**Практическое задание 1**

1. Рассчитайте основные статистики (меры центра и меры разброса) по распределениям всех переменных, имеющихся в файле данных.

2. Выберите наиболее интересный для вас количественный признак и охарактеризуйте его распределение при помощи соответствующих описательных статистик и графиков:

2.1. Какова форма распределения признака?

2.2. Можно ли говорить о том, что распределение признака согласуется с каким-либо теоретическим законом распределения?

2.3. Есть ли в данных «аномалии» / «выбросы»?

2.4. Какие меры центра и вариативности подходят для описания распределений лучше всего? Почему?

При выполнении этой части задания помните, что применимые меры центра и вариативности, а также графики, которые хорошо иллюстрируют распределения, зависят от типа шкалы измерения признака, а также от особенностей / формы распределения.

3. Сравните геозависимые и геонезависимые сферы (экспертная разметка, переменная IsGeo) по выбранному для анализа признаку. Есть ли отличия? В чем они состоят? (для ответа на вопрос используйте статистические и графические инструменты).

При выполнении этой части задания, учтите, пожалуйста, что важно представить все три компонента сравнения: цифры (соответствующие меры центра и разброса), график, иллюстрирующий различия, а также содержательный комментарий: как именно различаются сферы по выбранному признаку.

**Ответы на вопросы**

**Задание 1**

Основные статистики всех переменных из файла

[1] "AddressCount"

Min. 1st Qu. Median Mean 3rd Qu. Max.

9 81 371 1048 1195 9552

SDO: 1642.066

MAD: 471.4668

Q-Q: 1114

Razmah: 9543

-------------------------------------------------------------

[1] "CallsCount"

Min. 1st Qu. Median Mean 3rd Qu. Max.

20 346 931 3649 2458 48497

SDO: 8124.105

MAD: 1153.463

Q-Q: 2111.5

Razmah: 48477

-------------------------------------------------------------

[1] "ClicksCount"

Min. 1st Qu. Median Mean 3rd Qu. Max.

258 2055 6921 21826 30626 167155

SDO: 32474.96

MAD: 8309.973

Q-Q: 28570.5

Razmah: 166897

-------------------------------------------------------------

[1] "FirmsCount"

Min. 1st Qu. Median Mean 3rd Qu. Max.

14.0 71.5 185.0 305.1 402.5 2379.0

SDO: 382.0521

MAD: 203.1162

Q-Q: 331

Razmah: 2365

-------------------------------------------------------------

[1] "GeoPart"

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.09292 0.28153 0.32234 0.34264 0.41691 0.55618

SDO: 0.1036076

MAD: 0.1104339

Q-Q: 0.1353798

Razmah: 0.4632588

-------------------------------------------------------------

[1] "MobilePart"

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.0900 0.3573 0.4637 0.4457 0.5517 0.7373

SDO: 0.1461306

MAD: 0.1463676

Q-Q: 0.1943614

Razmah: 0.6472883

-------------------------------------------------------------

[1] "UsersCount"

Min. 1st Qu. Median Mean 3rd Qu. Max.

157 1168 2934 9753 13265 61127

SDO: 13927.3

MAD: 3522.658

Q-Q: 12097.5

Razmah: 60970

-------------------------------------------------------------

[1] "Distance"

Min. 1st Qu. Median Mean 3rd Qu. Max.

714.8 1562.1 2586.5 2669.4 3575.7 6292.2

SDO: 1427.785

MAD: 1540.14

Q-Q: 2013.589

Razmah: 5577.42

-------------------------------------------------------------

[1] "IsGeo"

0 1

51 28

Moda – 0 (геонезависимая сфера)

**Задание 2**

Для более детального анализа графическими методами я выбрал переменную – UsersCount: количество уникальных пользователей, которые совершали какие–то действия в данной сфере – открывали карточки организаций, тыкали во вход/адрес звонок и т.д.

Тип данных у переменной UsersCount – интервальный

Статистики

"UsersCount"

Min. 1st Qu. Median Mean 3rd Qu. Max.

157 1168 2934 9753 13265 61127

SDO: 13927.3

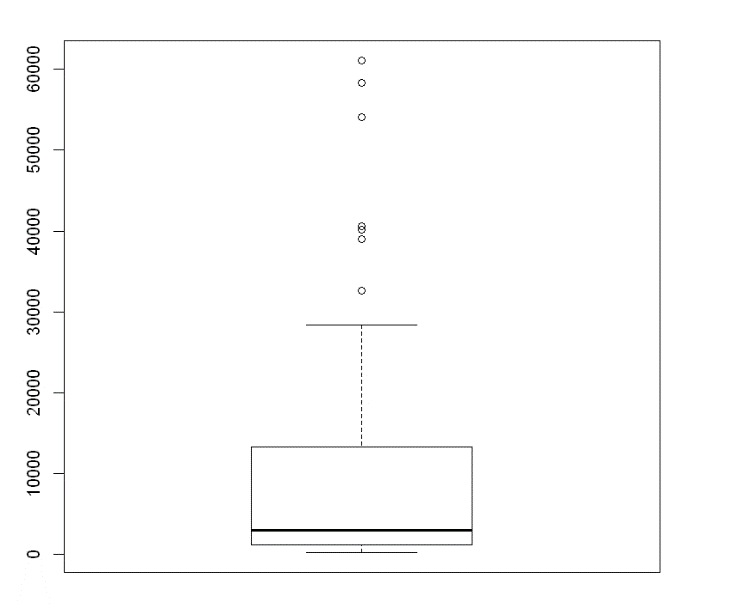
MAD: 3522.658

Q-Q: 12097.5

Razmah: 60970

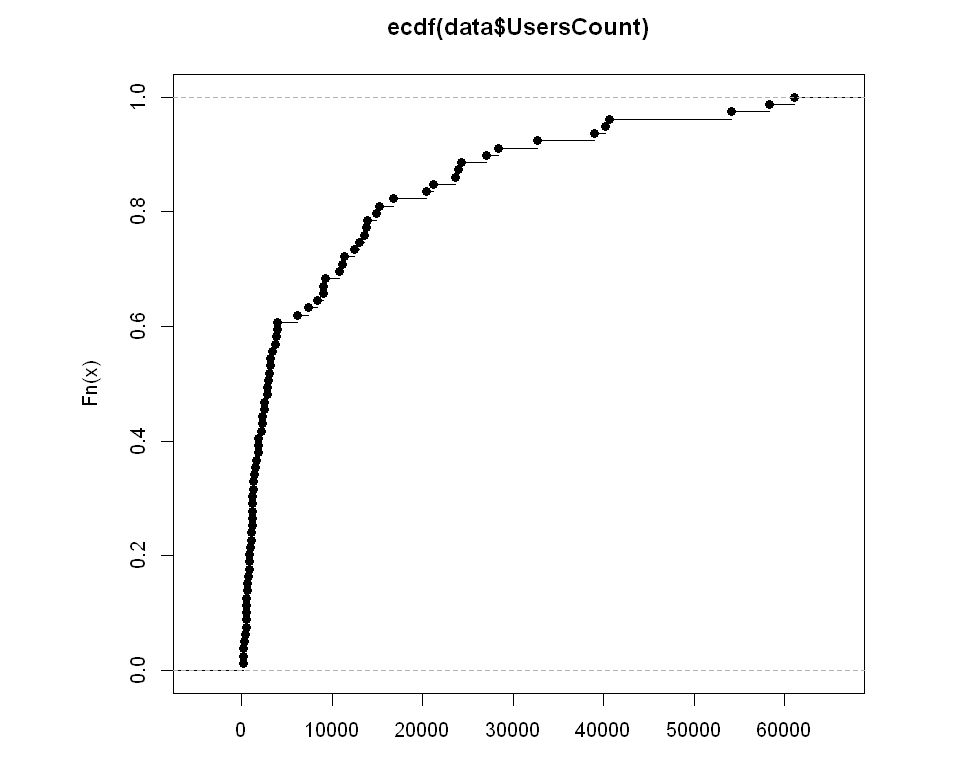
Исходя из посчитанных ранее мер центра и вариативности а также box plot, предварительно похоже, что в данных присутствуют выбросы, т.к. среднеквадратичное отклонение и среднее арифметическое достаточно сильно отличаются от медианы и MAD. Посмотрим подробнее на графики и подумаем с каким законом распределения оно согласуется

**Построим ящик с усами**

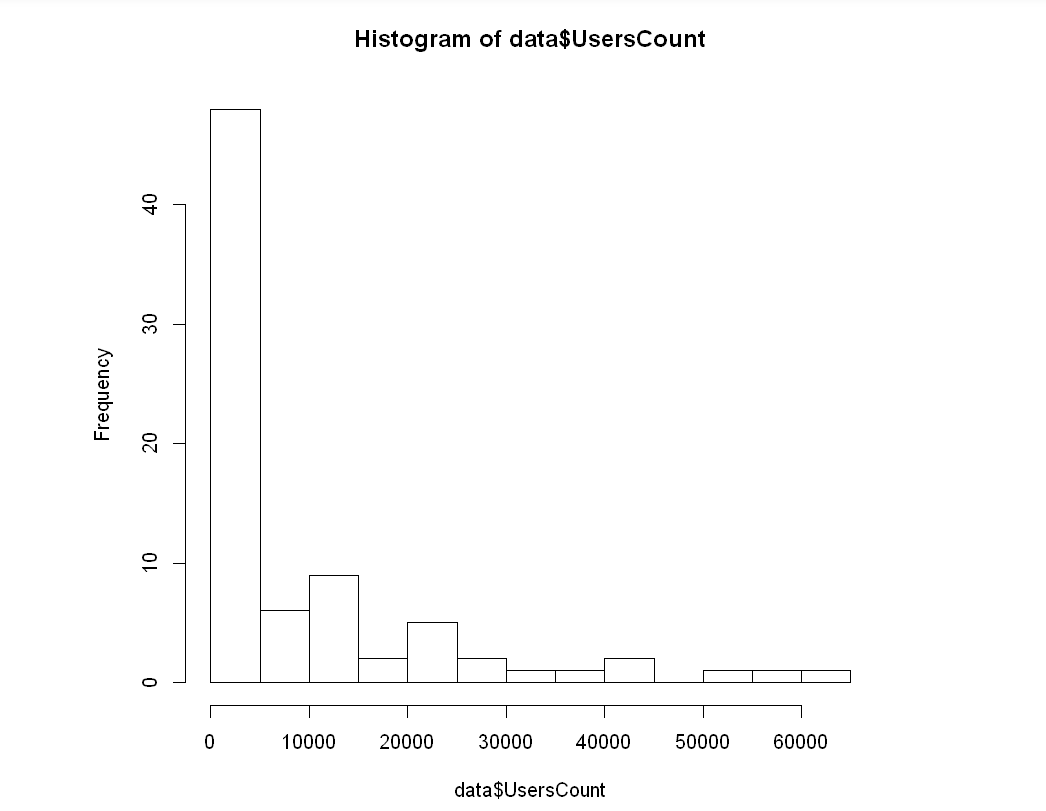


Из графика видно, что есть заметная асимметрия влево, а также выбросы справа. Что может указывать на логнормальный характер распределения в целом

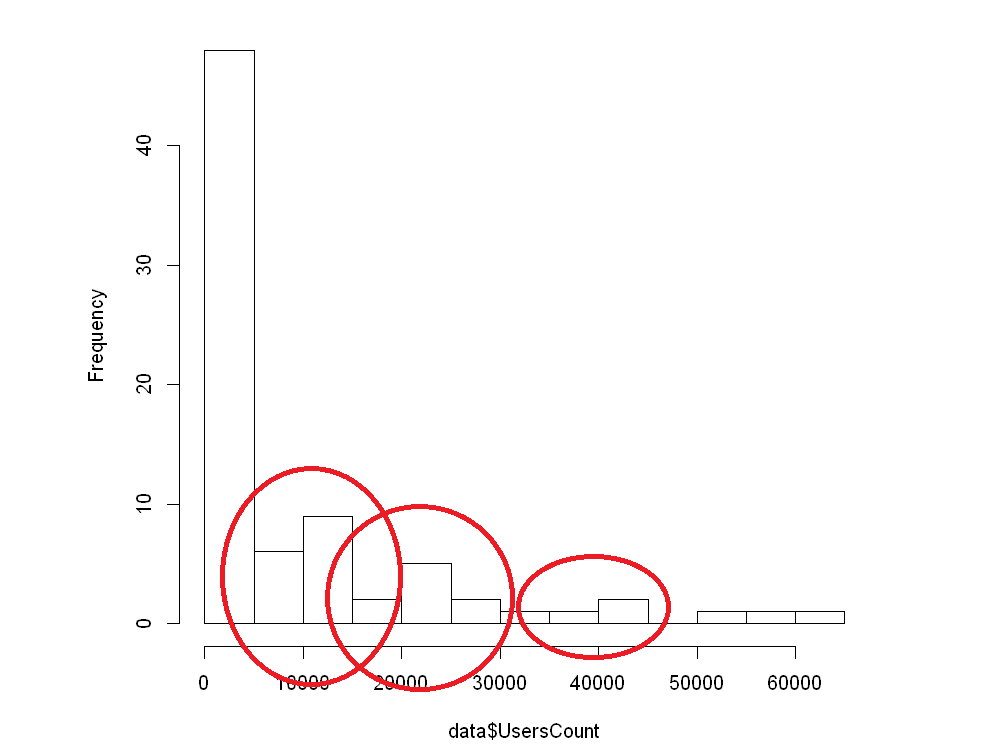
**Эмпирическая функция**



**Построим гистограмму методом Фридмана-Диакониса**



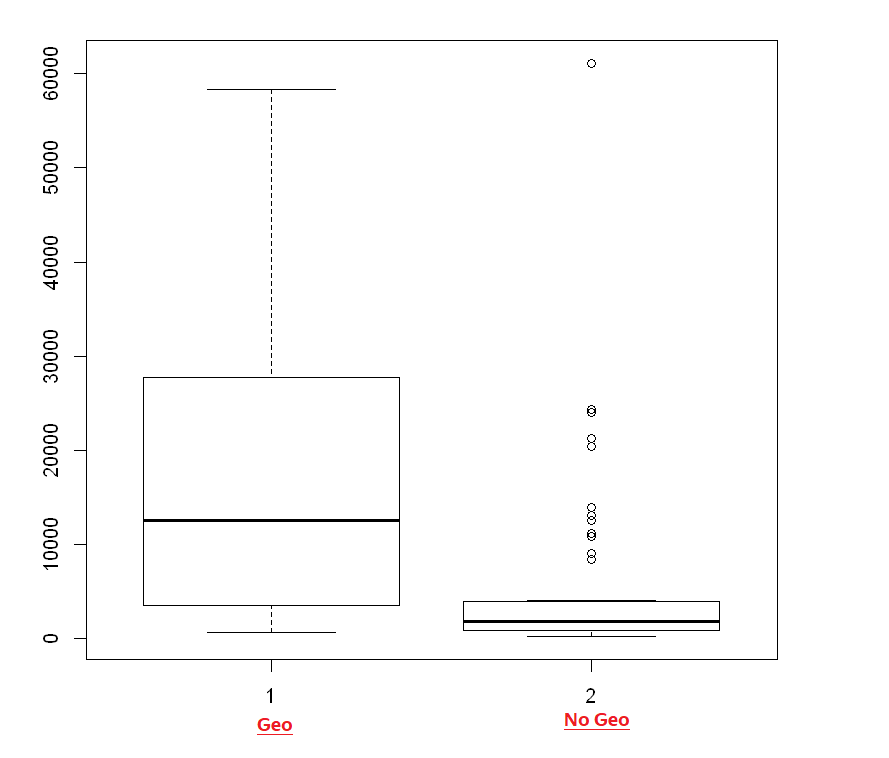
Исходя из гистограммы в целом распределение похоже на логнормальное (или даже экспоненциальное), при этом бросаются в глаза скачкообразные значения на отрезке от 10к до 40к, **похоже что может иметь место смесь распределений**



**Задача 3**

Сравним геозависимые и геонезависимые сферы

**Построим box plot для двух распределений**



**Меры центры и вариативности**

**GEO**

[1] "UsersCount"

Min. 1st Qu. Median Mean 3rd Qu. Max.

596 3687 12459 17197 27386 58351

SDO: 16718.43

MAD: 14559.13

Q-Q: 23699

Razmah: 57755

**NO GEO**

[1] "UsersCount"

Min. 1st Qu. Median Mean 3rd Qu. Max.

157.0 861.5 1821.0 5666.5 3855.5 61127.0

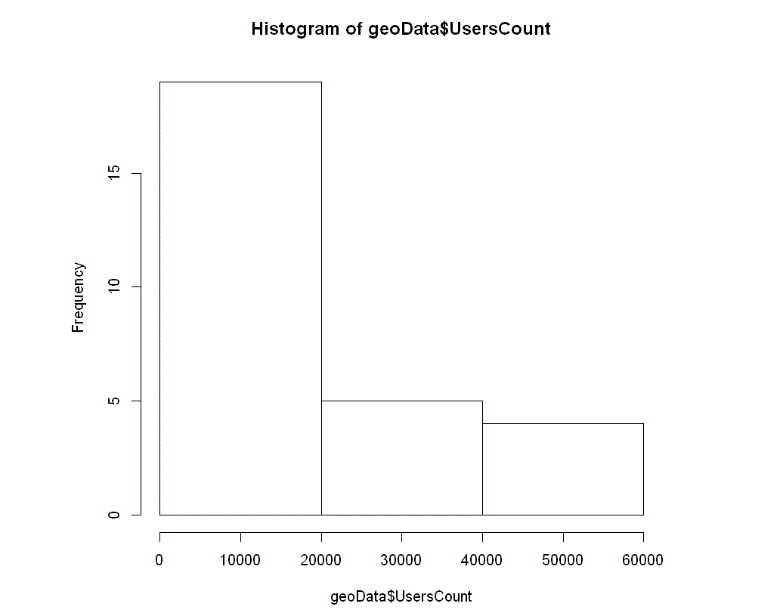
SDO: 10178.26

MAD: 1872.524

Q-Q: 2994

Razmah: 60970

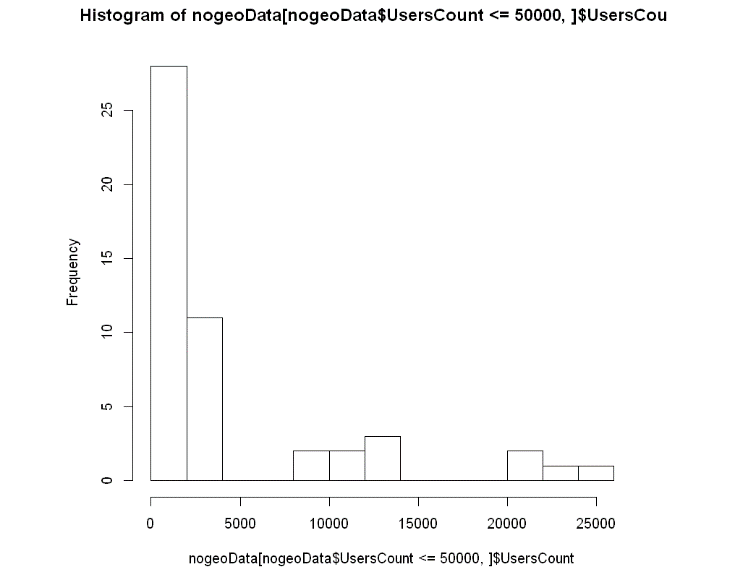
**Гистограмма геозависимых сфер**



Исходя из данных можно заметить, что:

* Геозависимые сферы по своему виду достаточно сильно согласовываются с логнормальным видом распределений, по графику с усами видно, что распределение смещено влево и не имеет выбросов. Из-за небольшого кол-ва данных гистограмма получилось не очень информативной, при этом все же тоже напоминает логнормальное распределение
* Негеозависимое распределение выглядит достаточно необычно, имеет очевидный аномальный выброс на графике с усами, посмотрим детальнее убрав выброс и применив другие методы графического анализа

**Гистограмма Фридмана-Диакониса для геонезависимой сферы**



Кажется, что с помощью этой гистограммы подтверждается гипотеза о смеси распределений разных случайных величин, и следовательно говорит о том,

А) что смеси присутсвуют в геонезависимых сферах,

Б) смешались целых три распределения.

**Меры центра и вариативности для описания полученных распределений:**

А) На мой взгляд для геозависимых сфер можно использовать все меры центра и вариативности, т.к. распределение близко к логнормальному (т.е. логарифм от нормального)

Б) В геонезависимых сферах присутствуют смеси, поэтому нужно в первую очередь их разделить и рассматривать отдельно (в данный момент при разделении текущего распределения на 3 у двух у смесей справа становится очень мало данных, поэтому на мой взгляд, для понимания характера их распределений и составлении каких либо выводов кроме всего прочего также следует их обогатить, т.е. – получить какую-то большую выгрузку для анализа). Что касается левой части похоже, что она не отличается от геозависимых сфер, т.е. распределена логнормально и для ее анализа в цифрах можно использовать все меры центра и вариативности