Задание 2

Соболькова Е., Запорожец А., Агаджанян Е.

МГУ имени М. В. Ломоносова

Москва, 2017

Описание задания

После оглушительного успеха в освобождении Астапора, Миэрина и Юнкая от власти работорговцев Дейенерис Бурерожденная открыла себе доступ к Летнему морю, а следовательно — путь в Вестерос. Для ведения войны с Семью Королевствами нужно оружие, а для оружия нужна сталь. Нет никаких сомнений в кузнечном искусстве Безупречных, однако поставщики стали не столь надежны. Два основных поставщика стали — это Westeros Inc. и Harpy & Co.

Описание задания

На протяжении нескольких месяцев мы закупаем сталь у обеих компаний, и каждая из них предлагает ощутимую скидку при заключении эксклюзивного договора на поставку. Советник королевы Тирион Ланнистер знает о твоем умении принимать взвешенные рациональные решения и просит помощи в объективном решении вопроса о том, с какой из компаний следует заключить эксклюзивный договор на поставку стали. У Тириона есть записи о производстве мечей каждым из кузнецов-безупречных, а также данные о количестве сломанных мечей в каждый из месяцев ведения боевых действий.

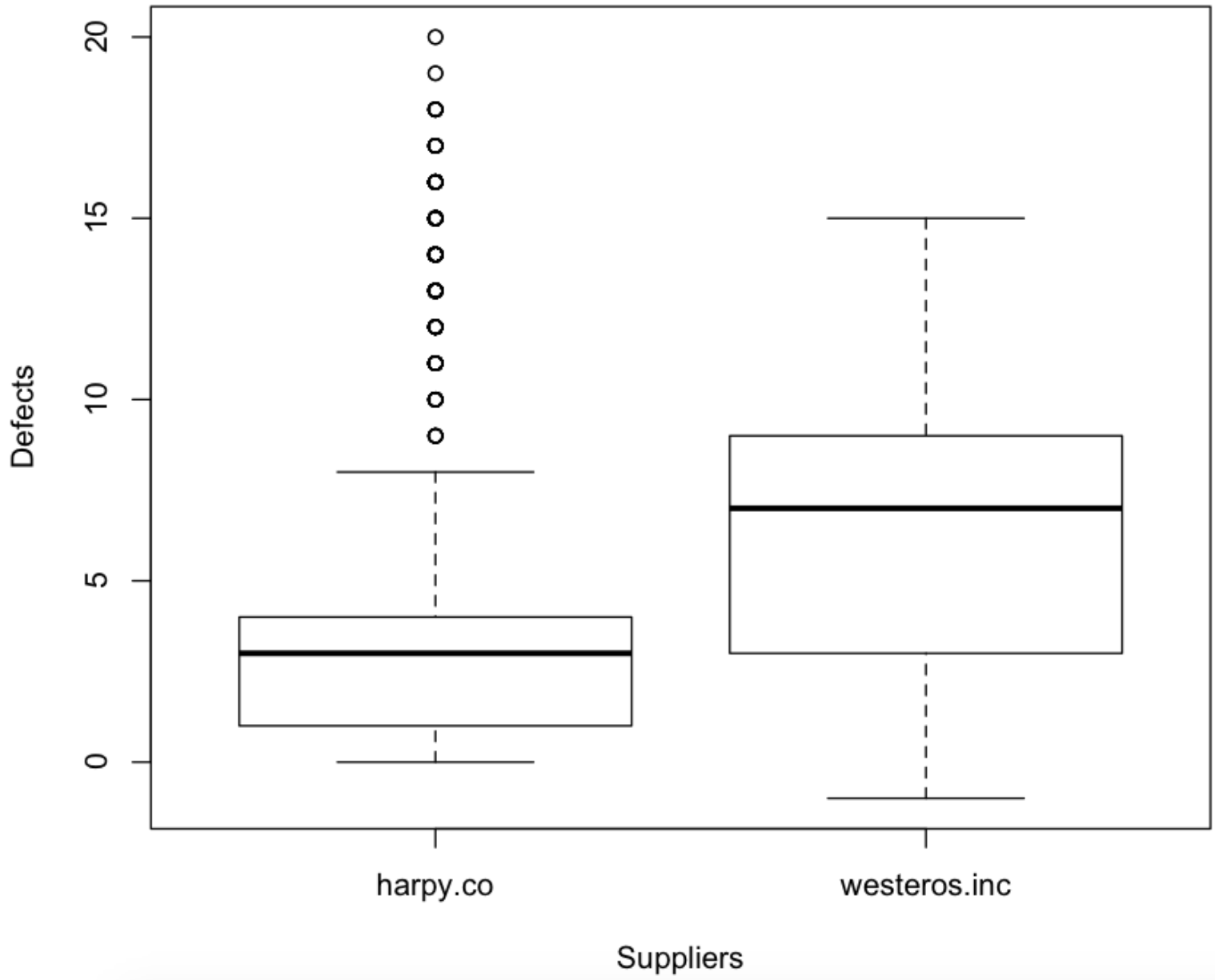
Выполнение задания

Необходимо провести разведывательный анализ данных с целью ответа на вопрос: "С каким из поставщиков стали следует заключить договор?"

BoxPlot

Проанализируем наши данные с помощью диаграммы размаха(ящик с усами), с помощью функции Boxplot(). Для этого сравним такие характеристики как Defects и Produced для двух компаний, которые занимаются производством оружия harpy.co и westeros.inc.

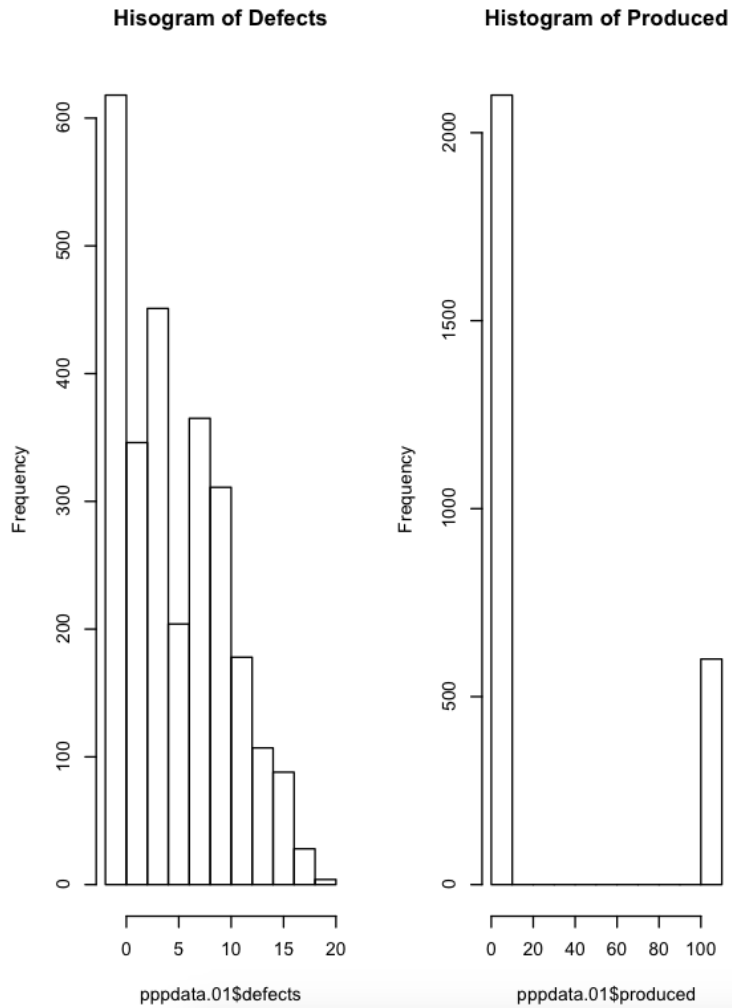
По данной диаграмме(представленной на следующем слайде) видно, что в основном количество дефектов у оружия, произведенного компанией harpy.co значительно меньше, чем westeros.inc.Так как медиана в первом случае ниже, чем медиана во втором случае. Однако, стоит заметить, что для компании westeros.inc нет характерных выбросов. А вот в harpy.co мы видим целых 12. Это говорит о том, что в какие-то единичные моменты может появляться большое количество дефектов.


Defects
0
5
10
15
20
harpy.co
westeros.inc
Suppliers

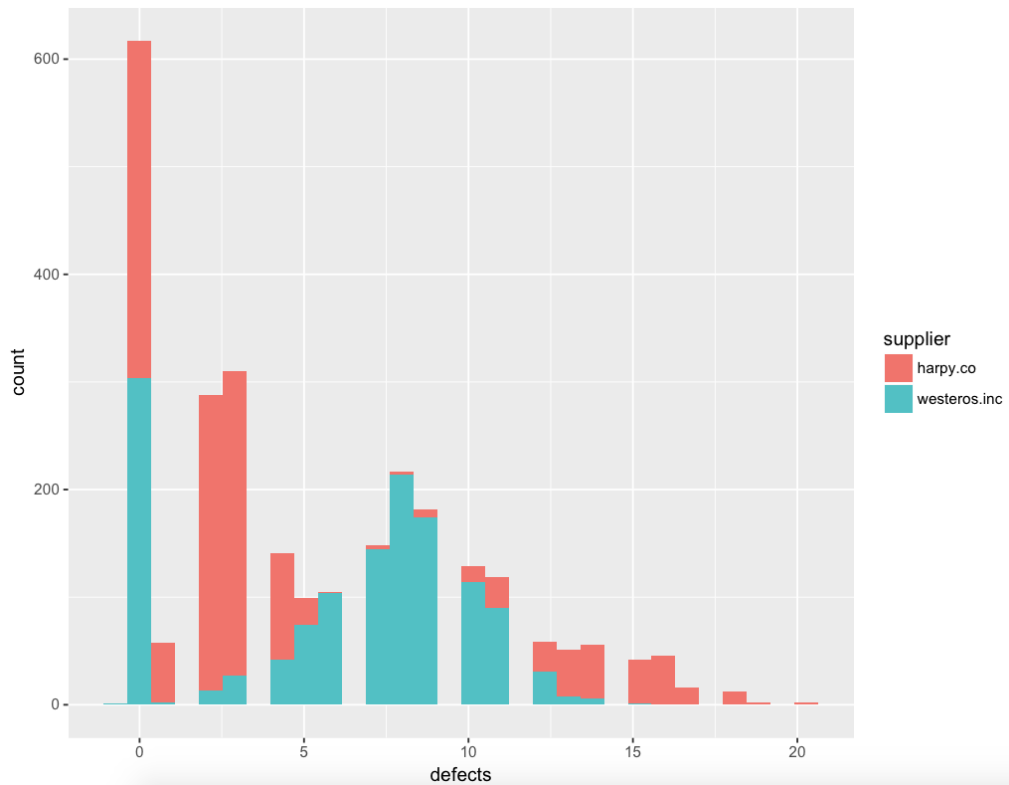
Следующая диаграмма нам показывает, что либо не производится ни одного оружия, либо они производятся уже в большем количестве единоразово. В данном случае от 100. Это верно для обеих компаний.

Histogram

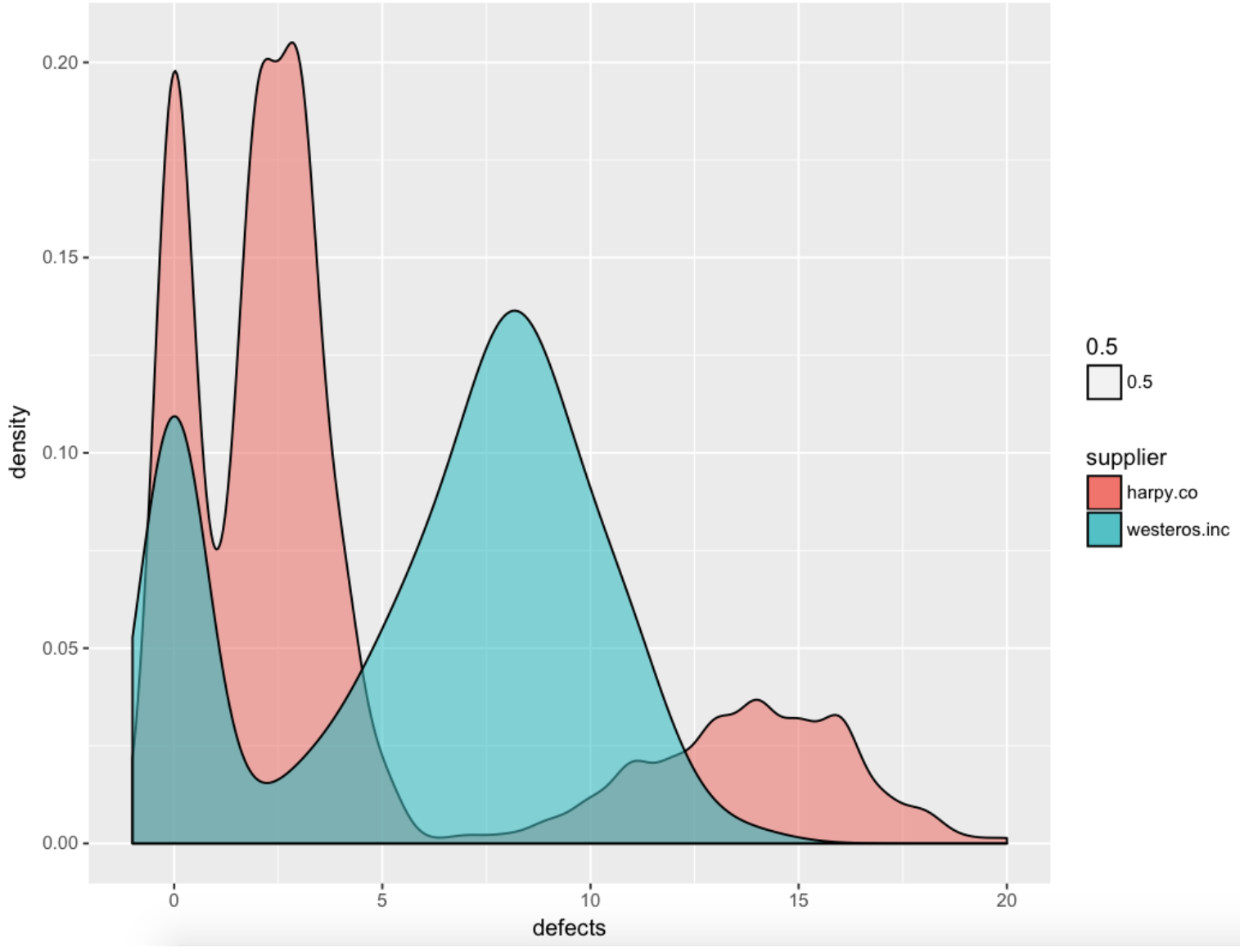
Проанализируем наши данные с помощью гистограммы. В этом нам поможет функция hist(). В результате выполнения этой команды, будут посчитаны частоты появления различных значений элементов передаваемого вектора.



Заметим, что эти данные были построены сразу для двух компаний, производящих оружие. Для того, чтобы провести более качественный анализ для каждой из компаний, воспользуемся функцией qplot().
Теперь мы видим, какое количество дефектов и сколько раз было у компании happy.co (персиковый цвет) и компании western.inc (мятный цвет).



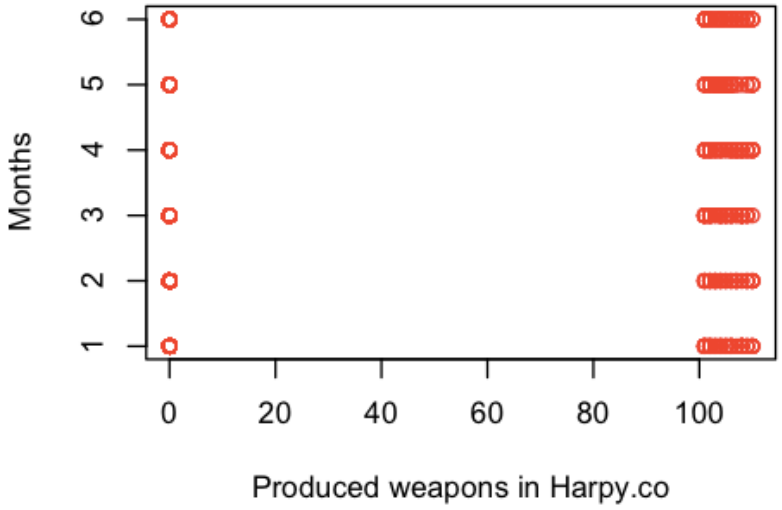
Для большей наглядности представим данные в виде

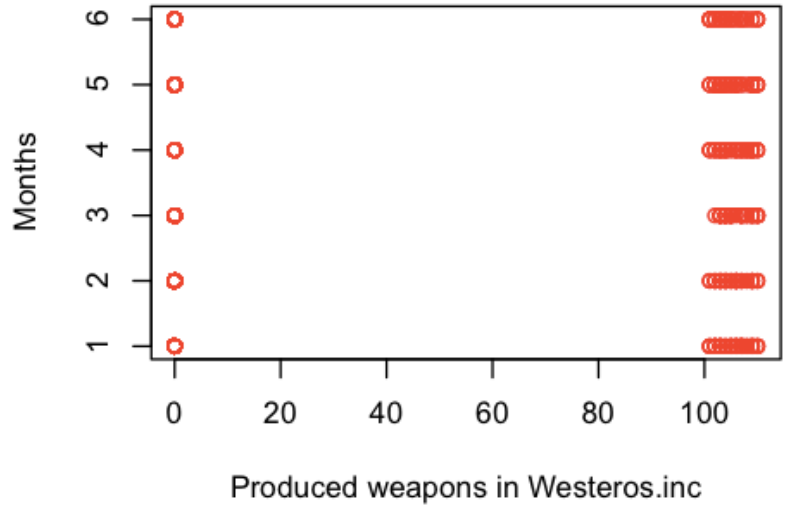


Основываясь на данном представлении мы можем сделать вывод, что в компании harpy.co очень много дефектов, которые принимают небольшие значения (примерно от 0 до 5), однако есть достаточно значимое количество дефектов с большими значениями(от 10 до 20), которыми нельзя пренебречь или сказать, что это случайные выбросы, так как частота их встречаемости ощутима. В компании westeros.inc основные значения дефектов колеблются от 4 до 12, что является весомым минусом.

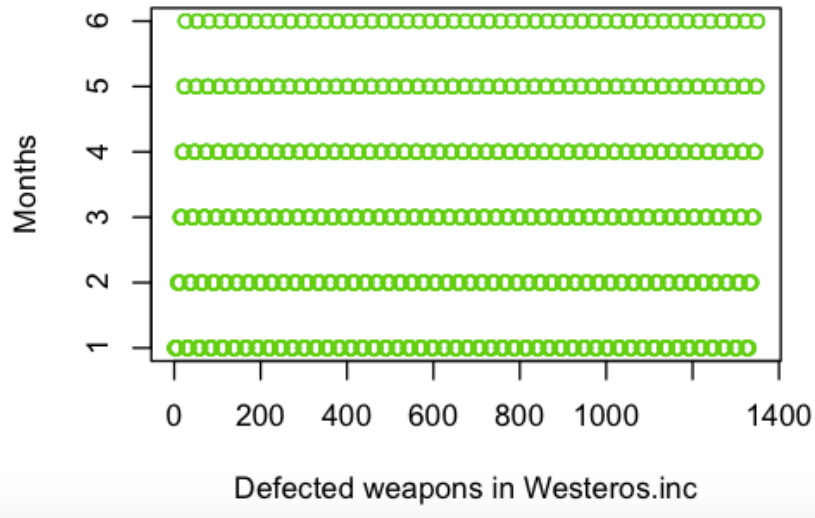
Scatter Plot

Посмотрим на график рассеивания для двух компаний harpy.co и Westeros.inc. Красным цветом изображено какое количество оружия в месяц было произведено в соответствующей компании, зеленым - количество сломанных мечей в данном месяце. Существенных различий между данными нету. Однако, видно, что в 3 месяце компания Westeros.inc произвела оружия меньше, чем в этом же месяце компания Harpy.co.


Months
1
2
3
4
5
6
0
20
40
60
80
100
Produced weapons in Harpy.co

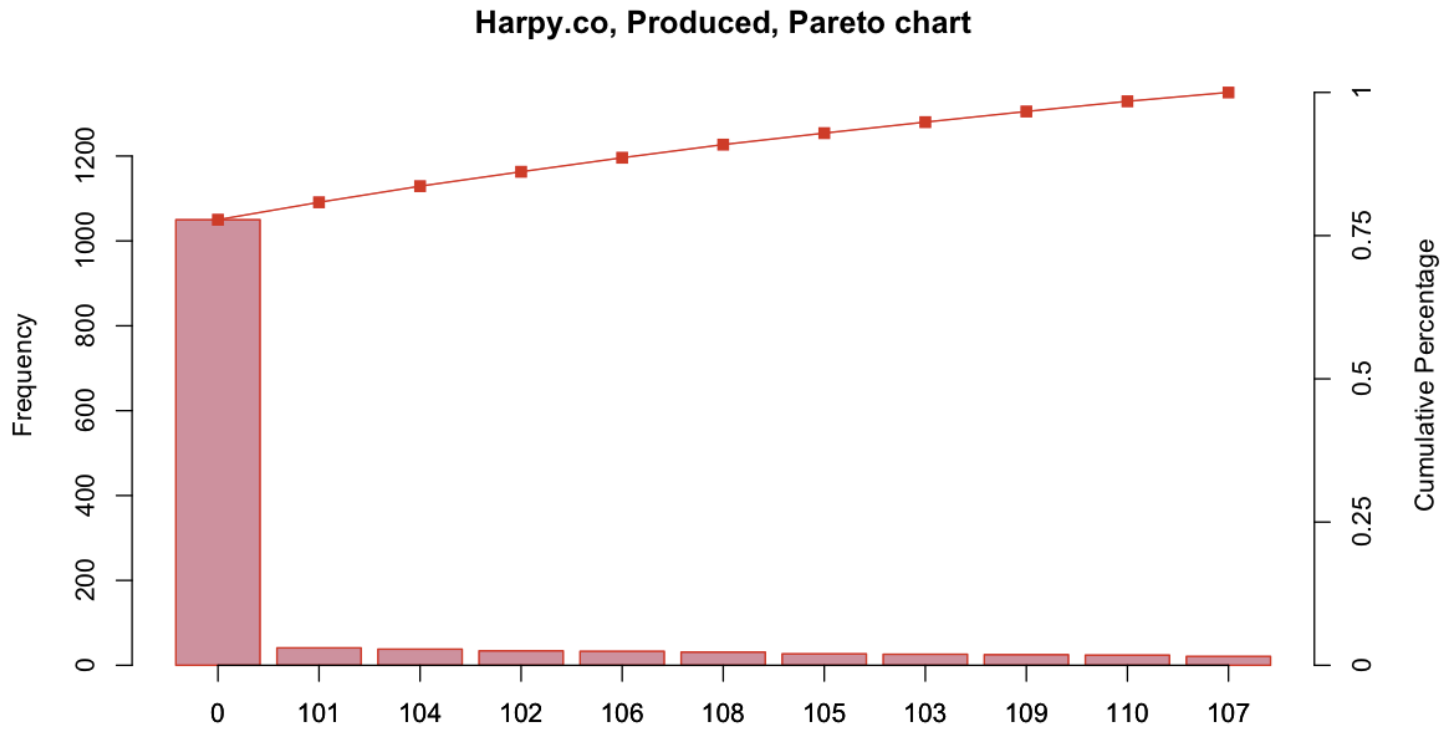

Months
1
2
3
4
5
6
0
20
40
60
80
100
Produced weapons in Westeros.inc


Months
1
2
3
4
5
6
0
200
400
600
800
1000
1400
Defected weapons in Harpy.co

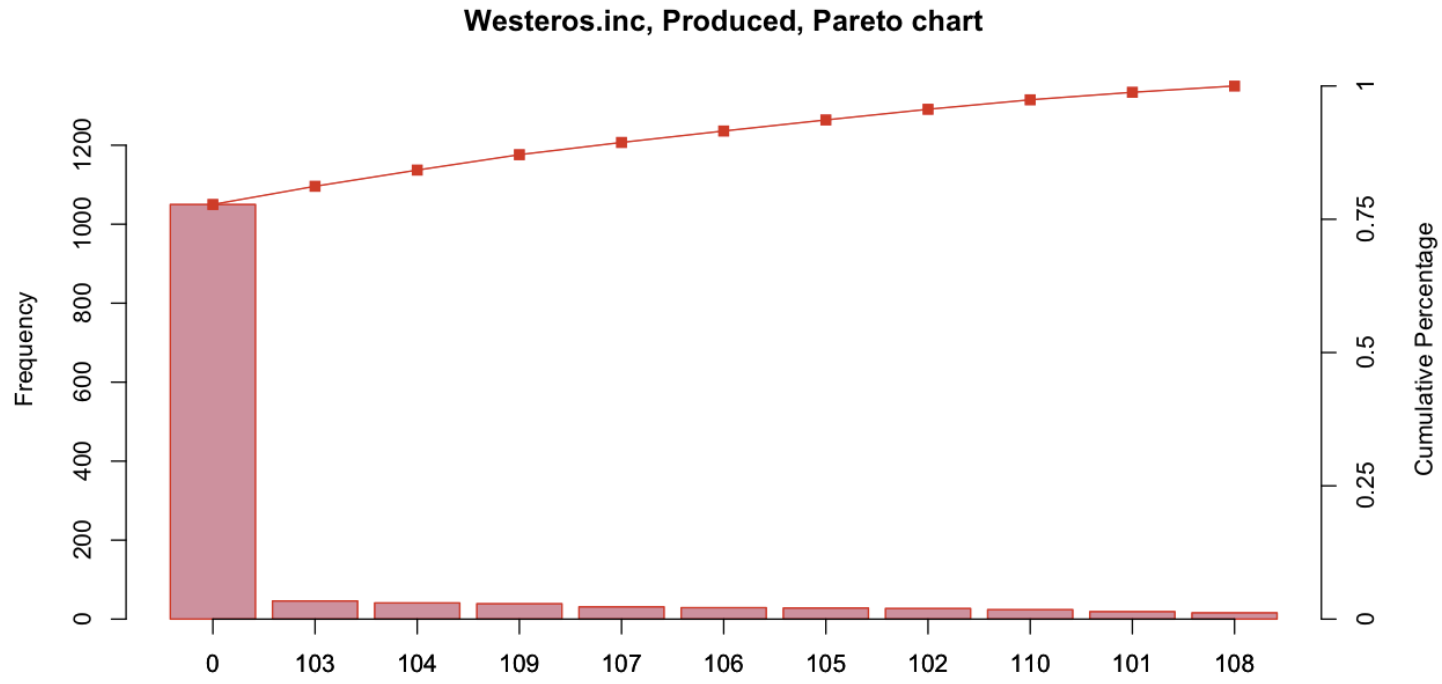

Months
1
2
3
4
5
6
0
200
400
600
800
1000
1400
Defected weapons in Westeros.inc

Pareto plot

Построим диаграмму Парето для двух компаний Westeros.inc и harpy.co



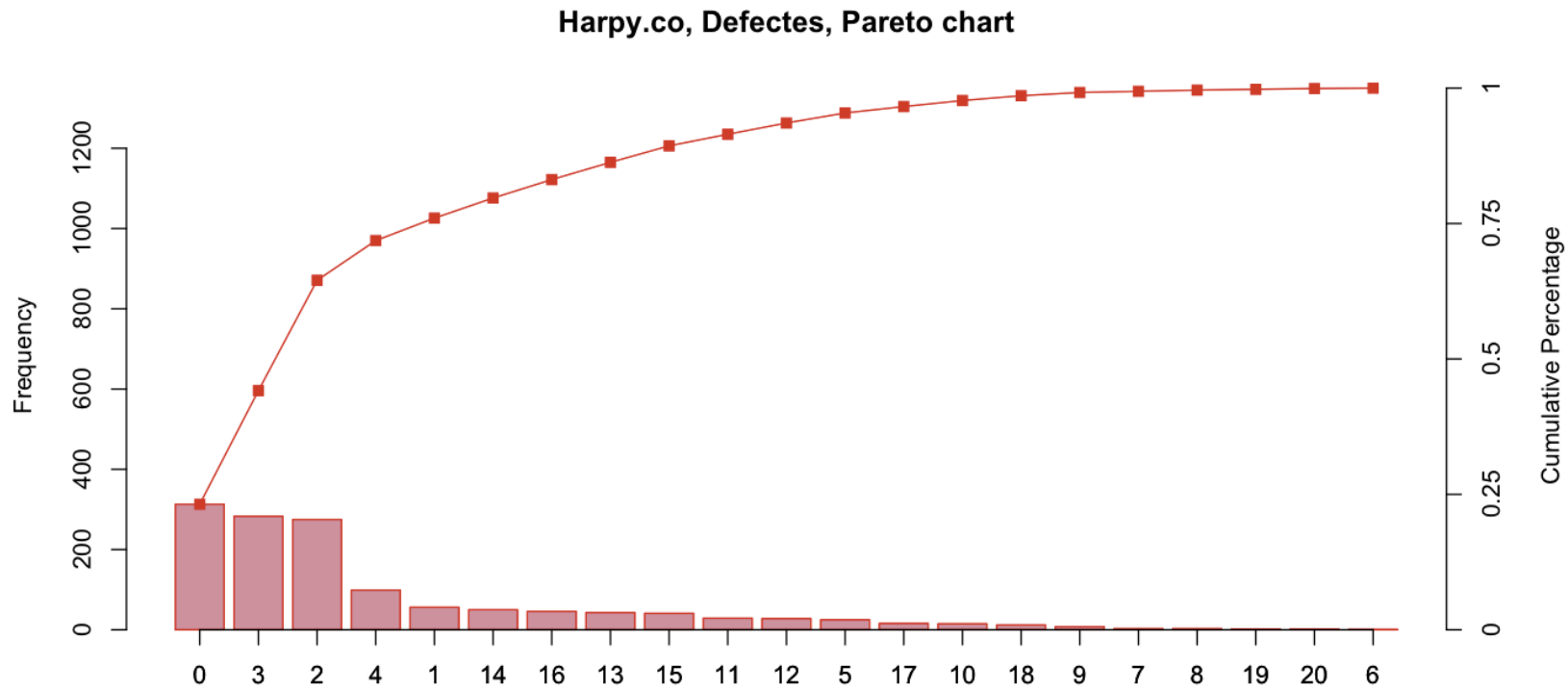
Frequency	1050	41	38	34	33	31	27	26	25	24	21
Cum. Frequency	1050	1091	1129	1163	1196	1227	1254	1280	1305	1329	1350
Percentage	78	3	3	3	2	2	2	2	2	2	2
Cum. Percentage	78	81	84	86	89	91	93	95	97	98	100



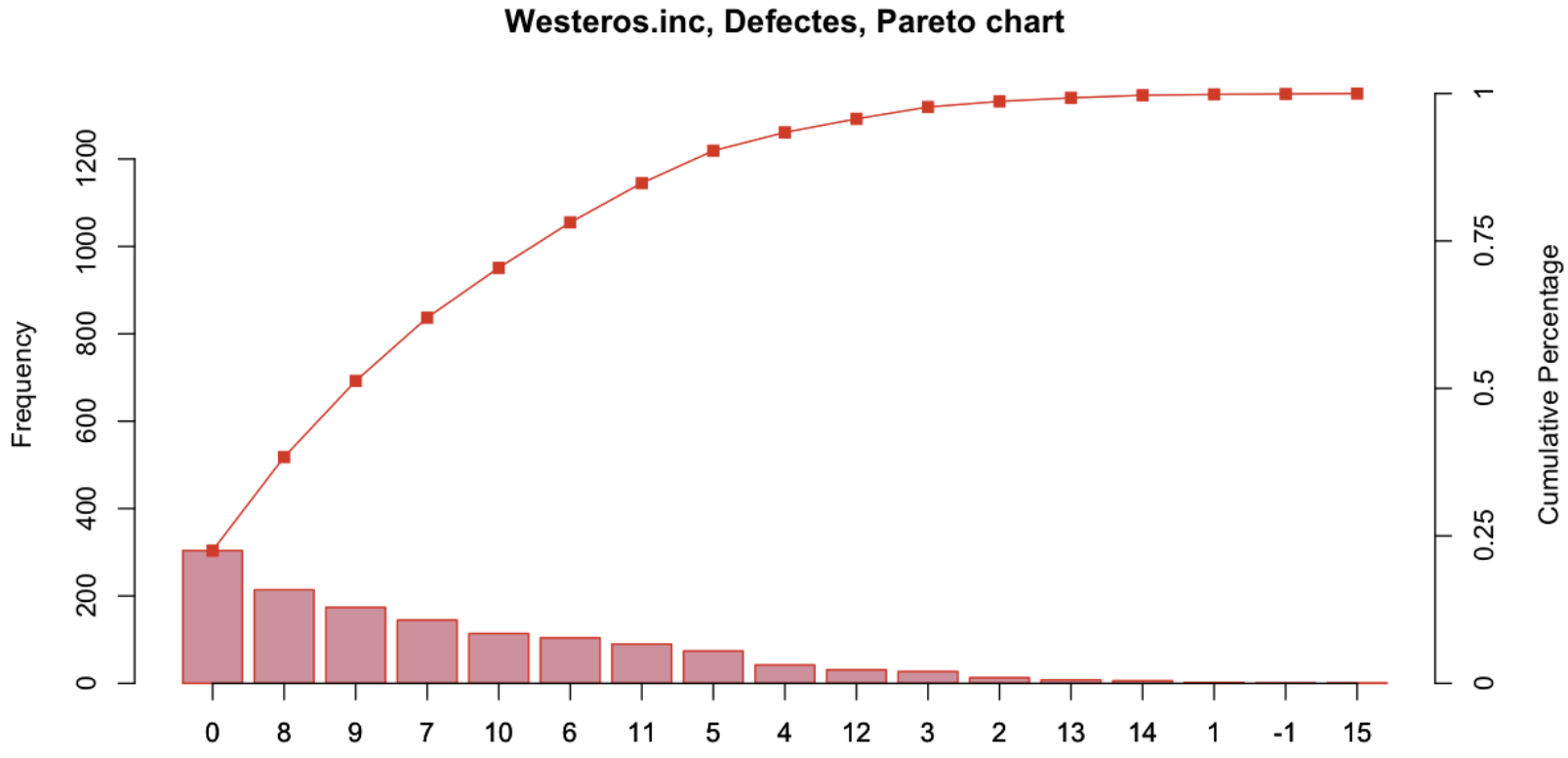
Frequency	1050	46	41	39	31	29	28	27	24	19	16
Cum. Frequency	1050	1096	1137	1176	1207	1236	1264	1291	1315	1334	1350
Percentage	78	3	3	3	2	2	2	2	2	1	1
Cum. Percentage	78	81	84	87	89	92	94	96	97	99	100

Для начала посмотрим статистику по производству оружия для этих компаний. По оси OX мы видим количество произведенного оружия, слева по оси OY - сколько раз встречалось данное наблюдение в нашей выборке, справа по оси OY - то же самое, только в процентном соотношении. Столбчатая диаграмма показывает нам распределение величин по всей выборке, а линия - суммирует предыдущие значения с текущим. Графики нам более наглядно представляют данные. Обратимся к числам, стоящим ниже для более подробного анализа. Как мы видим, количество раз производства оружия у этих компаний одинаково - 200(1350-1050). Для компании harpy.co по 3% составляют производство 101, 102, 104 деталей, по 2% - 103, 105, 106, 107, 108, 109, 110. Для компании Westeros.inc по 3% - 103, 104, 109, по 2% - 101, 102, 105, 106, 107, 108, 110. В общей сложности компании произвели примерно равное количество оружия.

Посмотрим на статистику по дефектам:



	0	3	2	4	1	14	16	13	15	11	12	5	17	10	18	9	7	8	19	20	6
Frequency	313	283	275	99	56	50	46	43	41	29	28	25	16	15	12	8	3	3	2	2	1
Cum. Frequency	313	596	871	970	1026	1076	1122	1165	1206	1235	1263	1288	1304	1319	1331	1339	1342	1345	1347	1349	1350
Percentage	23	21	20	7	4	4	3	3	3	2	2	2	1	1	1	1	0	0	0	0	0
Cum. Percentage	23	44	65	72	76	80	83	86	89	91	94	95	97	98	99	99	99	100	100	100	100



Frequency	304	214	174	145	114	104	90	74	42	31	27	13	8	6	2	1	1
Cum. Frequency	304	518	692	837	951	1055	1145	1219	1261	1292	1319	1332	1340	1346	1348	1349	1350
Percentage	23	16	13	11	8	8	7	5	3	2	2	1	1	0	0	0	0
Cum. Percentage	23	38	51	62	70	78	85	90	93	96	98	99	99	100	100	100	100

По данной статистике мы видим, что в компании Westeros.inc достаточно часто встречаются поломки в объеме от 5 до 11, наибольшее число неисправностей приходятся на 7, 8 и 9. Гораздо в меньшей степени встречаются от 2 до 4, однако есть 12 - 2% от общей выборки и 13 - 1% от общей выборки. Все остальные не столь значительны.
Если же посмотреть на статистику Harpy&co, то мы увидим, что значительная часть всех поломок приходится на объем от 2 до 4, что гораздо меньше, чем в Westeros.inc, однако мы видим, что пусть и не столь часто, но встречается большое число поломок: 14 - 4% всех поломок; 16, 13, 15 - 3%; 11, 12 - 2%, 17, 10, 18 - 1%. Таким образом, в компании Westeros.inc со средней периодичностью встречается среднее число дефектов(5-11), а в компании Harpy&co часто встречается небольшое число дефектов(1-4), но есть выбросы с количеством дефектов в размере от 12 до 18.

Stem-and-leaf

Воспользуемся описательной статистикой - диаграммой «Ветвей и листьев». Длина каждой строки соответствует количеству наблюдений, попадающих в определенный интервал. Кроме того здесь также отображено численное значение для каждого наблюдения. Для этой цели численное значения разбиваются на два компонента: ветвь, представляющую собой десятки и лист — единицы. Ветвь соответствует тем разрядам численного значения наблюдаемой переменной, которые не изменяются, а листья — разрядам, которые изменяются в пределах избранного интервала.

Посмотрим на результаты по производству оружия:

```
> stem(pppdata.01.westeros$produced)

  The decimal point is 1 digit(s) to the right of the |

   0 | 000000000000000000000000000000000000000000000000000000000000000000+970
   1 |
   2 |
   3 |
   4 |
   5 |
   6 |
   7 |
   8 |
   9 |
  10 | 11111111111111111112222222222222222222222222233333333333333333333333+196
  11 | 00000000000000000000
```

```
> stem(pppdata.01.harpy$produced)

  The decimal point is 1 digit(s) to the right of the |

   0 | 000000000000000000000000000000000000000000000000000000000000000000+970
   1 |
   2 |
   3 |
   4 |
   5 |
   6 |
   7 |
   8 |
   9 |
  10 | 11111111111111111111111111111111111111122222222222222222222222222+196
  11 | 00000000000000000000
```

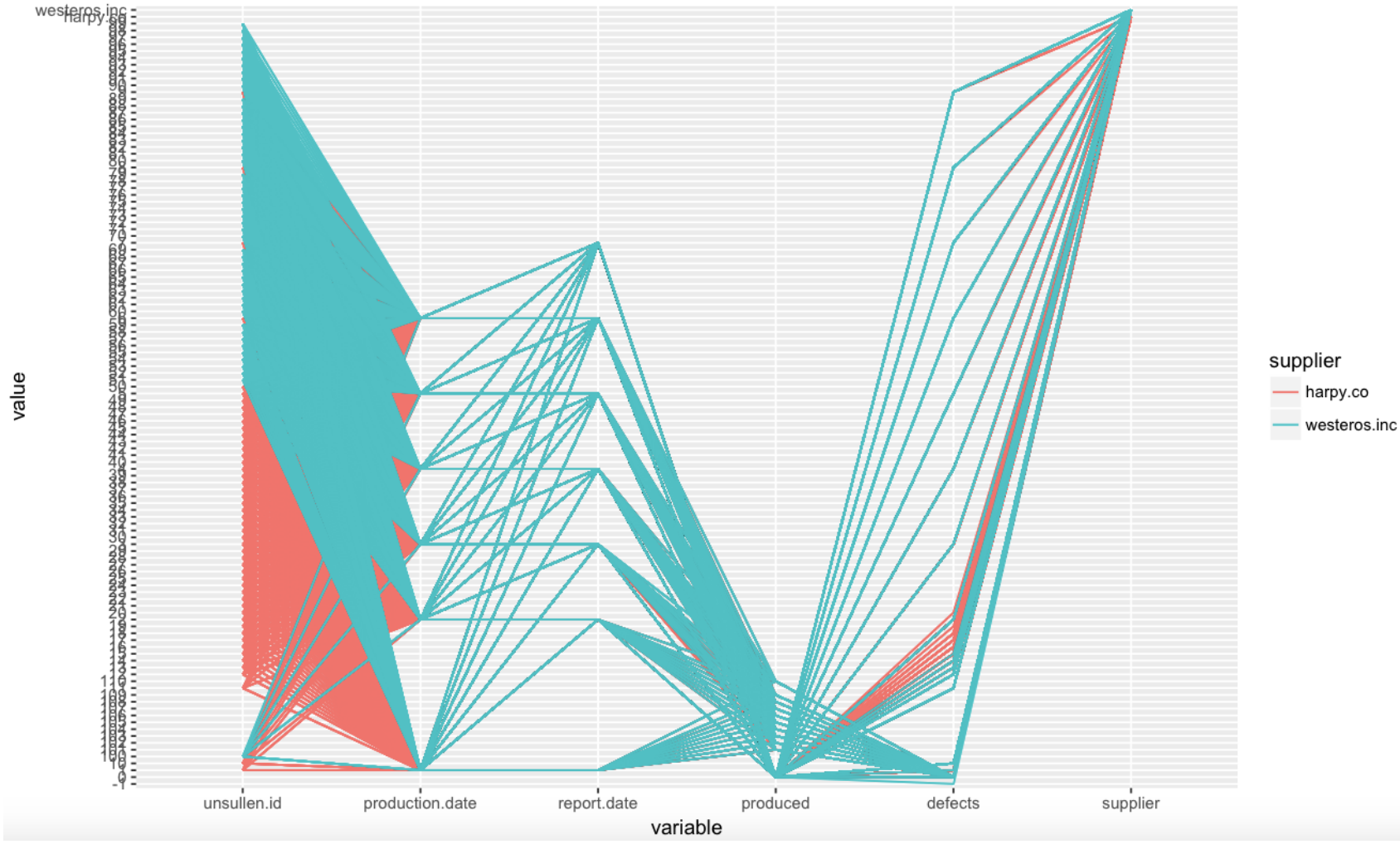
По данной диаграмме мы видим, что обе компании либо ничего не производят(строка из нулей), либо производят 101,102...110 оружия.
Теперь посмотрим на результаты дефектов оружия:
Первая диаграмма соответствует количеству дефектов для компании Westeros.inc, вторая - количеству дефектов для компании Harpy&Co.

```
> stem(pppdata.01.westeros$defects)

  The decimal point is at the |

  -1 | 0
  -0 |
   0 | 00000000000000000000000000000000000000000000000000000000000000000000000000000000+224
   1 | 00
   2 | 0000000000000
   3 | 000000000000000000000000000
   4 | 0000000000000000000000000000000000000000000
   5 | 00000000000000000000000000000000000000000000000000000000000000000000000000000000000
   6 | 00000000000000000000000000000000000000000000000000000000000000000000000000000000+24
   7 | 00000000000000000000000000000000000000000000000000000000000000000000000000000000+65
   8 | 00000000000000000000000000000000000000000000000000000000000000000000000000000000+134
   9 | 00000000000000000000000000000000000000000000000000000000000000000000000000000000+94
  10 | 00000000000000000000000000000000000000000000000000000000000000000000000000000000+34
  11 | 00000000000000000000000000000000000000000000000000000000000000000000000000000000+10
  12 | 000000000000000000000000000000
  13 | 00000000
  14 | 000000
  15 | 0


> stem(pppdata.01.harpy$defects)

  The decimal point is at the |

   0 | 00000000000000000000000000000000000000000000000000000000000000000000000000000000+233
   1 | 000000000000000000000000000000000000000000000000000000000000000000000000000000
   2 | 00000000000000000000000000000000000000000000000000000000000000000000000000000000+195
   3 | 00000000000000000000000000000000000000000000000000000000000000000000000000000000+203
   4 | 00000000000000000000000000000000000000000000000000000000000000000000000000000000+19
   5 | 0000000000000000000000000
   6 | 0
   7 | 000
   8 | 000
   9 | 00000000
  10 | 000000000000000
  11 | 00000000000000000000000000000
  12 | 000000000000000000000000000000
  13 | 0000000000000000000000000000000000000000000
  14 | 00000000000000000000000000000000000000000000000000000
  15 | 000000000000000000000000000000000000000000
  16 | 0000000000000000000000000000000000000000000000000
  17 | 0000000000000000
  18 | 000000000000
  19 | 00
  20 | 00
```

Судя по первой диаграмме, в компании Westeros.inc оружие либо не ломается, либо поломки происходят достаточно часто и их количество колеблется от 3 до 11-12. Числа достаточно весомые, что не очень хорошо говорит качестве продукции. Во второй диаграмме мы наблюдаем, что поломки делятся на 3 типа - либо совсем не ломаются, либо небольшое количество, но ломаются часто, либо большое количество дефектов(10-18), но случается относительно редко.

Parallel Coordinares

На графике параллельных координат отображается множество рядов данных в виде линий, проходящих по промежуточным осям. Каждая из осей отображает значения по выбранному показателю. Линии соединяют точки на промежуточных осях в соответствии со значениями элементов по нашим показателям.


value
variable
unsullen.id
production.date
report.date
produced
defects
supplier
supplier
harpy.co
westeros.inc

По данному графику мы видим, что для переменной production.date характерно выделены 6 точек. Каждая точка соответствует месяцу, в который было произведено оружие. Следующие 7 точек это месяцы, в которые пришел отчет о дефектах. Нумерация идет снизу. Можно сделать вывод, что по первому месяцу отчет приходит в 1 и 2 . . . 7 месяцы, по второму - во 2, 3 . . . 7 и тд. Для переменной produced у нас какрактерно выделен 0 и несколько точек, соответствующих значениям от 101 до 120, что говорит нам о том, что обе компании либо производят 0, либо сразу большое количество. Далее идет приличные разброс по дефектам для каждой компании и завершение на самих компаниях.

PCA

Прежде чем применять метод главных компонент, взглянем на наши данные:

```
> glimpse(pppdata.01.harpy)
Observations: 1,350
Variables: 5
$ unsullen.id     <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
$ production.date <int> 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4...
$ report.date     <int> 1, 2, 3, 4, 5, 6, 7, 2, 3, 4, 5, 6, 7, 3, 4, 5, 6, 7, 4, 5...
$ produced        <dbl> 103, 0, 0, 0, 0, 0, 0, 105, 0, 0, 0, 0, 0, 108, 0, 0, 0, 0...
$ defects         <dbl> 0, 2, 4, 5, 13, 11, 11, 0, 2, 2, 2, 13, 16, 0, 2, 3, 2, 16...
```

```
> glimpse(pppdata.01.westeros)
Observations: 1,350
Variables: 5
$ unsullen.id     <dbl> 51, 51, 51, 51, 51, 51, 51, 51, 51, 51, 51, 51, 51, 51, 51...
$ production.date <int> 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4...
$ report.date     <int> 1, 2, 3, 4, 5, 6, 7, 2, 3, 4, 5, 6, 7, 3, 4, 5, 6, 7, 4, 5...
$ produced        <dbl> 106, 0, 0, 0, 0, 0, 0, 104, 0, 0, 0, 0, 0, 106, 0, 0, 0, 0...
$ defects         <dbl> 0, 8, 9, 8, 12, 0, 5, 0, 9, 14, 4, 6, 9, 0, 8, 12, 5, 11, ...
```

Средние значения сильно разнятся между собой. Соответственно прежде чем применять метод главных компонент, необходимо стандартизировать переменные. Посмотрим описание результатов оценивания главных компонент:

```
> summary(pppdata.01.westeros.pca)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5
Standard deviation     1.4454 1.1858 1.0000 0.64213 0.30403
Proportion of Variance 0.4178 0.2812 0.2000 0.08247 0.01849
Cumulative Proportion  0.4178 0.6990 0.8991 0.98151 1.00000
```

Для компании Westeros.inc

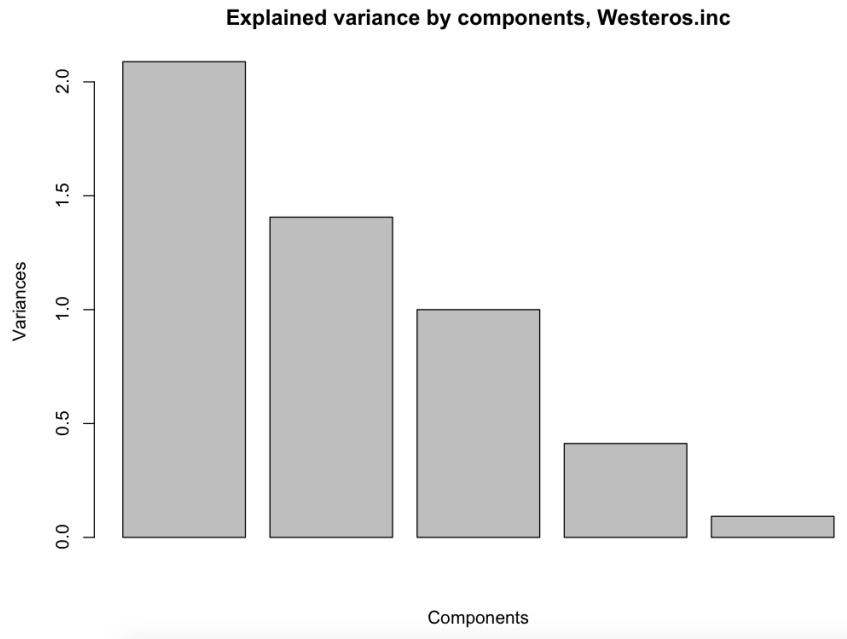
Посмотрим, как главные компоненты справляются с описанием наших данных и на сколько мы можем уменьшить нашу размерность, если изначально она равна 5(unsullen.id, production.date, report.date, produced, defects) PC1 описывает примерно 41% всех данных, PC2 - 28%, PC3 - 20%, PC4 - 8%, PCA5 - 1%. В сумме PC1 PC2 PC3 дают нам информацию о 89,9% информации. На практике берут те компоненты, которые содержат около 88% информации. В нашем случае это первые три компоненты.

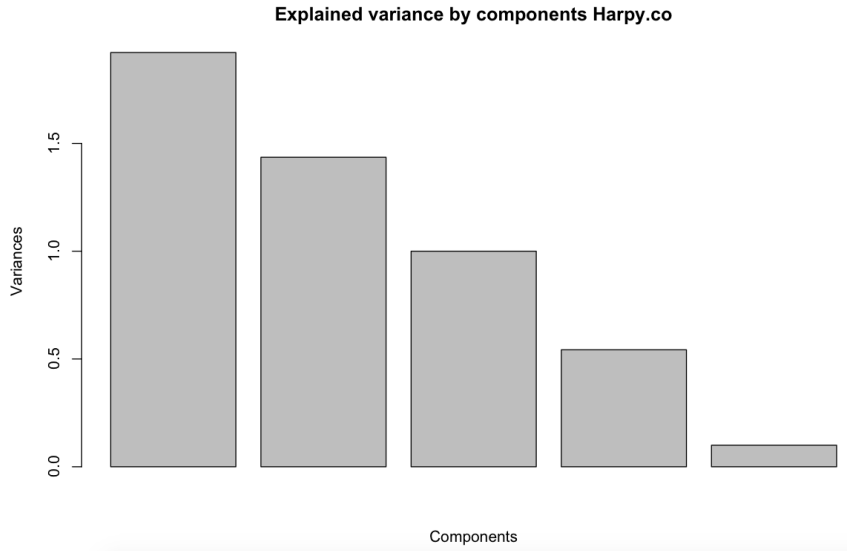
```
> summary(pppdata.01.harpy.pca)
Importance of components:
                          PC1    PC2    PC3    PC4     PC5
Standard deviation     1.3862 1.1982 1.0000 0.7367 0.31610
Proportion of Variance 0.3843 0.2872 0.2000 0.1085 0.01998
Cumulative Proportion  0.3843 0.6715 0.8715 0.9800 1.00000
```

Для компании Harpy&Co

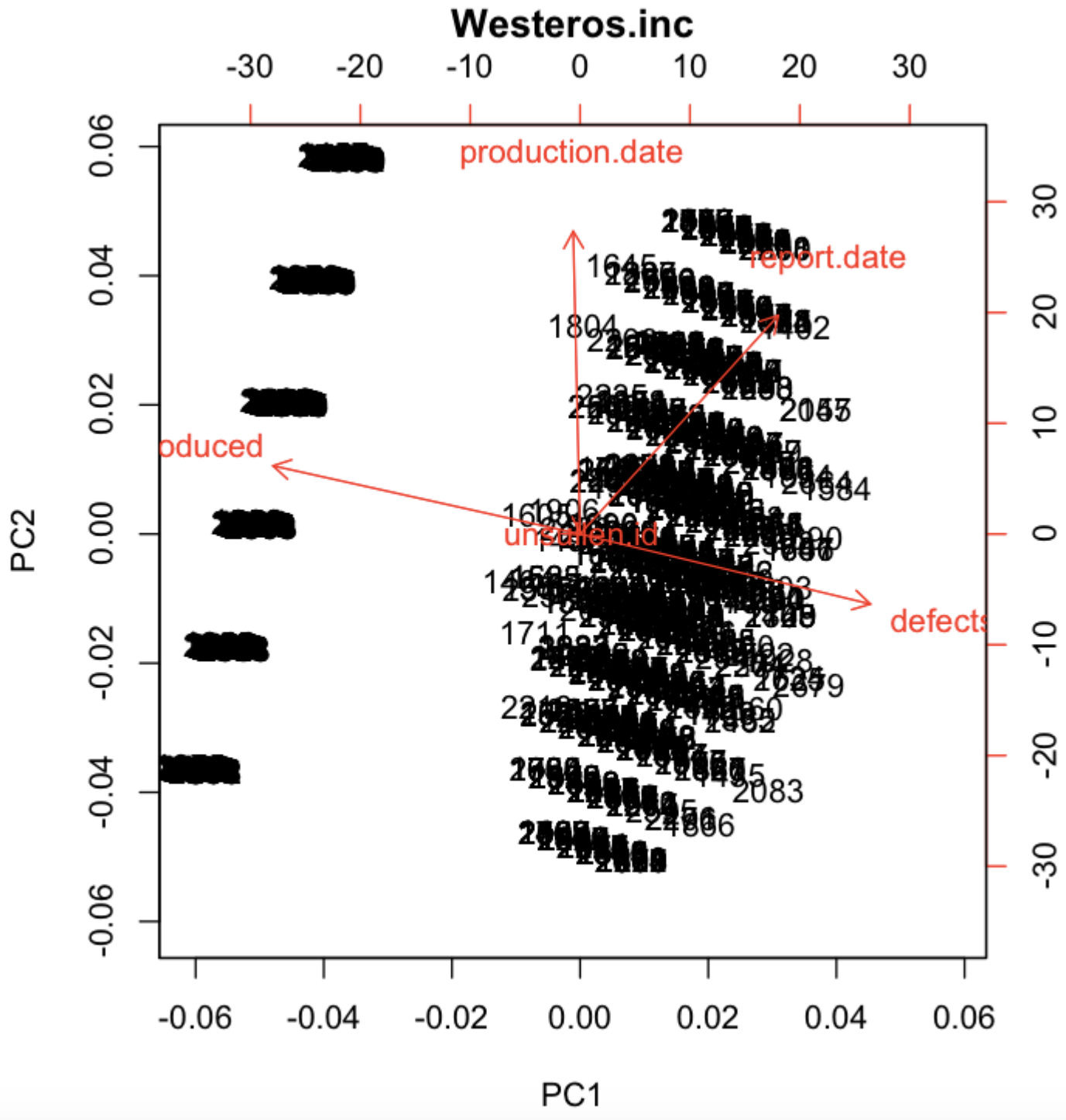
Посмотрим, как главные компоненты справляются с описанием наших данных и на сколько мы можем уменьшить нашу размерность, если изначально она равна 5(unsullen.id, production.date, report.date, produced, defects) PC1 описывает примерно 38% всех данных, PC2 - 28%, PC3 - 20%, PC4 - 10%, PCA5 - 1%. В сумме PC1 PC2 PC3 дают нам информацию о 87,1% информации. На практике берут те компоненты, которые содержат около 88% информации. В нашем случае это первые три компоненты. Итак, для обеих компаний достаточно первых трех главных компонент, чтобы описать основные данные. Для компании Westeros.inc мы смогли описать на 2% лучше.

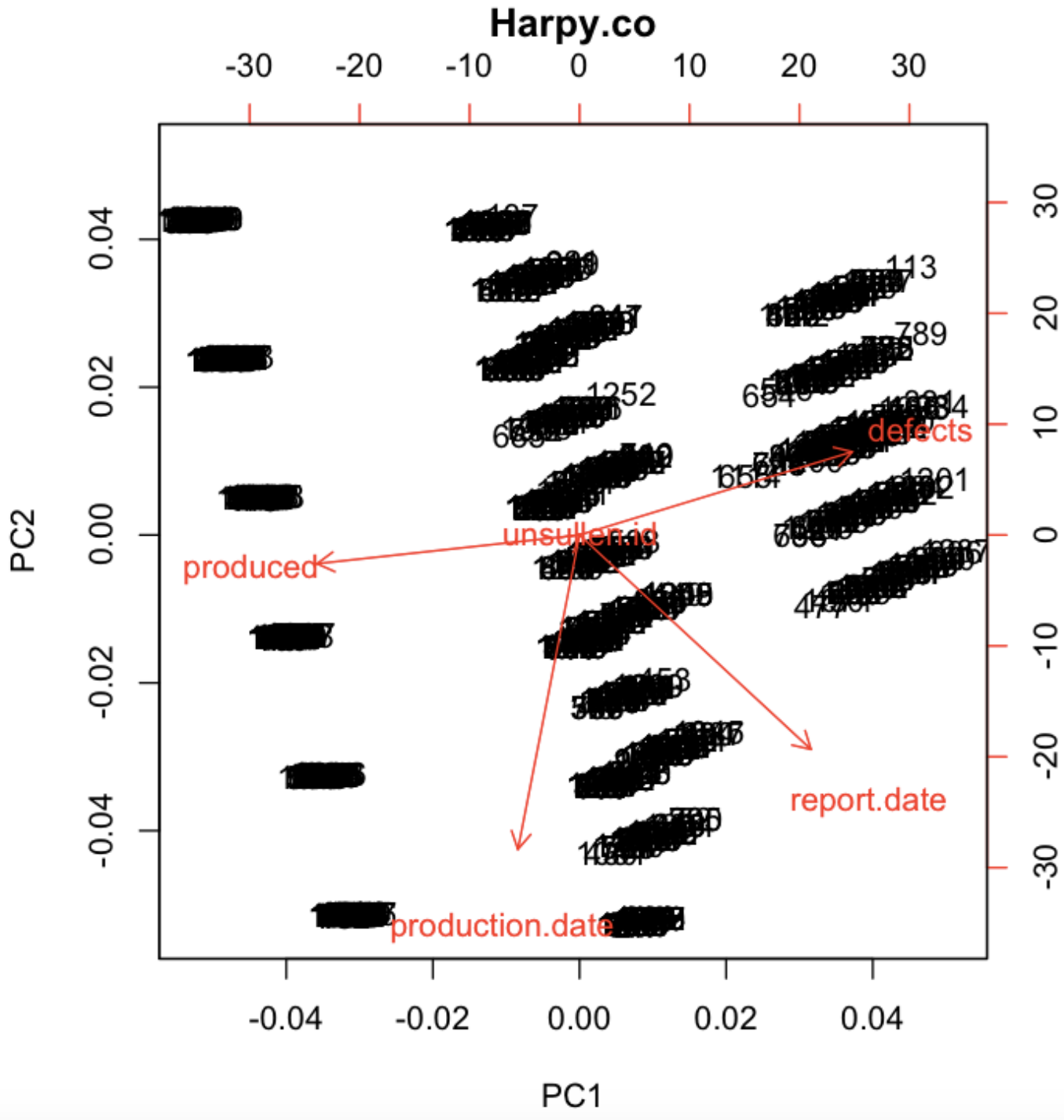
Посмотрим на график объяснений дисперсии:





Визуализируем наши данные для более наглядного представления. Построим график от первых двух компонент.
По оси OX будет PC1, по оси OY будет PC2
Красные векторы это отмеренные исходные переменные в координатах первых двух главных компонент. Соответственно, проецируя такой вектор на одну из осей, мы получим с каким весом она входит в соответствующую компоненту.


Westeros.inc
-30
-20
-10
0
10
20
30
production.date
report.date
oduced
unsullen.id
defects
1804
1645
1711
2083
0.06
0.04
0.02
0.00
-0.02
-0.04
-0.06
PC2
PC1


Harpy.co
-30
-20
-10
0
10
20
30
0.04
0.02
0.00
-0.02
-0.04
PC2
113
789
1252
defects
produced
unsullen.id
report.date
production.date
-0.04
-0.02
0.00
0.02
0.04
PC1

Для компании Westeros.inc: Мы видим, что вторая главная компонента с большим весом себе берет production.date и report.date, а первая главная компонента produced и defects. Анализируя полученные результаты, можно сказать, что первая главная компонента у нас отвечает за производство и дефекты, а вторая за время производства и получение отчетов. Для компании Harpy&co: Мы видим, что вторая главная компонента с большим весом себе берет production.date и report.date, а первая главная компонента produced и defects.

Анализируя полученные результаты, можно сказать, что первая главная компонента у нас отвечает за производство и дефекты, а вторая за время производства и получение отчетов. Таким образом, отличие между Westeros.inc и Hsrpy&co заключается лишь в том, что production.date и report.date смотрят в разные стороны(то есть противоположны по знаку), хотя проекции по модулю на вторую главную компоненту очень похожи между собой.

Для большей наглядности посмотрим численные характеристики: С какими коэффициентами входят исходные переменные в первую компоненту для компании Westeros.inc (w.v1) и Harpy&co (h.v1):

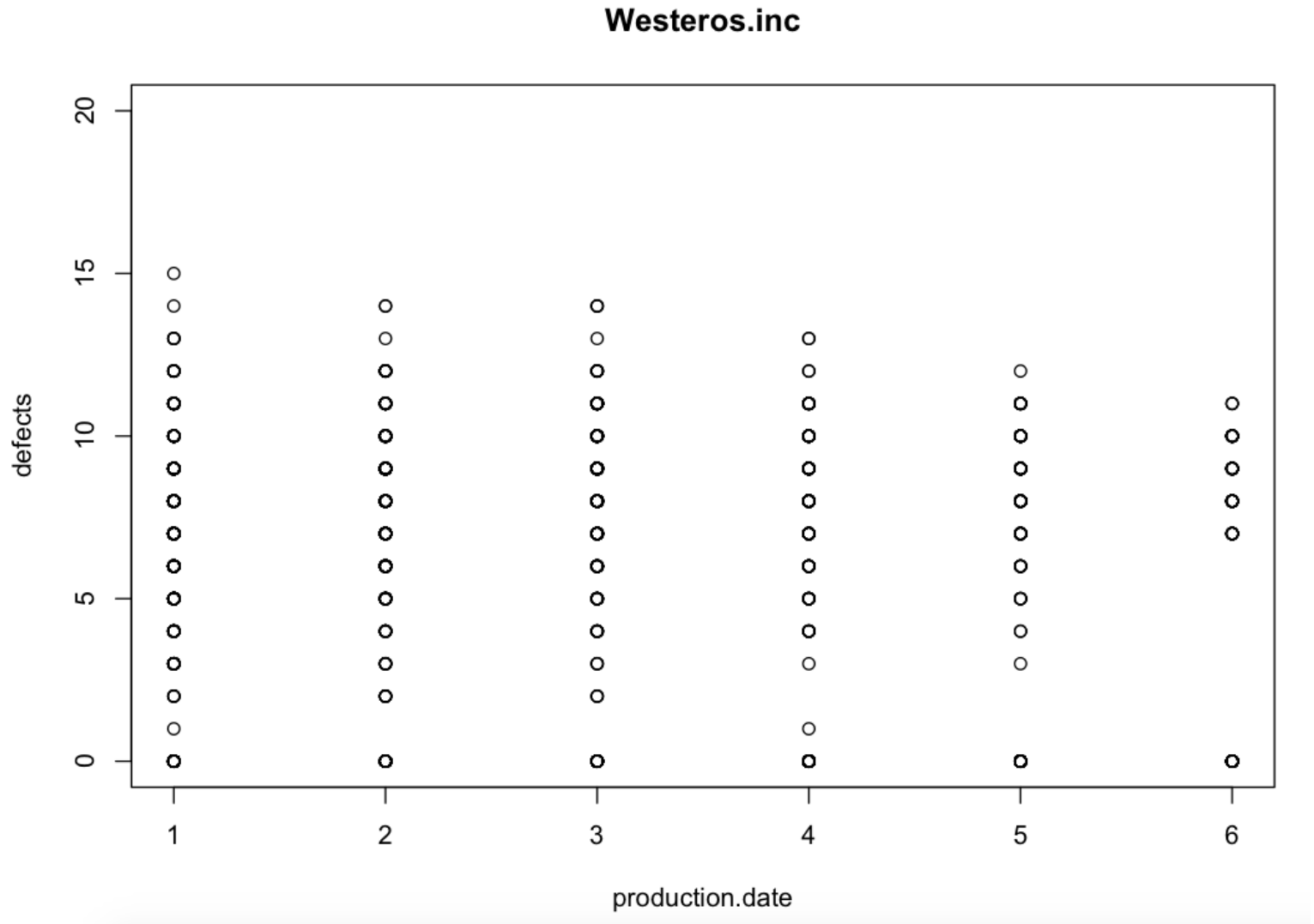
```
> w.v1
    unsullen.id production.date     report.date        produced         defects
    0.002063562    -0.014704681     0.424795323    -0.657778119     0.621817010
> h.v1
    unsullen.id production.date     report.date        produced         defects
     0.00165109     -0.13761584      0.51666342     -0.58590994      0.60895617
```

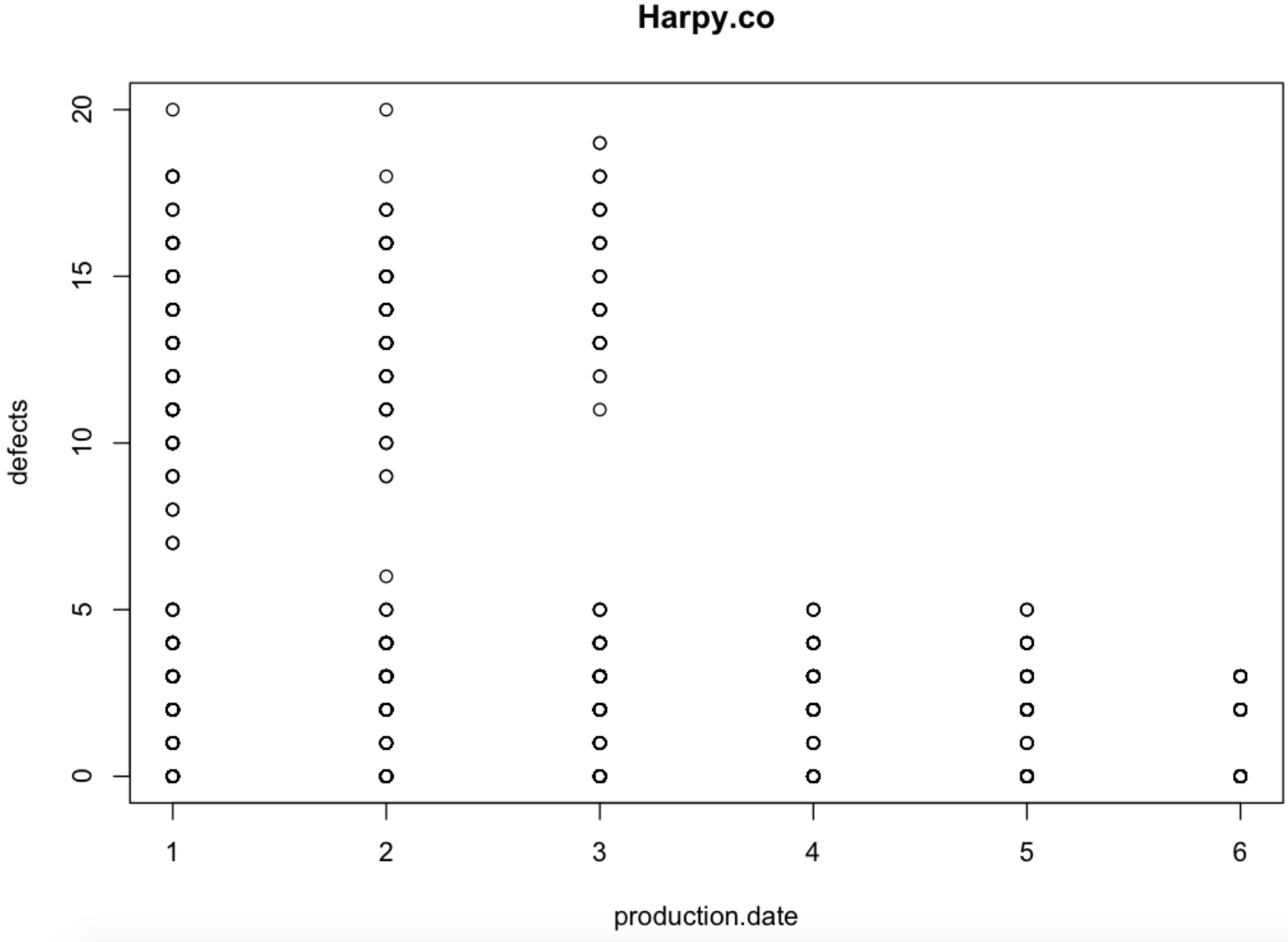
С какими коэффициентами входят исходные переменные в первую компоненту для компании Westeros.inc (w.v2) и Harpy&co (h.v2):

```
> w.v2
    unsullen.id production.date     report.date        produced         defects
   -0.001634682     0.784227126     0.566418017     0.176765415    -0.181410475
> h.v2
    unsullen.id production.date     report.date        produced         defects
    0.001534945    -0.804880959    -0.548936732    -0.073469943     0.213154681
```

Статистика по дефектам для компаний Westeros.inc и Harpy&co и Итог

Посмотрим, как распределены дефекты по месяцам в наших компаниях:




Harpy.co
defects
0
5
10
15
20
1
2
3
4
5
6
production.date

Для начала посмотрим на первые три месяца. В обеих компаниях показатели по дефектам очень большие, однако в Westeros.inc выше 15 поломок нет. В компании Harpy&co мы наблюдаем либо небольшие показатели в пределах 5, либо очень весомые дефекты вплоть до 20. Если же посмотреть на последующие 3 месяца, то в Westeros.inc регулярно встречаются поломки, которые колеблются от 5 до 12. В Harpy&co количество дефектов почти сведено на нет, все показатели в пределах 5.
Основываясь на результатах проведенного выше анализа, мы делаем вывод, что в будущем лучше сотрудничать с компанией Harpy&co.

Задание выполняли

- Запорожец Анастасия, студентка 412 группы.
  Занималась написанием программы.
- Собольковa Екатерина, студентка 412 группы.
  Анализировала данные и писала отчет.
- Агаджанян Елизавета, студентка 412 группы.
  Составляла презентацию и писала отчет.