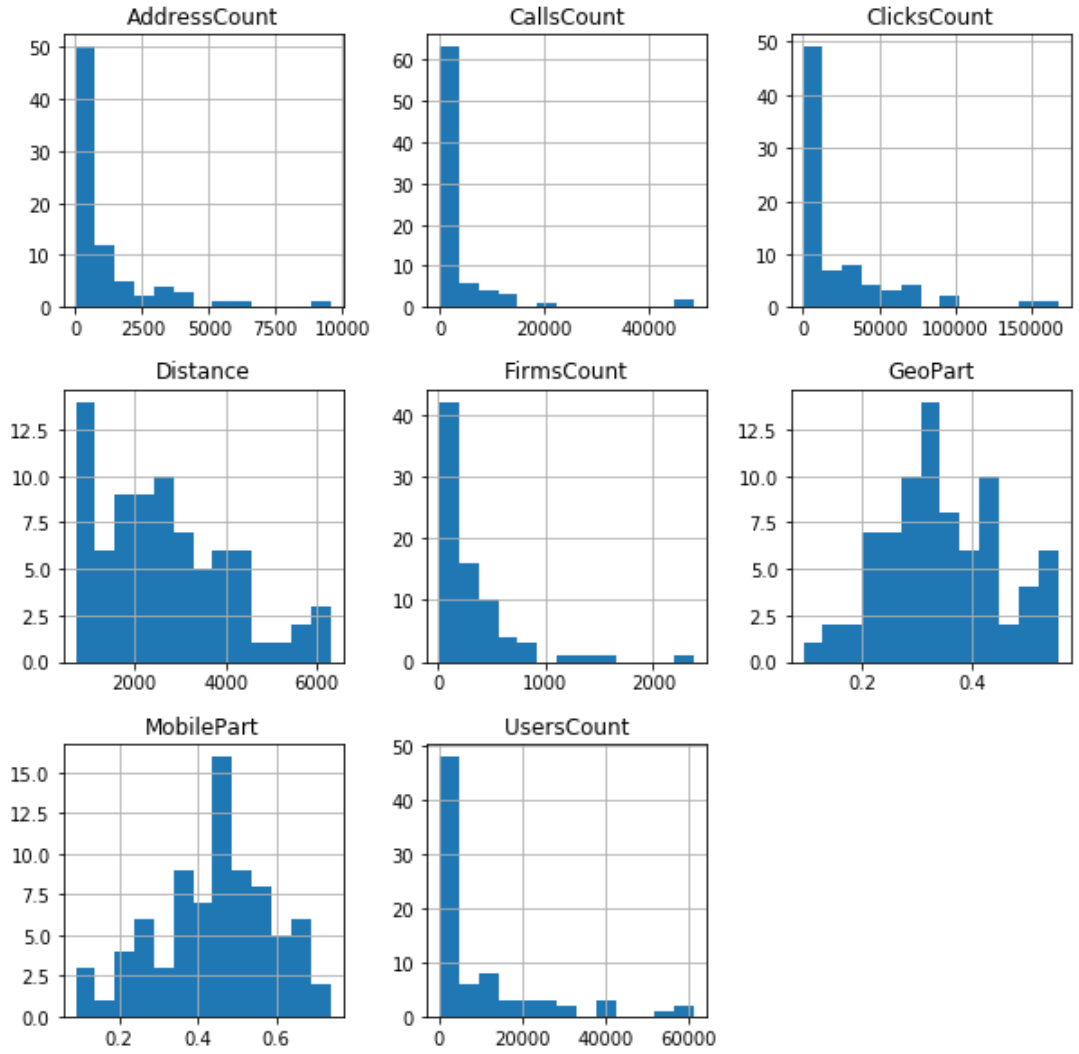
Seilkhan Abilmansur

Анализ данных

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

Гистограмма всех данных:



Арифметическая середина:

```
AddressCount      1048.037975
CallsCount        3648.683544
ClicksCount      21826.012658
FirmsCount         305.088608
GeoPart              0.342645
MobilePart           0.445746
UsersCount        9753.126582
Distance          2669.426352
IsGeo                0.354430
dtype: float64
```

Медиана:

```
AddressCount      371.000000
CallsCount        931.000000
ClicksCount      6921.000000
FirmsCount        185.000000
GeoPart             0.322342
MobilePart          0.463744
UsersCount       2934.000000
Distance         2586.503274
IsGeo               0.000000
dtype: float64
```

Дисперсия со всей совокупностью:

```
IsGeo= 0.22880948565934947
AddressCount= 2662249.7327351384
CallsCount= 65165631.76061528
ClicksCount= 1041273337.227688
FirmsCount= 144116.15670565615
GeoPart= 0.01059865558619776
MobilePart= 0.021083853669105496
UsersCount= 191514255.12321743
Distance= 2012766.1791736197
```

Дисперсия:

```
IsGeo= 0.2317429406037001
AddressCount= 2696381.139565076
CallsCount= 66001088.57805907
ClicksCount= 1054622995.3972737
FirmsCount= 145963.79974034405
GeoPart= 0.010734535786020809
MobilePart= 0.021354159485376077
UsersCount= 193969566.08633563
Distance= 2038570.8737784098
```

Мода:

	AddressCount	CallsCount	ClicksCount	FirmsCount	GeoPart	MobilePart	UsersCount	Distance	IsGeo
0	175.0	20	258	19.0	0.092917	0.090000	157	714.787236	0.0

Стандартное отклонение:

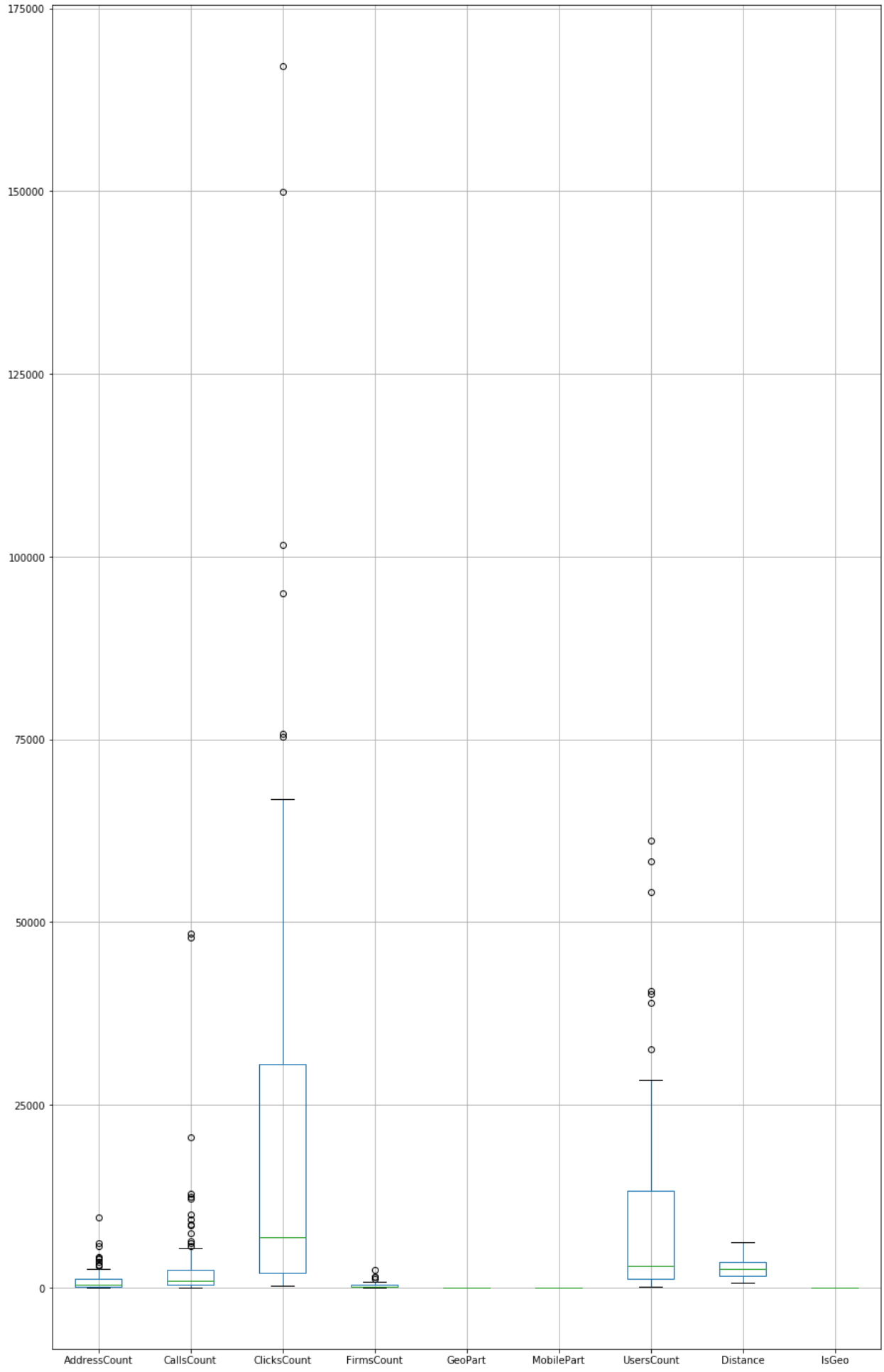
```
AddressCount     1642.066119
CallsCount       8124.105402
ClicksCount     32474.959513
FirmsCount        382.052090
GeoPart             0.103608
MobilePart          0.146131
UsersCount      13927.295721
Distance         1427.785304
IsGeo               0.481397
```

Общее описание датасета:

	AddressCount	CallsCount	ClicksCount	FirmsCount	GeoPart	MobilePart	UsersCount	Distance	IsGeo
count	79.000000	79.000000	79.000000	79.000000	79.000000	79.000000	79.000000	79.000000	79.000000
mean	1048.037975	3648.683544	21826.012658	305.088608	0.342645	0.445746	9753.126582	2669.426352	0.354430
std	1642.066119	8124.105402	32474.959513	382.052090	0.103608	0.146131	13927.295721	1427.785304	0.481397
min	9.000000	20.000000	258.000000	14.000000	0.092917	0.090000	157.000000	714.787236	0.000000
25%	81.000000	346.000000	2055.000000	71.500000	0.281527	0.357293	1167.500000	1562.103740	0.000000
50%	371.000000	931.000000	6921.000000	185.000000	0.322342	0.463744	2934.000000	2586.503274	0.000000
75%	1195.000000	2457.500000	30625.500000	402.500000	0.416907	0.551654	13265.000000	3575.692331	1.000000
max	9552.000000	48497.000000	167155.000000	2379.000000	0.556175	0.737288	61127.000000	6292.207311	1.000000

Count- количество ненулевых значений.
Mean- арифметическая середина
Std- стандартное отклонение
Min- минимальный элемент
25%- нижний процентиль
50%-медиана
75%- верхний процентиль
max- максимальный элемент

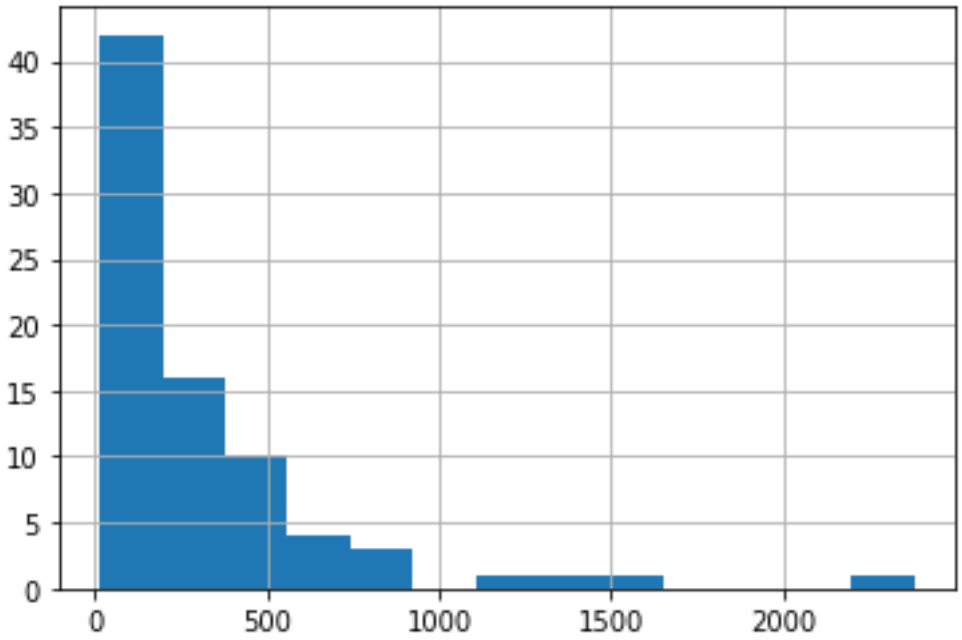
Боксплот всего датасета:


175000
150000
125000
100000
75000
50000
25000
0
AddressCount
CallsCount
ClicksCount
FirmsCount
GeoPart
MobilePart
UsersCount
Distance
IsGeo

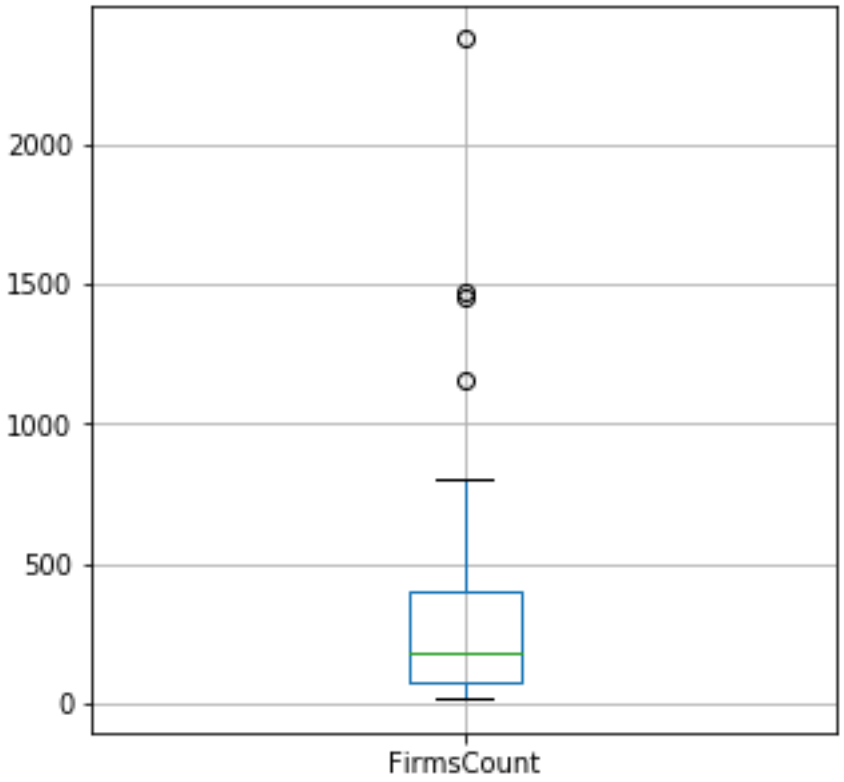
Как мы видим, в данных много выбросов.

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

Для описание было взято столбец FirmsCount.



Как видно на гистограмме, есть правосторонняя асимметрия, есть 2 выброса (аномалии). Форма распределения - геометрическое.



По boxplot отчетливо видны 4 выброса, по медиане можно сказать что у данных есть правосторонняя асимметрия.

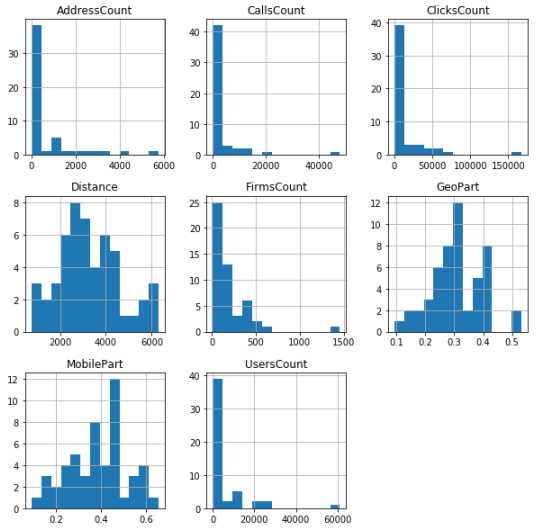
Основные меры центральной тенденции:

```
арифметическая середина= 305.0886075949367
дисперсия= 145963.79974034405
медиана= 185.0
```

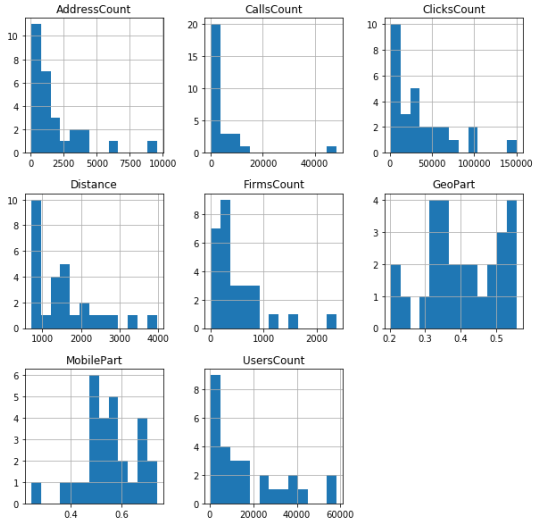
Наилучшим методом описание данного столбца, я считаю boxplot в качестве графической визуализацией, а для описание арифметическую середину, дисперсию и медиану, так как по ним отчетливо видны разбросы и другие характеристики.

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).

Геонезависимые сферы



Геозависимые сферы



Арифметическая середина геонезависимых(0) и геозависимых(1) данных

IsGeo	AddressCount	CallsCount	ClicksCount	FirmsCount	GeoPart	MobilePart	UsersCount	Distance
0	656.450980	3134.196078	14460.803922	199.058824	0.309498	0.387760	5666.470588	3263.639505
1	1761.285714	4585.785714	35241.214286	498.214286	0.403019	0.551364	17196.678571	1587.109538

Медиана геонезависимых(0) и геозависимых(1) данных

IsGeo	AddressCount	CallsCount	ClicksCount	FirmsCount	GeoPart	MobilePart	UsersCount	Distance
0	656.450980	3134.196078	14460.803922	199.058824	0.309498	0.387760	5666.470588	3263.639505
1	1761.285714	4585.785714	35241.214286	498.214286	0.403019	0.551364	17196.678571	1587.109538

Стандартное отклонение геонезависимых(0) и геозависимых(1) данных

IsGeo	AddressCount	CallsCount	ClicksCount	FirmsCount	GeoPart	MobilePart	UsersCount	Distance
0	1159.881258	7488.023659	27787.407922	241.356617	0.088584	0.131829	10178.260318	1341.851808
1	2116.836544	9242.558438	36437.636542	503.867881	0.103071	0.107644	16718.427766	820.415216

Как мы видим, через гистограммы и метрик данные сильно друг от друга отличаются. У геонезависимых данных выбросов относительно меньше и более виднеются законы распределения нежели у геозависимых данных. Виды распределения геонезависимых данных: геометрическое распределение, нормальное (смежное) распределение и гипергеометрическое распределение. По метрикам тоже видим сильное различие, особенно в арифметических серединах и в медианах.