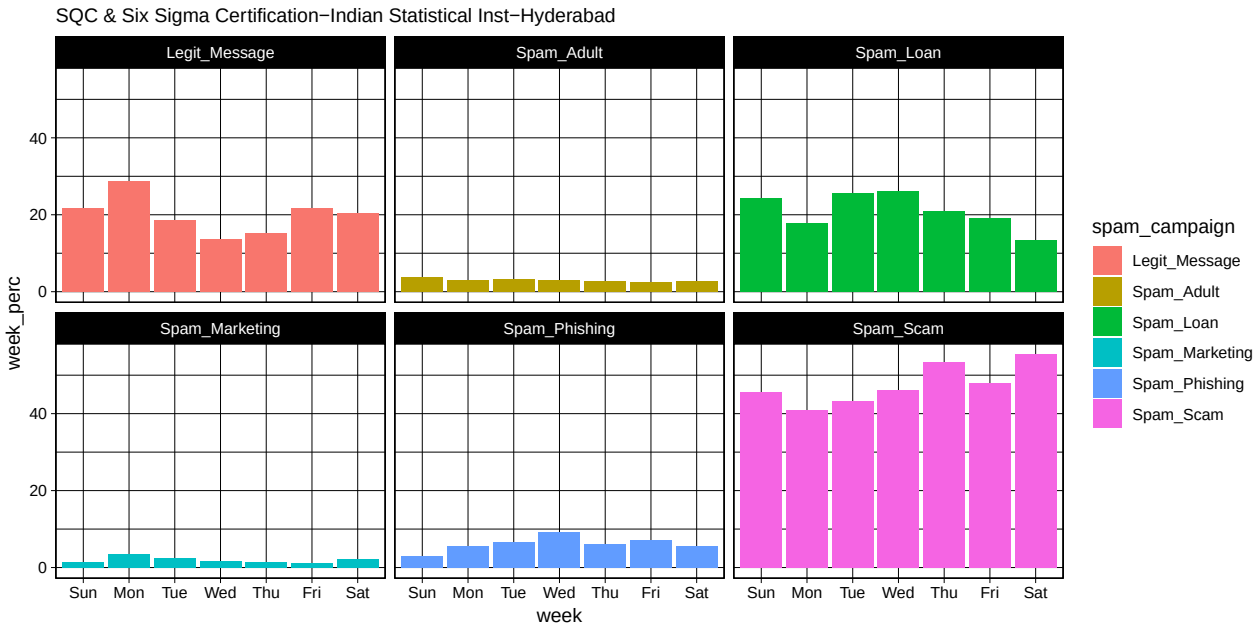
In last 10 months of messaging traffic, we have been able to classify about 1 Million messages in our security center, it's clearly known/fairly true that email messaging traffic(4%) volumes are lower than sms messaging traffic(96%),we provide anti-spam service solutions for controlling both email and sms spams

EDA2: What percent of messaging traffic has been sent in and which messaging status has got more count?



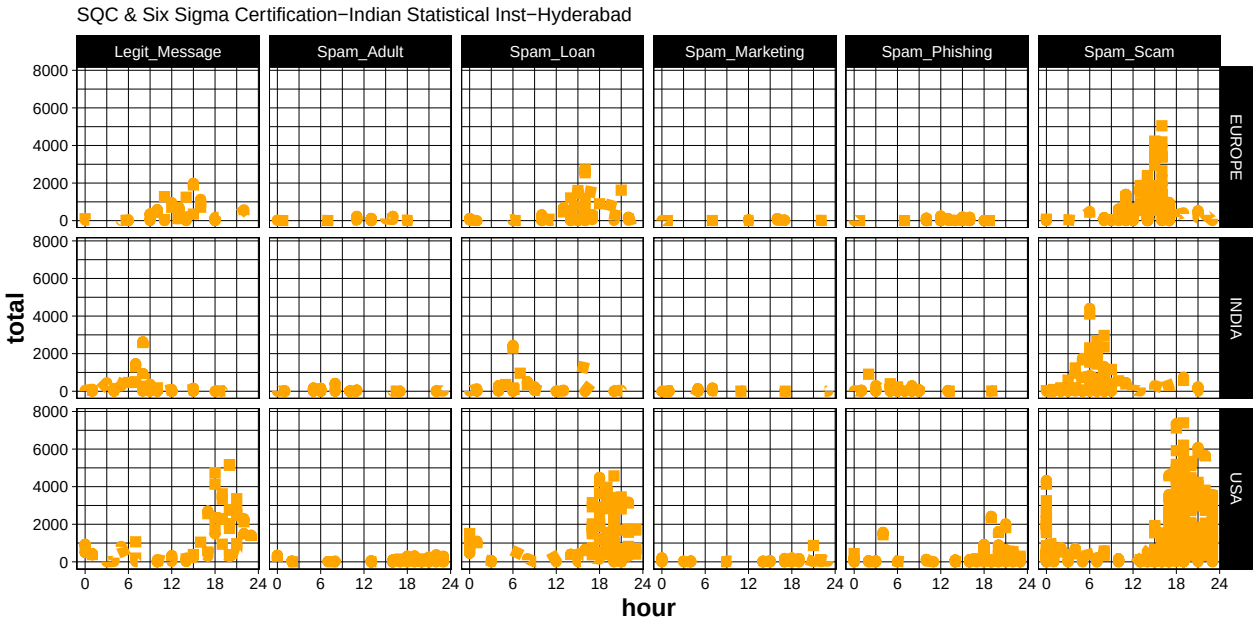
SFS refers to a message which is classified as spam and sent to blocking, similarily CS indicates to a message which is classified as Legit message and no blocking shoud happen on it. SFS status messages found to be higher(75-80%) than the CS classified messages(20-25%), and The spam traffic have got larger volumes during the mid of weekdays,

EDA3: What percent of messaging traffic has been sent in and blocked out per each spam category?



The each and every classified spam messages are assigned a spam category to which it belongs, here we have 5 typical spam campaign categores, phishing spams would be more effect to subscribers when it's not controlled on time.

EDA4 : How has been Messaging traffic per hour during the three different timezones ?

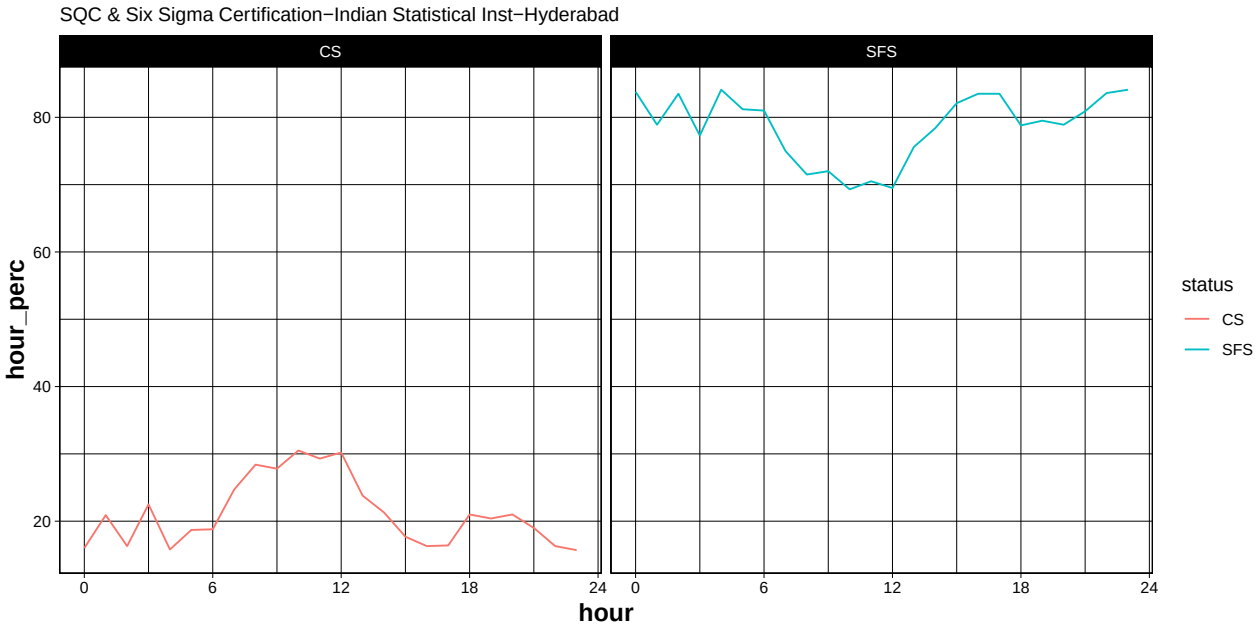


During US Shiff, the spam traffic flow of scams/loans/phishing's found to be coming in higher volumes in between afternoon to midnight hours i.e 12-23 and even the legitimate traffic is being sent in more in these hours. since the huge traffic flows in this shift lead analyst should focus on marking as soon as they reach to Security Center.

During INDIA shift, there are only 2 major spam campaign's flows found to be a moderate level, from the midnight hours to early morning hours spam traffic is in active for about 2-3 hours only. because of the non busy hours lead analysts would be able to do other task apart from marking messages in SC.

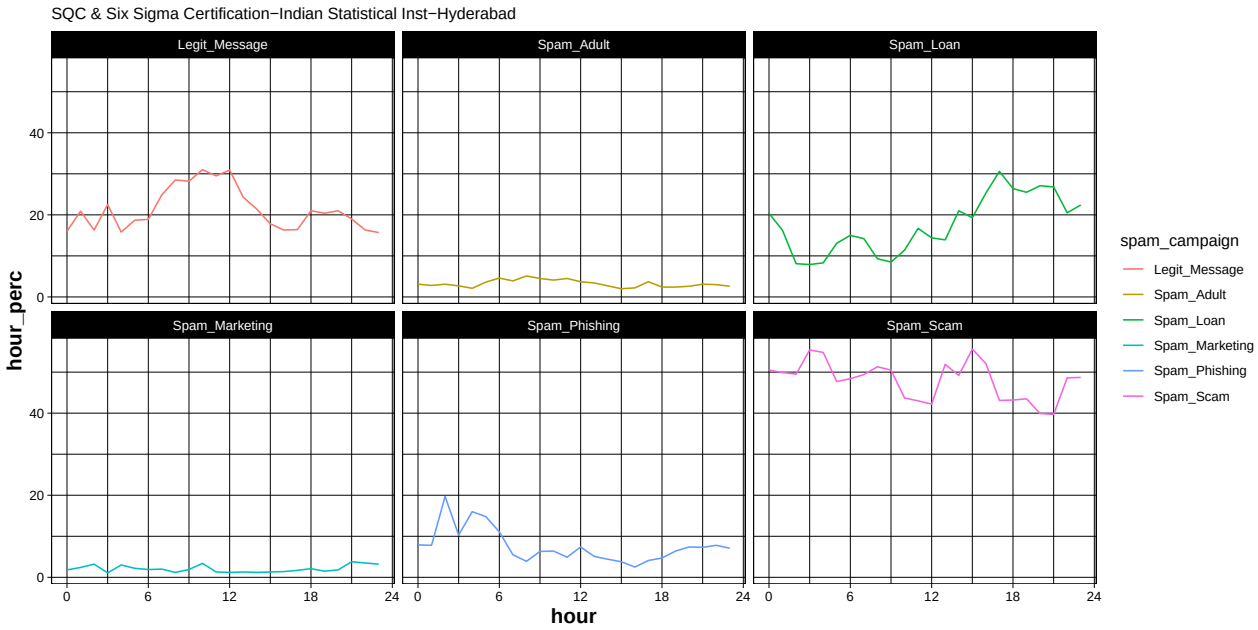
During Europe Shift, ere there are only 2 major spam campaign's(scam/loan) flows found to be a higher levels in early morning hours for about 4-5hours, so lead analyst would also get free to concentrate on doing other tasks.

EDA5 : How has been the Messaging classifcation per hour?



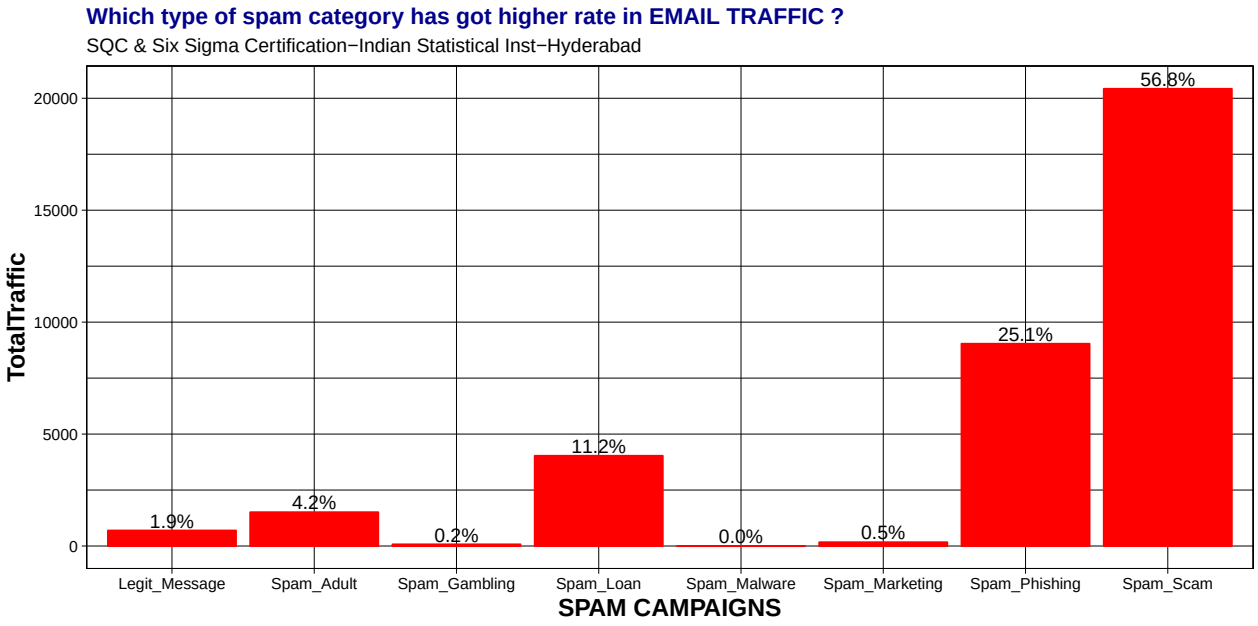
The spam traffic(SFS) has been classified in between 70-90%, and the Legitimate traffic(CS) is in between 10-30% through out the day, It seems the spam marking has come down from hour 6-12, in contrast the legitimate messages has significanlty gone up, here in these hours Europe team takes care of controlling spams.

EDA6 : What type of spam have been getting caught up per hour?



Spam-Loan, and Spam-Scam are seen getting blocked out at higher rate during US Shift hours only, interestingly Spammers targeted on subscribers by sending out phishing messages heavily in the midnight hours, spam adults and marketings are found to be lower levels thought the hours.

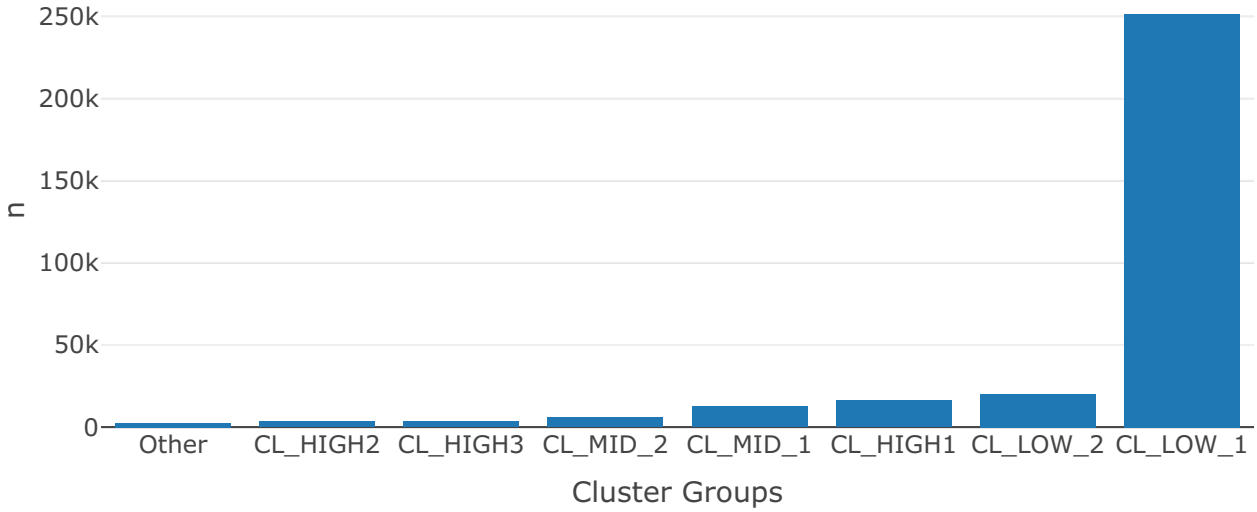
EDA7 : Which type of spam category has got higher rate in EMAIL TRAFFIC?



Here this graph gives us an insights of only EMAIL traffic, ss we already know spam scam and spam_loan are being classifed at higher rate in whole traffic, however in case of email top 3 major spam campaigns are drawn as spam_scam(60%), spam_phishing(25%) and

spam_loan(11%)

EDA8: In Which cluster group the messages get clustered higher?

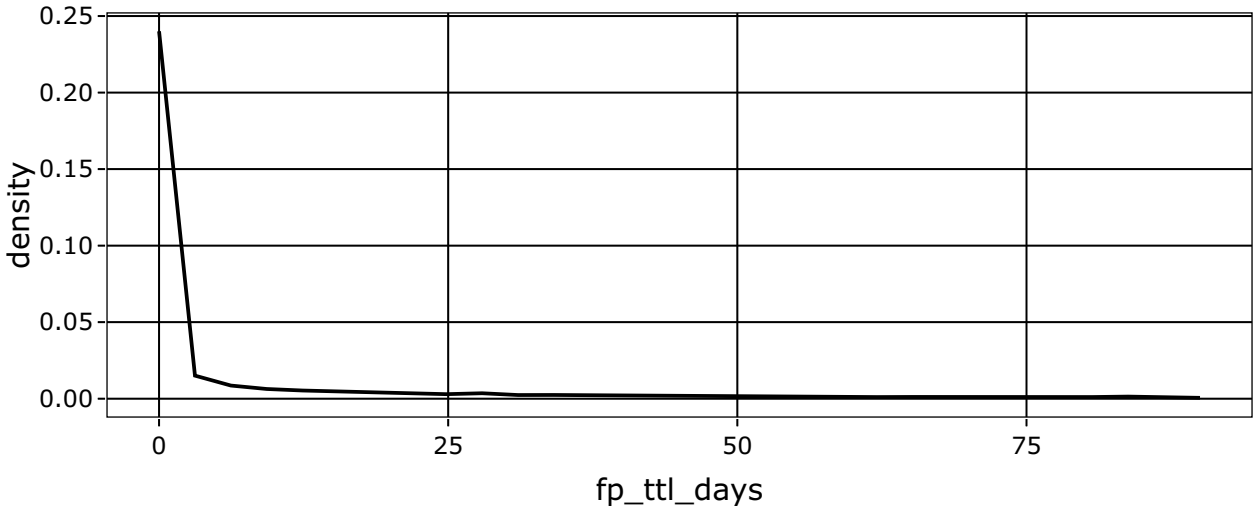


Here it is clearly graphed that about 80% of messaging traffic has been under cluster group id cl_low-1, It means that each messages would get clussted only for 1-2 times as their patterns go on changing. the messages which don't have more patterns have got added up in to other cluster groups where it's total are minimal.

Cluster group low 1 and low 2 have consisted of messages which have got clustered only for 1-2-3-4 times, lets assume that a message like "your account has been locked, please get it activated here http//:xxx.abb" has come into security center, and classified as Spam, at this time it's cluster size is 1, after a while the same message has appeared in Security centre, and got clustered to earlier fingerprint, now it's cluster size is 2 and it goes on increasing if the same message is repeatdly sent in. If a message is sent out only one time it's cluster size would always be 1, similarily if a message is sent out for 100times, it gets clustered and cluster size would be 100.

In a conclusion we have been receving messaging traffic with variuos pattern.

EDA9: How is the time to live of fingerprints in clsuter group low- 1 and low-2?



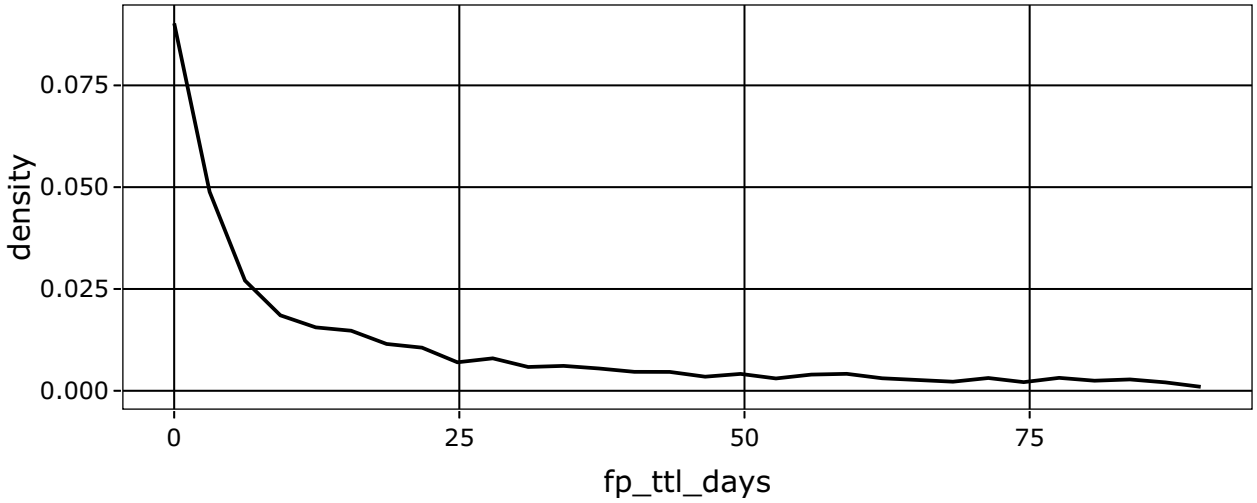
Here we are trying to understand how long a classified fingerprint is in active state, it's named as fp_ttl_days and it gets caluclated as the difference between time on which a message is in and time on which a message stop coming in.

As we can seen in graph,the time to live(fp_ttl) of fingerprints in cluster groups low1 and low2has stranded in between 0 and 1 days, there is no use of a fingerprint if it's not actively being blocked out. the reason for this lower ttl time could be that spammers have kept on changing the message patterns that wouldn't be matched with similarity percent.

if the fp_ttl_days doen't get better we are unnecessarily spending time for manual messaging classification.

Here we are trying to make use of message contents sent from spammers for further implementations in our security center.

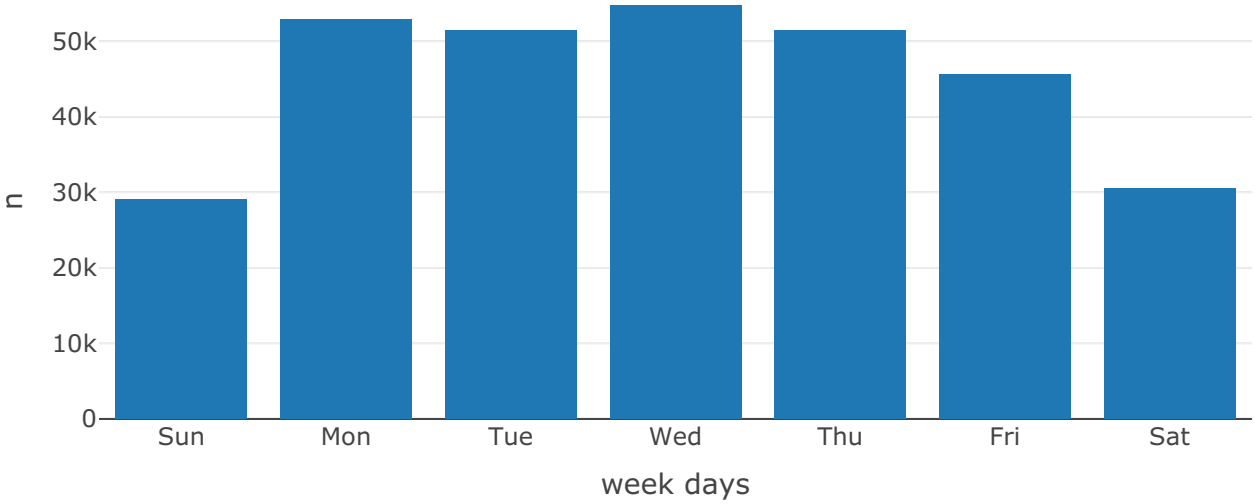
EDA10: How is the time to live of fingerprints in clsuter group mid- 1, mid-2 and high1?



Here we are seeing the fp_ttl_day of the messages who have got cluster size in between 5 to

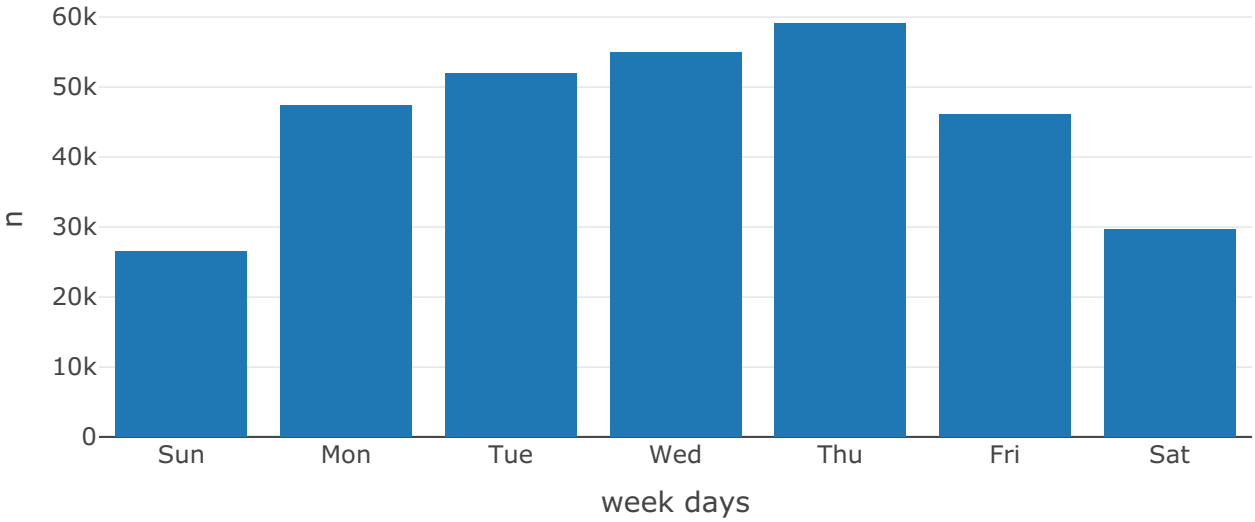
50 i.e under categories mid-1,mid-2 and high 1, fp_ttl days of messages has been significantly coming down from 0 to about 10 days, and 10-14th days onwards it's has got flatten and constant. it's reasonably OK that the more cluster group a message has the more days it's in active state.

EDA11: On which days the messages(first message received) are being sent in to US SC ?



We can find out that about 80% of the messaging traffic have been received from days monday to friday, and during weekends saturday and sunday it's about 20%. keeping it in a note we can assign classification tasks to 1-2 analysts in weekends, and 5 analysts requiered for weekdays.

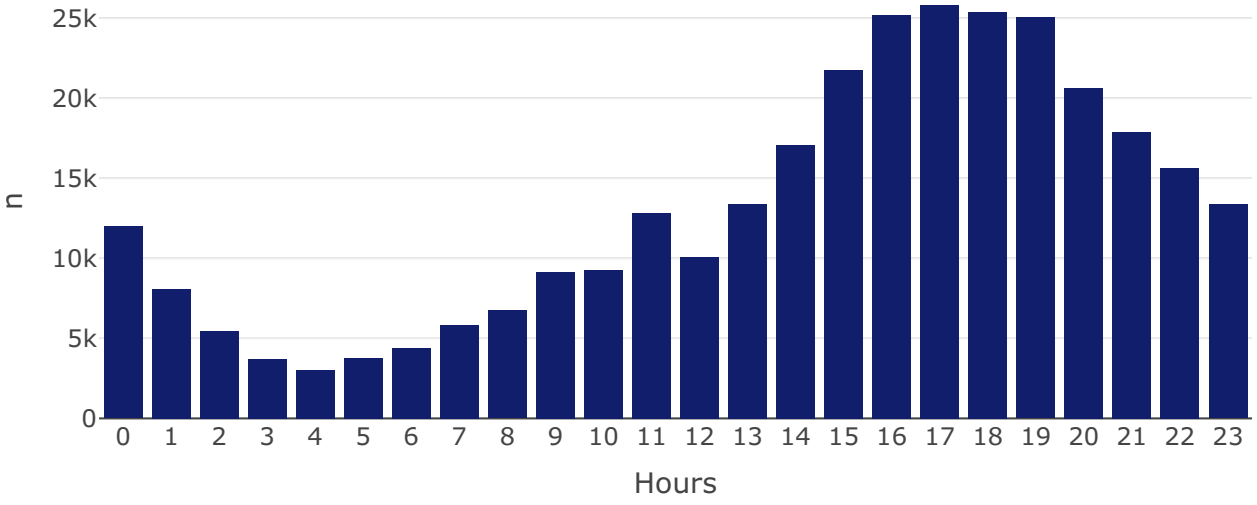
EDA12: On which days the messages(last message received) are being sent and clustered in US SC ?



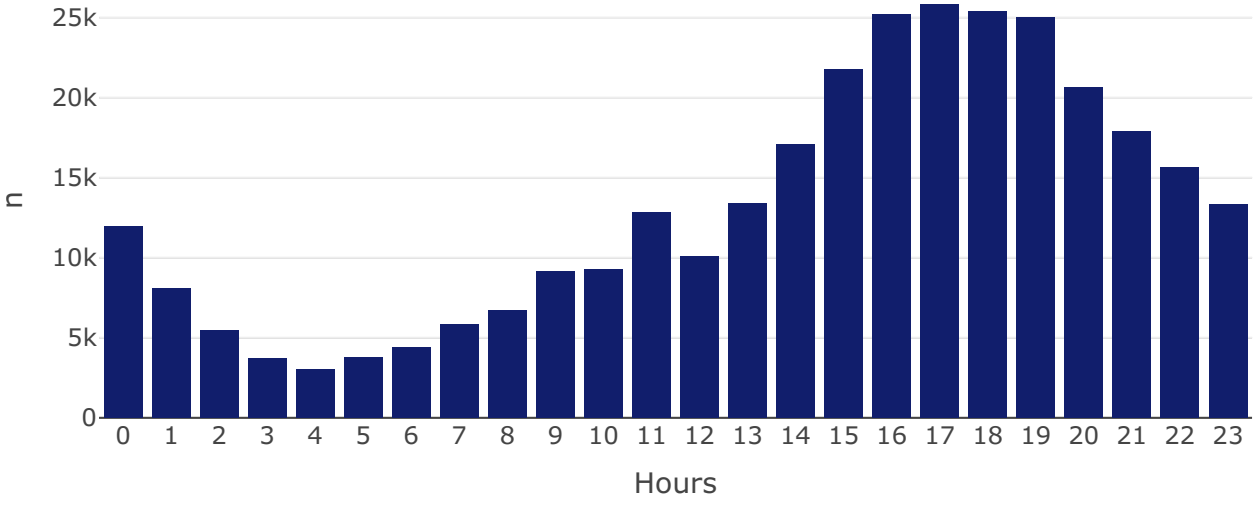
Here also about 80% of the messaging traffic have been updated from days monday to friday,

and it's helping us to understand the earlier analysis question of determining ttl_days of message, if we compare this graph with above graph on most of the days the counts seems to be matched, it means that the messages are not active for longer days.

EDA13: How is a flow of the messages(first message received) in US SC per hour?

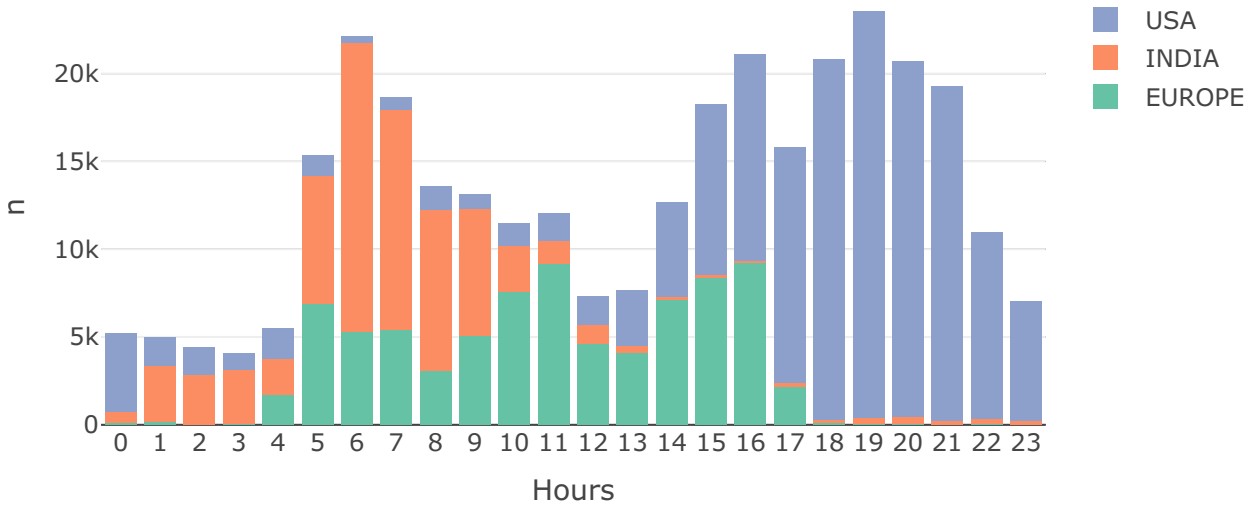


EDA14: How is a flow of the messages(last message received) in US SC per hour?

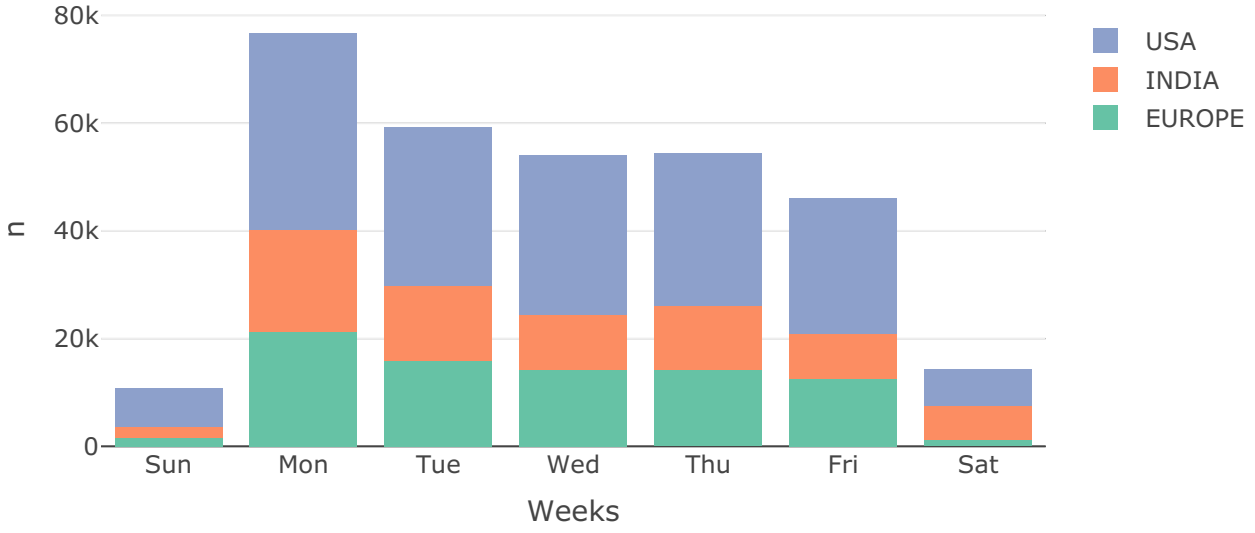


The messaging traffic inflow starts getting clustered to similiar messages in from morning 9 to midnight 0Hrs as well when USA shift/India team is ON, last night hours are covered by india team only.

EDA15: How are the messages are being classified in US SC per Shift?

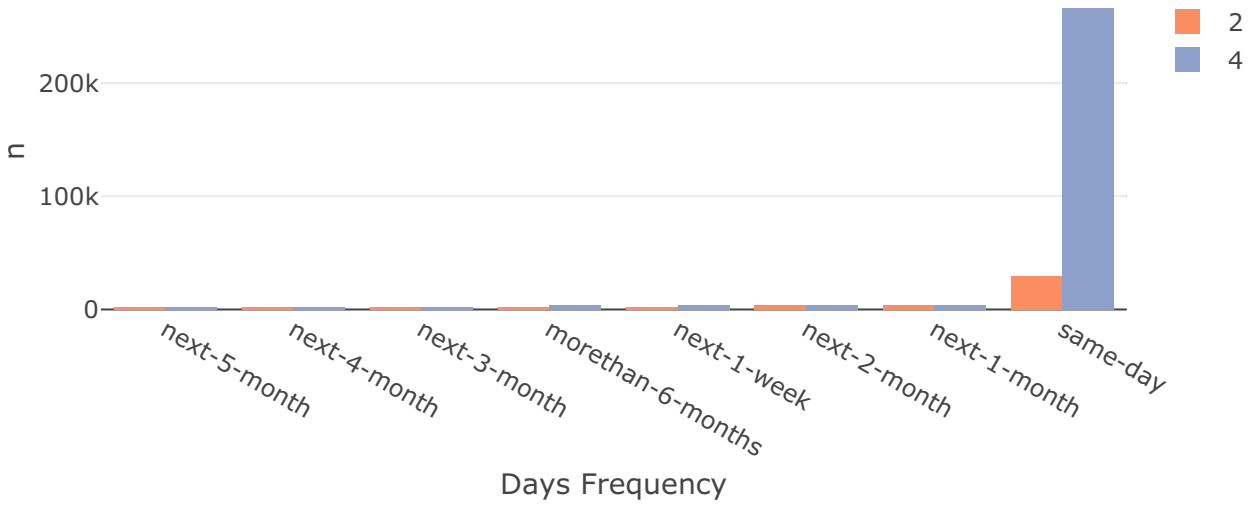


From afternoon 13 to midnight 23 all the messaging traffic has been handled by USA team members, the full night and early morning hours traffic were found at moderate levels and they were controlled by India/Europe Teams.



Like said eariler classification of messages have been lower during weekends in all the 3 shifts, Europe shift team have got very few amount of traffic in saturday and sunday, and the traffic found to be huge volumes in weekdays i.e monday to friday in all the shift.

EDA16: How long a classified fingerprint is in active blocking mode ?

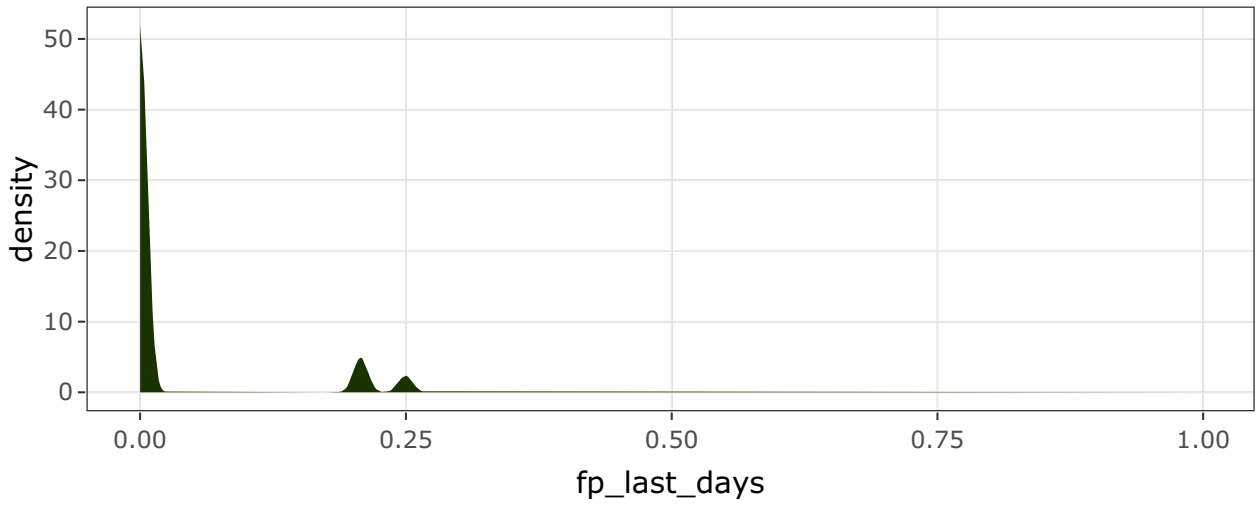


Here,status-4 refers to a message which is classified as Spam, and status-2 refers to a message classified a clean.

fp_last_days is caluclated as the time difference between processing date and last_update date of a fingerprint message, for example, a message has entered into SC on a date 2018-12-01:09:00, next its processed by classifying into Status-4(Spam) at 2018-12-01:09:05, after a while a similar message got clustered to this message and it's updated time at 2018-12-01:10:00, from here onwards no message has been clusted to it, fp_last_days will be estimated keeping the menioned formula in mind. Here fp_last_days has been binned into a different categories based on caluclated fp_last_days range.

About 80% of classified spam messages were in active for one day only (same-day refers to it in above graph), very few amount of spam fingerprint were in a state of active blocking for more than a week. and no fingerprint has last for more than 4 months.

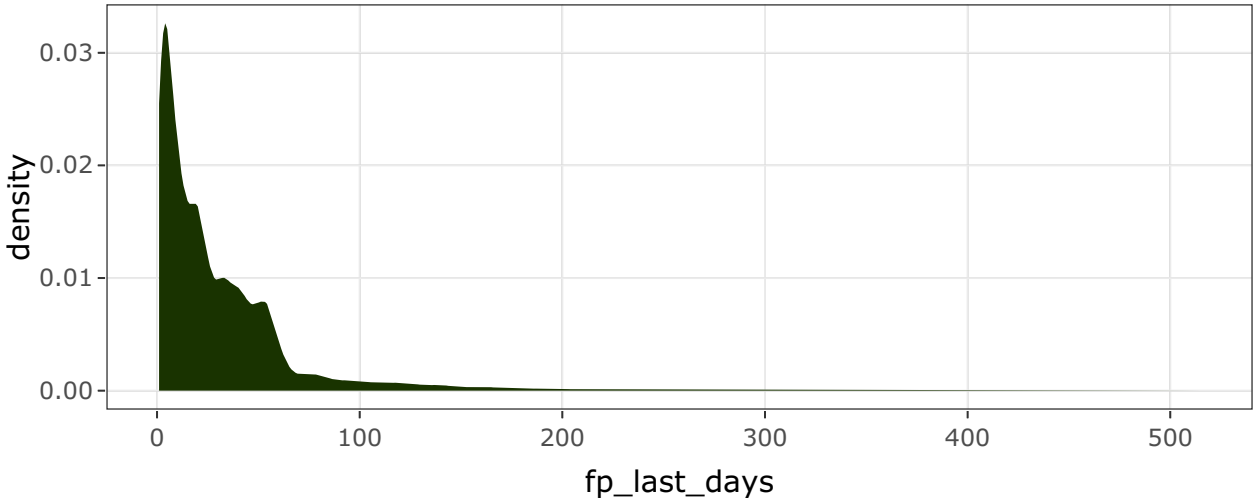
EDA16 Case 1: Trends of classified messages which have last for only one day



It is an evidance that majority of one day fingerprints have been in active for only few minutes

i.e 1hr(0.00), there has also been a little trend at 0.25 as we can see it on graph.

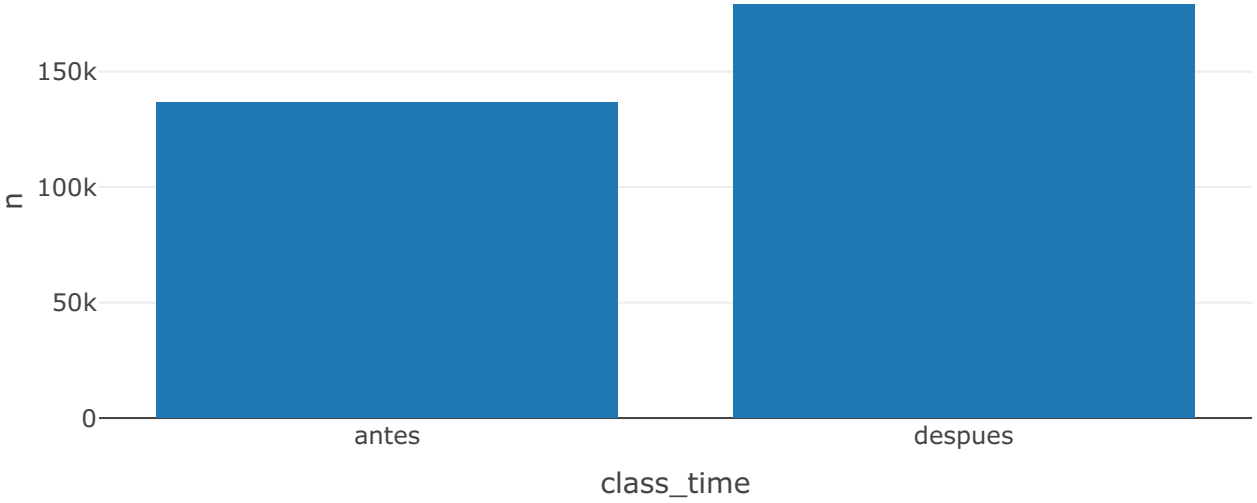
EDA16 Case 2: Trends of classified messages which have last for more than one day



Here as days go by, fp_last_day's of messages has been drastically falling down and no message has crossed over 5-6months

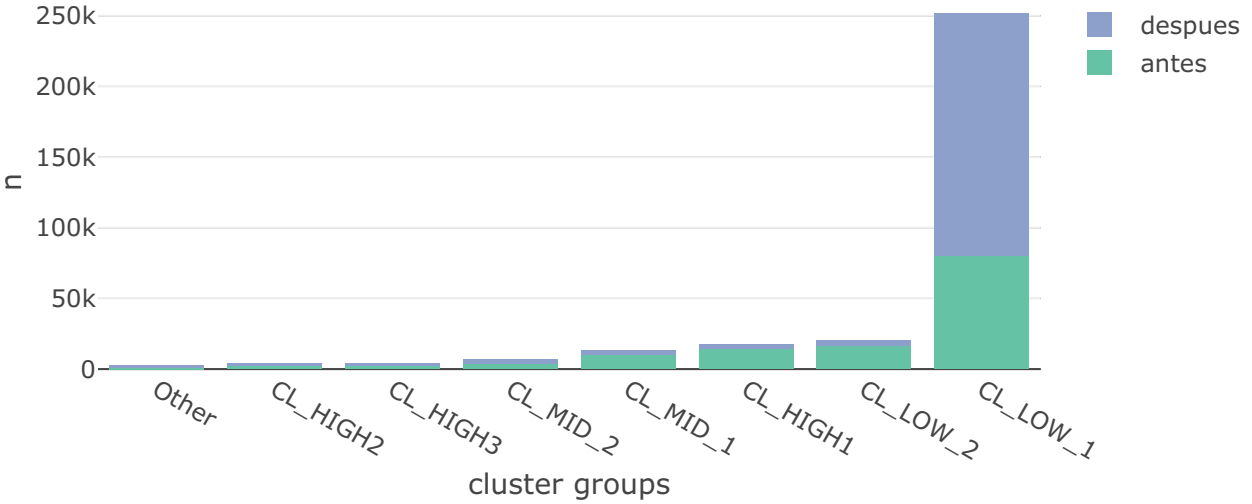
EDA17: Are the messages classified on time ?.

EDA17-Case-1: When a message is classified?



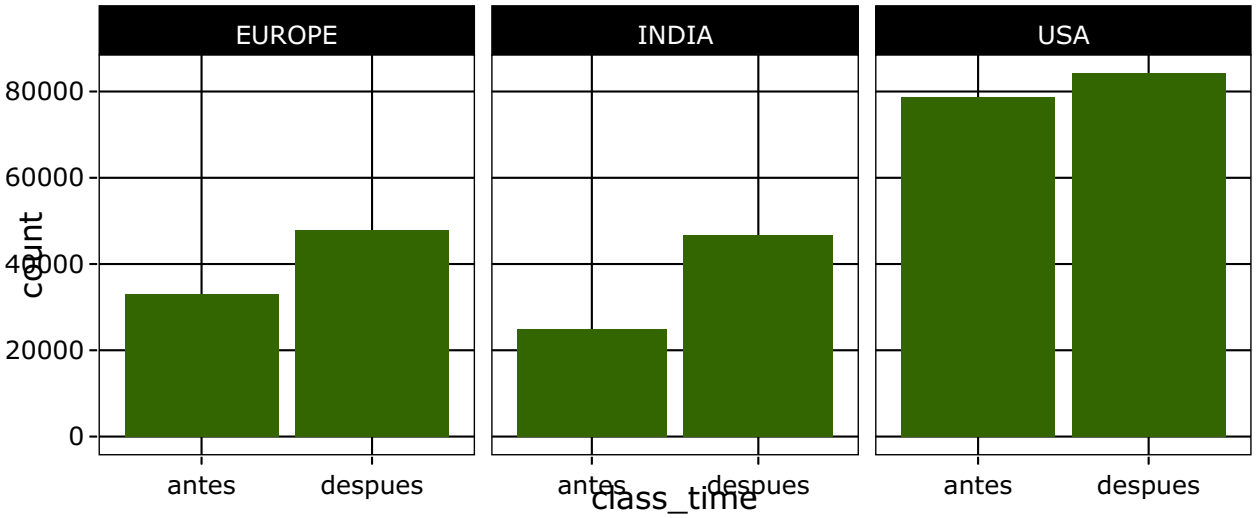
Here its found that about 60% of messaging traffic had been classified/clustered after they were processed in security center, and about 40% of traffic had been received before they were processed in security center.

EDA17-Case-2: Classification time across cluster groups



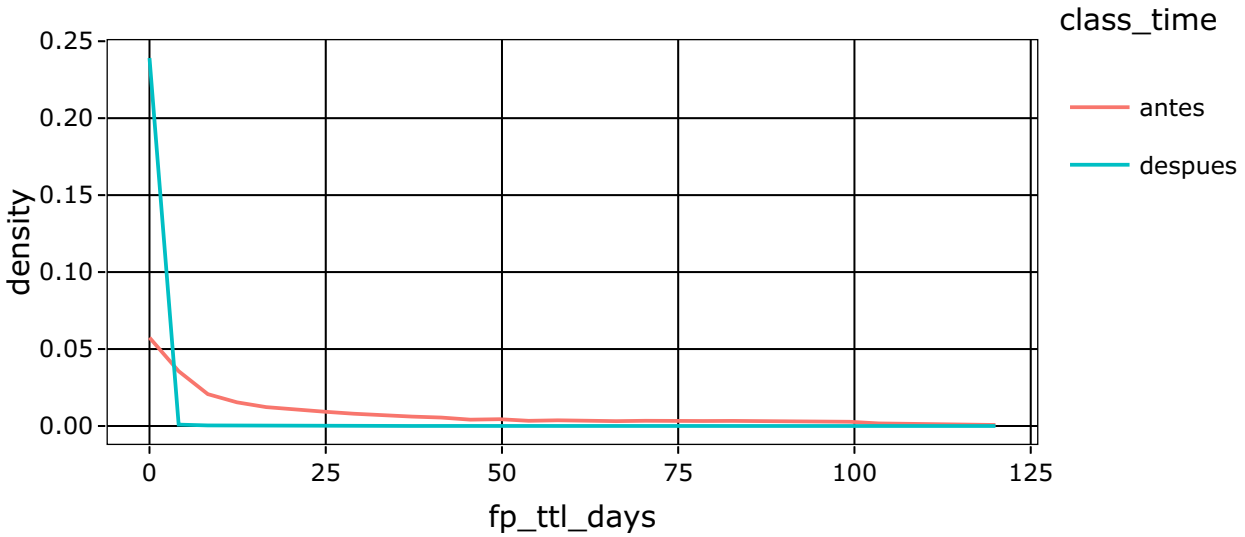
Here CL_LOW_1 cluster group is mixed up with the two classes(antes,despues-higher proportion), it means that the many messages which have got clustered for only 1 or 2 times had received after they were processed. and in the remaining cluster groups, "antes" classification category messages were seen higher, it means the larger cluster group messages were processed after they received.

EDA17-Case-3: Classification time over the three different shifts



The classification time counts of categories(antes and despues) during the india and europe shifts seems to be equal, where as in USA shift there has been a little difference these two category counts, by looking at antrs category count we can understood that there was a delay in classifying messages during US shift.

EDA17-Case-4: Classification time versus fp_ttl_days of fingerprints



Here it's found that the fingerprints that have been applied to a classification category-despues were not in active state for longer days. and they have last more for only 1 day. the antes category fingerprint were some what better than despues, they have been in active for fewer days.

Part 2: Statistical Inferences on SMS-EMAIL Messaging Traffic Datasets

Statistical inference is a process of drawing conclusions about population parameters based on a sample taken from the population.

Hypothesis:

In a hypothesis test, we will use data from a sample to help us decide between two competing hypotheses about a population. We make these hypotheses more concrete by specifying them in terms of at least one population parameter of interest. We refer to the competing claims about the population as the null hypothesis, denoted by
H0, and the alternative (or research) hypothesis, denoted by Ha The roles of these two hypotheses are NOT interchangeable.

Randomization:

We can use hypothesis testing to investigate ways to determine,for example, whether a treatment has an effect over a control and other ways to statistically analyze if one group performs better than, worse than, or different than another.

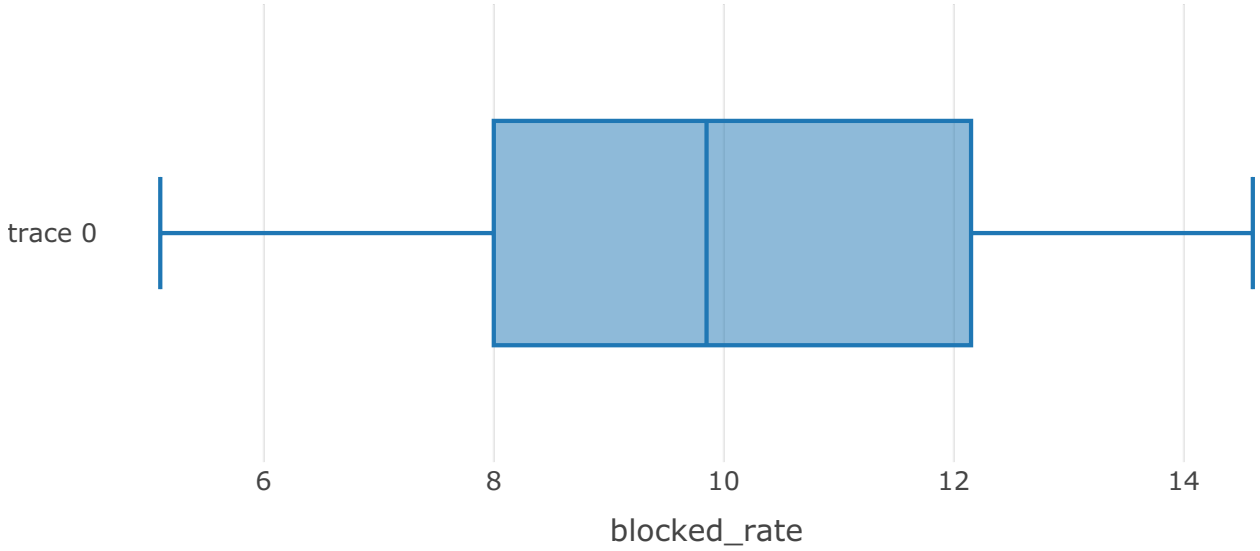
We are interested here in seeing how we can use a random sample of spam traffic and a random sample of non spam traffic from email-sms messaging traffi to determine if a statistical difference exists in the proportion traffic of each group.

Research Question 0:

Null Hypothesis(H_o): The variables total no.of classified spam messages and blocking rate(%) in a day aren't linearly associated in positive direction.

Alternate Hypothesis (H_a): The variables total no.of classified spam messages and blocking rate(%) in a day are linearly associated in positive direction.

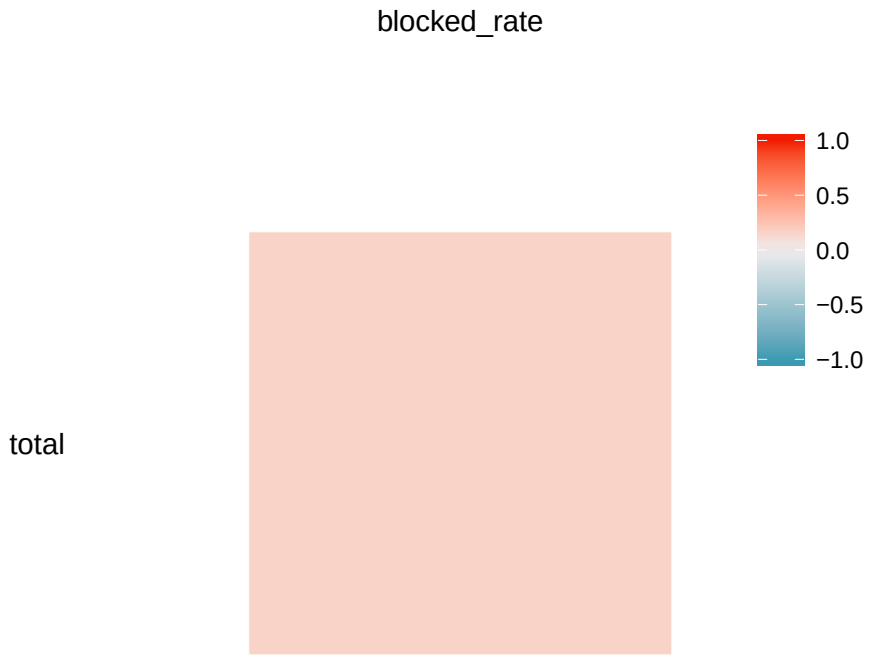
```
#Visualize how are the blocking rates
trafico_de_email_sfs_brate_df_sin_out %>%
  plot_ly(x=~blocked_rate) %>%
  add_boxplot()
```



It's observed that the blocking rates data are non-normal,and left skewed. and there are aslo few outliers

1) 25% are lower than 8%
2) 25% of observations are between 8-11
3) 25% of observations are between 11-12.5
4) 25% of observations are higher than 12.5

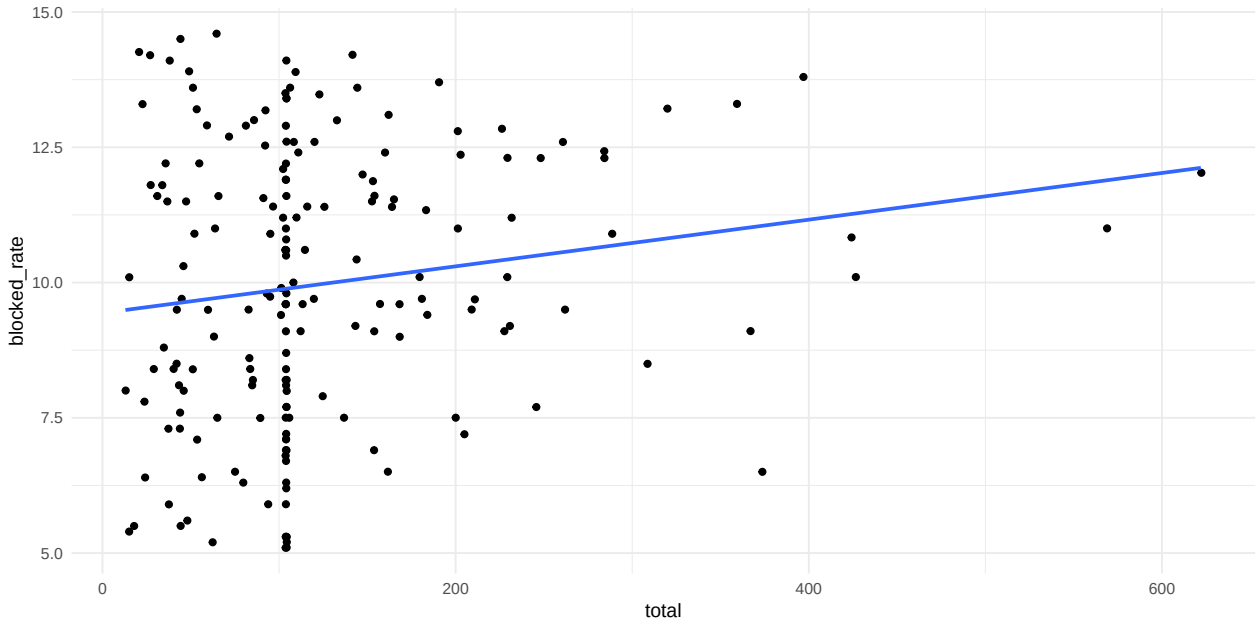
```
# checking correlation between the spam total and blocked rate
GGally::ggcorr(trafico_de_email_sfs_brate_df_sin_out)
```



The correlation coefficient of 0.125 indicates that the relationship between day

total and blocking rate is “weakly positive.”

```
#how has been a change in blocked rate per total in a day understood in a scatter plot
trafico_de_email_sfs_brate_df_sin_out %>%
  ggplot(.,aes(total,blocked_rate)) + geom_jitter() + geom_smooth(method = "lm", se = F
```



Specifying a null hypothesis blocked rate(Y) is a function of total(X)

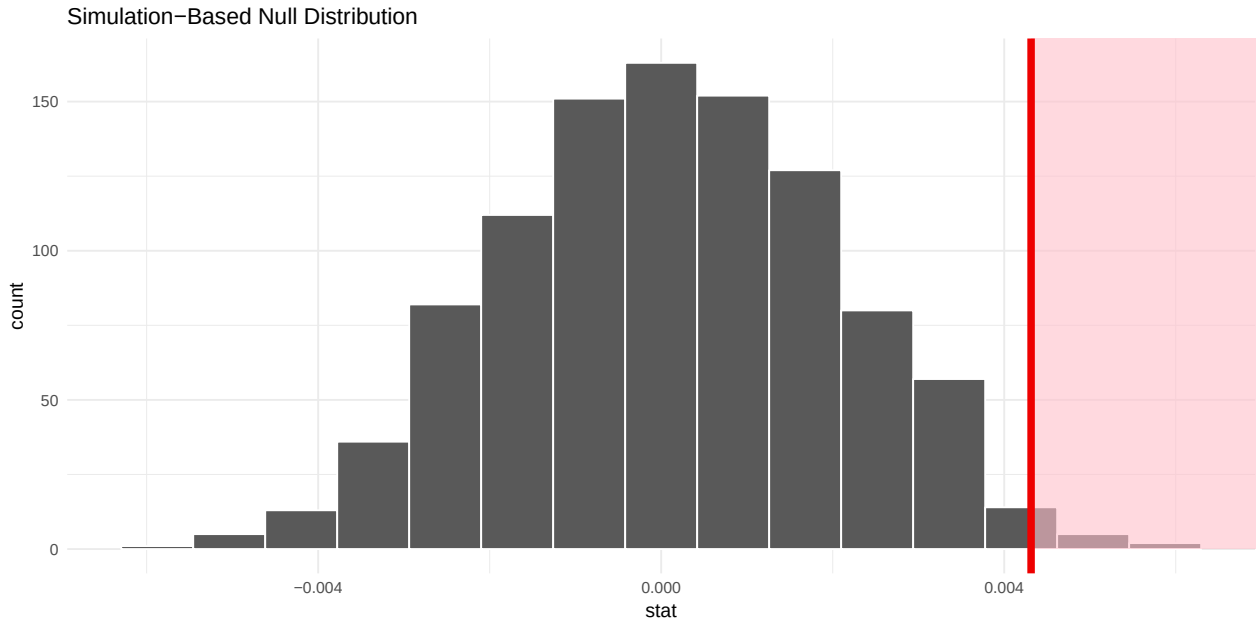
```
set.seed(100)
email_sfs_slopes <- trafico_de_email_sfs_brate_df_sin_out %>%
  specify(blocked_rate ~ total) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000) %>%
  calculate(stat = "slope")
```

Caluclating observed statistic - slope

```
email_sfs_obs_slope <- lm(blocked_rate ~ total,data = trafico_de_email_sfs_brate_df_sin_
  tidy() %>%
  filter(term=="total") %>%
  pull(estimate)
```

Visualizing null distribution

```
email_sfs_slopes %>%
  visualise() + shade_p_value(obs_stat =email_sfs_obs_slope, direction = "right" )
```



Caluclating p-value and taking a decision

```
email_sfs_slopes %>%
  get_p_value(obs_stat = email_sfs_slopes,direction = "right")
```

p_value
0

Since 0.004 falls far to the right of this plot beyond where any of the histogram bins have data, we can say that we have a p-value of 0. We,thus,have evidence to reject the null hypothesis in support of there being a positive association between total no.of classified messages and blocking rate

Glance at model diagnostics

```
lm(blocked_rate ~ total,data = trafico_de_email_sfs_brate_df_sin_out) %>%
  tidy()
```

term	estimate	std.error	statistic	p.value
(Intercept)	9.4365013	0.3064242	30.795544	0.0000000
total	0.0043133	0.0019652	2.194895	0.0294093

Regression line equation as follows,

$$E(y) = \alpha + \beta x$$

$$blockedrate = 10.17 + 0.004 * total$$

$slope(\beta)$ is a numerical quantity that summarizes the relationship between the outcome and explanatory variables. Note that the sign is positive, suggesting a positive relationship between blocking rate and total, meaning as day total goes up, so also do day blocking rate go up.

For every increase of 1 unit in total, there is an associated increase of, on average, 0.004(0.4%) units of blocking rate.

Research Question 1:

Null Hypothesis(H_o): The true proportion of spam traffic is about 80%, i.e π=p_o and p_o=80

Alternate Hypothesis(Ha): The true proportion of spam traffic is greater than 80% $\pi >$ 80

Observed statistic

```
p_hat_status_one_prop<- hyp_status_prop %>%
  specify(response = status,success = "SFS") %>%
  calculate(stat = "prop")
```

Observed statistic of spam traffic is caluclated as *0.801*

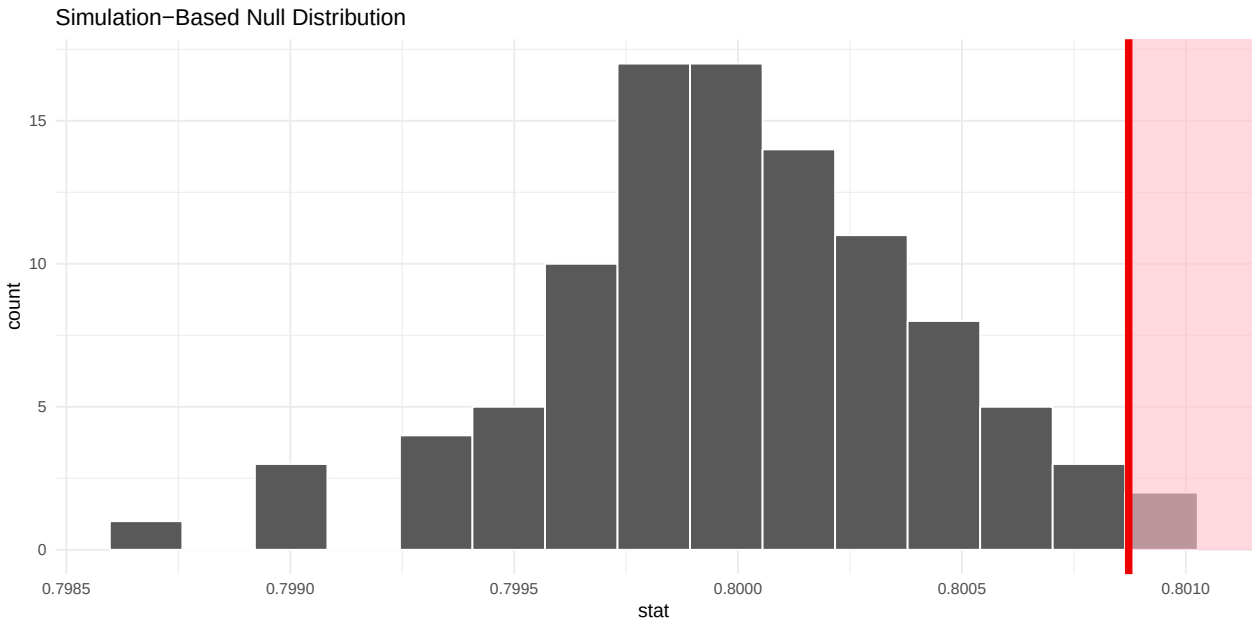
Specifying hypothesis

```
null_dist_status_one_prop <- hyp_status_prop %>%
  specify(response = status,success = "SFS") %>%
  hypothesize(null = "point",p=.80) %>%
  generate(reps = 100) %>%
  calculate(stat = "prop")
```

> In order to look to see if 0.801 is statistically different from 0.80, we need to account for the sample size. We also need to determine a process that replicates how the original sample of size 900K was selected, We can use the idea of an unfair coin to simulate this process. We will simulate taking out messaging traffic with probability of success 0.81 matching the null hypothesis for 100 times. Then we will keep track of how many spam traffic messages have come up in those 100 trials. Our simulated statistic matches with how we calculated the observed statistic, and create the null distribution looking at the simulated proportions of successes as showed below.

Visualizing null distribution

```
null_dist_status_one_prop %>%
  visualize() + shade_p_value(obs_stat = p_hat_status_one_prop,direction = "greater")
```



We can next use this distribution to observe our p-value. as this is a right-tailed test we will be looking for values that are greater than or equal to 0.801 for our p value

Understanding p-value and decision making with it

```
null_dist_status_one_prop %>%
  get_p_value(obs_stat = p_hat_status_one_prop,direction = "greater")
```

p_value
0.02

So our p-value is 0.05 and we reject the null hypothesis at the 5% level. We can also see this from the histogram above that we are far into the tail of the null distribution. Here it's concluded that the true proportion of spam traffic is higher than 80%

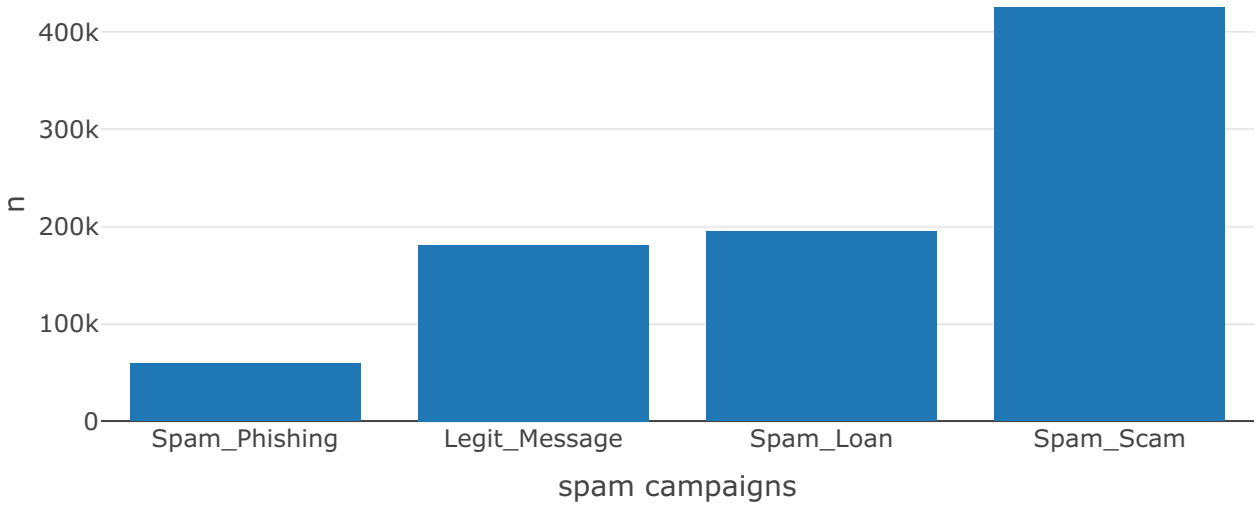
Research Question 2:

Null Hypothesis(H_o): The true proportions of spam campaigns are Spam_Scam (0.5), Spam_Loan (0.25),Legit_Message(0.2),Spam_Phishing(0.05)

Alternate Hypothesis (H_a): The distributions of messaging spam campaign proportions

are different from the specified ones

Checking how are the total of different spam campaigns



Observed statistic

```
p_hat_spam_campaign_prop <-hyp_spam_campaign_prop %>%
  specify(spam_campaign ~ NULL) %>% # alt: response = cyl
  hypothesize(null = "point", p = c("Spam_Scam" = .5, "Spam_Loan" = .25, "Legit_Message'
  calculate(stat = "Chisq")
```

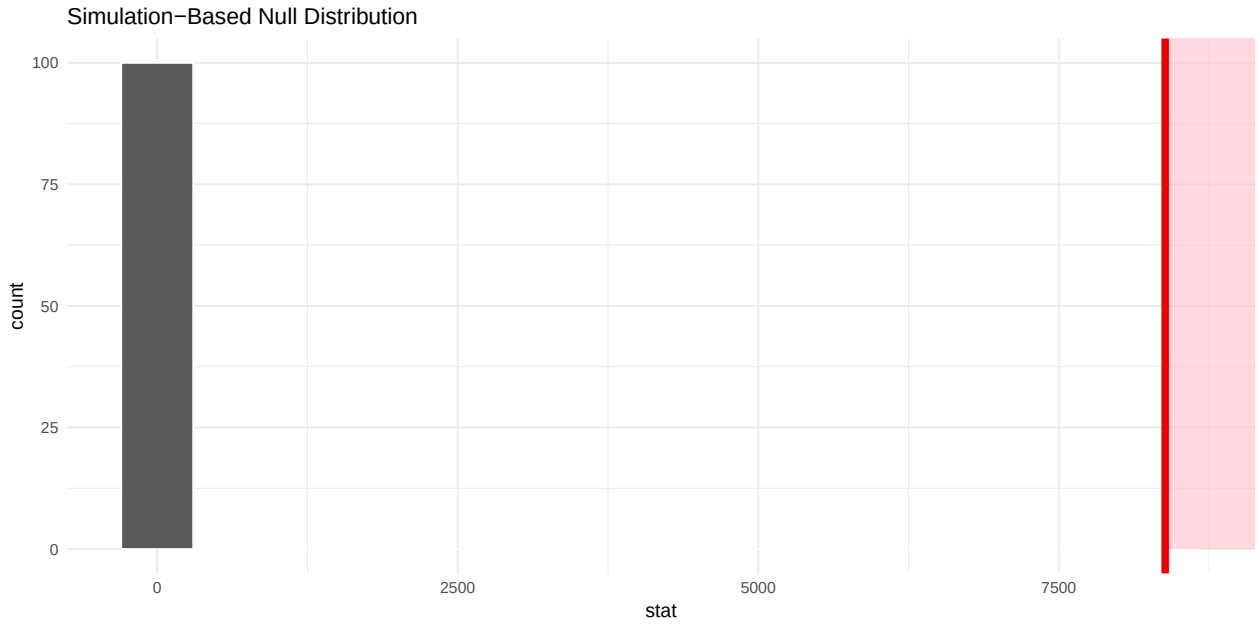
Observed statistic $\tilde{\chi}^2_{obs}$ is 8385 with degrees of freedom 3

Specifying a null hypothesis

```
null_dist_spam_campaign_prop <- hyp_spam_campaign_prop %>%
  specify(spam_campaign ~ NULL) %>%
  hypothesize(null = "point", p = c("Spam_Scam" = .5, "Spam_Loan" = .25, "Legit_Message'
  generate(reps = 100, type = "simulate") %>%
  calculate(stat = "Chisq")
```

Visualizing null distribution

```
null_dist_spam_campaign_prop %>%
  visualize() + shade_p_value(obs_stat = p_hat_spam_campaign_prop,direction = "right")
```



Understanding p-value and decision making with it

```
null_dist_spam_campaign_prop %>%
  get_p_value(obs_stat = p_hat_spam_campaign_prop,direction = "right")
```

p_value
0

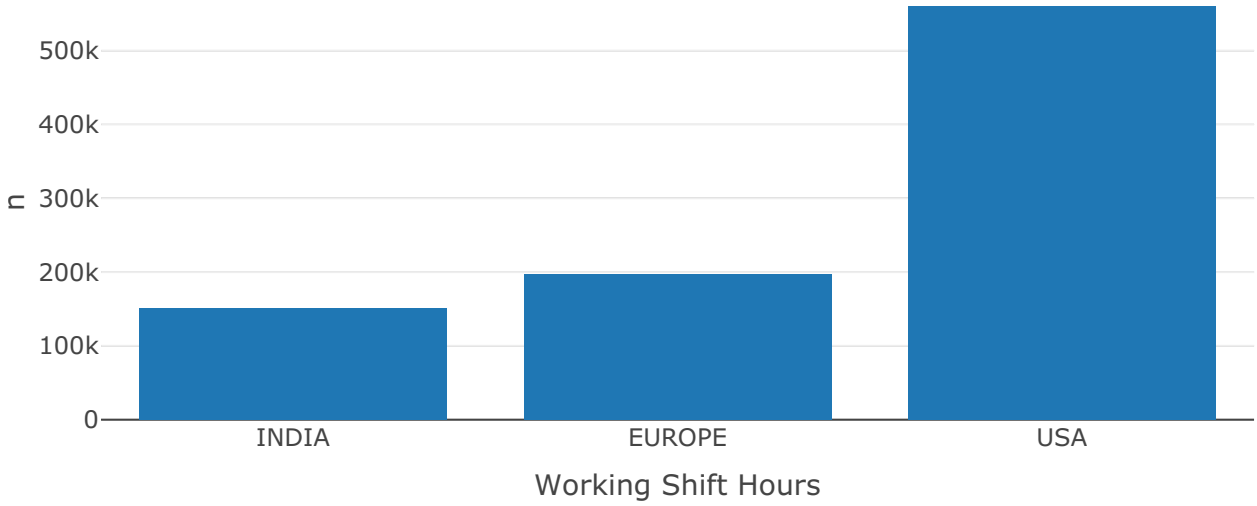
Here our p-value is 0 and we reject the null hypothesis at the 5% level.We can also see this from the histogram above that we are far into the tail of the null distribution. Here it's stated that the distributions of messaging spam campaign proportions are different from the specified ones such as Spam_Scam (0.5), Spam_Loan (0.25),Legit_Message(0.2), and Spam_Phishing(0.05).

Research Question 3:

Null Hypothesis(H_o): The true proportions of spam marking for each shift are USA p_{usa}=(.62), EUROPE p_{europe}=(0.22), and INDIA p_{india}=(.16)

Alternate Hypothesis (H_a): The distributions of messaging spam campaing proportions are different from the specified ones

Getting to know how spam marking is carried out in each shift hours



Observed statistic

```
p_hat_user_prop <-hyp_user_prop %>%
  specify(username ~ NULL) %>% # alt: response = cyl
  hypothesize(null = "point", p = c("USA" =0.62, "EUROPE" = 0.22, "INDIA" = 0.16)) %>%
  calculate(stat = "Chisq")
```

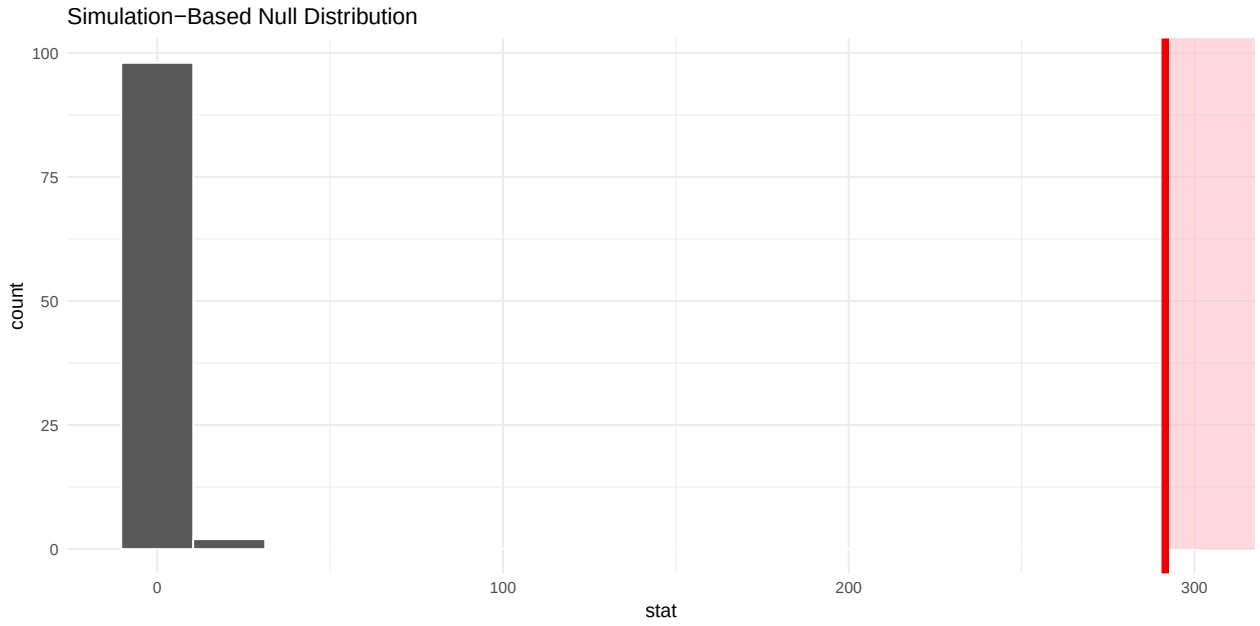
Observed statistic $\tilde{\chi}^2_{obs}$ is 292 with degrees of freedom 2.

Specifying a null hypothesis

```
null_dist_user_prop <- hyp_user_prop %>%
  specify(username ~ NULL) %>% # alt: response = cyl
  hypothesize(null = "point", p = c("USA" =0.62, "EUROPE" = 0.22, "INDIA" = 0.16)) %>%
  generate(reps = 100, type = "simulate") %>%
  calculate(stat = "Chisq")
```

Visualizing null distribution

```
null_dist_user_prop %>%
  visualize() + shade_p_value(obs_stat = p_hat_user_prop,direction = "greater")
```



Understanding p-value and decision making with it

```
null_dist_user_prop %>%
  get_p_value(obs_stat = p_hat_user_prop,direction = "right")
```

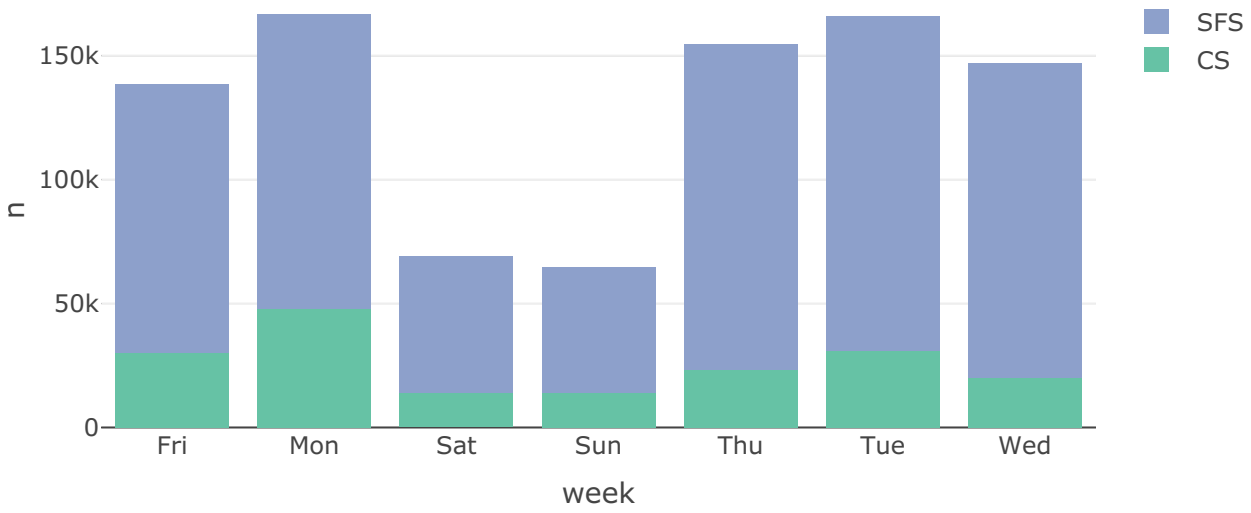
p_value
0

Here our p-value is 0 and we reject the null hypothesis at the 5% level.We can also see this from the histogram above that we are far into the tail of the null distribution. Here it's stated that the distributions of messaging spam campaign proportions are different from the specified ones such as USA (0.62), EUROPE(0.22), and iNDIA(0.16).

Research Question 4:

Null Hypothesis(H_o): Spam traffic classification is same across all the the days. i.e $P_{mon} = P_{tue} = P_{wed} = P_{thu} = P_{fri} = P_{sat} = P_{sun}$ where p represents the long-run probability a message will be a spam.

Alternate Hypothesis (H_a): Spam traffic classification is different atleast on one day



Observed statistic

```
p_hat_week_status <- hyp_day_status_prop %>%
  chisq_stat(status ~ week)
```

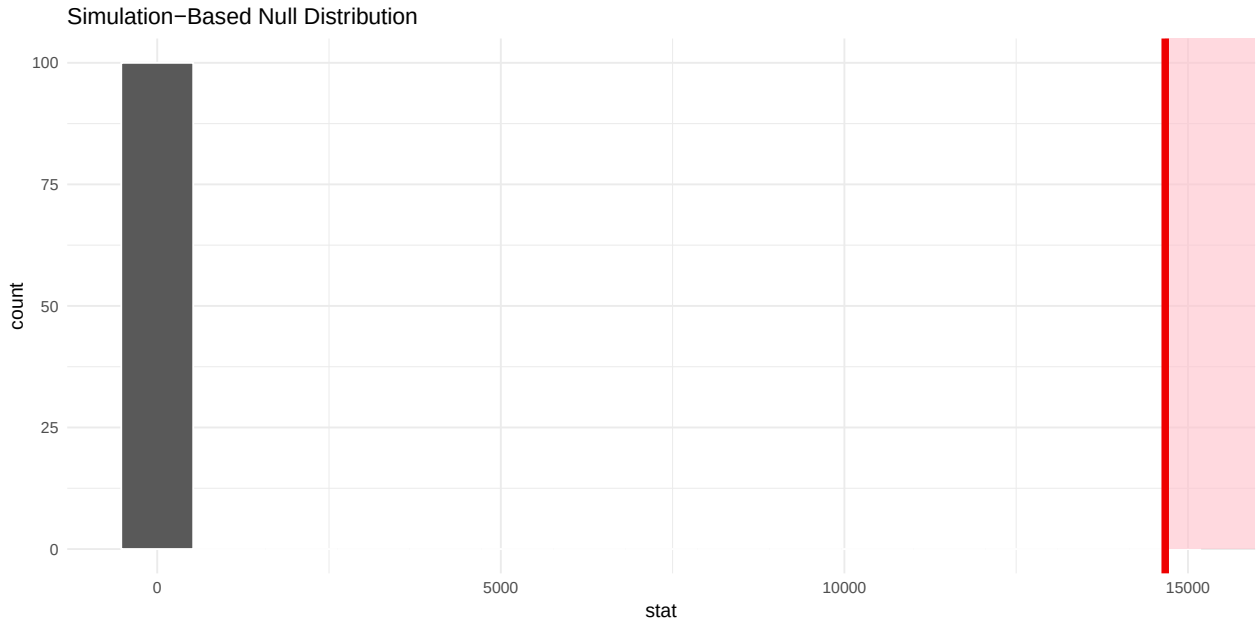
Observed statistic $\tilde{\chi}^2_{obs}$ is 14669 with degrees of freedom 6

Specifying a null hypothesis

```
null_dist_week_status <- hyp_day_status_prop %>%
  specify(status ~ week) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 100, type = "permute") %>%
  calculate(stat = "Chisq")
```

Visualizing null distribution

```
null_dist_week_status %>%
  visualize() + shade_p_value(obs_stat = p_hat_week_status,direction = "greater")
```



Understanding p-value and decision making with it

```
null_dist_week_status %>%
  get_p_value(obs_stat = p_hat_week_status,direction = "greater")
```

p_value
0

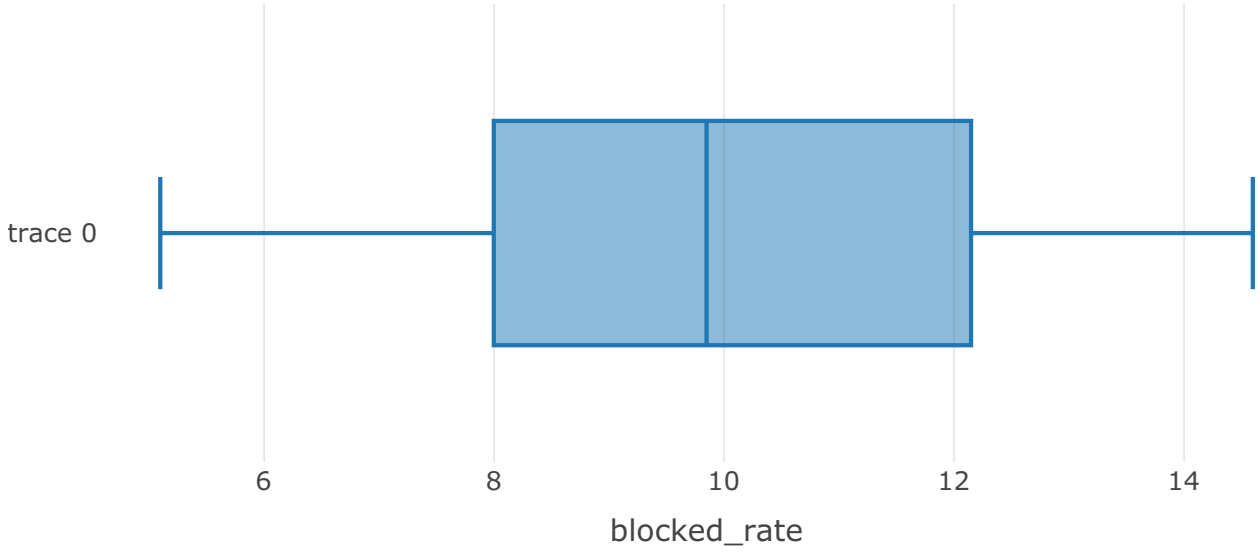
Here our p-value is 0 and we reject the null hypothesis at the 5% level.We can also see this from the histogram above that we are far into the tail of the null distribution. Here it's understood that spam traffic classification is different at least on one day.

Research Question 5

Null Hypothesis(H_o): The median blocking rate of email traffic is about 11 i .e ($\bar{X}_{blockingrate}$)=11

Alternate Hypothesis (H_a): The median blocking rate of email traffic is not euqal to 11 i.e $\bar{X}_{blockingrate} \neq 11$

```
trafico_de_email_sfs_brate_df_sin_out %>%
  plot_ly(x=~blocked_rate) %>%
  add_boxplot()
```



Observed statistic

```
p_hat_brate <- trafico_de_email_sfs_brate_df_sin_out %>%
  specify(response = blocked_rate) %>%
  calculate(stat = "median")
```

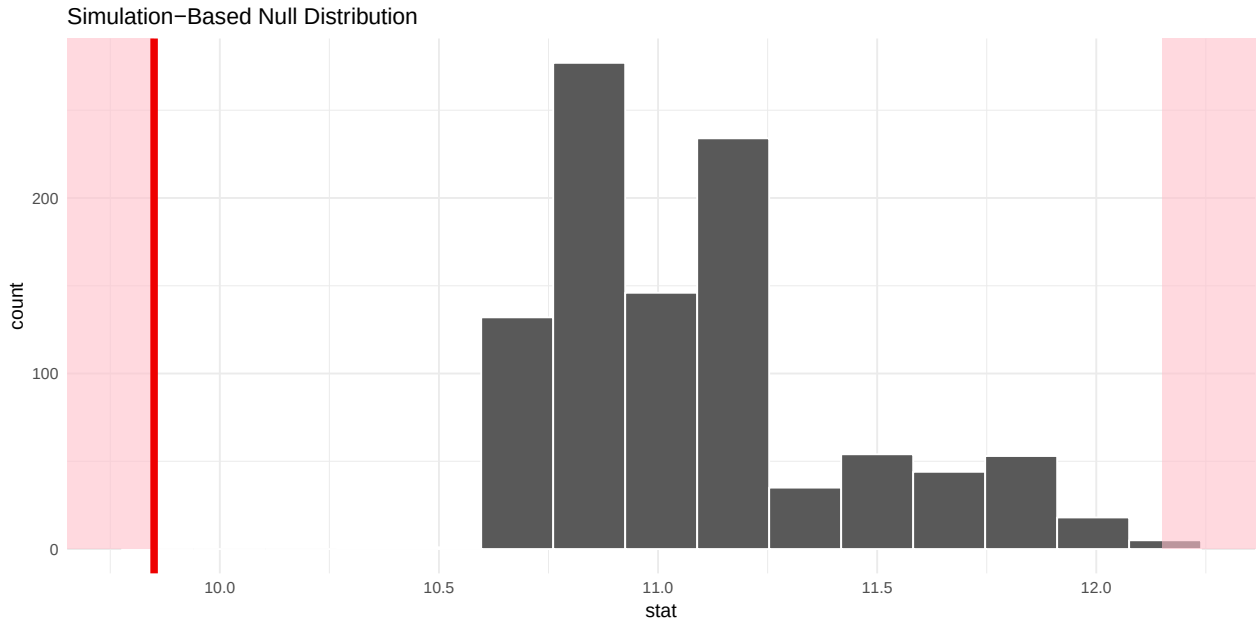
Observed statistic of blocking rate is caluclated as $\bar{X}_{obsblockingrate}$ 11.2

Specifying a null hypothesis

```
null_dist_blocked_rate <- trafico_de_email_sfs_brate_df_sin_out %>%
  specify(response = blocked_rate) %>%
  hypothesize(null = "point",med=11) %>%
  generate(reps = 1000)%>%
  calculate(stat = "median")
```

Visualizing a null distribution

```
null_dist_blocked_rate %>%
  visualize() + shade_p_value(obs_stat = p_hat_brate, direction = "two_sided")
```



We can next use this distribution to observe our p-value. as this is a two-tailed test we will be looking for values that are 11-11.2 = -0.02 away from 11 in BOTH directions for our p-value

Understanding p-value and decision making with it

```
null_dist_blocked_rate %>%
  get_p_value(obs_stat = p_hat_brate, direction = "two_sided")
```

p_value
0

our p-value is 0 and we reject the null hypothesis at the 5% level. we can also see this from the histogram above that we are very far into the tail of the null distribution.

Constructing confidence intervals

```
null_dist_blocked_rate %>%
  get_ci(point_estimate = p_hat_brate)
```

2.5%	97.5%
10.7	11.85037

Research Question 6:

Case 1:

Null Hypothesis(H_o): The median time to live of fingerprints that are in cluster group low 1 and low 2 is about 0.5 i.e ($\bar{X}_{ftldays}$)=0.5

Alternate Hypothesis (H_a): The median time to live of fingerprints that are in cluster group low 1 and low 2 is not 0.5 $\bar{X}_{ftldays} \neq 0.5$

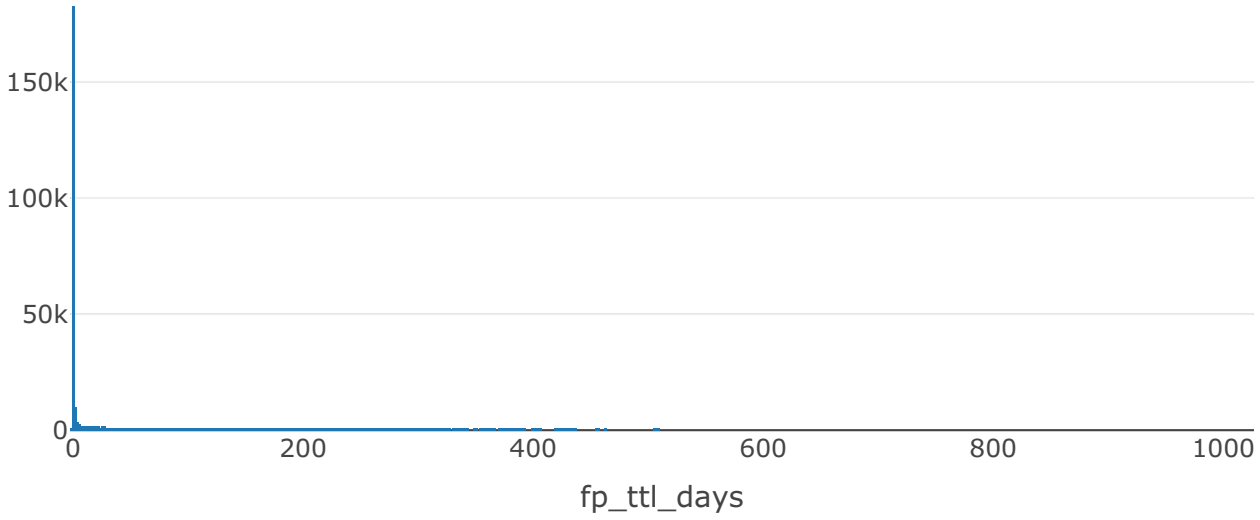
Case 2:

Null Hypothesis(H_o): The median time to live of fingerprints that are not into cluster group low 1 and low 2 is about 15 i.e ($\bar{X}_{ftldays}$)=15

Alternate Hypothesis (H_a): The median time to live of fingerprints that are not into cluster group low 1 and low 2 is not 15 i.e $\bar{X}_{ftldays} \neq 15$

View how fingerprint filters time to live are distributed which are into cluster group low 1 and low 2 ?

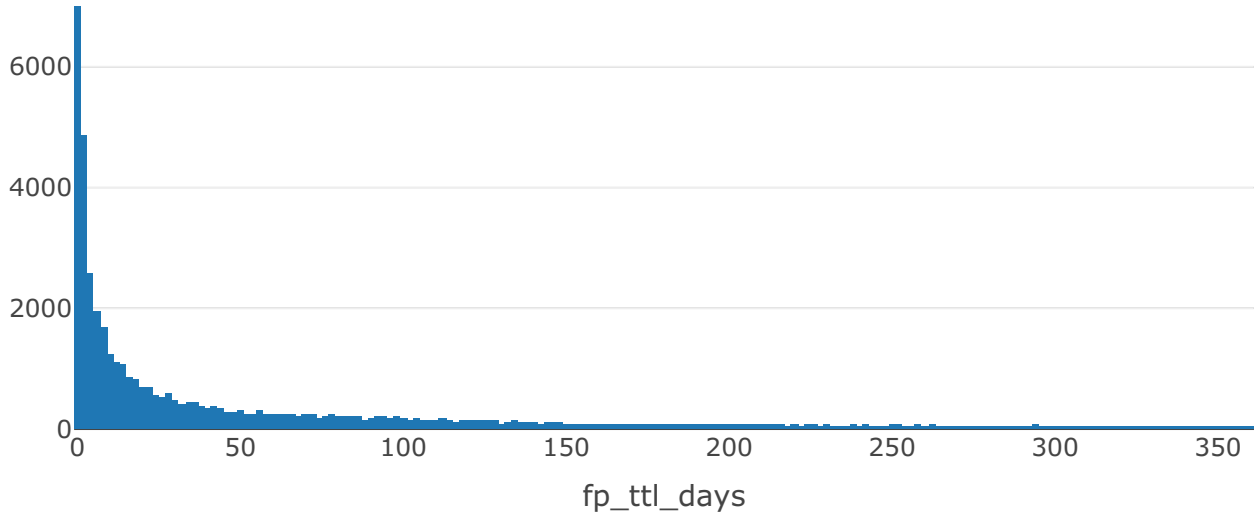
```
mensajes_de_sc_process_days %>%
    filter(fp_process_days>0) %>%
  filter(cluster_group %in% c("CL_LOW_1","CL_LOW_2")) %>%
  mutate(fp_ttl_days=as.numeric(fp_ttl_days)) %>%
  plot_ly(x=~fp_ttl_days) %>%
  add_histogram()
```



View how fingerprint filters time to live are distributed which are not into cluster group low 1 and low 2 ?

```
#hist2
mensajes_de_sc_process_days %>%
```

```
    filter(fp_process_days>0) %>%
  filter(!cluster_group %in% c("CL_LOW_1","CL_LOW_2"), between(fp_ttl_days,0,360)) %>%
  mutate(fp_ttl_days=as.numeric(fp_ttl_days)) %>%
  plot_ly(x=~fp_ttl_days) %>%
  add_histogram()
```



Caluclating observed statistics

```
d_hat_low <- mensajes_de_sc_process_days %>%
    filter(fp_process_days>0) %>%
  filter(cluster_group %in% c("CL_LOW_1","CL_LOW_2")) %>%
  mutate(fp_ttl_days=as.numeric(fp_ttl_days)) %>%
  specify(response = fp_ttl_days) %>%
  calculate(stat = "median")

d_hat_not_normal <- mensajes_de_sc_process_days %>%
    filter(fp_process_days>0) %>%
  filter(!cluster_group %in% c("CL_LOW_1","CL_LOW_2"),between(fp_ttl_days,0,360)) %>%
  mutate(fp_ttl_days=as.numeric(fp_ttl_days)) %>%
  infer::specify(response = fp_ttl_days) %>%
  calculate(stat = "median")
```

Case 1: Observed statistic of filter time to live is caluclated as $\bar{X}_{obsblockingrate}$ 0

Case 2: Observed statistic of filter time to live is caluclated as $\bar{X}_{obsblockingrate}$ 14.3

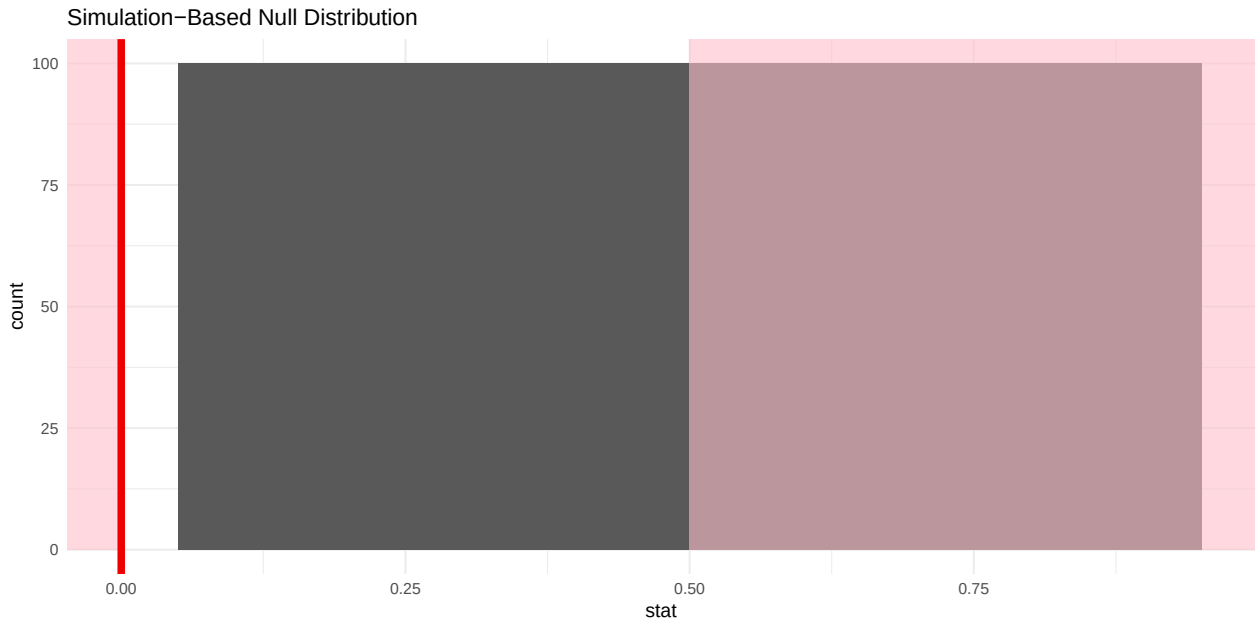
Framing hypothesis

```
null_dist_fp_ttl_normal <- mensajes_de_sc_process_days %>%
    filter(fp_process_days>0) %>%
  filter(cluster_group %in% c("CL_LOW_1","CL_LOW_2")) %>%
```

```
  mutate(fp_ttl_days=as.numeric(fp_ttl_days)) %>%
  specify(response = fp_ttl_days) %>%
  hypothesize(null = "point",med=0.5) %>%
  generate(reps = 100) %>%
  calculate(stat = "median")
```

Visualizing null distribution with P-Value

```
null_dist_fp_ttl_normal %>%
  visualize() + shade_p_value(obs_stat = d_hat_low, direction = "two_sided")
```



We can make use of this distribution to observe our p-value. as this is a two-tailed test we will be looking for values that are 0.5-0= 0.5 away from 0.5 in BOTH directions for our p-value

Looking at P-Value

```
null_dist_fp_ttl_normal %>%
  get_p_value(obs_stat = d_hat_not_normal, direction = "two_sided")
```

p_value
0

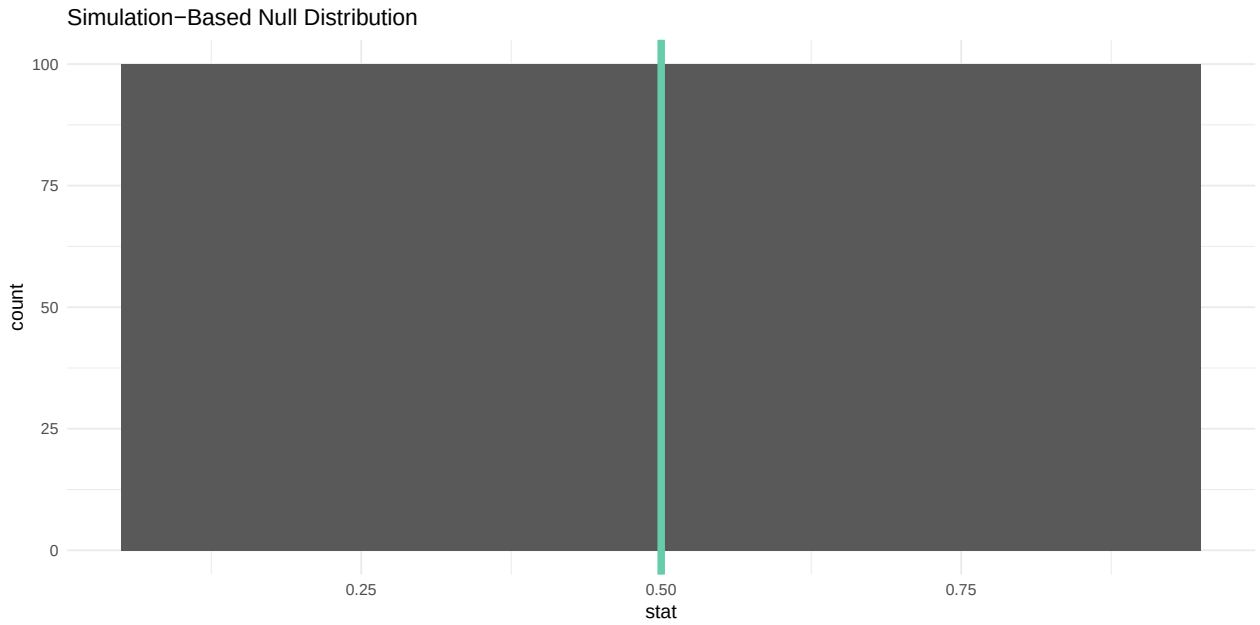
our p-value is 0 and we reject the null hypothesis at the 5% level. we can also see this from the histogram above that we are very far into the tail of the null distribution.

Looking at confidence intervals

```
percentile_ci_normal <- get_confidence_interval(null_dist_fp_ttl_normal)
```

Visualizing confidence intervals

```
visualize(null_dist_fp_ttl_normal) + shade_confidence_interval(endpoints = percentile_c
```



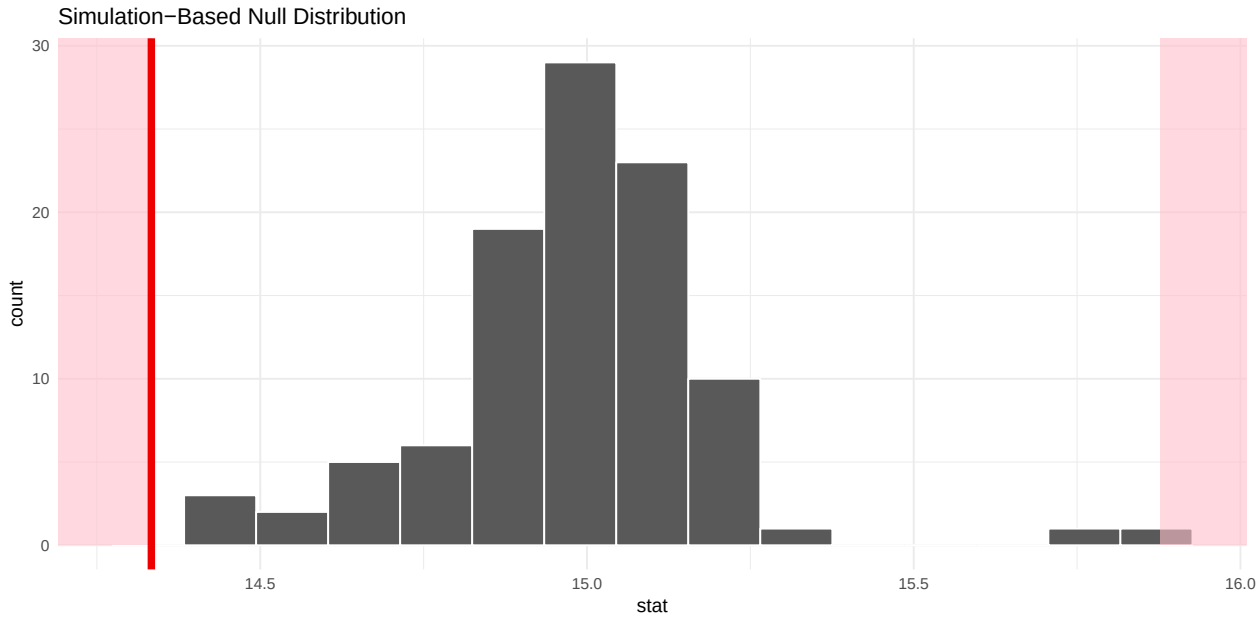
Here it's suggested that there is a 95% chance that the range of median values [0.14, 0.82] encompasses the true population median.

Framing hypothesis for cluester groups not in low_1 and low_2

```
null_dist_fp_ttl <- mensajes_de_sc_process_days %>%
    filter(fp_process_days>0) %>%
  filter(!cluster_group %in% c("CL_LOW_1","CL_LOW_2")) %>%
  mutate(fp_ttl_days=as.numeric(fp_ttl_days)) %>%
  specify(response = fp_ttl_days) %>%
  hypothesize(null = "point",med=15) %>%
  generate(reps = 100) %>%
  calculate(stat = "median")
```

Visualizing null distribution with P-Value

```
null_dist_fp_ttl %>%
  visualize() + shade_p_value(obs_stat = d_hat_not_normal, direction = "two_sided")
```



We can make use of this distribution to observe our p-value. as this is a two-tailed test we will be looking for values that are 15-14.3= 0.7 away from 15 in BOTH directions for our p-value

Looking at P-Value

```
null_dist_fp_ttl %>%
  get_p_value(obs_stat = d_hat_not_normal, direction = "two_sided")
```

p_value
0

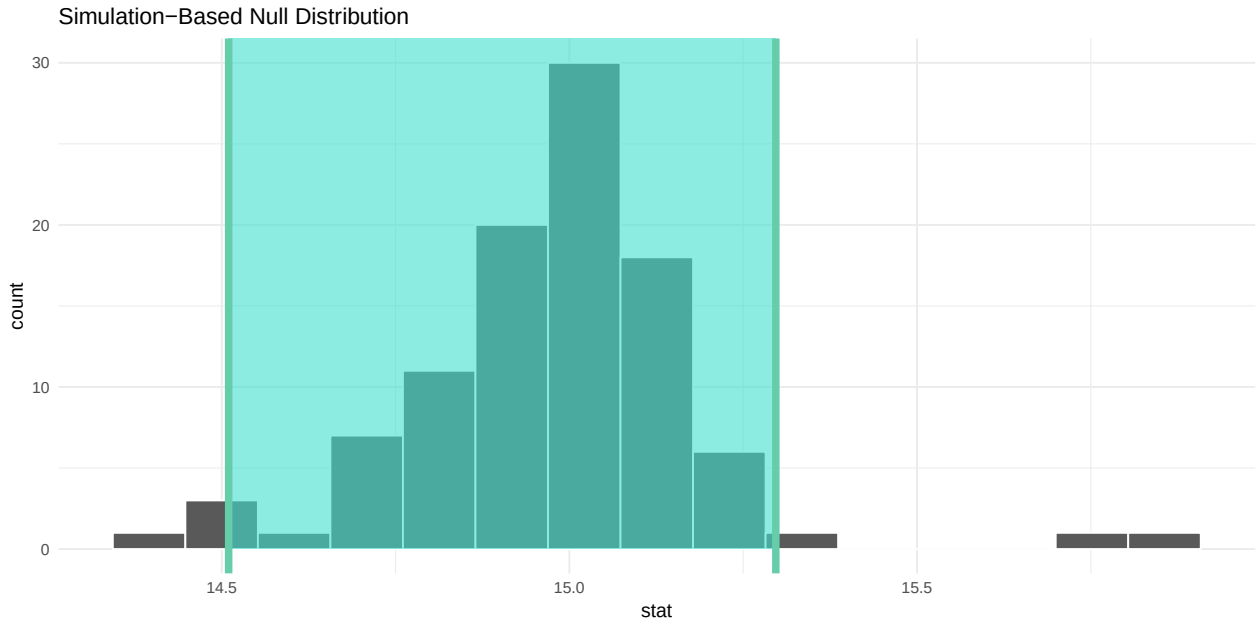
Since p value is 0 we can reject the null hypothesis at the 5% level. we can also see this from the histogram above that we are very far into the tail of the null distribution

Looking at confidence intervals

```
percentile_ci_low <- get_confidence_interval(null_dist_fp_ttl)
```

Visualizing confidence intervals

```
visualize(null_dist_fp_ttl) + shade_confidence_interval(endpoints = percentile_ci_low)
```



Here it is seen that there is a 95% chance that the range of median values [14.6, 15.3] encompasses the true population median.

Research Question 7:

Null Hypothesis(H_o): There is no difference between the median time of first message received and last message received that are not into cluster low-1 and low-2 groups i.e

$(\bar{X}_{fmsgrec})$ - $(\bar{X}_{lmsgrec}) = 0$

Alternate Hypothesis (H_a): There is a difference between the median time of first message received and last message that are not into cluster low-1 and low-2 groups i.e $(\bar{X}_{fmsgrec})$ - $(\bar{X}_{lmsgrec}) \neq 0$

Observerd statistic

```
d_hat_non_low <- mensajes_de_sc_fl_hours_non_low %>%
  specify(dias ~ type_of_time_hour) %>%
  calculate(stat = "diff in medians",order = c("fmsg_rec_h","lmsg_rec_h"))
```

Observed statistic $\bar{X}_{fmsgrec})$ - $(\bar{X}_{lmsgrec}$ is 1

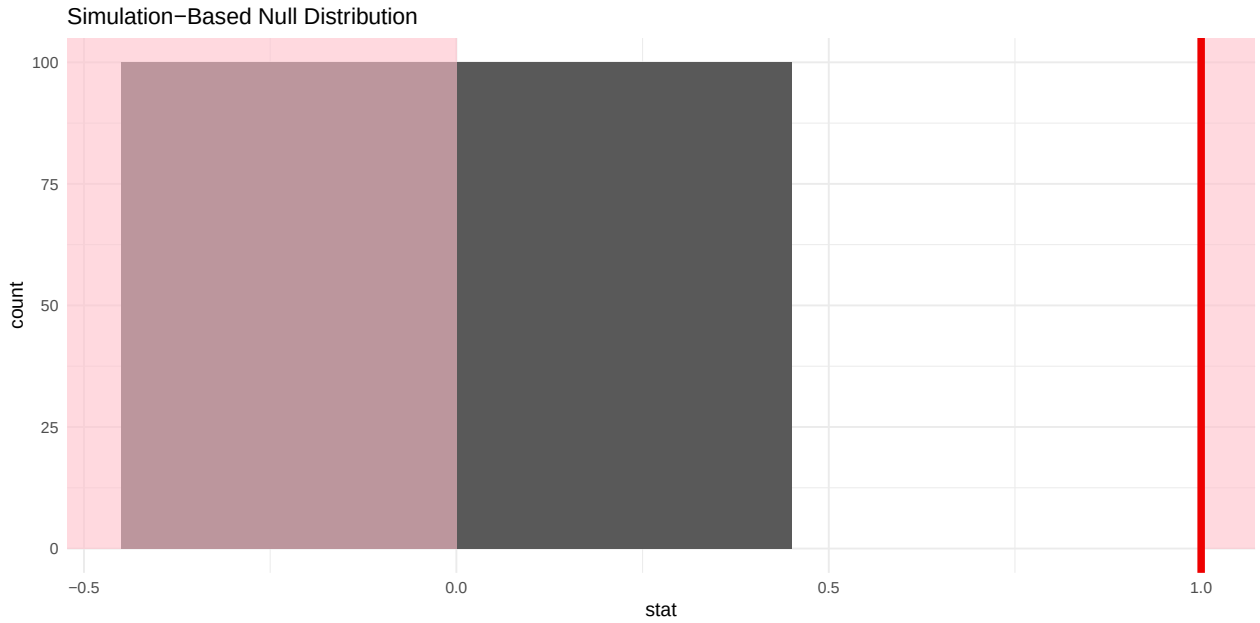
Framing hyothesis

```
null_dist_fl_hours_non_low <- mensajes_de_sc_fl_hours_non_low %>%
  specify(dias ~ type_of_time_hour) %>%
  hypothesize(null="independence") %>%
```

```
  generate(reps = 100) %>%
  calculate(stat = "diff in medians",order = c("fmsg_rec_h","lmsg_rec_h"))
```

Visualizing null distribution

```
null_dist_fl_hours_non_low%>%
  visualize() + shade_p_value(obs_stat = d_hat_non_low, direction = "two_sided")
```



We can next use this distribution to observe our p-value. since this is a two-tailed test we will be looking for values that are greater than or equal to 1 or less than or equal to -1 for our p-value

Understanding p-value and decision making with it

```
null_dist_fl_hours_non_low %>%
  get_p_value(obs_stat = d_hat_non_low, direction = "two_sided")
```

p_value
0

Since p value is 0 we can reject the null hypothesis at the 5% level. we can also see this from the histogram above that we are very far into the tail of the null distribution

Research Question 8:

Null Hypothesis(H_o): There is no difference between the median time of first message received and last message received that are into cluster low-1 and low-2 groups i.e ($\bar{X}_{fmsgrec}$) - ($\bar{X}_{lmsgrec}$) = 0

Alternate Hypothesis (H_a): There is a difference between the median time of first message received and last message that are into cluster low-1 and low-2 groups i.e ($\bar{X}_{fmsgrec}$) - ($\bar{X}_{lmsgrec}$) $\neq$ 0

Observerd statistic

```
d_hat_low_fl <- mensajes_de_sc_fl_hours %>%
  specify(dias ~ type_of_time_hour) %>%
  calculate(stat = "diff in medians",order = c("fmsg_rec_h","lmsg_rec_h"))
```

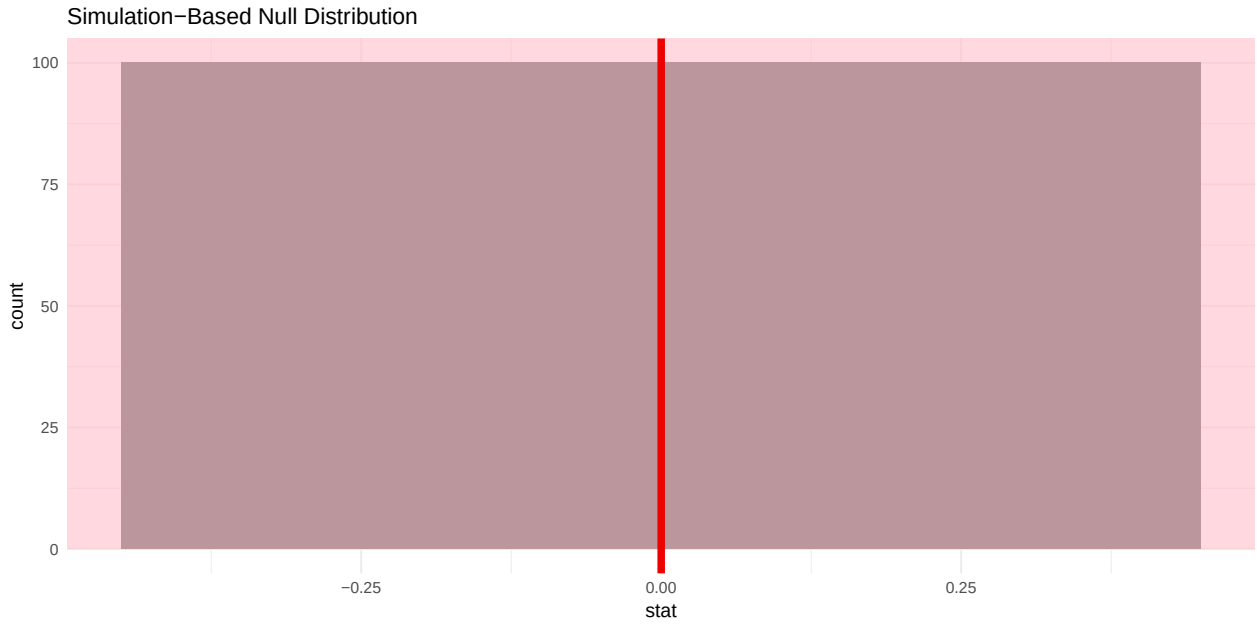
Observed statistic is $\bar{X}_{fmsgrec}$) - ($\bar{X}_{lmsgrec}$ 0

Framing hyothesis

```
null_dist_fl_hours_low <- mensajes_de_sc_fl_hours %>%
  specify(dias ~ type_of_time_hour) %>%
  hypothesize(null="independence") %>%
  generate(reps = 100) %>%
  calculate(stat = "diff in medians",order = c("fmsg_rec_h","lmsg_rec_h"))
```

Visualizing null distribution

```
null_dist_fl_hours_low%>%
  visualize() + shade_p_value(obs_stat = d_hat_low_fl, direction = "two_sided")
```



We can next use this distribution to observe our p-value. since this is a two-tailed test we will be looking for values that are greater than or equal to 0 or less than or equal to -0 for our p-value

Understanding p-value and decision making with it

```
null_dist_fl_hours_low %>%
  get_p_value(obs_stat = d_hat_low_fl, direction = "two_sided")
```

p_value
1

Our p value caluclated as 1, here it's greater than the significance level 0.05 hence we fail to reject the null hypothesis and we can also see this from the histogram above that we are not far into the tails of the null distribution

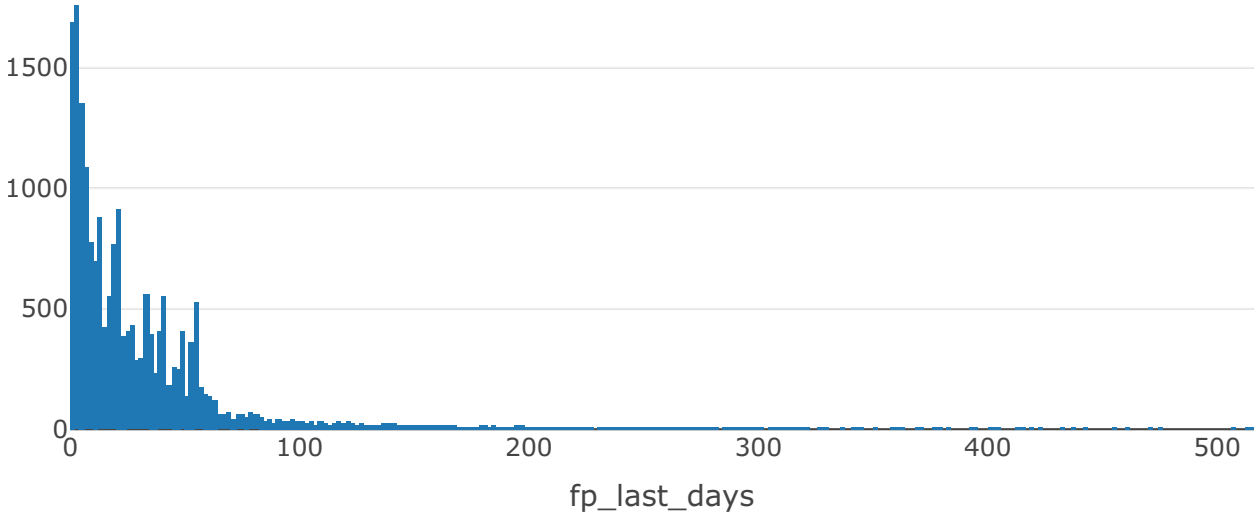
Research Question 9:

Null Hypothesis(H_o): The median time of fingerprints that are last for more than a day is about 18days i.e ($\bar{X}_{fplastdays}$)=18

Alternate Hypothesis (H_a): The median time of fingerprints that are last for more than a day is not 18days i.e $\bar{X}_{fplastdays} \neq 18$

```
mensajes_de_sc_change_fp_other_days <-mensajes_de_sc_process_days %>%
  mutate(fp_last_days=as.numeric(fp_last_days)) %>%
  filter(!st_change_dias=="same-day")
```

```
mensajes_de_sc_change_fp_other_days%>%
  plot_ly(x=~fp_last_days) %>%
  add_histogram()
```



Observerd statistic

```
d_hat_st_change_other_days<- mensajes_de_sc_change_fp_other_days %>%
  specify(response = fp_last_days) %>%
  calculate(stat = "median")
```

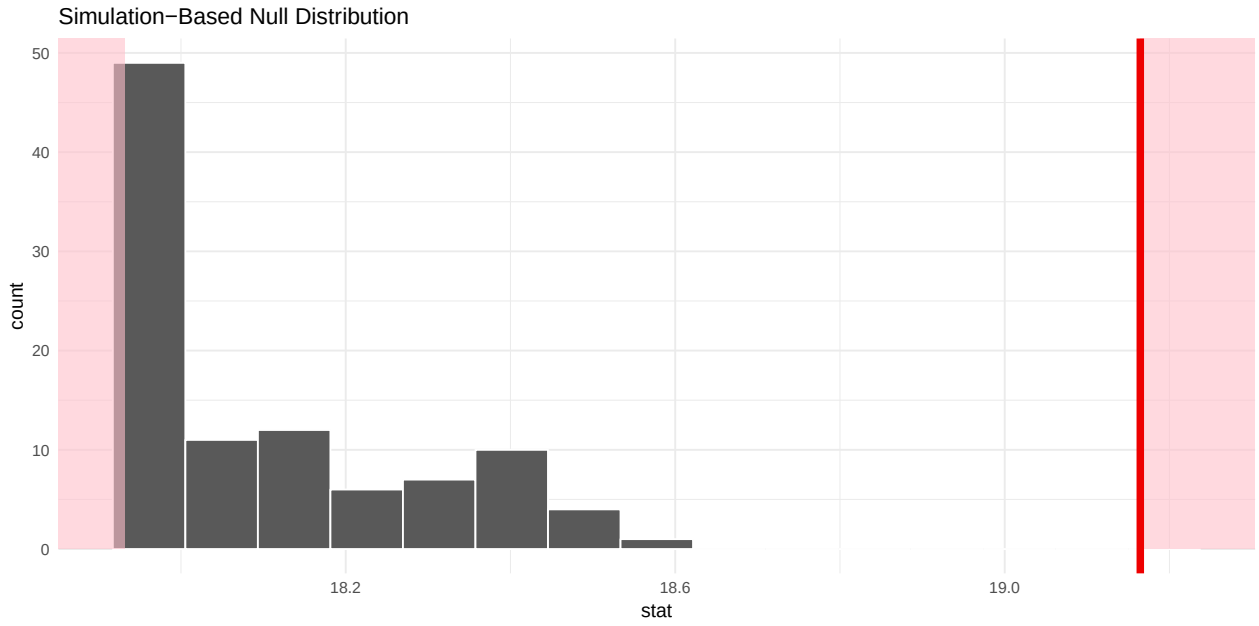
Observed statistic $\bar{X}_{obs}$ is 19.2

Specifying null hypothesis

```
null_dist_st_change_other_days <- mensajes_de_sc_change_fp_other_days %>%
  specify(response = fp_last_days) %>%
  hypothesize(null="point",med=18) %>%
  generate(reps = 100) %>%
  calculate(stat = "median")
```

Visualiing null distribution

```
null_dist_st_change_other_days %>%
  visualize() + shade_p_value(obs_stat = d_hat_st_change_other_days,direction = "two_si
```



We can make use of this distribution to observe our p-value. as this is a two-tailed test we will be looking for values that are 18-19.2= -1.2 away from 18 in BOTH directions for our p-value

Understanding p-value and decision making with it

```
null_dist_st_change_other_days %>%
  get_p_value(obs_stat = d_hat_st_change_other_days,direction = "two_sided")
```

p_value
0

Our p value caluclated as 0, here it's lesser than the significance level 0.05 hence we can reject the null hypothesis.

Research Question 10:

Null Hypothesis(H_o): The true proportion of messaging traffic that have been received after having got processed is about 50% i.e i.e π=p_o and p_o=.5

Alternate Hypothesis (H_a): The true proportion of messaging traffic that have been received after having got processed is higher than 50% i.e $\pi > 0.5$

Observerd statistic

```
d_hat_class_prop <- mensajes_de_sc_process_days %>%
  mutate(class_time=as.factor(class_time)) %>%
```

```
  filter(class_time !="Other") %>%
  specify(response = class_time,success = "antes") %>%
  calculate(stat = "prop")
```

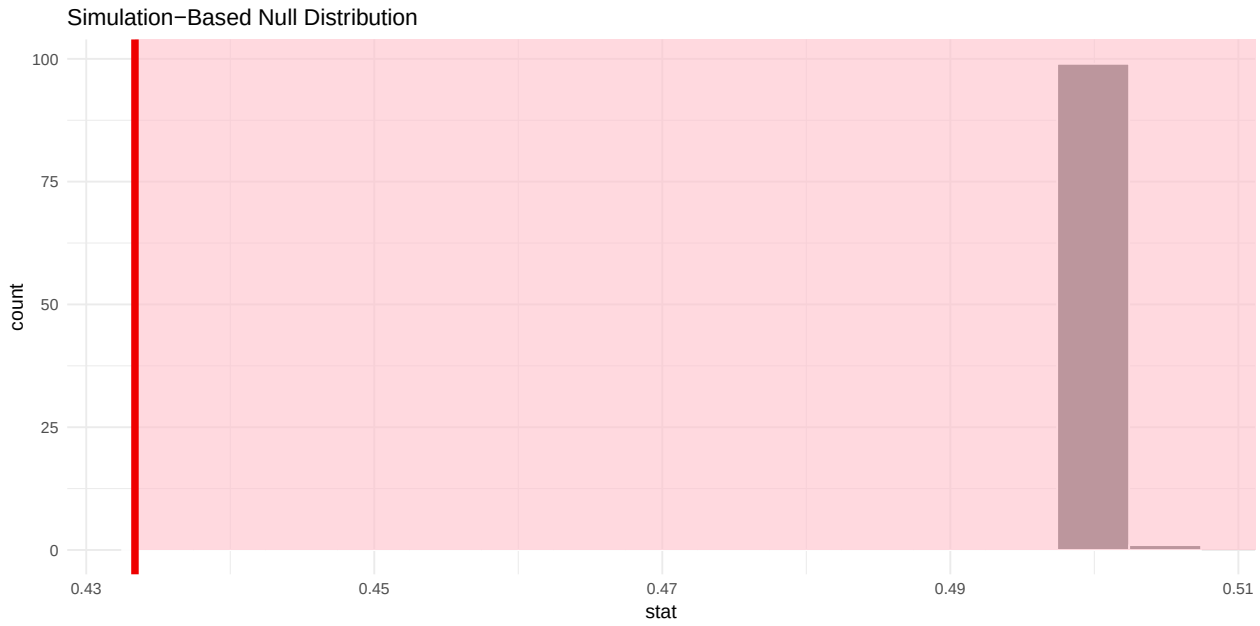
Observed statistic π_{obs} is 0.43

Framing hyothesis

```
null_dist_class_prop <- mensajes_de_sc_process_days %>%
  mutate(class_time=as.factor(class_time)) %>%
  filter(class_time !="Other") %>%
  specify(response = class_time,success = "antes") %>%
  hypothesize(null = "point",p=.5) %>%
  generate(reps = 100) %>%
  calculate(stat = "prop")
```

Visualizing null distribution

```
null_dist_class_prop %>%
  visualize() + shade_p_value(obs_stat = d_hat_class_prop,direction = "right")
```



We can next use this distribution to observe our p-value. as this is a right-tailed test we will be looking for values that are greater than or equal to 0.43 for our p-value

Understanding p-value and decision making with it

```
null_dist_class_prop %>%
  get_p_value(obs_stat = d_hat_class_prop,direction = "right")
```

p_value
1

Our p value caluclated as 0.93, here it's greater than the significance level 0.05 hence we failt to reject the null hypothesis and we can also see this from the histogram above that we are not far into the tails of the null distribution

Research Question 11:

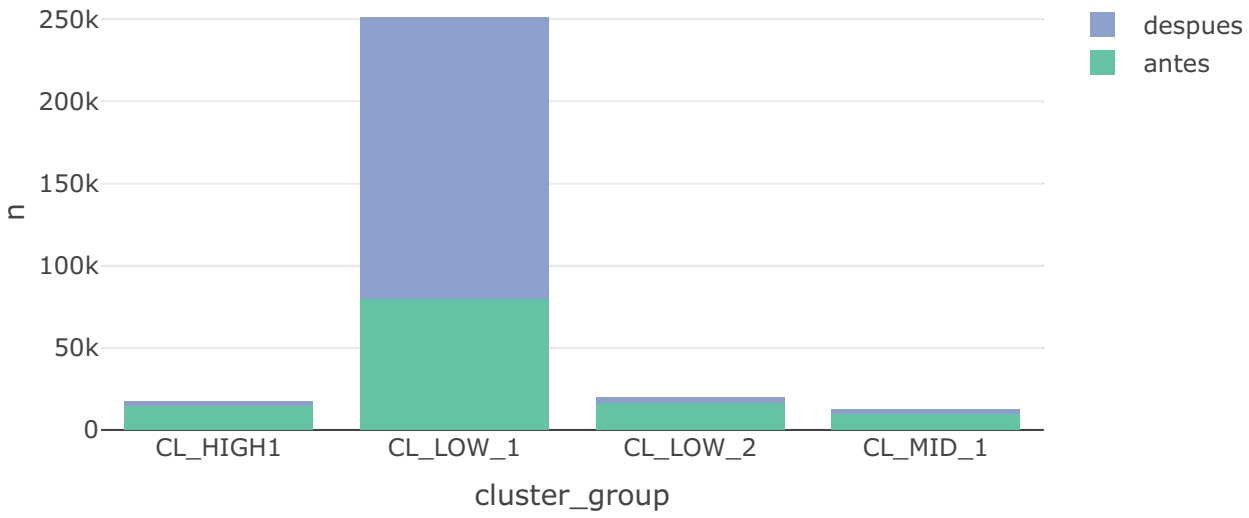
Null Hypothesis(H_o): $P_{clow1} = P_{cllow2} = P_{clmid1} = P_{clhigh1}$ where p represents the long-run probability a message will be received after it's processed.

Alternate Hypothesis (H_a): At least one of these parameter probabilities is different from the others

cluster groups data per messaging classification time

```
mensajes_de_sc_process_days %>%
  filter(cluster_group %in% c("CL_LOW_1","CL_LOW_2","CL_MID_1","CL_HIGH1")) %>%
  janitor::tabyl(class_time,cluster_group)
```

class_time	CL_HIGH1	CL_LOW_1	CL_LOW_2	CL_MID_1
antes	15158	79890	16634	11032
despues	1130	171623	3608	1473



Observerd statistic

```
d_hat_class_user <-  mensajes_de_sc_process_days %>%
  filter(cluster_group %in% c("CL_LOW_1","CL_LOW_2","CL_MID_1","CL_HIGH1")) %>%
  specify(class_time ~ cluster_group ) %>%
  calculate(stat = "Chisq")
```

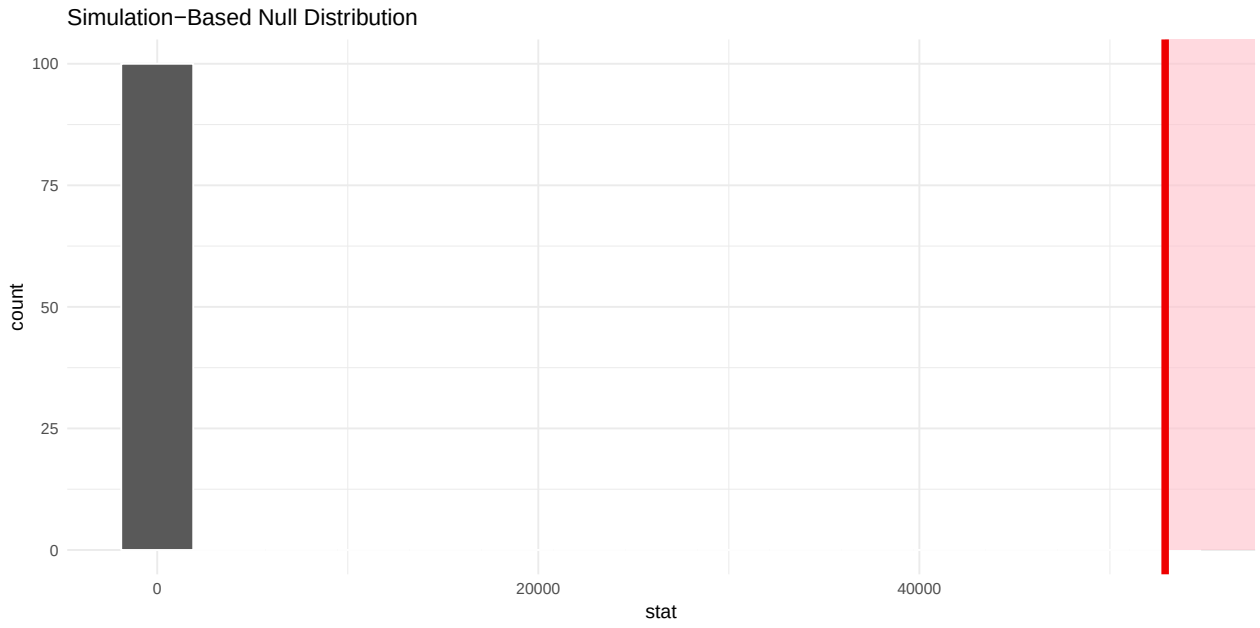
Observed statistic $\tilde{\chi}^2_{obs}$ is 3883 with degrees of freedom 3

Framing hyothesis

```
null_dist_class_prop <- mensajes_de_sc_process_days %>%
  filter(cluster_group %in% c("CL_LOW_1","CL_LOW_2","CL_MID_1","CL_HIGH1")) %>%
  specify(class_time ~ cluster_group ) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 100, type = "permute") %>%
  calculate(stat = "Chisq")
```

Visualizing null distribution

```
visualize(null_dist_class_prop) + shade_p_value(obs_stat = d_hat_class_user,direction =
```



Understanding p-value and decision making with it

```
null_dist_class_prop %>% get_p_value(obs_stat = d_hat_class_user,direction = "greater")
```

p_value
0

Here our p-value is 0 and we can reject the null hypothesis at the 5% level.We can also see this from the histogram above that we are far into the tail of the null distribution.

Research Question 12:

Null Hypothesis(H_o): $P_{USA} = P_{INDIA} = P_{EUROPE}$ where p represents the long-run probability a message will be received after it's processed.

Alternate Hypothesis (H_a): At least one of these parameter probabilities is different from the others

User data per messaging classification time

```
mensajes_de_sc_process_days %>%
  janitor::tabyl(class_time,user)
```

class_time	EUROPE	INDIA	USA
antes	33143	25026	78629
despues	47861	46692	84301

Observerd statistic

```
d_hat_class_user_time <-  mensajes_de_sc_process_days %>%
  specify(class_time ~ user ) %>%
  calculate(stat = "Chisq")
```

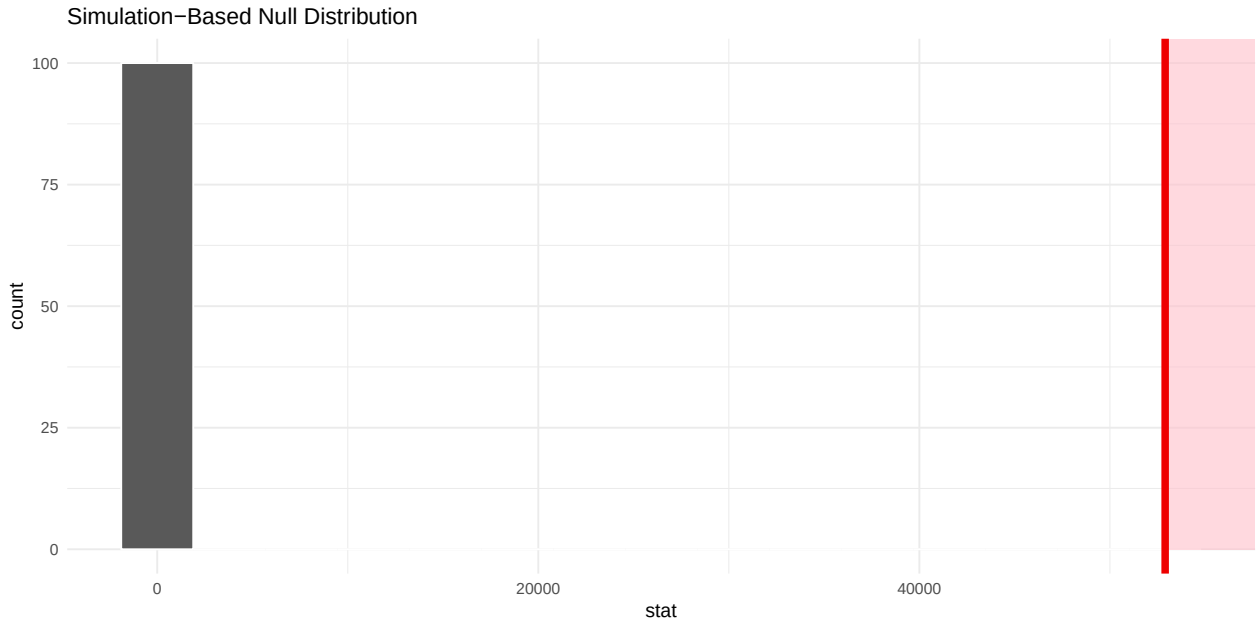
Observed statistic $\tilde{\chi}^2_{obs}$ is 3883 with degrees of freedom 2

Framing hyothesis

```
null_dist_class_user_prop <- mensajes_de_sc_process_days %>%
  dplyr::select(class_time,user) %>%
  specify(class_time ~ user ) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 100, type = "permute") %>%
  calculate(stat = "Chisq")
```

Visualizing null distribution

```
null_dist_class_user_prop %>%
  visualize() + shade_p_value(obs_stat = d_hat_class_user,direction = "greater")
```



Understanding p-value and decision making with it

```
null_dist_class_user_prop %>% get_p_value (obs_stat = d_hat_class_user,direction = "grea
```

p_value
0

Here our p-value is 0 and we can reject the null hypothesis at the 5% level.We can also see this from the histogram above that we are far into the tail of the null distribution

Research Question 13:

Null Hypothesis(H_o): There is no difference between the median time of message received after it's processed i.e (antes) and a message received before it's processed i.e (despues) ($\bar{X}_{antes}$) - ($\bar{X}_{despues}$) = 0

Alternate Hypothesis (H_a): There is a difference between the median time of message received after it's processed i.e (antes) and a message received before it's processed i.e (despues) ($\bar{X}_{antes}$) - ($\bar{X}_{despues}$) $\neq$ 0

Observerd statistic

```
d_hat_class_fp_ttl <- mensajes_de_sc_process_days %>%
  mutate(fp_ttl_days=as.numeric(fp_ttl_days)) %>%
  specify(fp_ttl_days ~ class_time  ) %>%
  calculate(stat = "diff in medians",order = c("antes","despues"))
```

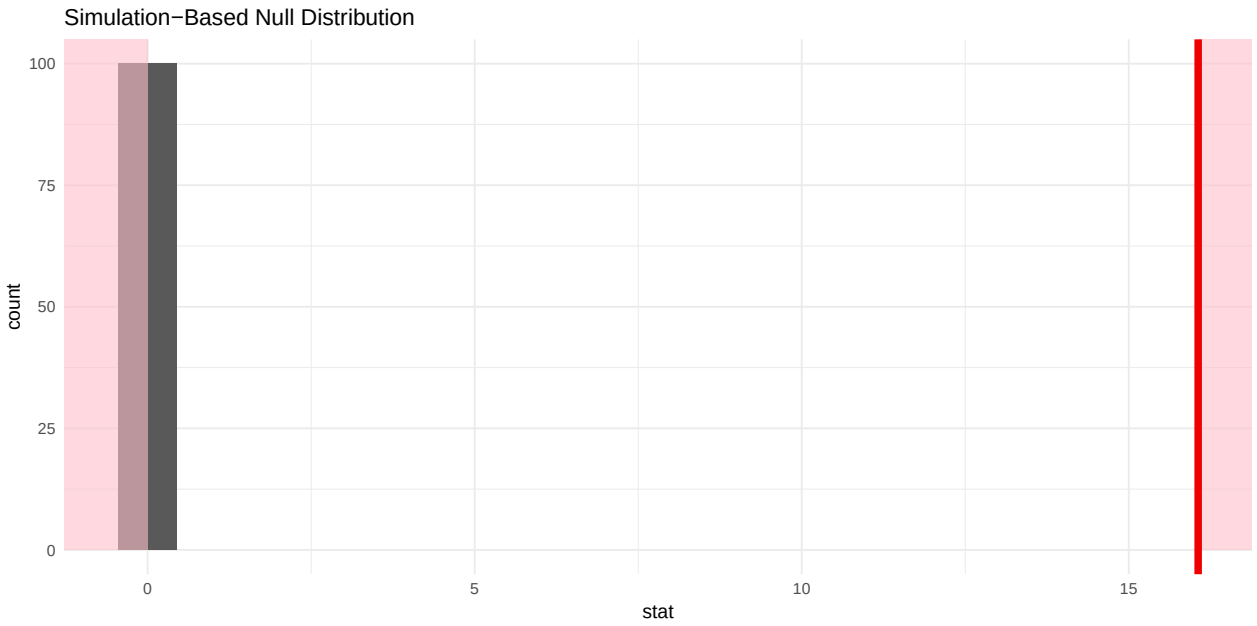
Observed statistic $\bar{X}_{antes}$) - ($\bar{X}_{despues}$ is 16.1

Framing hyothesis

```
null_dist_class_fp_ttl <- mensajes_de_sc_process_days %>%
  mutate(fp_ttl_days=as.numeric(fp_ttl_days)) %>%
  specify(fp_ttl_days ~ class_time  ) %>%
  hypothesize(null="independence") %>%
  generate(reps = 100) %>%
  calculate(stat = "diff in medians",order = c("antes","despues"))
```

Visualizing null distribution

```
null_dist_class_fp_ttl%>%
  visualize() + shade_p_value(obs_stat = d_hat_class_fp_ttl, direction = "two_sided")
```



We can next use this distribution to observe our p-value. since this is a two-tailed test we will be looking for values that are greater than or equal to 16.1 or less than or equal to -16.1 for our p-value.

Understanding p-value and decision making with it

```r
null_dist_class_fp_ttl %>%
  get_p_value(obs_stat = d_hat_class_fp_ttl, direction = "two_sided")
```

p_value
0

Here our p-value is 0 and we can reject the null hypothesis at the 5% level.We can also see this from the histogram above that we are far into the tail of the null distribution

Part 3. Text analytics on SMS-EMAIL Messaging Traffic Datasets

The process of distilling actionable insights from text, it has got a workflow as illustrated below

1. Problem definition and Specific goals
2. Identify text to be collected
3. Text Organization
4. Feature extraction
5. Analysis
6. Reach an insight

Text datasets are diverse and ubiquitous, and sentiment analysis provides an approach to understand the attitudes and opinions expressed in these texts.

In text mining, we often have collections of documents, such as blog posts or news articles, that we'd like to divide into natural groups so that we can understand them separately. Topic modeling is a method for unsupervised classification of such documents, similar to clustering on numeric data, which finds natural groups of items even when we're not sure what we're looking for.

Latent Dirichlet allocation (LDA) is a particularly popular method for fitting a topic model. It treats each document as a mixture of topics, and each topic as a mixture of words. This allows documents to "overlap" each other in terms of content, rather than being separated into discrete groups, in a way that mirrors typical use of natural language

Latent Dirichlet allocation is one of the most common algorithms for topic modeling.

1. Every document is a mixture of topics.

We imagine that each document may contain words from several topics in particular proportions. For example, in a two-topic model we could say "Document 1 is 90% topic A and 10% topic B, while Document 2 is 30% topic A and 70% topic B."

2. Every topic is a mixture of words.

For example, we could imagine a two-topic model of American news, with one topic for "politics" and one for "entertainment." The most common words in the politics topic might be "President", "Congress", and "government", while the entertainment topic may be made up of words such as "movies", "television", and "actor". Importantly, words can be shared between topics; a word like "budget" might appear in both equally.

LDA is a mathematical method for estimating both of these at the same time: finding the mixture of words that is associated with each topic, while also determining the mixture of topics that describes each document.

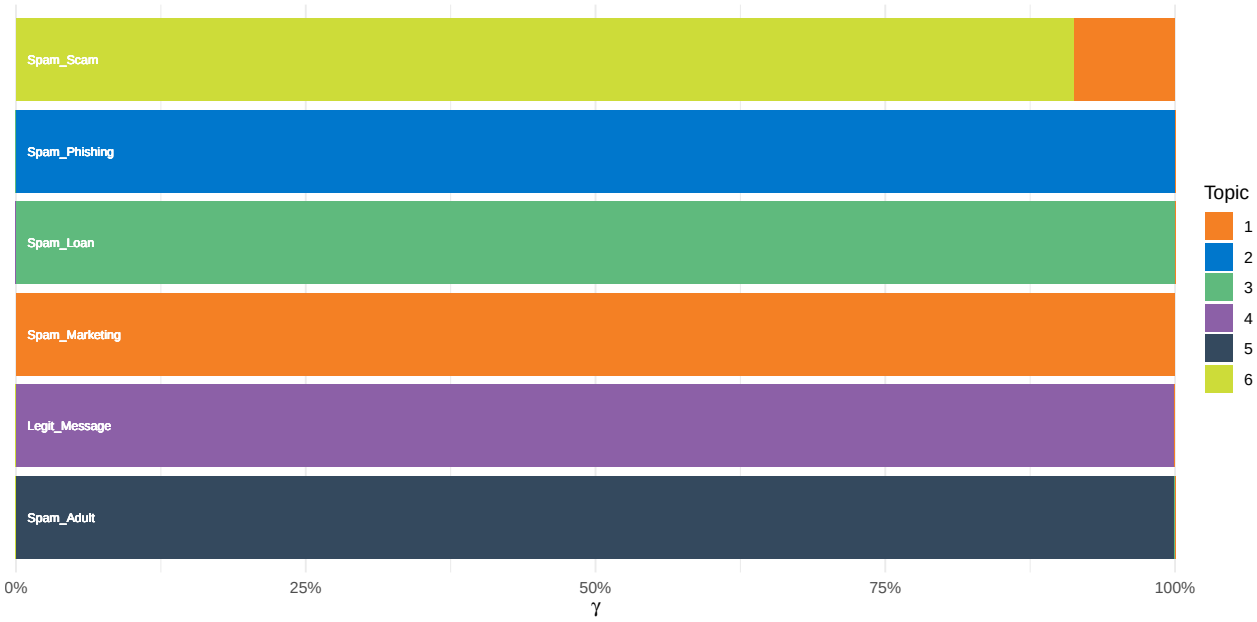
Word-topic probabilities: The tidytext package provides this method for extracting the per-topic-per-word probabilities, called β-"beta".

Document-topic probabilities: Besides estimating each topic as a mixture of words, LDA

also models each document as a mixture of topics. We can examine the per-document-per-topic probabilities, called γ-"gamma".

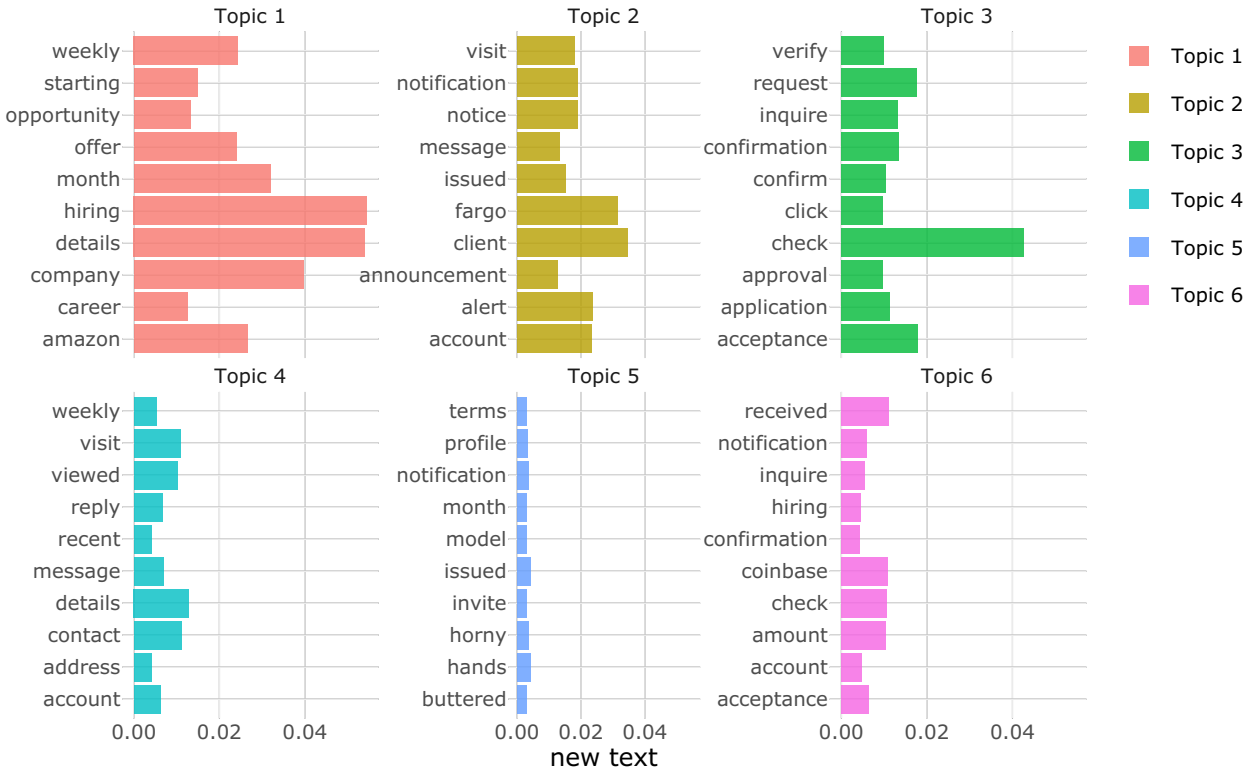
Making use of above derived concepts the topic modeling has been applied on to the email messaging traffic data to look through messaging patterns being sent out by spammmers, here we have asked ourselves the below given questions to understand the patterns and how they have been handled by fingerprinting filter .

Text EDA1: Documents by topic,here we can see the per-document-per-topic probabilities, called γ-"gamma".



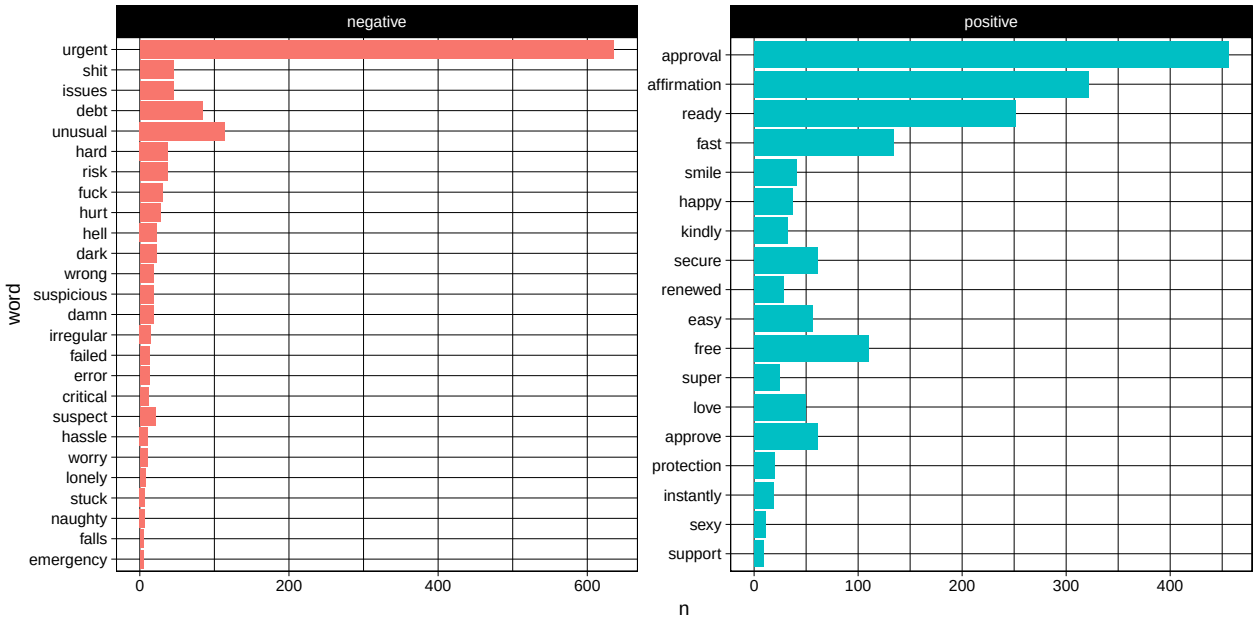
Here it is observed that most of spam campaing categories(such spam_phishing(topic-2),spam_loan(topic-3)„spam_adult(topic-5) and legit_messages(topic-4)) have consisted of unique their unique words in documents, and on the other hand, spam_scam and spam_marketing(topic 1 and topic 6) were found to have got unique words in their messaging traffic. in the below eda words per each dopic can be reviewed.

Text EDA2: Words by topic, here we can see the per-topic-per-word probabilities, called β-"beta"



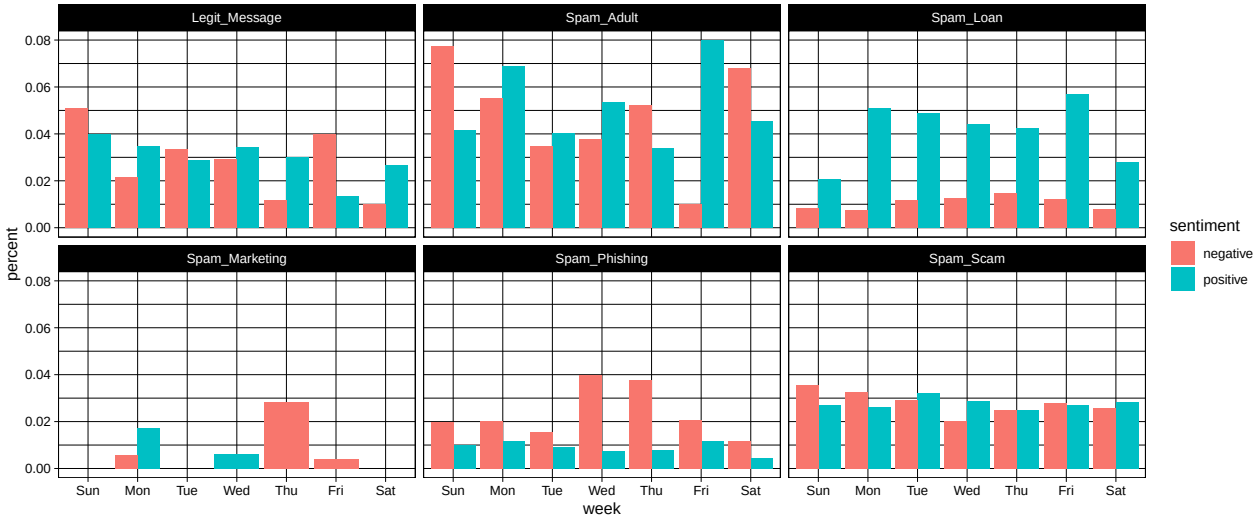
As said the topic 1(spam_scam) and topic 6(spam_marketing) have got commong word framings in the messaging traffic , and a message can be formed with these words for example: scams:- amazon is hiring for call center exucutives please contact to xxxx@gmail.com, you can earn money weekly by clicking on this link https://eee.eeeee. marketing:- XYZ Company is selling/offering products, check them out on https://xxxx.cccc. From the above diagram we can come to know how the messages are being framed with different kind of words,

Text EDA3 :Visualizing Positive and Negative words of EMAIL Traffic



An abjective of this eda is to understand the users/subscribers are being attacked/asked with messages that could have positive and negative kind of meanings. as seen here positive words would be framed in a sentance as your loan approval has been done, kindly send your details to a mail id, secure your account, so on and so forth. and similarily negative words would be used as there is a suspicious activity done on your account, please check it out, feeling lonely there chat with me here on app. so on and do forth.

Text EDA4: How the sentiment has been getting changed over the weeks in EMAIL Traffic?



The negative messaging content was found to be more in spam_adult category, and through out the days it's levels were at same figure, spam_phishing was also hit with negative wording. and spam_loans and spam_scams have been sent out

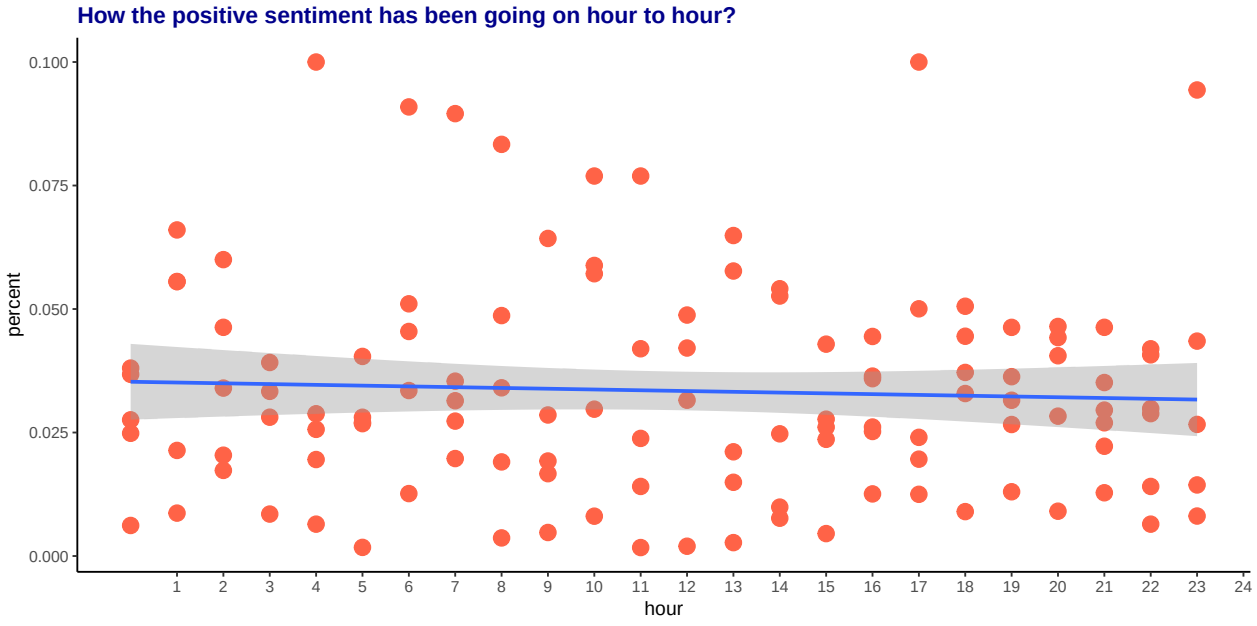
with more of positive contents.

Text EDA5: Sentiment analysis with the "bing" lexicon to look at how the positive sentiment is getting changed over the hours in email traffic?

```
# Use summary to see the results of the model fitting
email_hour_model_positive %>%
  tidy()
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.0352757	0.0038722	9.1100053	0.0000000
hour	-0.0001567	0.0002830	-0.5535165	0.5809483

We are trying to see if the positive content messaging do come in certain hours or not, a simple linear regression model:positive word percentage as a function of hour has been fitted and it's results as, a time goes on increasing by 1hr there is an decrement of 0.01% in the positive sentiment messaging traffic.

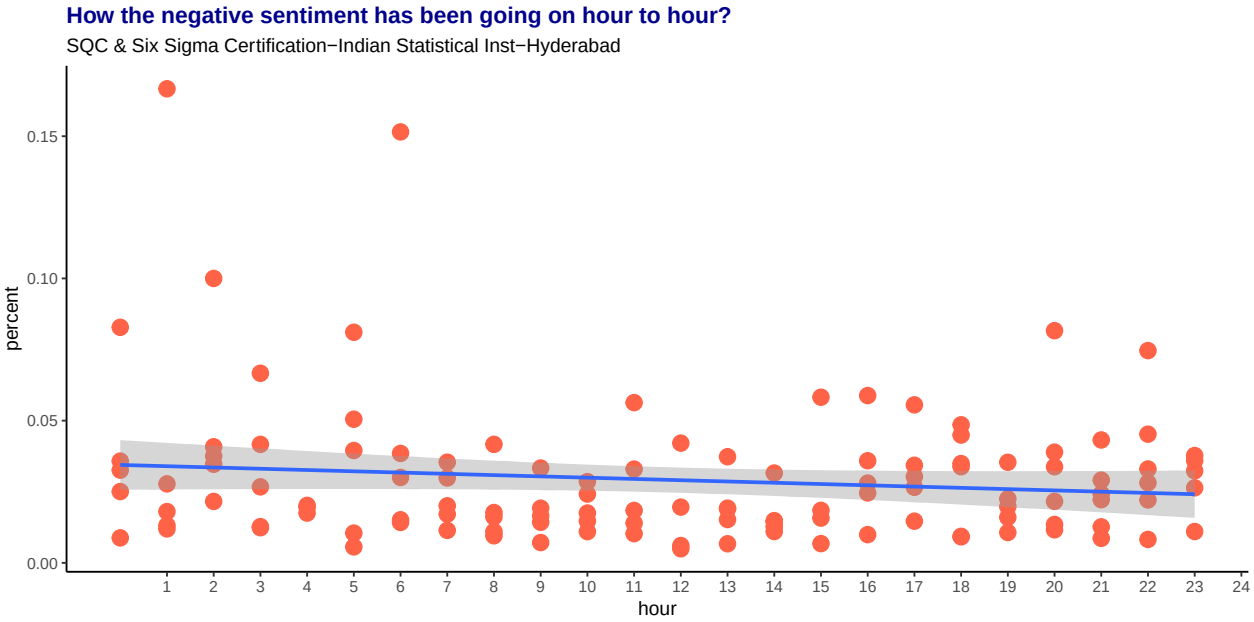


Text EDA6: Sentiment analysis with the "bing" lexicon to look at how the negative sentiment is getting changed over the hours in email traffic?

```
# Use summary to see the results of the model fitting
email_hour_model_negative %>%
  tidy()
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.0344353	0.0044062	7.815156	0.0000000

term	estimate	std.error	statistic	p.value
hour	-0.0004478	0.0003192	-1.402828	0.1632939



We are trying to see if the negative content messaging do come in certain hours or not, a simple linear regression model:negative word percentage as a function of hour has been fitted and it's results as, a time goes on increasing by 1hr there is an decrement of 0.01% in the negative sentiment messaging traffic.

6. DMAIC Improvement

As analized causes eariler the filter fingerprinting had been given an inefficient algorithm that works on older messaging similarity matching theory, it requires a lot of manual interactions by analyst to go on controllig spams,even after classification was done via this filter there were still spam leakages/lower spam blocking happened because of it's time to live was for minutes/hours only and not for days, here as spammers keep on changing the messaging patterns this fingerprinting filter had failed to figure out and block the spams.There would definitely be misclassication/delayes in classification occured unless analysts in a team should be getting their knowledge updated themselves on on-going spam trends,here a Knowledge transformation is absolutely requeired in team.There were days on when a huge amount of spam traffic had been coming in to networks,it could n't be controlled by 1 or 2 analysts during those peak hours,and it had lead to happend spam leakages towards users.

Spam blocking rates are required to be improved as higher as posible without having any spam leakages,to get succeded in it we have carried out couple of researches and come up with an idea of implementing natural language processing techniques in fingerprinting filter,this technique is called word2vec which would be acting as

- 1. Make fingepriting filter to autoclassify and block messages out more efficiently/accurately/quickly.
- 2. Messages start getting blocked as soon as it's autoclassification done, hence no delays happen in classification like we had in manual task.
- 3. Avoid misclassifications, however there is a very less probability to happen misclassifications which can be easily identified/rectified.
- 4. It learns itself from the given training data, cathes out spam patterns easily and no analyst would have to get their knowledge updated himself.
- 5. It eliminates work duplications and the manual interactions done by analysts,so there is no scope for allocating more number of analysts in shift hours.

Modeling- Word2Vec and Gradient Boosting Machines using H2O on EMAIL Messaging Traffic

Word2Vec

The Word2vec algorithm takes a text corpus as an input and produces the word vectors as output. The algorithm first creates a vocabulary from the training text data and then learns vector representations of the words.

The vector space can include hundreds of dimensions, with each unique word in the sample corpus being assigned a corresponding vector in the space. In addition, words that share similar contexts in the corpus are placed in close proximity to one another in the space.

GBM

Gradient boosted machines (GBMs) are an extremely popular machine learning algorithm

that have proven successful across many domains.Whereas random forests build an ensemble of deep independent trees, GBMs build an ensemble of shallow and weak successive trees with each tree learning and improving on the previous. When combined, these many weak successive trees produce a powerful "committee" that are often hard to beat with other algorithms.

It's model implementation given below with suitable examples.

```
#"Splitting the data into Train/Validation/Test set"
data_email_split <- h2o.splitFrame(data_email,c(0.8,0.1), seed = 123)

# Test Data-SMS-EMAIL
test_email_traffic <- data_email_split[[3]]

#"Break Message content into sequence of words"
words <- tokenize(traffico_de_email_content_en_h2o$content_txt)


#"Build word2vec model"
w2v_model_traffic <- h2o.word2vec(words, sent_sample_rate = 0, epochs = 10)


#"Build a basic GBM model"
email_gbm_model <- h2o.gbm(x = names(spam_cat_vec), y = "spam_campaign",
                     training_frame = data_email_split[[1]],
                     validation_frame = data_email_split[[2]])
```

Case 1: Here let's check what word vectors are more similar to the word "send"

```
#"Sanity check - find synonyms for the word 'send'"
print(h2o.findSynonyms(w2v_model_traffic, "send", count = 10))
```

```
##     synonym     score
## 1      pays 0.6903180
## 2     reply 0.6867326
## 3     promo 0.6828830
## 4     decal 0.6780915
## 5      drop 0.6702445
## 6    dollar 0.6688312
## 7  wireless 0.6650684
## 8       pay 0.6619691
## 9   sticker 0.6595681
## 10 wrapping 0.6532398
```

a word **send** has got a list of 10 synonyms those were learned from the given training data by a model, here for example word send associates with reply/pays/dollar/dollars/kindly in this context, they could be spam_scam categories:

- kindly send your cv to this email we will provide you an opportunity
- reply to this email we will send money to your account
- you have won a lottery we send dollars to you

Case 2: Here let's check what word vectors are more similar to the word "locked"

```
#"Sanity check - find synonyms for the word 'locked'"
print(h2o.findSynonyms(w2v_model_traffic, "locked", count = 10))
```

```
##       synonym     score
## 1   suspended 0.7689240
## 2     blocked 0.7639100
## 3    disabled 0.7125040
## 4       debit 0.7121428
## 5      unlock 0.7048428
## 6         pnc 0.6508078
## 7     recover 0.6449929
## 8  restricted 0.6411480
## 9      closed 0.6247987
## 10    resolve 0.6201200
```

a word **locked** has got a list of 10 synonyms those were learned from the given training data by a model, here for example word "locked" associates with suspended/blocked/unlock/disabled/debit/recover/visa in this context,they could be spam_phishing categories:

- your visa account has been locked please unlock it here http:///xxxxx.xxxx
- we have suspended your bank account please registered here http:///xxxxx.xxxx
- your visa debit card got locked recover it here http:///xxxxx.xxxx

We are giving in new unclassified messages to word2vec model to see how it starts classifying them accuratley as demonstrated below five cases.

6.ex-case-1

```
print(predict("John Find cute grls nearby At getfindone.org",
              w2v_model_traffic, email_gbm_model))
```

```
##
  |
  |                                                                      |   0%
  |
  |======================================================================| 100%
##
  |
```

```
  |                                                                      |   0%
  |
  |======================================================================| 100%
##         predict Legit_Message Spam_Adult Spam_Gambling  Spam_Loan
## 1 Spam_Phishing    0.01950904 0.02256289   0.001006233 0.01591703
##   Spam_Malware Spam_Marketing Spam_Phishing Spam_Scam
## 1   0.07910605    0.001587938     0.5356576 0.3246532
##
## [1 row x 9 columns]
```

Here the given message *John Find cute grls nearby At getfindone.org* has been correctly classified to spam_scam category.

6.ex-case-2

```
print(predict("whatever it is we need to say. Kisses
              like that cant lingerhttp://goo.gl/K6MQ4S",
              w2v_model_traffic, email_gbm_model))
```

```
##
  |
  |                                                                      |   0%
  |
  |======================================================================| 100%
##
  |
  |                                                                      |   0%
  |
  |======================================================================| 100%
##     predict Legit_Message Spam_Adult Spam_Gambling  Spam_Loan Spam_Malware
## 1 Spam_Scam   0.007302804 0.02848359  0.0007848102 0.03695989   0.06201926
##   Spam_Marketing Spam_Phishing Spam_Scam
## 1    0.001131732    0.04672656 0.8165913
##
## [1 row x 9 columns]
```

Here the given message "whatever it is we need to say. Kisses like that cant lingerhttp://goo.gl/K6MQ4S" has been correctly classified to spam_scam category.

6.ex-case-3

```
print(predict("Your BMOMontréal online access has been disabled.
              Please login and verify your info to restore:
              http://www.12bm0-action45.com/1",
              w2v_model_traffic, email_gbm_model))
```

```
##
  |
  |                                                                      |   0%
  |
  |======================================================================| 100%
##
  |
  |                                                                      |   0%
  |
  |======================================================================| 100%
##     predict Legit_Message  Spam_Adult Spam_Gambling  Spam_Loan
## 1 Spam_Scam     0.0206519 0.005436774  0.0008310246 0.01511046
##   Spam_Malware Spam_Marketing Spam_Phishing Spam_Scam
## 1   0.06822108    0.001443289     0.4269917 0.4613138
##
## [1 row x 9 columns]
```

Here the given message "Your BMOMontréal online access has been disabled. Please login and verify your info to restore: http://www.12bm0-action45.com/1" has been correctly classified to spam_phishing category.

6.ex-case-4

```
print(predict("work from home as a PA/Errand for a reasonable weekly fee
of $500 if interested contact georgepw212@gmail.com for more information",
              w2v_model_traffic, email_gbm_model))
```

```
##
  |
  |                                                                      |   0%
  |
  |======================================================================| 100%
##
  |
  |                                                                      |   0%
  |
  |======================================================================| 100%
##     predict Legit_Message  Spam_Adult Spam_Gambling  Spam_Loan
## 1 Spam_Scam    0.07725171 0.008036234  0.0007247574 0.02798183
##   Spam_Malware Spam_Marketing Spam_Phishing Spam_Scam
## 1   0.06025254    0.002078699    0.03656411 0.7871101
##
## [1 row x 9 columns]
```

Here the given message *work from home as a PA/Errand for a reasonable weekly fee of $500 if interested contact georgepw212@gmail.com for more information* has been correctly classified

to spam_scam category.

6.ex-case-5

```
print(predict("Card Message MSG:http://chase.com.online.odbetxnpgpjupu
              qzqkqjmgcejqztvpawnewbwemi@nalcrni.org",
              w2v_model_traffic, email_gbm_model))
```

```
##
  |
  |                                                                      |   0%
  |
  |======================================================================| 100%
##
  |
  |                                                                      |   0%
  |
  |======================================================================| 100%
##         predict Legit_Message  Spam_Adult Spam_Gambling   Spam_Loan
## 1 Spam_Phishing   0.007862514 0.002106489  0.0003644812 0.004312733
##   Spam_Malware Spam_Marketing Spam_Phishing  Spam_Scam
## 1   0.02952538   0.0009653113     0.8777885 0.07707459
##
## [1 row x 9 columns]
```

Here the given message "Card Message MSG:http://chase.com.online.odbetxnpgp jupuqzqkqjmgcejqztvpawnewbwemi@nalcrni.org" has been correctly classified to spam_phishing category.

Glance at 2 weeks latest spam locking rates after introducing new algorithm in fingerprinting filter.

day	blocked_rate
2019-01-07	20
2019-01-08	24
2019-01-09	23
2019-01-10	21
2019-01-11	24
2019-01-12	21
2019-01-13	25
2019-01-14	26
2019-01-15	30
2019-01-16	32

Hypothesis testing on spam blocking rates of manual and automation classification process.

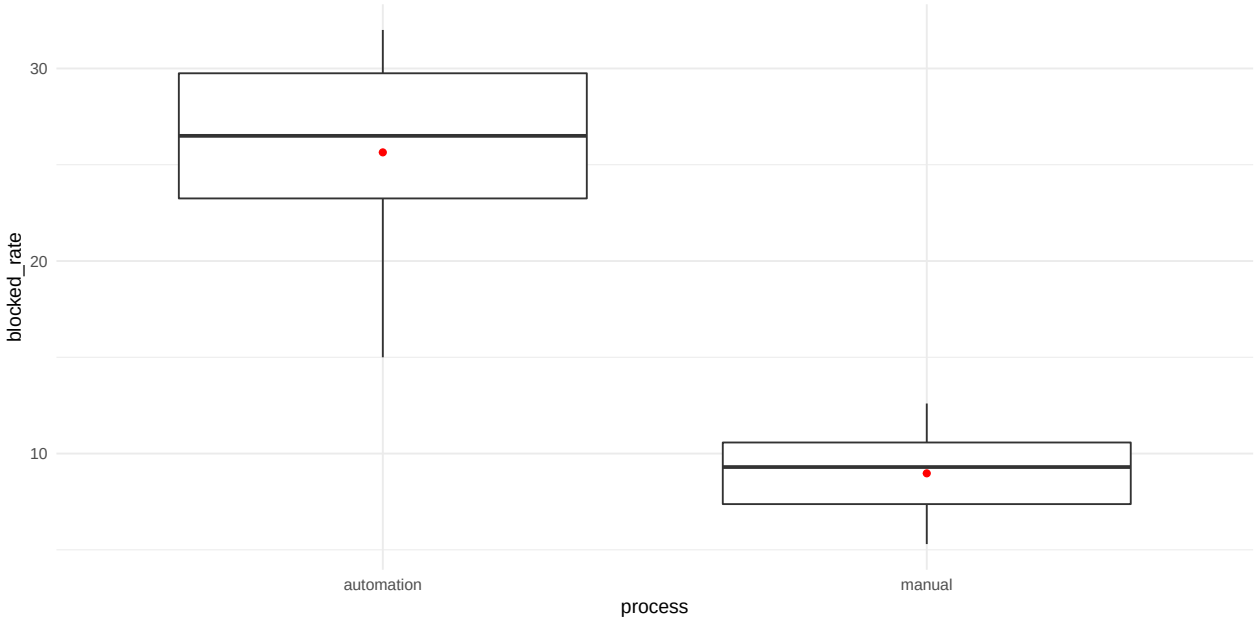
Null Hypothesis(H_o): There is no difference between the avg spam blocking rates of automation and manual spam classification.

μ_{diff}:(μ_{manual}) - (μ_{auto}) = 0

Alternate Hypothesis (H_a): The avg spam blocking rates in automated spam classification process is higher than the manual classification process

μ_{diff}: > 0

Taking a look at spam blocking rates recoderd for automation and manual process



Observerd statistic

```
d_hat_auto_man <- spam_block_rates_auto_man %>%
  mutate(process=as.factor(process)) %>%
  specify(blocked_rate ~ process) %>%
  calculate(stat = "diff in means",order = c("automation","manual"))
```

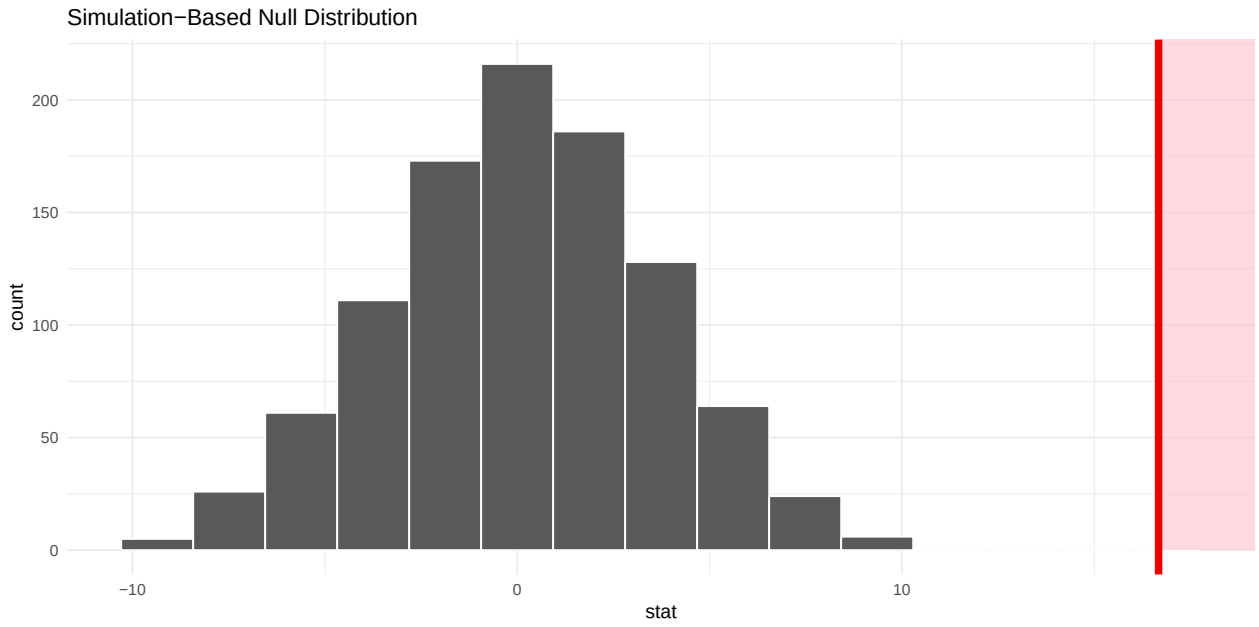
specifying null hypothesis

```
null_distn_blocked_auto_man <- spam_block_rates_auto_man %>%
  mutate(process=as.factor(process)) %>%
  specify(blocked_rate ~ process) %>%
  hypothesize(null = "independence") %>%
```

```
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means",order = c("automation","manual"))
```

Visualizing null distribution with p value

```
visualize(null_distn_blocked_auto_man) +
  shade_p_value(obs_stat = d_hat_auto_man, direction = "right")
```



Understanding p-value and decision making with it

```
null_distn_blocked_auto_man %>%
  get_p_value(obs_stat = d_hat_auto_man, direction = "right")
```

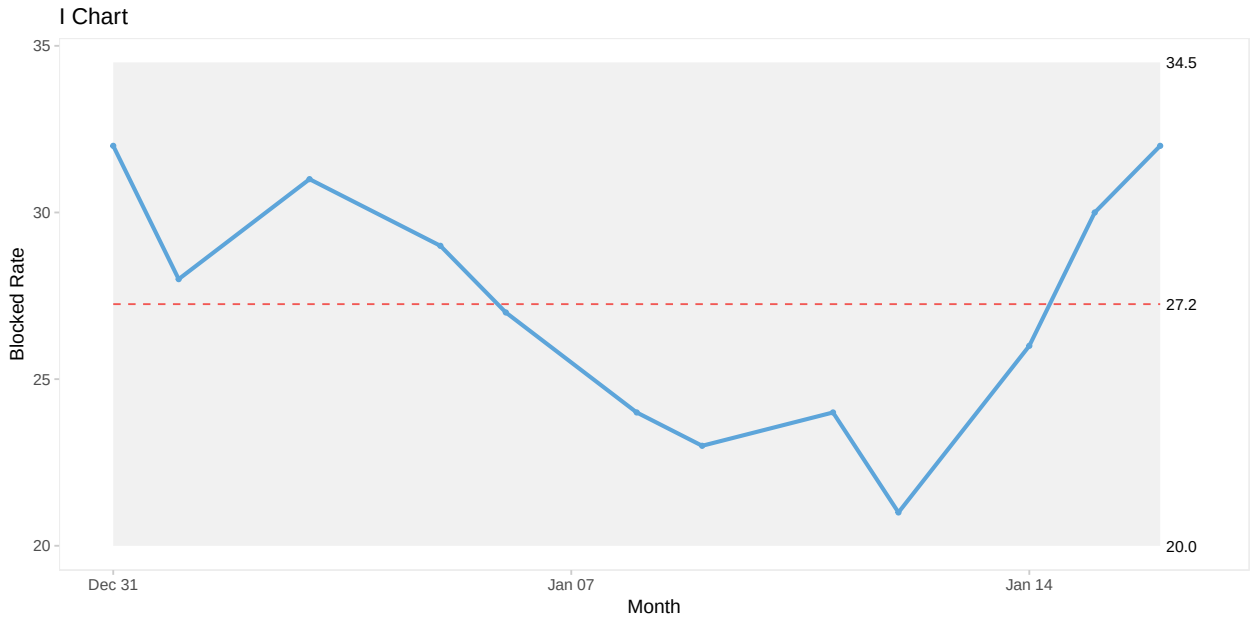
p_value
0

Here our p-value is 0 and we can reject the null hypothesis at the 5% level.We can also see this from the histogram above that we are far into the tail of the null distribution.

Here it's proved that after implementing a new algorithm in fingerprinting filter for classification and blocking of spams in messaging traffic the blocked rates have gone increasing and they are fairly higher than the rates of manual classification process.

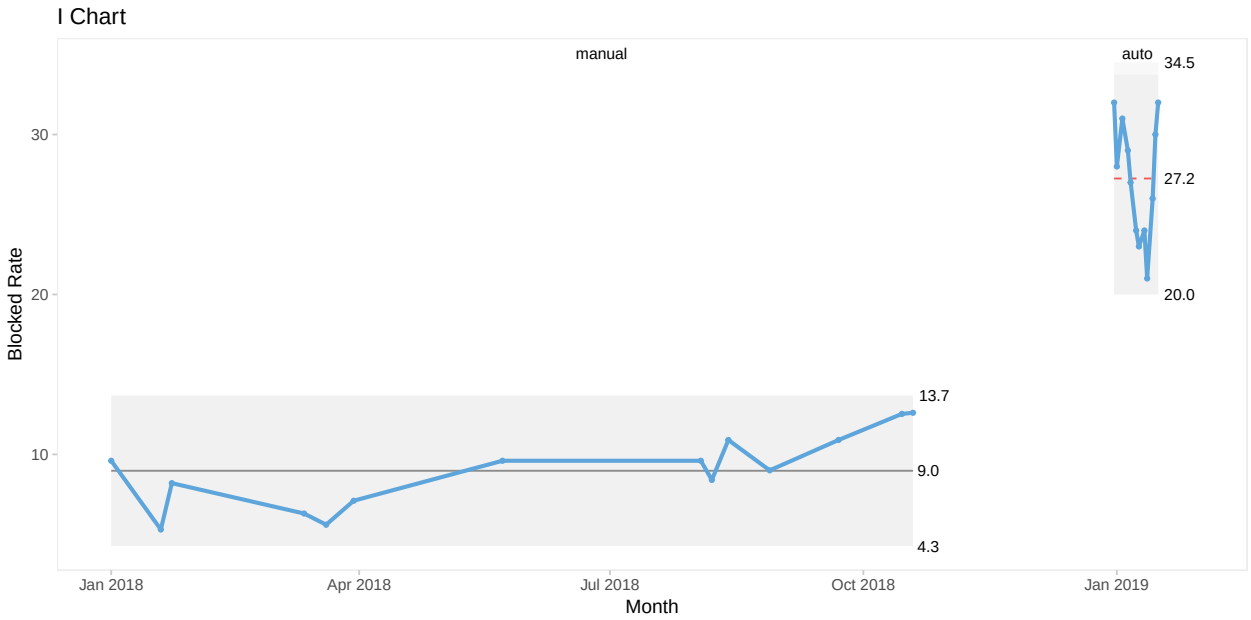
7. DMAIC Control

Process Control:I Chart of spam blocking rates-automated spam classification process



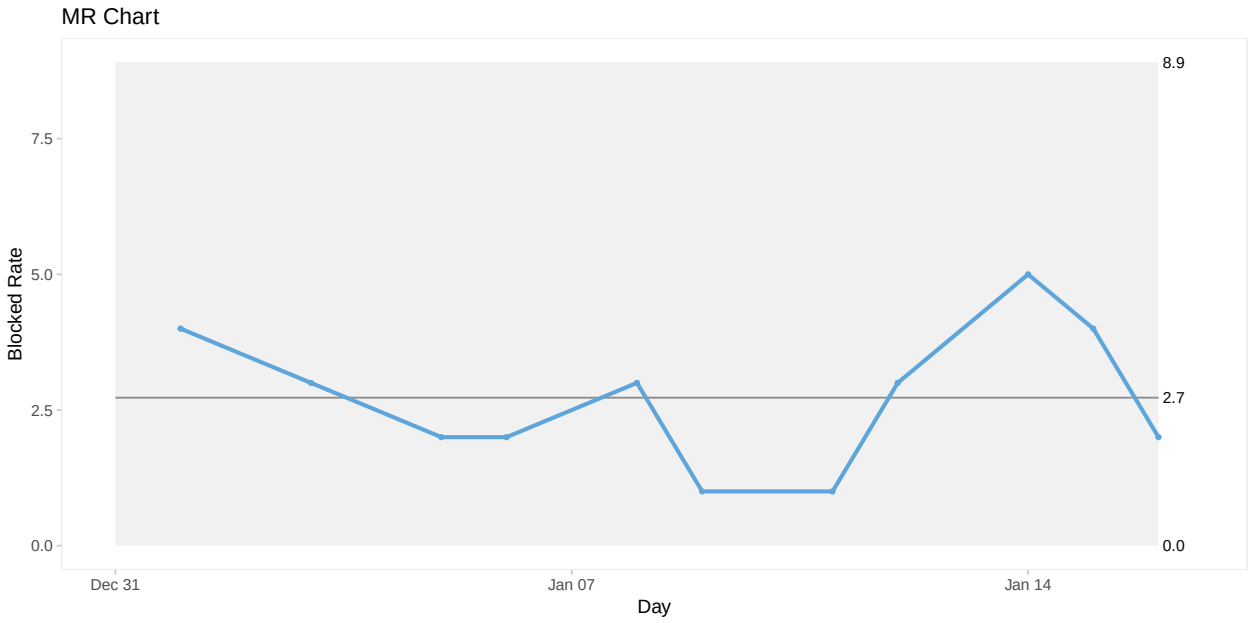
Here I-chart plots spam blocking rates of each day processed by automated classification system the center line is an estimate of the process average, and it's found to be **27** and The control limits **(UCL:34.5,LCL:20)** on the I chart which are set at a distance of 3 standard deviations above and below the center line, show the amount of variation that is expected in the individual sample values.Here all the process observations have stranded in between the control limits, no observation has gone beyond the control limits, hence the said process is found to be in control.

Process Control:I Chart-Manual Vs Auto Spam classification process



Here I-chart plots spam blocking rates of each day processed by automated classification system Vs Manual classification system, the center line is an estimate of the process averages, and it's observed that the automated process average **27** is signifacntly higher than the avg manual process.

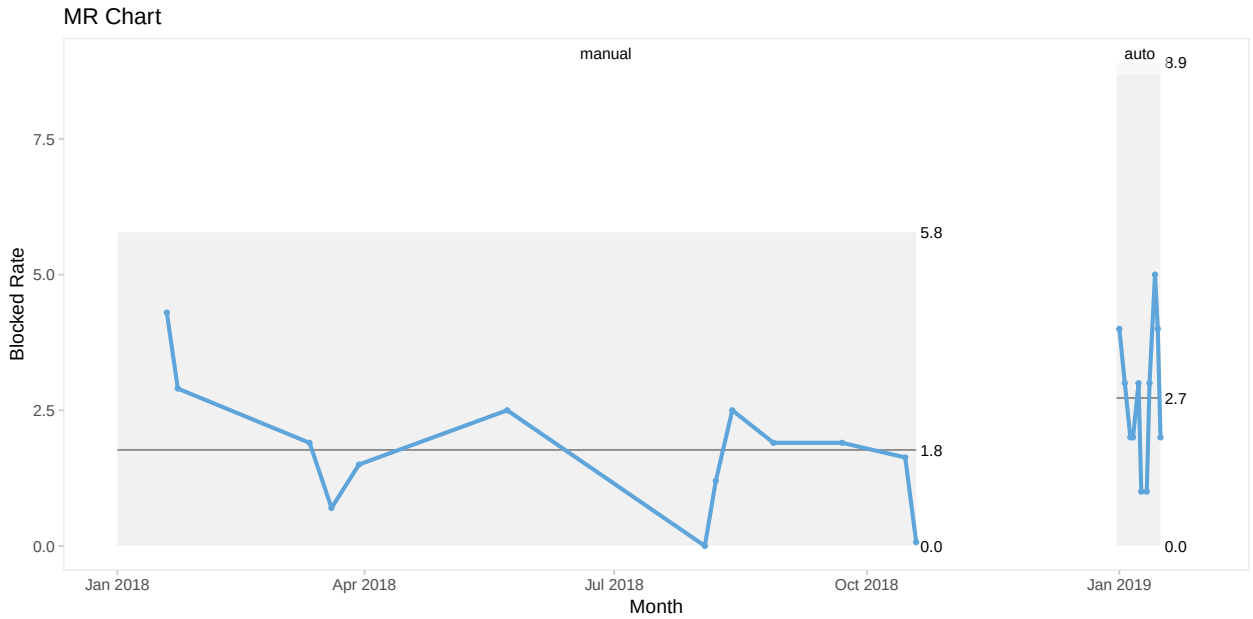
Process Control:MR Chart of spam blocking rates-automated spam classification process



Here MR-Chart plots the moving ranges of spam blocking rates of automated classification system,The center line is the average of all moving ranges i.e **2.7**.The control limits on the moving range chart **(UCL:8.9,LCL:0.0)**, which are set at a distance of 3 standard deviations

above and below the center line, show the amount of variation that is expected in the moving ranges of the standardized data..

Process Control:MR Chart- Manual Vs Auto Spam classification process



CONTROL SUBJECT	FREQUENCY OF CHECKS	BY WHOM	WHEN TO TAKE ACTION	WHAT ACTION TO TAKE	BY WHOM2
Training Word2Vec algorithm in Fingerprint Filter	every 2-3Months	Senior Analyst	Blocking rates found to be lower or new spam patterns seen	Feed the new data	Lead Analyst
Do review of messages classified by new fingerprint algorithm	every 1 Week	Junior/Senior Analyst	False positive or False Negatives occurred more	Reclassify messages in US SC	Senior Analyst
Update catridges	every 1 Week	Junior/Senior Analyst	Spam leakages happen	Manually update catridges	Senior Analyst

Figure 4: Process Control Plan

Process Control Plan

The Control Plan is a document that describes the actions (measurements, inspections, quality checks or monitoring of process parameters) required at each phase of a process to assure the process outputs will conform to pre-determined requirements.

In a new process of automated messaging classification the following mentioned process should be taken care of to keep it in control,

- 1. Training Word2Vec algorithm in Fingerprint Filter: this check should be executed every 2-3Months, if these were observed- Lower Blocking rates or new spam patterns seen, Feed the new data into word2vec model, these can be handled by Senior/Lead Analysts.
- 2. Do review of messages classified by new fingerprint algorithm : Automated Classified messages should be reviewed once in a week in case of finding more and more false positives reclassify them in Security Center, and this can be carried out by Senior/Junior Analysts
- 3. Update catridges : Fingerprinting filter catridges are confifured to gets updated regularly in a time interval of 2-5mins, if they are found to be not getring updated propery there could be a chance to have spam leakages even after they are classified to spam, in this case manual updation of catridges should be required to avoid spam leakages, and this can be carried out by Senior/Junior Analysts

8. Conclusion

The process- Messaging classification system that was in Fingerprint filter i.e manual classification have been reimplemented with a newer algorithm after investigating that which were causing to have recorded lower spam blocking rates/Improper blockings and the new process system have also met business requierements as illustrated below,

- 1. Manual classification/monitoring tasks by analysts are no more required to be carried out.
- 2. Messages are correctly classified into their respective categories and false positive and false negatives rates are also controlled and minimized.
- 3. Tracing out spam messaging traffic have become more easier and efficient that makes Spam blocking rates to go on getting improved within a business specifications.

9. Bibliography

- 1. R for DS Book
- 2. Text Mining with R Book
- 3. Tidyverse R packages and QiCharts Package Online Documentation
- 4. Statistical Quality Control- Montgomery Book, and Six Sigma HandBook
- 5. OpenIntro Statistic Books and Online references