Задание банка Открытие

# Подготовка данных

Для начала считаем данные из файла и взглянем на них.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Bank | Parameter | M1 | M2 | M3 | M4 | M5 |
| А | Кол-во принятых кредитных заявок | 7357.1012019 | 1.047682e+04 | 1.054394e+04 | 1.070278e+04 | 7656.7701397 |
| NA | Уровень одобрения | 0.1576522 | 1.571524e-01 | 1.581591e-01 | 1.605418e-01 | 0.1640736 |
| NA | Кол-во выданных кредитов | 890.7747526 | 1.185449e+03 | 1.427483e+03 | 1.580785e+03 | 934.6679990 |
| Б | Кол-во принятых кредитных заявок | 6660.2318842 | 9.544878e+03 | 9.484121e+03 | 9.343363e+03 | 6399.5652352 |
| NA | Уровень одобрения | 0.3678551 | 3.666888e-01 | 3.690379e-01 | 3.745975e-01 | 0.3828385 |
| NA | Кол-во выданных кредитов | 2156.0000000 | 3.276000e+03 | 2.744000e+03 | 2.660000e+03 | 1960.0000000 |

Приведём данные к *чистому* виду. Каждое наблюдение будет в своём ряду, каждый показатель - в своём столбце.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Bank | Month | Requests | Approval | Issues |
| A | 1 | 7357 | 0.158 | 891 |
| A | 2 | 10477 | 0.157 | 1185 |
| A | 3 | 10544 | 0.158 | 1427 |
| A | 4 | 10703 | 0.161 | 1581 |
| A | 5 | 7657 | 0.164 | 935 |
| A | 6 | 11229 | 0.168 | 1695 |

Данные отсортированы по месяцам внутри каждого отдельного банка. В конце таблицы будут данные за последние месяцы 5 банка, который мы обозначили за **E**.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Bank | Month | Requests | Approval | Issues |
| E | 19 | 5325 | 0.068 | 186 |
| E | 20 | 4891 | 0.067 | 168 |
| E | 21 | 2090 | 0.066 | 70 |
| E | 22 | 5455 | 0.065 | 180 |
| E | 23 | NA | NA | NA |
| E | 24 | NA | NA | NA |

Тут же мы видим, что некоторые данные в таблице отсутствуют. Всего строк с пропущенными значениями в файле 8. Это не так много, поэтому взглянем на них.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Bank | Month | Requests | Approval | Issues |
| D | 1 | NA | NA | NA |
| D | 2 | NA | NA | NA |
| E | 1 | NA | NA | NA |
| E | 2 | NA | NA | NA |
| E | 3 | NA | NA | NA |
| E | 4 | NA | NA | NA |
| E | 23 | NA | NA | NA |
| E | 24 | NA | NA | NA |

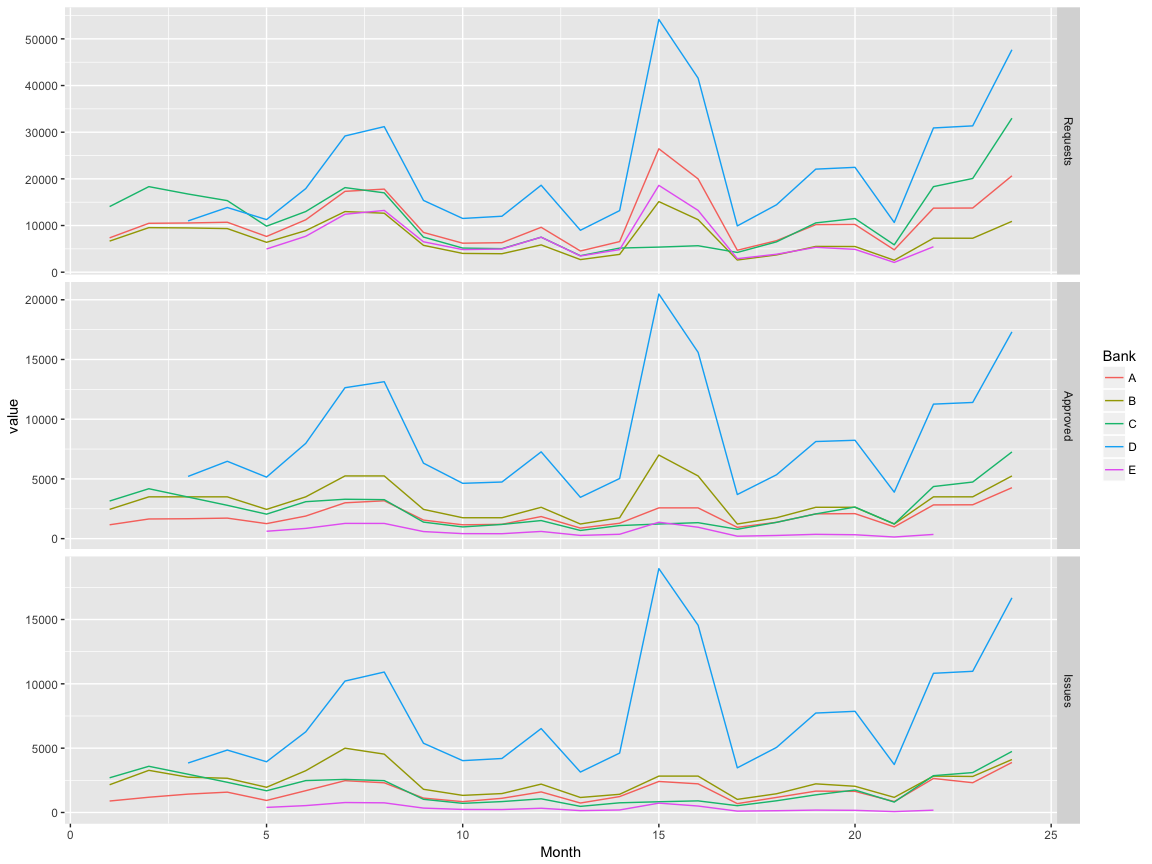
Примечательно, что от банков **D** и **E** нет информации за первые месяцы - возможно, они недавно открылись. Более примечательно, что от банка **E** нет информации за последние месяцы. Возникает предположение, что это отделение закрылось. Посмотрим, есть ли какие-то очевидные поводы для такого решения.

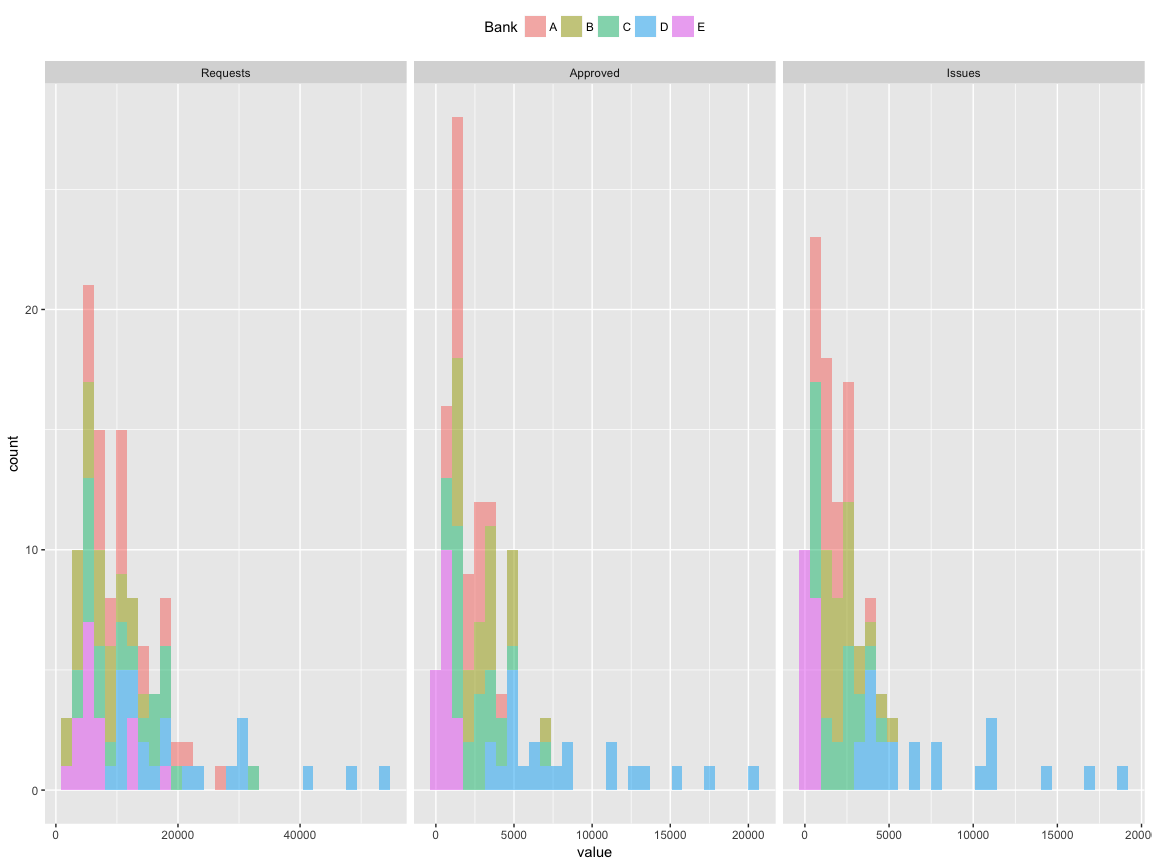
# Изучим данные

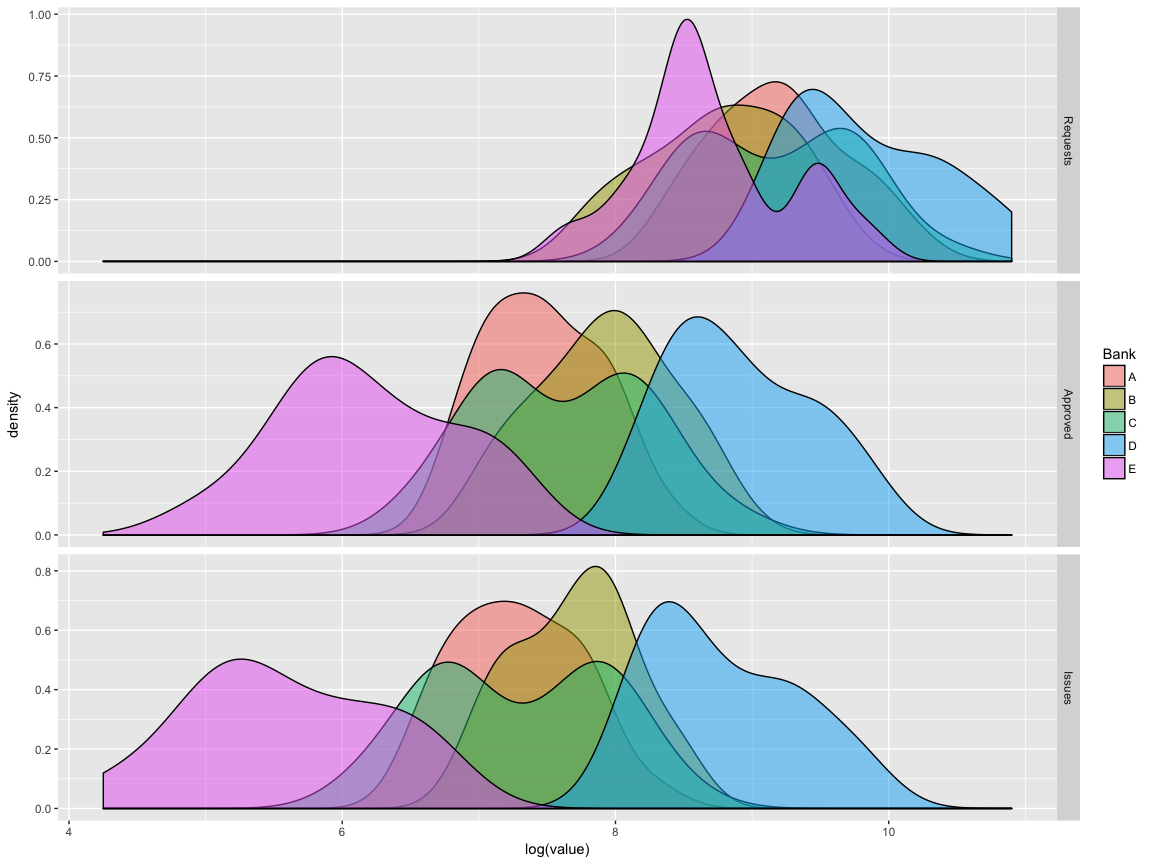
Для начала добавим в нашу таблицу столбцы с абсолютным количеством одобренных заявок, а так же "уровень согласия" - то есть процент одобренных заявок, на которые согласились клиенты.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Bank | Month | Requests | Approved | Issues | Approval | Acceptance |
| A | 1 | 7357 | 1160 | 891 | 0.158 | 0.768 |
| A | 2 | 10477 | 1646 | 1185 | 0.157 | 0.720 |
| A | 3 | 10544 | 1668 | 1427 | 0.158 | 0.856 |
| A | 4 | 10703 | 1718 | 1581 | 0.161 | 0.920 |
| A | 5 | 7657 | 1256 | 935 | 0.164 | 0.744 |
| A | 6 | 11229 | 1891 | 1695 | 0.168 | 0.896 |

Построим графики абсолютных показателей:

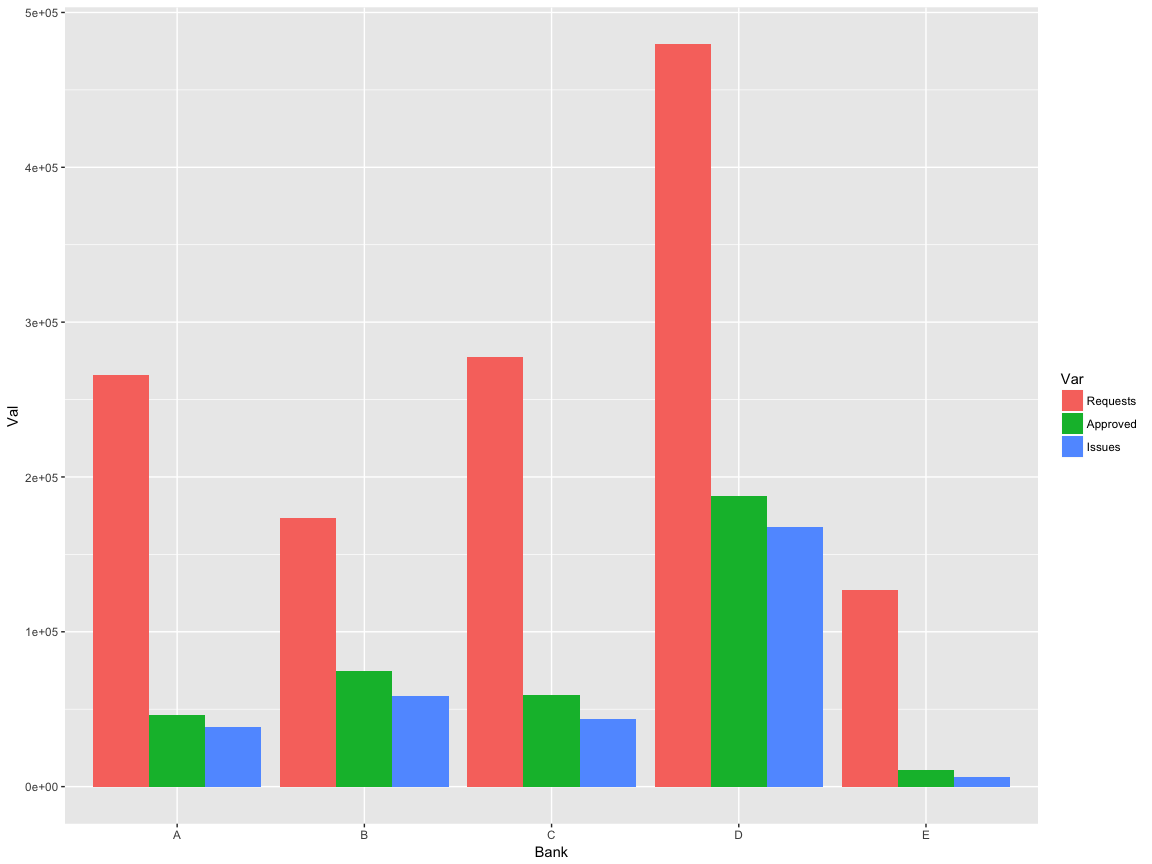


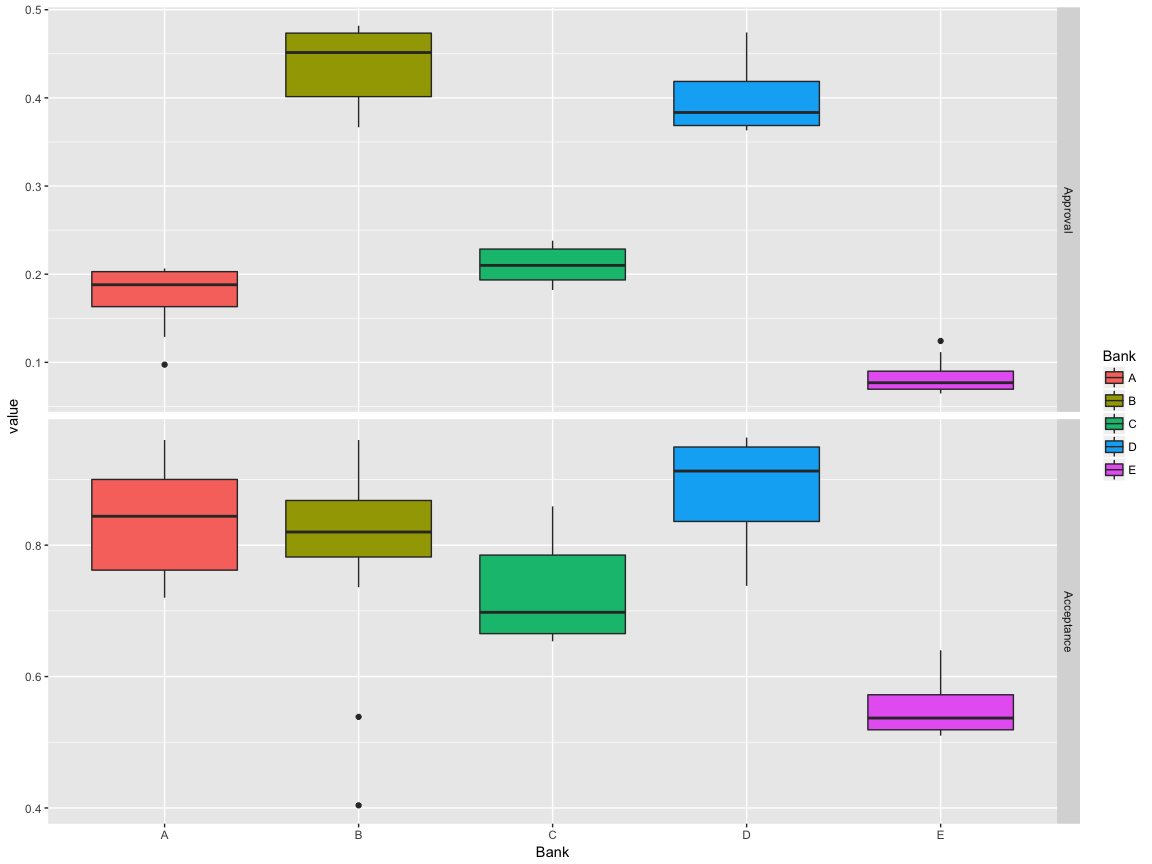
Действительно, в банк **E** долгое время подавалось наименьшее количество заявок, почти всегда там одобрялось наименьшее количество заявок и всегда там выпускалось наименьшее количество кредитов. Возможно, он находился в районе с малым числом потенциальных клиентов. Так выглядят гистограммы наших абсолютных показателей: 

Видно, что они имеют далеко не нормальное распределение. Простое логарифмирование тоже нормализует слишком грубо: 

Чтобы не нарушать предпосылок таких тестов, как ANOVA и текст Тьюки, воспользуемся непараметрическим тестом Данна, чтобы узнать, в какие банки подавалось статистически больше заявок.

## Kruskal-Wallis rank sum test  
##   
## data: x and group  
## Kruskal-Wallis chi-squared = 37.0577, df = 4, p-value = 0  
##   
##   
## Comparison of x by group   
## (Bonferroni)   
## Col Mean-|  
## Row Mean | A B C D  
## ---------+--------------------------------------------  
## B | 2.231156  
## | 0.1284  
## |  
## C | -0.011111 -2.242268  
## | 1.0000 0.1247  
## |  
## D | -3.063733 -5.245847 -3.052866  
## | 0.0109 0.0000 0.0113  
## |  
## E | 2.366033 0.300383 2.376320 5.166562  
## | 0.0899 1.0000 0.0874 0.0000

Сильно от всех остальных отличается банк **D**. А ближе всего друг к другу находятся пары **A** и **C**, а так же **B** и **E**. Возможно, другие банки проводили более гибкую кредитную политику. 

Действительно, в банк **B** за все время было подано немногим больше заявок, чем в банк **E** (то же мы увидели и в результате теста), и меньше, чем в банки **A** и **C**, однако число одобренных и выданных кредитов даже больше последних. Взглянем на относительные уровни одобрения и согласия. 

В самом деле, банк **B** одобрял в среднем даже больше самого успешного банка **D**. Возможно, последний находится в густонаселенном районе и подобная политика привела к лояльности клиентов к банку, что и стало причиной его успеха. Тогда банку **B** стоит продолжать придерживаться данной стратегии, если это не приводит к росту невыплаченных задолжностей. Продолжим анализ и посмотрим, какие из этих различий статистически значимые.

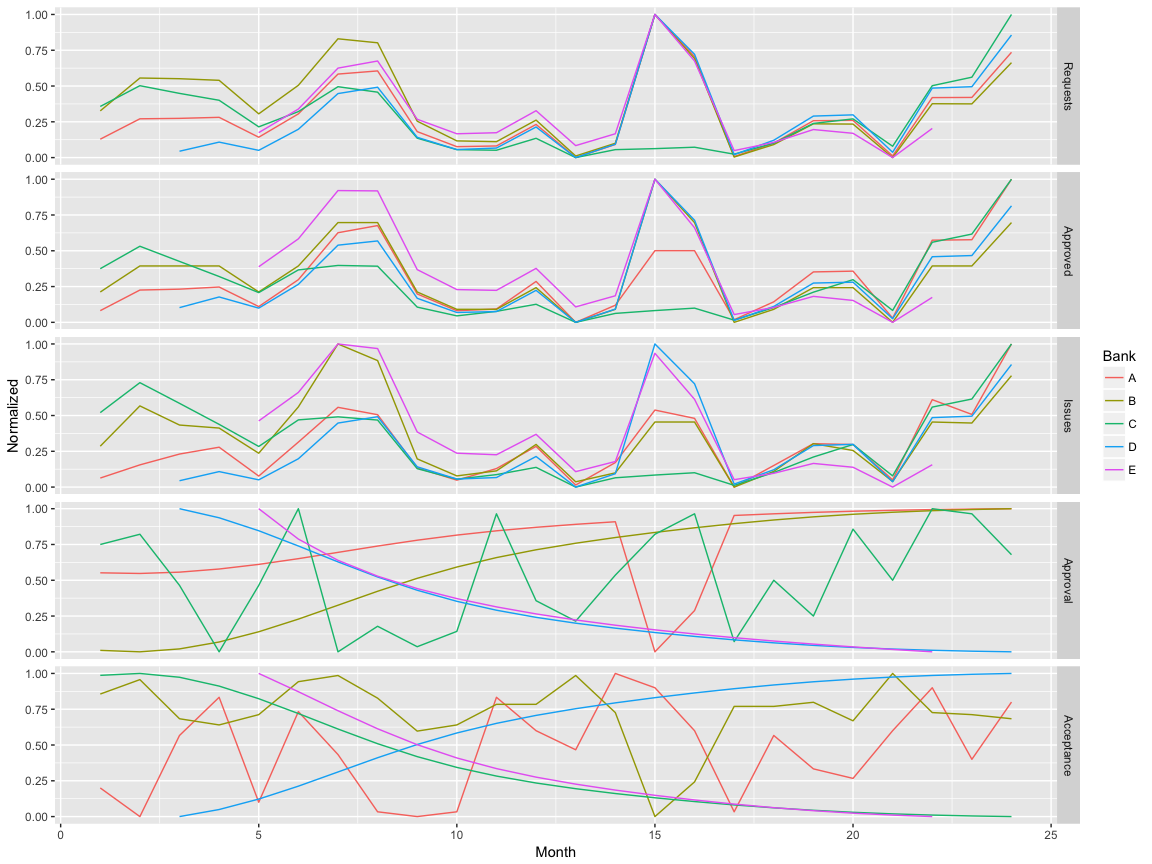
## Kruskal-Wallis rank sum test  
##   
## data: x and group  
## Kruskal-Wallis chi-squared = 98.3324, df = 4, p-value = 0  
##   
##   
## Comparison of x by group   
## (Bonferroni)   
## Col Mean-|  
## Row Mean | A B C D  
## ---------+--------------------------------------------  
## B | -6.475743  
## | 0.0000  
## |  
## C | -1.666715 4.809028  
## | 0.4779 0.0000  
## |  
## D | -5.060952 1.272450 -3.430872  
## | 0.0000 1.0000 0.0030  
## |  
## E | 2.464810 8.460184 4.007888 7.118304  
## | 0.0685 0.0000 0.0003 0.0000

Банки **B** и **D** действительно похожи, а предположительно закрытый банк **E** сильно отличается от всех остальных. Больше всего с ним статистически схож банк **A**.

## Kruskal-Wallis rank sum test  
##   
## data: x and group  
## Kruskal-Wallis chi-squared = 64.0496, df = 4, p-value = 0  
##   
##   
## Comparison of x by group   
## (Bonferroni)   
## Col Mean-|  
## Row Mean | A B C D  
## ---------+--------------------------------------------  
## B | 0.404470  
## | 1.0000  
## |  
## C | 2.997973 2.593502  
## | 0.0136 0.0475  
## |  
## D | -1.672621 -2.068201 -4.604697  
## | 0.4720 0.1931 0.0000  
## |  
## E | 5.787782 5.413315 3.012198 7.231602  
## | 0.0000 0.0000 0.0130 0.0000

По уровню согласия на одобренные условия банк **E** уступает всем остальным. Однако и в банке **C** клиенты берут одобренные кредиты статистически меньше, чем в банках **A**,**B** и **D**. Эти замечания являются поводом для пересмотра кредитной политики в банках **A** и **C**. Такая строгость при одобрении может послужить причиной потери лояльности клиентов и повторения судьбы банка **E**.

# Примечание

Так как данные представлены за каждый месяц, было бы полезно проанализировать их, как временные ряды. Возможно, исходя из анализа сезонности, мы смогли бы объяснить эти пики и их разнонаправленность (особенно на 15 месяце): 

Однако, так как в банках **D** и **E** отсутствует информация за несколько месяцев, это представляется бессмысленным, так как нет двух полных периодов.

**Стоит заметить, что этих данных не хватает, чтобы делать никакие выводы. Любая кредитная политика может быть эффективна и нет, её выбор должен происходить из анализа параметров района, макроэкономической ситуации и даже самого банка.**

**Так же было бы интересно сделать попарные сравнения первых и последних 12 месяцев для каждого банка, но у меня уже больше 500 слов.**