Курсов проект

По дисциплина „Вероятности и Статистика“

Летен семестър 2019/2020

Тема:

SPAM E-mail Prediction Analysis

Изготвил:

Мария Балева, Компютърни науки, 3 курс, 81659

15.05.2020г. **гр. София**

1. Описание на данните

За целите на проекта са използвани данни от проучване, проведено от Mark Hopkins, Erik Reeber, George Forman, Jaap Suermondt от Hewlett-Packard Labs. Информацията, използвана в изследването, е предоставена от George Forman. Данните включват 4 601 имейла, сред които работни и лични писма, категоризирани като спам (1 813) и неспам (2 788). Проучването съдържа над 50 велични, представящи честотата на употребата на определени низове или символи в мейлите. Проектът съдържа:

a) Основни въпроси, на които се отговаря чрез анализа на данните

Целта на анализа е да се провери каква значимост имат изброените фактори за маркирането на даден имейл като спам или неспам и дали съществуват съществени зависимости между отделните величини.

b) Променливи, чрез който са представени данните

- **class** -> от категорийен тип, обозначава дали имейлът е класифициран като спам или не
- **capital_run_length_average** -> от числов тип, средна дължина на непрекъснати поредици от главни букви
- **char_freq_CHAR** -> от числов тип, процент от символи в имейла, съвпадащи с CHAR, където CHAR е "**!**"
- **word_freq_WORD** -> от числов тип, процент от думи в имейла, съвпадащи с WORD, където WORD е "**our**", "**free**", "**you**"

c) Използвани статистически методи

- Дескриптивна статистика – средна стойност, медиана, мода, стандартно отклонение и дисперсия
- Графично представяне чрез хистограми и бокс плотове
- Тест на Shapiro-Wilcoxon с нулева хипотеза H0 „Разпределението е нормално“ и алтернативна хипотеза „Разпределението не е нормално“ с равнище на значимост p = 0,05
- Изследване за логистична регресия – установяване как влияят отделните числови величини върху категорийната величина class

- Корелационен анализ за установяване на зависимост между отделните числови непрекъснати величини

2. Дескриптивна статистика

Данните са разделени в две групи според класа си:

email_dataset_spam и email_dataset_nonspam.

Descriptive Statistics of capital run length average					
meanValue	medianValue	modeValue	stanDeviationValue	variationValue	Type
5.19151511	2.276	1	31.72944874	1006.757917	All
9.51916492	3.621	1	49.84618579	2484.642238	Spam
2.37730093	1.857	1	5.113684695	26.14977116	Non-spam

Descriptive Statistics of word frequency YOU					
meanValue	medianValue	modeValue	stanDeviationValue	variationValue	Type
1.6620995	1.31	0	1.775480665	3.152331591	All
2.2645394	2.18	0	1.566885161	2.455129106	Spam
1.2703407	0.51	0	1.793636259	3.217131028	Non-spam

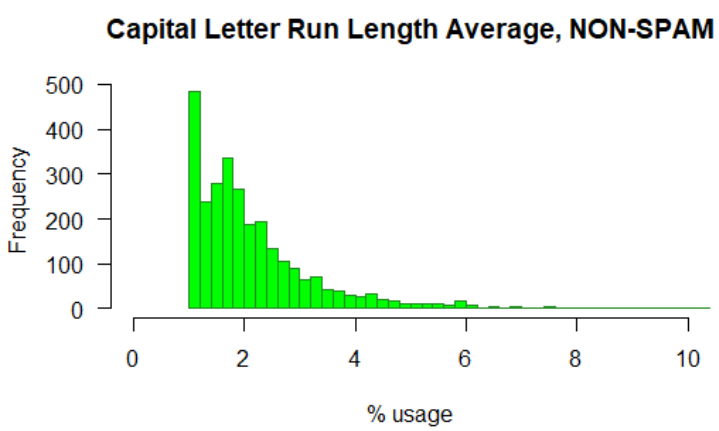
Descriptive Statistics of word frequency OUR					
meanValue	medianValue	modeValue	stanDeviationValue	variationValue	Type
0.3122234	0	0	0.672512769	0.452273425	All
0.5139548	0.29	0	0.707194947	0.500124693	Spam
0.1810402	0	0	0.614521083	0.377636162	Non-spam

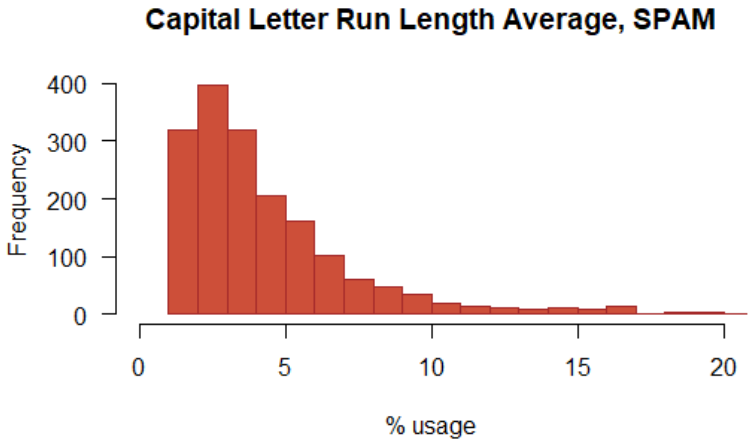
Descriptive Statistics of word frequency FREE					
meanValue	medianValue	modeValue	stanDeviationValue	variationValue	Type
0.2488481	0	0	0.825791701	0.681931934	All
0.5183618	0.14	0	1.013169854	1.026513154	Spam
0.0735868	0	0	0.616573902	0.380163377	Non-spam

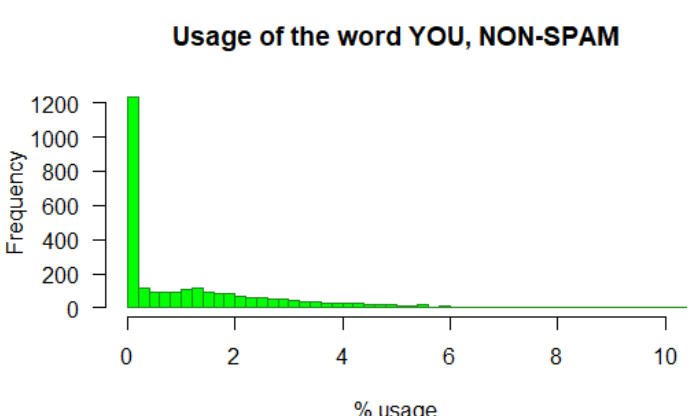
Descriptive Statistics of char frequency "!"					
meanValue	medianValue	modeValue	stanDeviationValue	variationValue	Type
0.269070854	0	0	0.815671631	0.66532021	All
0.513712631	0.331	0	0.74418252	0.553807623	Spam
0.109983501	0	0	0.82085862	0.673808875	Non-spam

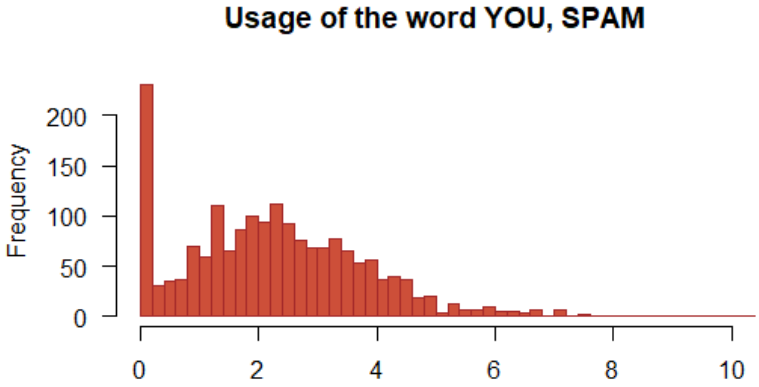
Забелязва се, че най-честата стойност при модата е 0 при анализа на честотата на думите и употребата на възклицателен знак, но като цяло средните стойности са значително по-ниски при имейлите, които са отбелязани като non-spam.

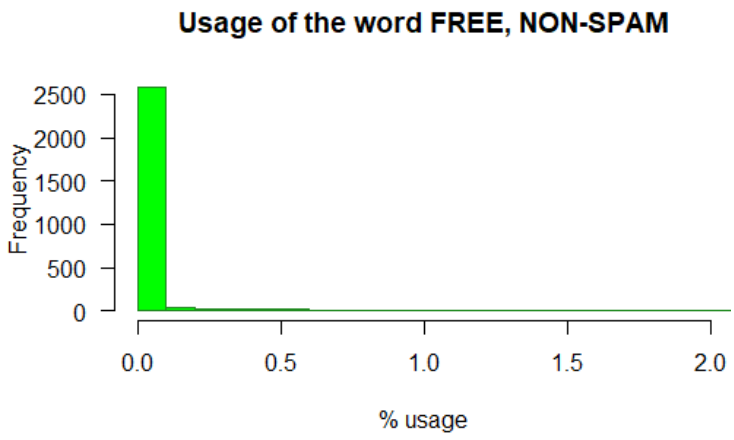
3. Графично представяне

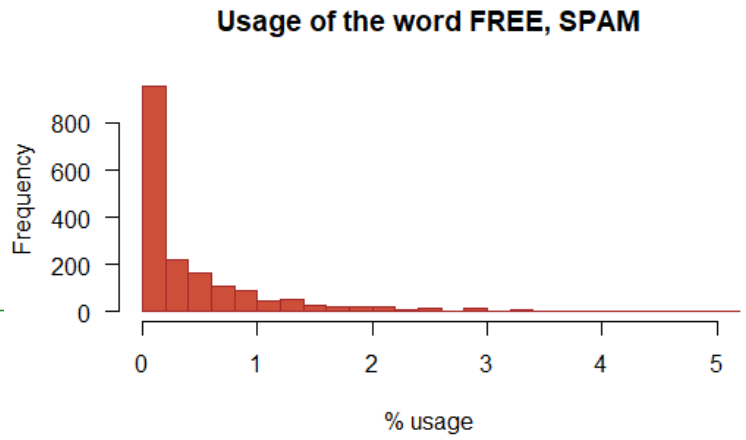


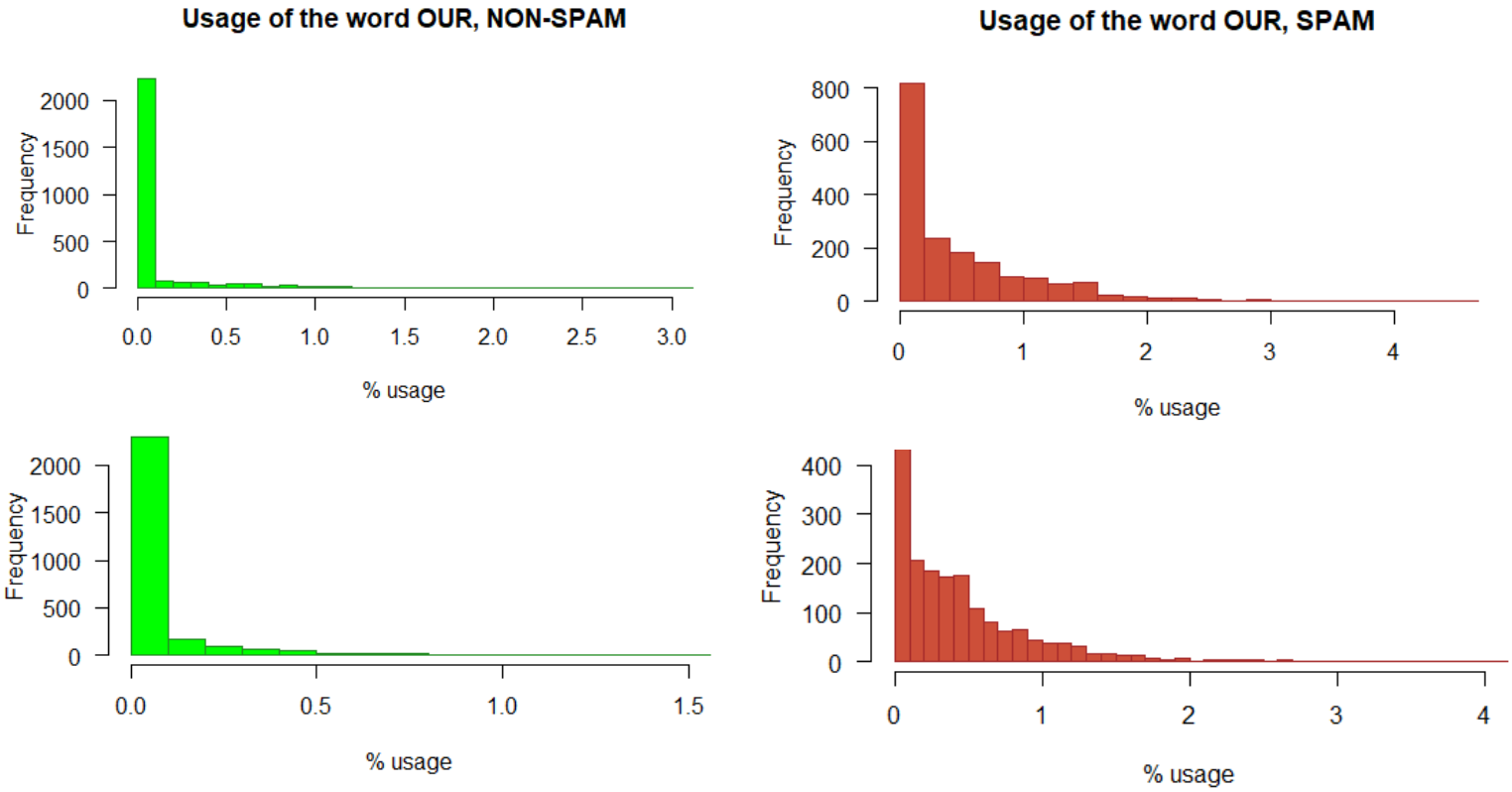












Хистограмите представят

неравномерно разпределение на стойности с ясно изразено струпване в ляво (около нулата). От разгледаните плотове се придобива визуална представа, че принципно spam имейлите имат повече стойности, които не попадат в нулевия хистограмен интервал.

4. Определяне вида на разпределенията

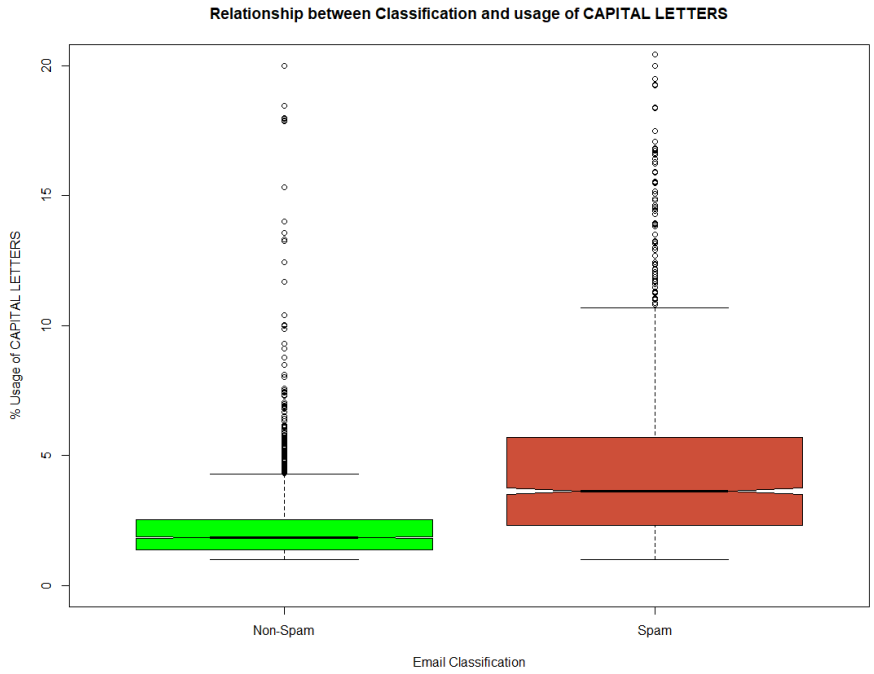
- Тест на Shapiro-Wilcoxon с нулева хипотеза H_0 „Разпределението е нормално“ и алтернативна хипотеза „Разпределението не е нормално“ с равнище на значимост p = 0,05. При p < 0.05 отхвърляме H_0 и приемаме, че разпределението не е нормално. Таблицата по-долу показва, че нито едно от разпределенията не е нормално.
- След провеждане на непараметричен тест за сравняване на независими извадки на Mann-Whitney-Wilcoxon върху извадките по-долу е получено, че има статистически значима разлика между всеки две от тях.

Фактор	p-value	Нормалност на разпределението
email_dataset_spam$word_freq_our	$< 2.2e^{-16}$	не
email_dataset_spam$word_freq_free	$< 2.2e^{-16}$	не
email_dataset_spam$word_freq_you	$< 2.2e^{-16}$	не
email_dataset_spam$char_freq_.21	$< 2.2e^{-16}$	не
email_dataset_nonspam$capital_run_length_average	$< 2.2e^{-16}$	не
email_dataset_nonspam$word_freq_our	$< 2.2e^{-16}$	не
email_dataset_nonspam$word_freq_free	$< 2.2e^{-16}$	не
email_dataset_nonspam$word_freq_you	$< 2.2e^{-16}$	не
email_dataset_nonspam$char_freq_.21	$< 2.2e^{-16}$	не

5. Анализ на взаимодействието между категорийна и числова променлива:

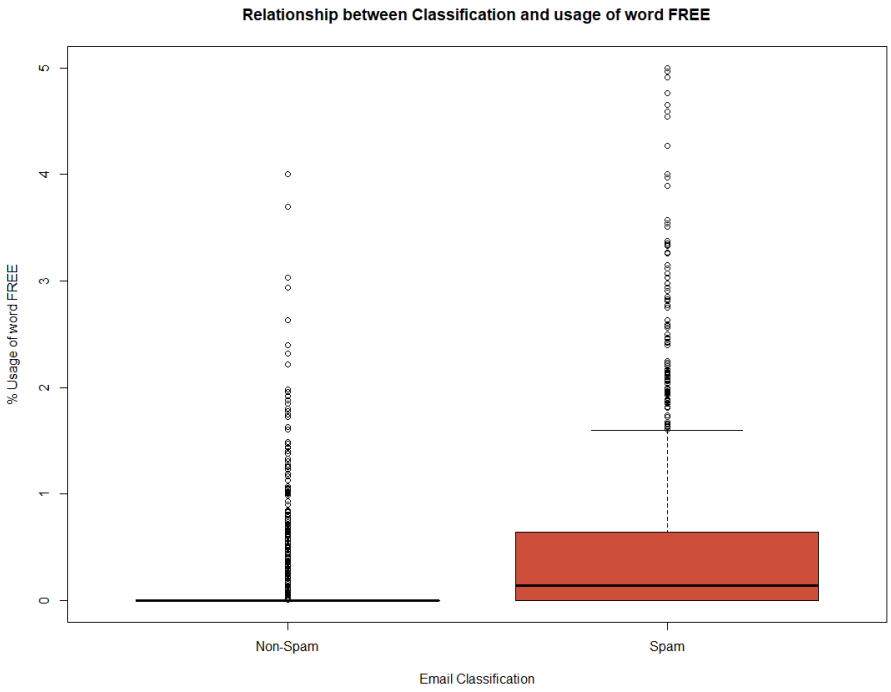
а) <u>категорийна vs числова</u>

- **Boxplot**



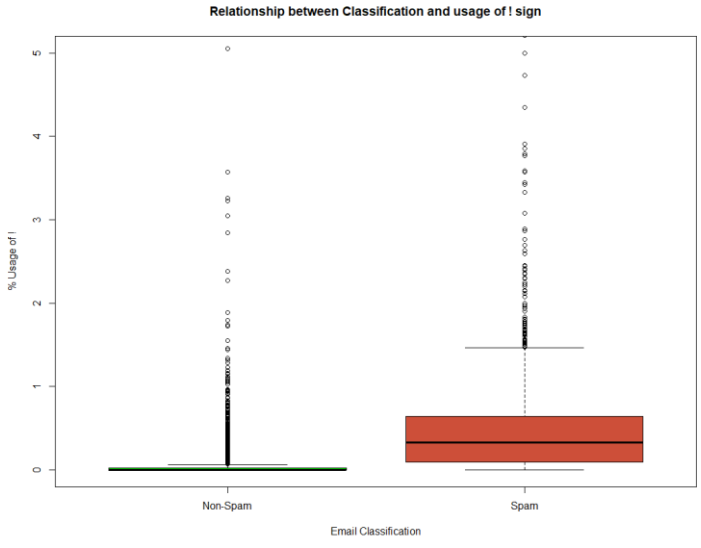
➢ Забелязва се, че има струпване в долния край на разпределението, което сочи, че при non-spam мейлите има доста по-малко срещания на дълги непрекъснати поредици от главни букви. Очевидно размахът при spam е по-обширен, като максималната стойност при non-spam-а попада в

размаха на spam, докато минималните им стойности съвпадат. Може да се заключи, че стойностите на тази величина са по-консистентни при мейлите, категоризирани като non-spam.



- При честотата на употреба на думата FREE отново се наблюдава неравномерно разпределение с още по-явно струпване около нулата при non-spam имейлите, до степен, че дори няма ясно разграничение между миинимална, максимална стойност, първи и трети квартил и медиана, което сочи за още по-силно изразена консистентност. При spam имейлите също се забелязва струпване на първите 50% от стойностите около нулата, но употребата на думата FREE при тях е доста по-застъпена, като има значително по-високи стойности на 3-тия квартил и максималната стойност. Интересно е, че и при двете категоризации има доста на брой изключения (outliers), което до известна степен би могло да попречи на machine learning модели за категоризация на типа имейл.

- Подобни са заключенията и при честотата на употреба на удивителен знак в имейлите.



- **Логистична регресия**

За целта на изследването на логистичната регресия данните се разделят на тестови и тренировъчни сетове от данни, като тренировъчните данни се използват да тренират модела. Подават се тестовите данни на модела и на база на резултата се заключава точността, прецизността и др. на модела.

Ще използвам библиотеката caTools.

Съотношение на разделението: 80% ще са тренировъчни данни, 20% - тестови.

Зависимата променилва ще е class, независимите - capital_run_length_average, word_freq_free, word_freq_our , word_freq_you, char_freq_.21 (т.е. class ще е функция на тези величини). Използва се функцията glm() - general model function, като за параметри data задавам тренировъчите данни, за family - binomial, което индикира, че е логистична регресия. Изпълняват се данните в модела и остава да се провери дали предсказаните отговори са верни. Моделът се валидира с матрица на грешката (confusion table).

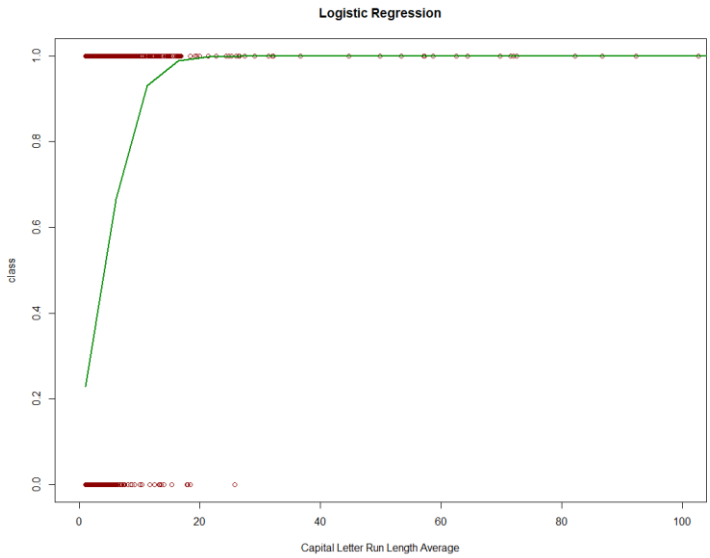
Един примерен резултат е следната табличка:

	Predicted value	
Actual value	FALSE	TRUE
0	1714	145
1	398	810

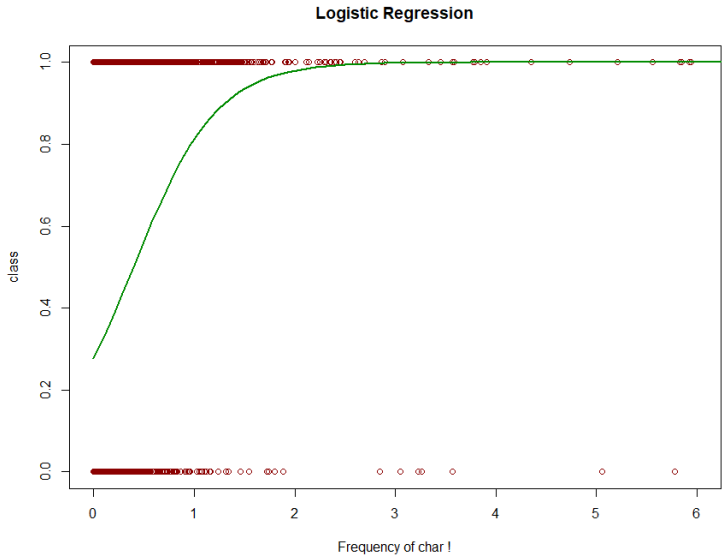
В нея се вижда, че 1 714 от очакваните да не бъдат spam наистина не са, а 810 от очакваните да бъдат spam наистина са spam. Забелявзва се, обаче, че има 398 мейла, за който моделът е предположил, че няма да са spam, а те всъщност са, както и 145 spam мейла, които моделът е сметнал за non-spam.

Точността на модела в този пример е 82,3%, но тя може да варира от около 81,6 % до 82,7 % от което можем да заключим, че има изразена закономерност между изследваните фактори и класифицикацията като spam и non-spam, но все още има място за грешка в рамките на около 20%, което сочи за това, че дори с този модел, съдържащ 5 сметнати за най-релевантни фактора, не могат да бъдат достигнати 100%-ви резултати за класифициране на мейли като spam.

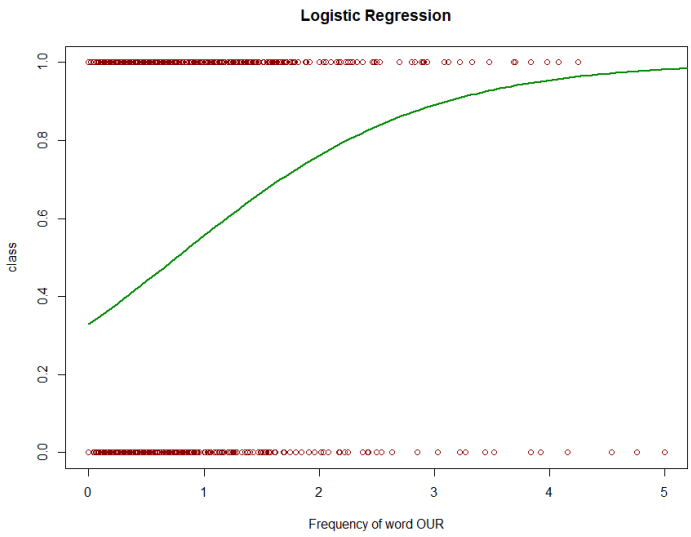
Ще онагледим как всяка от отделните величини сама по себе си влияае върху категоризацията на мейл като spam, като проведем отново експеримента по-горе, но за всяка от величините по отделно, като ще подкрепим резултатите с графики.



➢ При стойност от 0 има почти сигурно предположение за случай на non-spam, при 1 – spam.
С покачването на средната употреба на поредици от главни букви значително се учвеличава вероятността такъв мейл да е spam.
При честота от 20 и нагоре се забелязва, че вероятността мейлът да бъде категоризирам като spam става почти 1.



➢ Отново се наблюдава същото явление – при повече от 2% срещане на символа „!“ вероятността мейл да е spam расте дори още по-бързо.



➢ При честотата на срещане на думата “our” също може да се отбележи, че вероятността расте с покачването ѝ, но не толкова бързо.
Все пак можем да се заключи, че величините, които бяха избрани, са съществени и оказват влияние за класифицирането на даден мейл като spam.

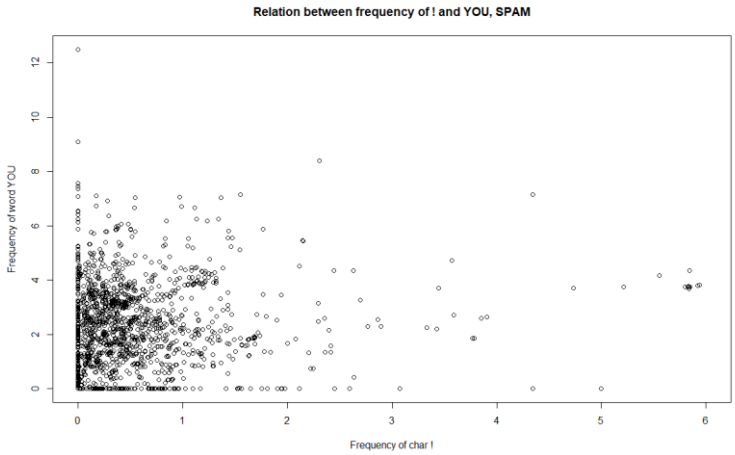
б) числова vs числова

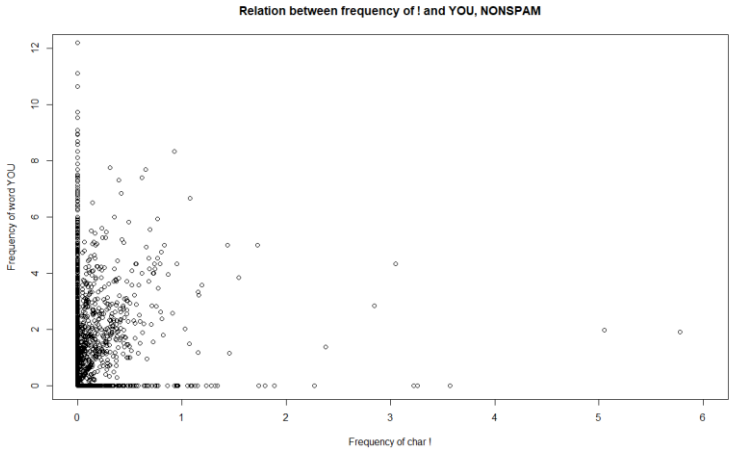
- **Dotplot и корелационен анализ**

  Направен е корелационен анализ чрез функцията cor() на всеки две от разглежданите велични. Най-значителните резултати са в следните случаи:

```
> cor(email_dataset_spam$char_freq_.21, email_dataset_spam$word_freq_you) #
[1] 0.101854
> cor(email_dataset_spam$word_freq_our, email_dataset_spam$word_freq_you) #
[1] 0.1282942
```

Корелацията между тези две велични е много слаба (под 15%). В останалите случаи процентът на корелация не надвишава 10%. Следователно няма как да се говори за линейна регресия.





И на двете графики се забелязва струпване на най-много стойности около нулата, като при non-spam това е дори още по-ясно изразено. Можем да заключим, че не съществува линейна зависимост между тези две велични. Подобно заключение може да се изведе и от графиките на останалите велични. Това отново доказва, че за целите на този анализ не е подходящо изследването за линейна регресия.

6. Заключение

Създадените с помощта на Rstudio графики и проведените тестове свидетелстват за това, че проучените данни за класификация на мейл като spam имат явно неравномерно разпределение. Проведените анализи върху отделните велични онагледяват, че между

отделните числови непрекъснати величини няма статистически значима разлика и корелация. По тази причина достигнахме до извода, че изследване за линейна регресия не би било подходящо за анализирането на дадените данни. Резултатите от изследването за логистична регресия върху модела, пресмятащ вероятността мейл да спада към категория spam или non-spam според стойностите на 5те избрани величини, сочат за това, че моделът работи сравнително добре с точност около 82%. Моделите върху всяка величина по отделно, обаче, дават значително по-лоши резултати по отношение на точност на предвиждането. Оттук може да се стигне го заключението, че ефективността на алгоритъм за класифициране на мейли като spam е много по-висока, когато той работи на базата на множество от релевантни фактори, а не отделни величини. Все пак, в алгоритъма, чрез който са получени изследваните данни, има място за подобрение, тъй като точността му все още е далеч от 100%.