

Пример из лекции

1. Импортируем файл с исходными данными в проект (обращаем внимание на то, что заголовков в исходном файле нет. Это нужно учесть при импорте).

Upload a new dataset

SELECT THE DATA TO UPLOAD:

Выбрать файл | digits.csv

☐ This is the new version of an existing dataset

ENTER A NAME FOR THE NEW DATASET:

digits.csv

SELECT A TYPE FOR THE NEW DATASET:

Generic CSV File With no header (.nh.csv)

PROVIDE AN OPTIONAL DESCRIPTION:

2. В окне эксперимента перенесем на рабочую область датасет и блок **Principal Component Analysis**. Определим следующие параметры

Выбор колонок

Количество ГК

Нормализация

Principal Component Analysis

Selected columns

Selected columns:
Launch the selector tool to make a selection

Launch column selector

Number of dimensions to ...

2

☒ Normalize dense colu...

Web Service Parameters

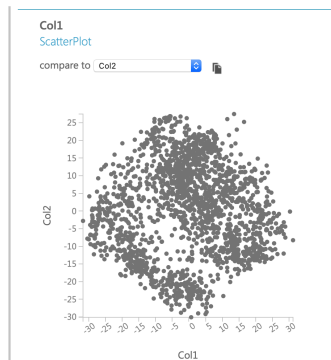
undefined

Нам нужны все колонки, строить будем две главные компоненты и будем использовать центрирование данных (центрирование включено в нормализацию).

3. Запустим эксперимент и проанализируем полученные результаты при помощи пункта контекстного меню visualize. Новые координаты (первые две ГК) отображаются в соответствующих столбцах. Представим объекты в виде точек на плоскости

rows	columns
1797	2

Col1	Col2
1.259466	-21.274883
-7.957611	20.768699
-6.991923	9.955986
15.906105	-3.332464
-23.306867	-4.269061
14.087086	-7.914448
-21.36341	-5.28834
2.952607	21.071664
5.255134	-1.18336
5.480199	-8.076324
-11.215079	-16.919862
-3.009182	11.99481
2.414193	4.851664
73.755618	1.906755



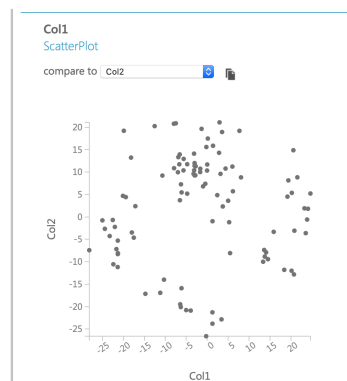
Из-за того, что объектов почти 1800, визуализация не самая удачная.

4. Отберем, например, первые 100 объектов при помощи блока **Apply SQL Transformation**, которым напомним SQL запрос вида:
- ```
select * from t1 limit 0, 100
```

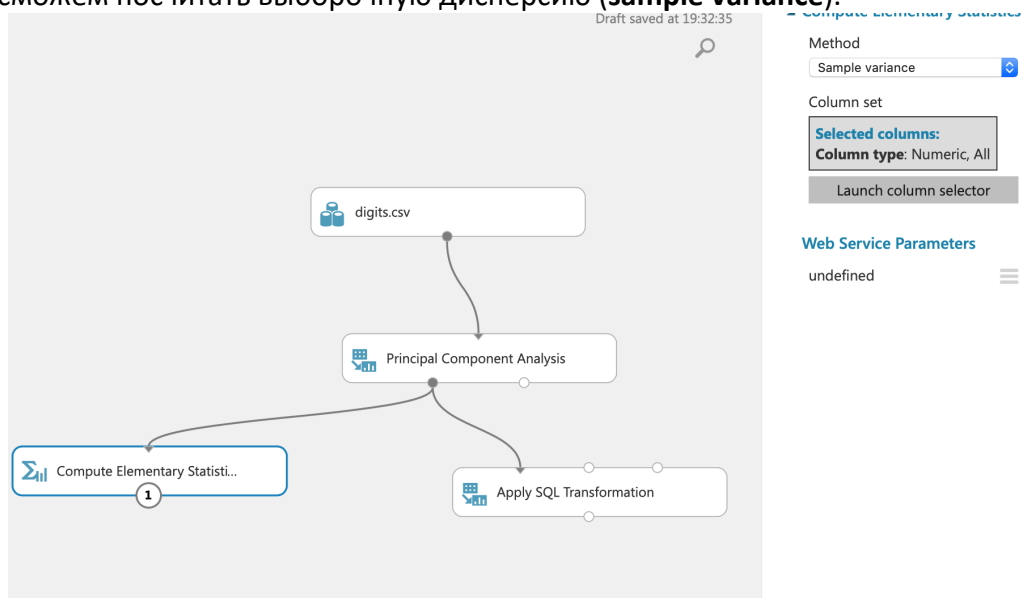
| rows | columns |
|------|---------|
| 100  | 2       |

| Col1       | Col2       |
|------------|------------|
| 1.259466   | -21.274883 |
| -7.957611  | 20.768699  |
| -6.991923  | 9.955986   |
| 15.906105  | -3.332464  |
| -23.306867 | -4.269061  |
| 14.087086  | -7.914448  |
| -21.36341  | -5.28834   |
| 2.952607   | 21.071664  |
| 5.255134   | -1.18336   |
| 5.480199   | -8.076324  |
| -11.215079 | -16.919862 |
| -3.009182  | 11.99481   |
| 2.414193   | 4.851664   |
| 23.255618  | 1.906755   |



5. Для нахождения доли объясненной дисперсии нам удобнее будет построить все 64 ГК и использовать блок **Compute Elementary Statistics**, при помощи которого мы сможем посчитать выборочную дисперсию (**sample variance**).



6. К сожалению, дальнейшие вычисления в Azure не выполнить, поэтому осуществим их, например, в MS Excel. Скопируем значения выборочных дисперсий, заменим точки на запятые и расположим значения в виде столбца. Найдем относительную выборочную дисперсию по каждой ГК как частное выборочной дисперсии каждой

ГК и суммы всех выборочных дисперсий. Найдем долю объясненной дисперсии в зависимости от количества используемых ГК.

| Номер | Выб. Дисперсия | Часть | Доля объясненной дисперсии | сумма выб. Дисперсий |
|-------|----------------|-------|----------------------------|----------------------|
| 1     | 179,007        | 0,149 | 0,149                      | 1202,147712          |
| 2     | 163,718        | 0,136 | 0,285                      |                      |
| 3     | 141,788        | 0,118 | 0,403                      |                      |
| 4     | 101,100        | 0,084 | 0,487                      |                      |
| 5     | 69,513         | 0,058 | 0,545                      |                      |
| 6     | 59,109         | 0,049 | 0,594                      |                      |
| 7     | 51,885         | 0,043 | 0,637                      |                      |
| 8     | 44,015         | 0,037 | 0,674                      |                      |
| 9     | 40,311         | 0,034 | 0,707                      |                      |
| 10    | 37,012         | 0,031 | 0,738                      |                      |
| 11    | 28,519         | 0,024 | 0,762                      |                      |
| 12    | 27,321         | 0,023 | 0,785                      |                      |
| 13    | 21,901         | 0,018 | 0,803                      |                      |
| 14    | 21,324         | 0,018 | 0,821                      |                      |

7. Построим график зависимости доли объясненной дисперсии в зависимости от количества используемых ГК.

