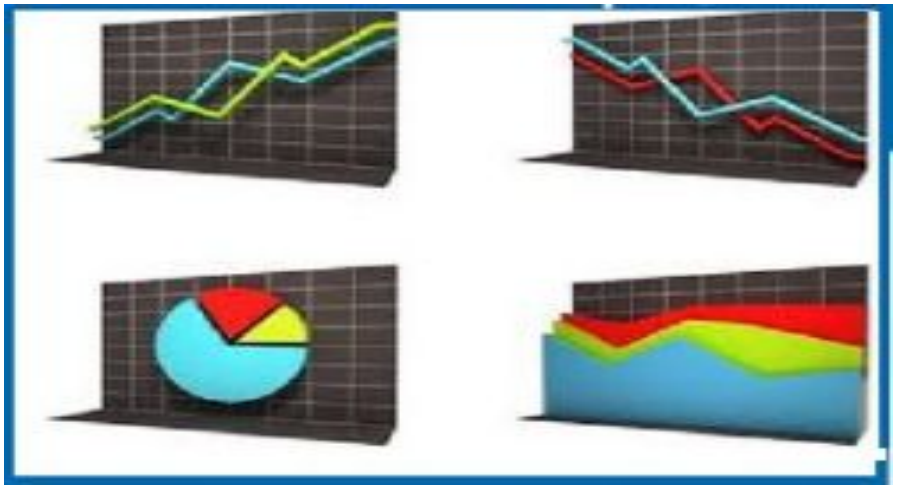
Визуализация данных

Графеева Н.Г.

2017

Визуализация данных

- К способам визуального или графического представления данных относят графики, диаграммы, схемы, карты и т.п.
- Визуализация традиционно рассматривалась как вспомогательное средство при анализе данных, однако в последнее время все больше исследований говорят о ее самостоятельной роли при анализе данных.



Применение методов визуализации позволяет:

- Представлять пользователю информацию в наглядном виде.
- Компактно описывать закономерности, присущие набору данных.
- Сжимать информацию.
- Обнаруживать пропуски в данных.
- Обнаруживать шумы и выбросы в данных.

Методы визуализации

Методы визуализации в зависимости от количества используемых измерений принято делить на две группы:

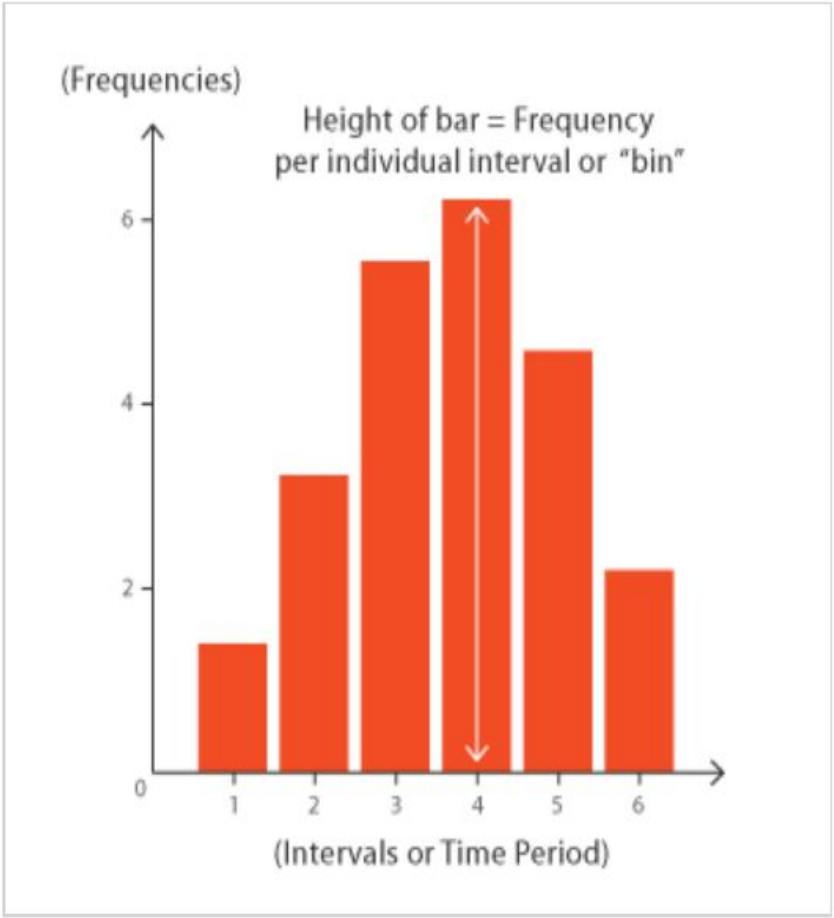
- Методы визуализаций для одного, двух и трех измерений.
- Методы визуализации для измерений больше трех.

Методы визуализации для одного, двух и трех измерений

К первой группе относятся достаточно хорошо известные способы визуализации. Однако среди них особо следует отметить двумерные изображения, как наиболее естественно воспринимаемые человеческим глазом.

Histogram

Гистограмма отображает частоту появления данных. Позволяет установить где концентрируются основные данные и увидеть выбросы.



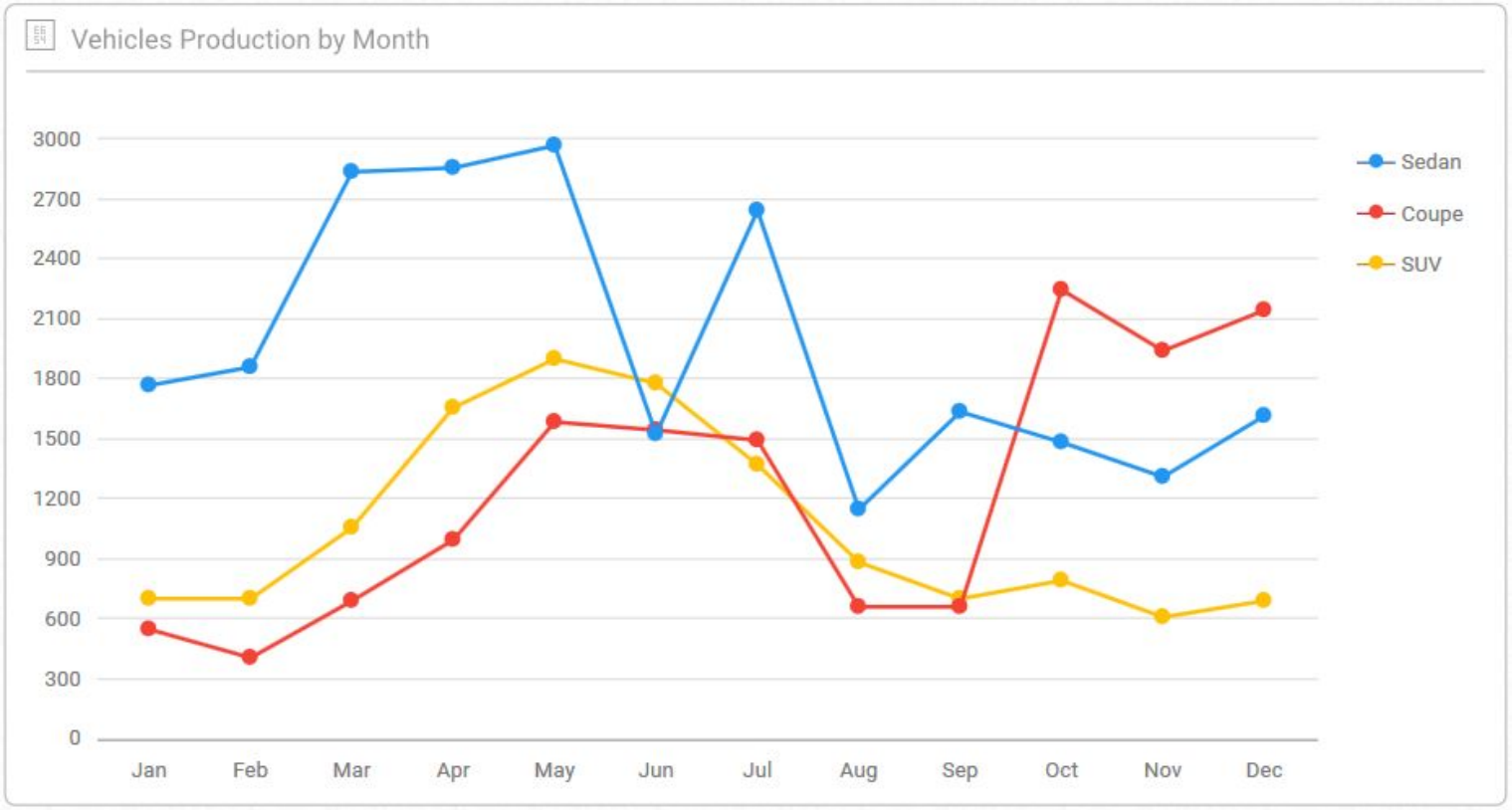
Word Cloud

Plato Republic
Yes
good
one
State
may
true
man
like
replied
must
justice
just
another
men
Certainly
knowledge
whether
now
nature
things
soul
mean
True
also
life
know
evil
unjust
think
sort
injustice
saying
nothing
city
every
better
greatest
honour
Thrasymachus
friend
hear
argument
always
youth
principle
see
tell
children
believe
little
kind
rulers
wisdom
much
anything
education
philosophy
desires
still
speaking
great
reason
found
mind
well
order
part
Well
world
neither
case
even
place
use
art
spirit
quite
made
thing
power
right
let
law
old
ever
Socrates
answer
far
suppose
whole
guardians
light
go
Now
yet
money
son
Let
human
away
pleasure
less
likely
women
among
body
words
make
therefore
virtue
pleasures
come
take
first
interest
never
two
class
certainly
consider
ask
else
either
others
opinion
agree
give
might
view
rather
truth
understand
many
way
gods
desire
end
without
able
manner
war
three
best
question
time
call
citizens
Glaucon
greater

Word Cloud

- ***Tag Clouds – синоним.***
- Метод визуализации, позволяющий отобразить частоту использования слов в тексте. Цвет может использоваться для разбивки слов на категории (по частоте использования). Не отображает точные значения, однако весьма удобен для восприятия.

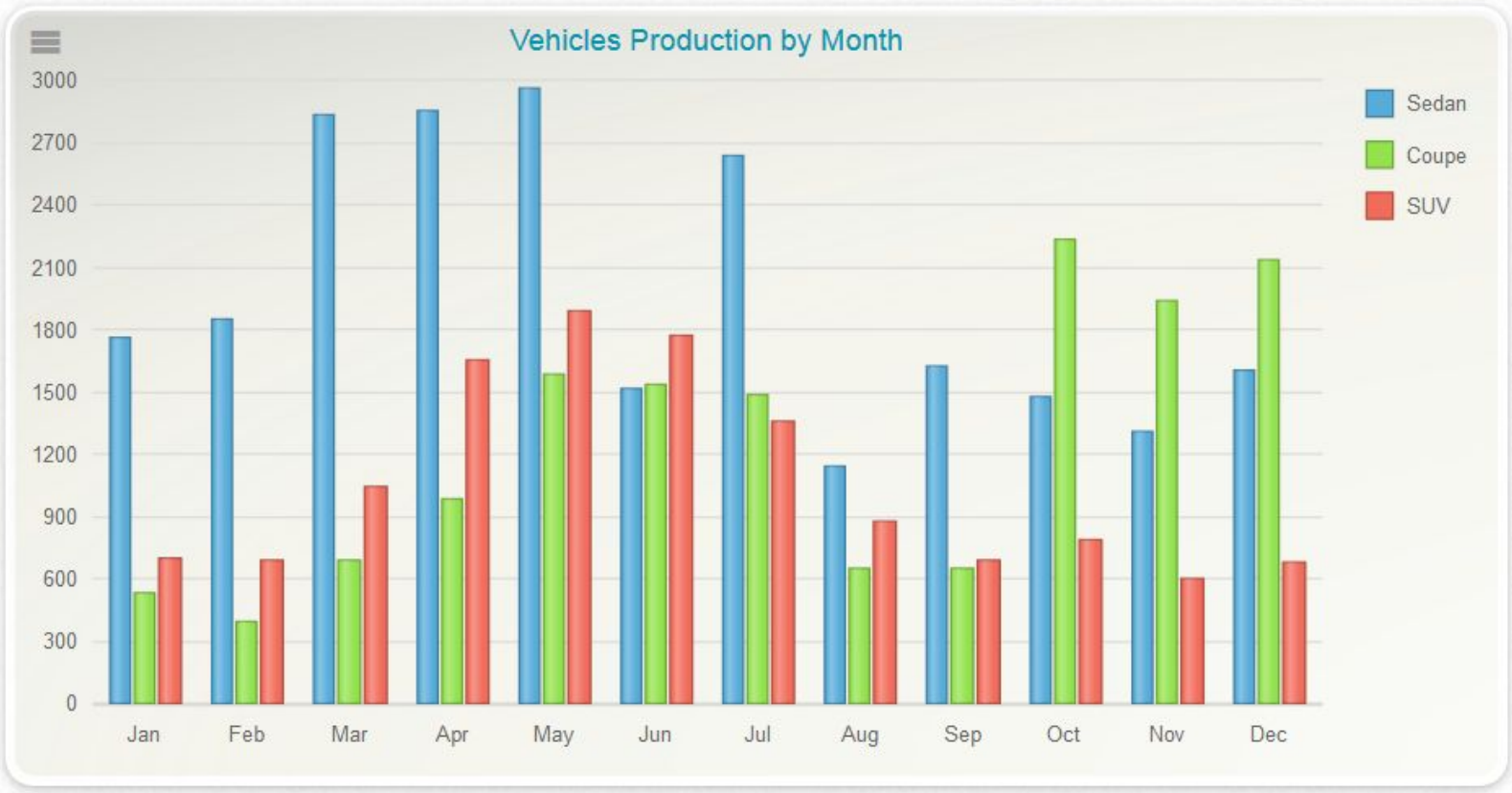
Line Graph


Vehicles Production by Month
3000
2700
2400
2100
1800
1500
1200
900
600
300
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Sedan
Coupe
SUV

Line Graph

- Линейные графики используются для отображения количественного значения в течение непрерывного интервала. Чаще всего он используется для отображения тенденций и отношений между категориями (при группировании с другими линиями). Линейные графики также помогают отобразить "картину в целом" за промежуток времени, чтобы увидеть, как она развивалась за этот период.
- При группировке нескольких линий необходимо отображать линии разными цветами и указывать в легенде какая линия чему соответствует.

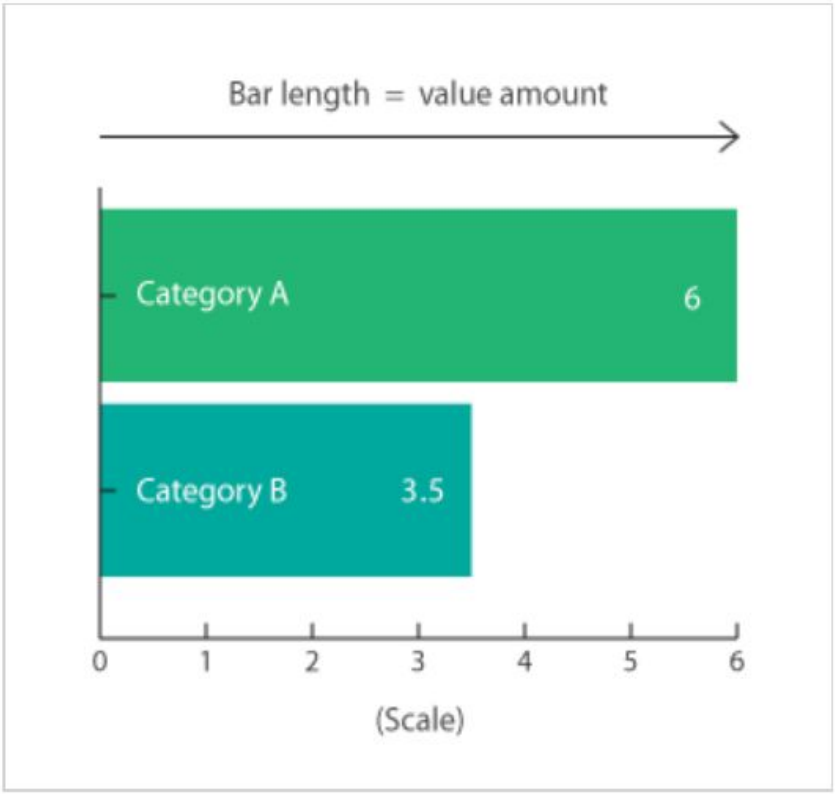
Bar chart


Vehicles Production by Month
3000
2700
2400
2100
1800
1500
1200
900
600
300
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Sedan
Coupe
SUV

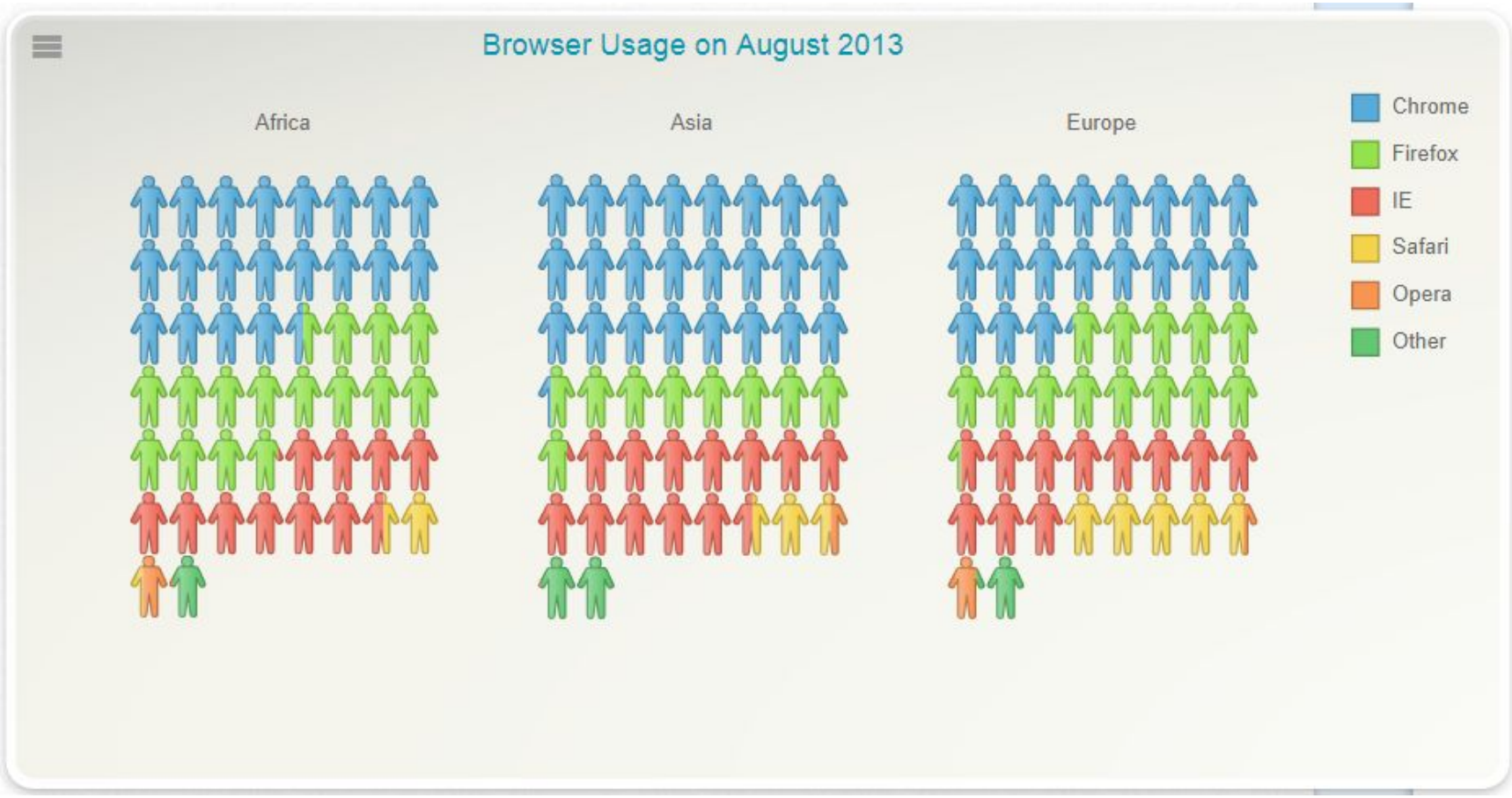
Bar Chart

- *Column Graph - синоним.*

- Bar Chart отображает различные категории (выделяя их цветом) и отвечает на вопрос “Как много” для каждой категории.
- Есть два варианта отображения категорий – вертикальная и горизонтальная.
- Категории выделяются цветом и идентифицируются легендой.

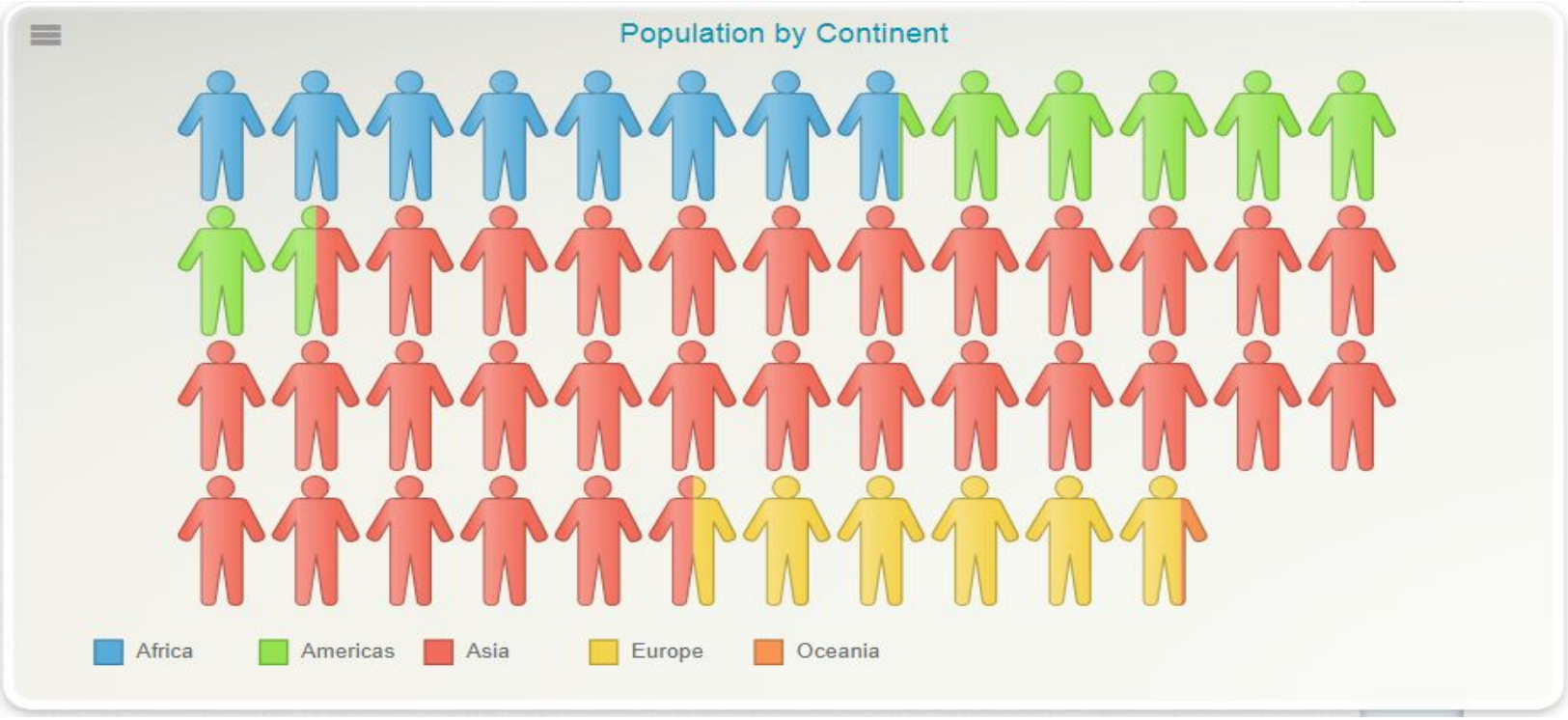


Pictograph

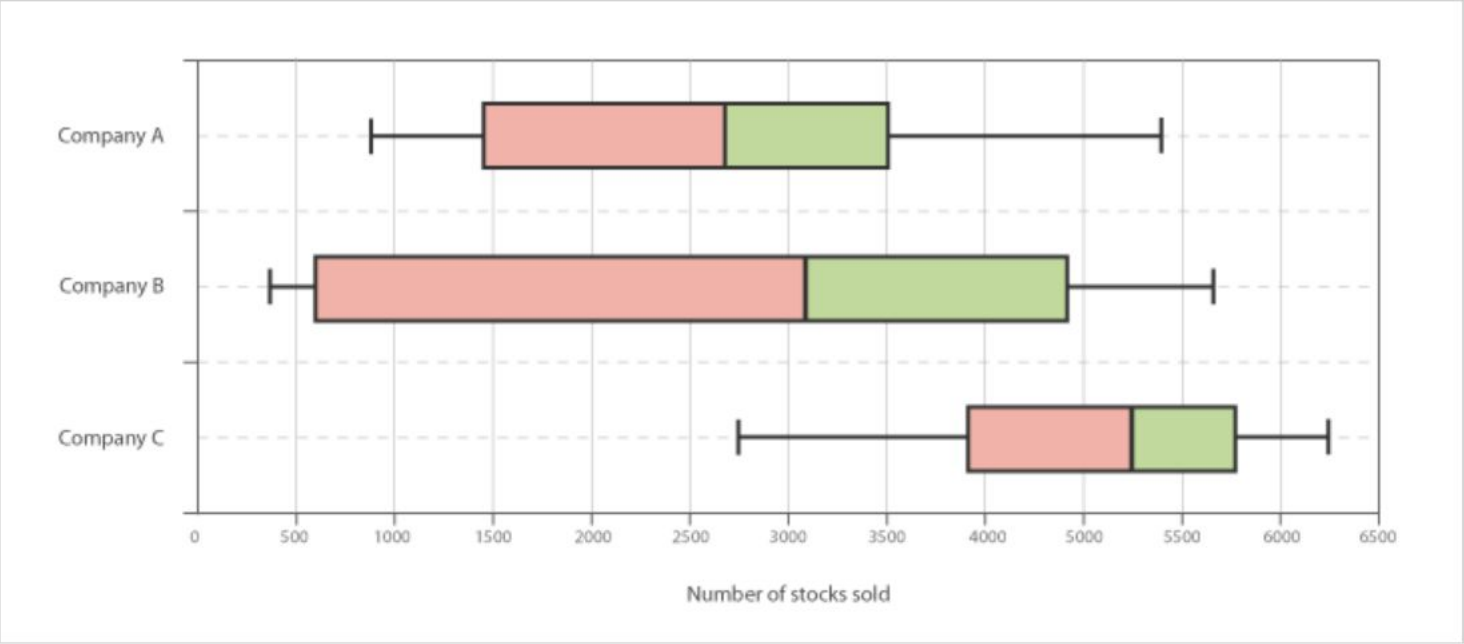

Browser Usage on August 2013
Africa
Asia
Europe
Chrome
Firefox
IE
Safari
Opera
Other

Pictograph

- Pictograph - это график, в котором значки, также известные как пиктограммы, представляют собой числа, чтобы сделать их более интересными и понятными. Все значки должны быть одинакового размера, а дроби обычно представляются частью значка. Каждый значок представляет процент от общего значения.

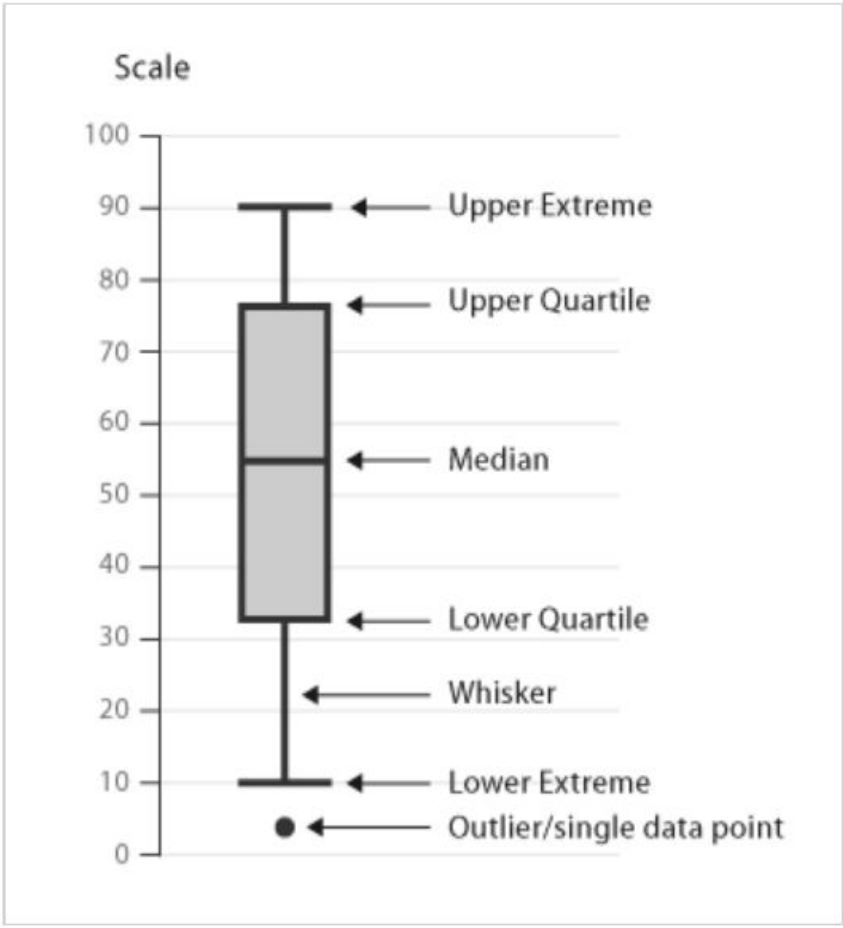


Box Plot

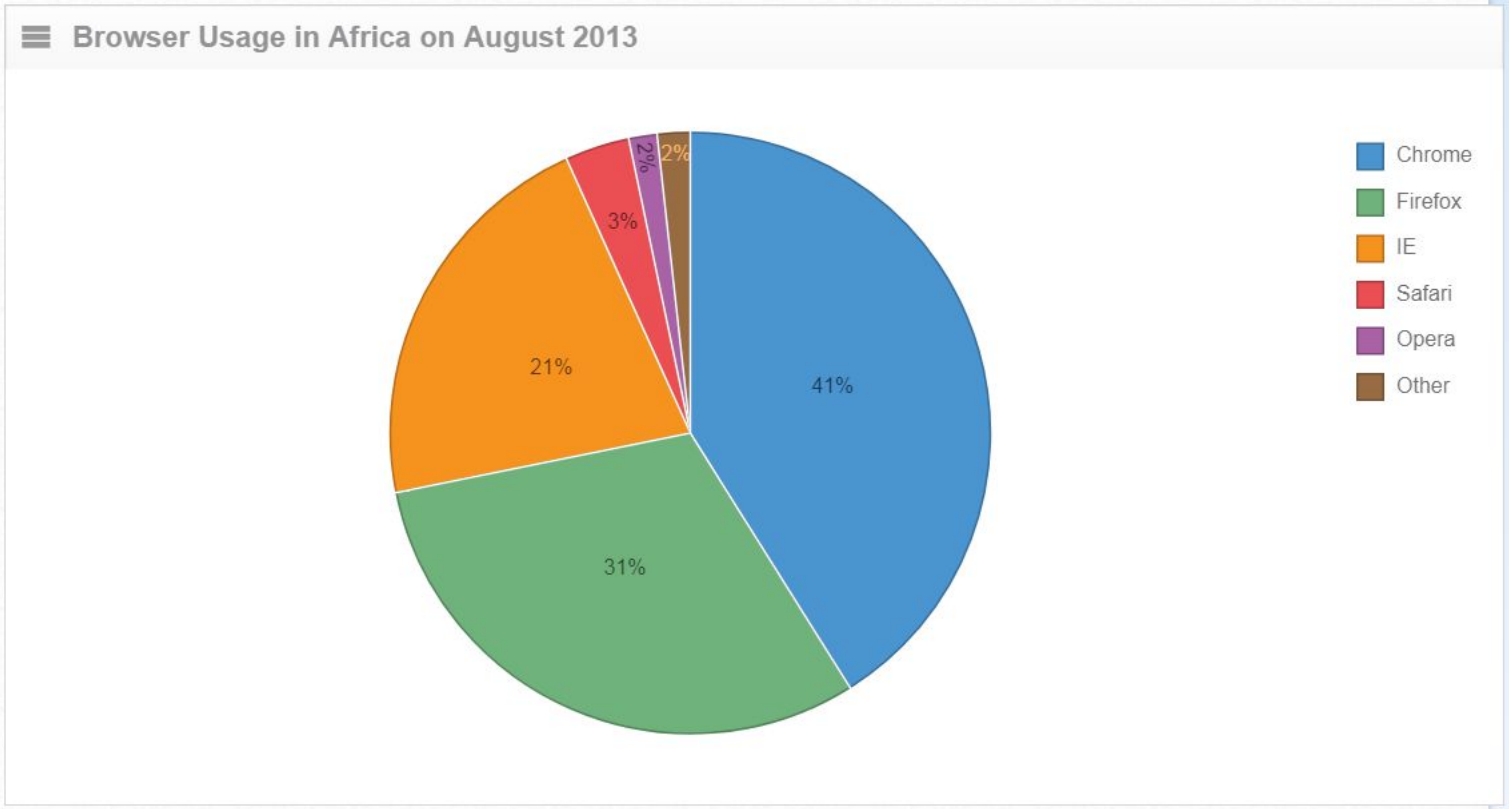

Company A
Company B
Company C
0
500
1000
1500
2000
2500
3000
3500
4000
5000
5500
6000
6500
Number of stocks sold

Box Plot

A Box Plot - удобный способ наглядного отображения групп числовых данных с помощью квартилей. Линии, идущие параллельно от коробок, известны как" усы", которые используются для обозначения изменчивости вне верхней и нижней квартилей. Окрестности иногда прорисовываются как отдельные точки, которые находятся на линии с усами. Коробки с усами могут быть нарисованы вертикально или горизонтально.

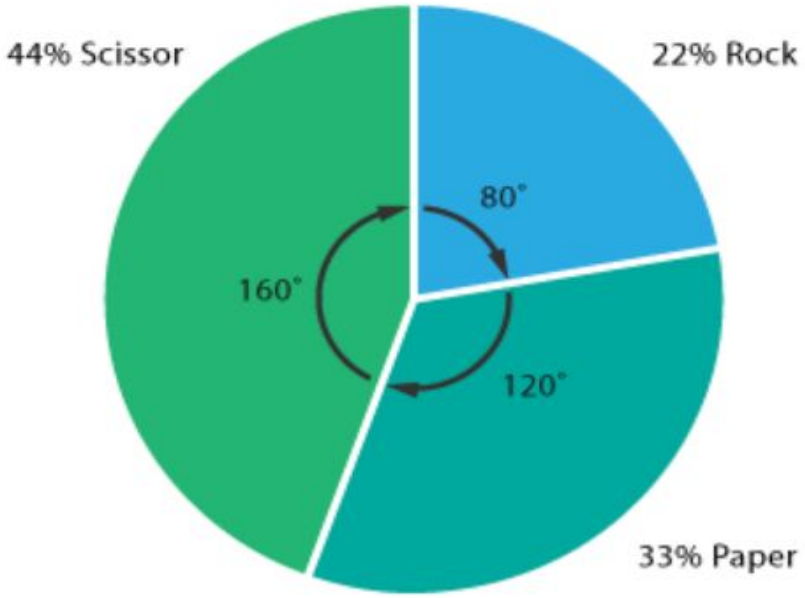


Pie Charts


Browser Usage in Africa on August 2013
41%
31%
21%
3%
2%
2%
Chrome
Firefox
IE
Safari
Opera
Other

Pie Charts

Pie диаграммы помогают показать пропорции и процентные доли между категориями, разделяя круг на пропорциональные сегменты. Каждая длина дуги представляет собой долю каждой категории, а весь круг представляет собой сумму всех данных, равную 100%. Круговые диаграммы идеально подходят для представления о пропорциональном распределении данных. Основным недостатком круговых диаграмм можно считать то, что они не могут отображать больше, чем несколько значений, потому что по мере увеличения числа показанных значений размер каждого сегмента/среза становится меньше. Это делает их непригодными для больших объемов данных.



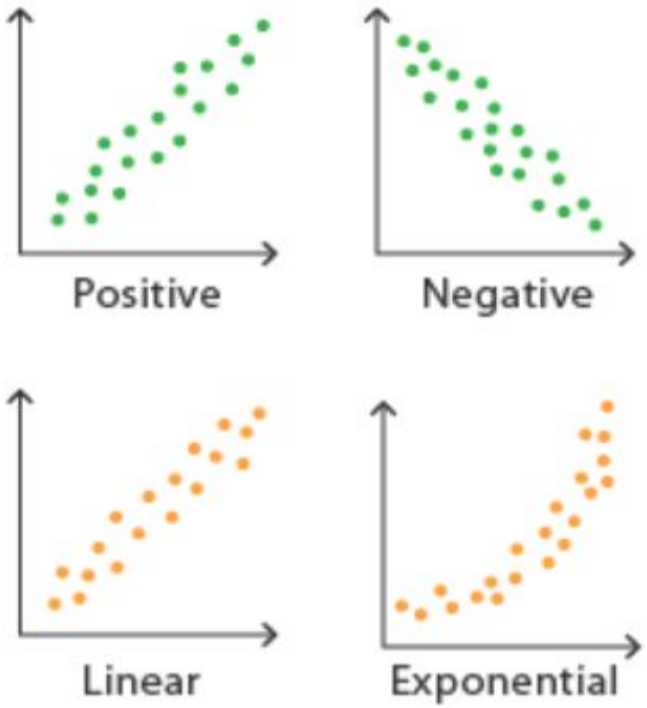
Data			
Rock	Paper	Scissor	TOTAL
2	3	4	9
To calculate percentages			
2/9=22%	3/9=33%	4/9=44%	100%
Degrees for each "pie slice"			
(2/9) x 360 = 80°	(3/9) x 360 = 120°	(4/9) x 360 = 160°	360°

Scatter plot

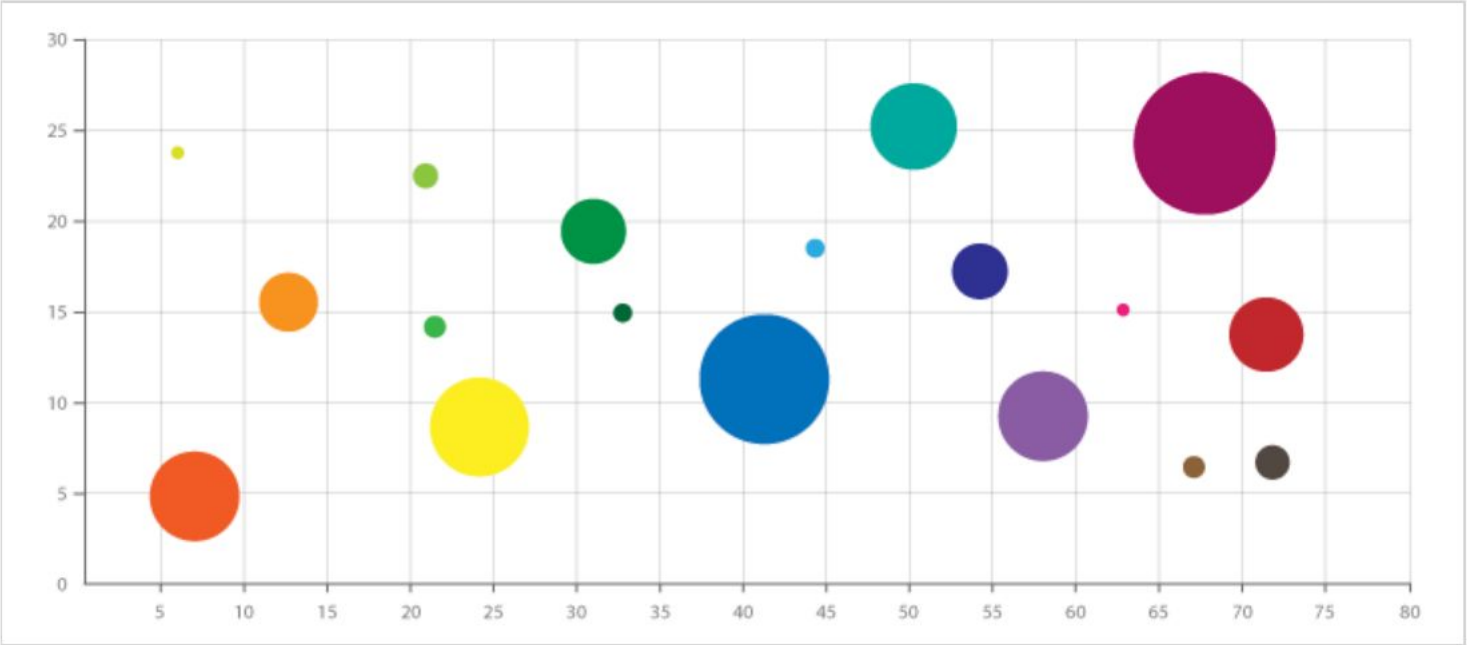
Scatter plot

Синонимы - *Scatter Graph, Point Graph, X-Y Plot, Scatter Chart или Scattergram*.

Диаграммы рассеивания используют декартовы координаты для отображения значений двух переменных. Такое отображение переменных по каждой оси позволяет визуально предположить, существует ли связь или корреляция между двумя переменными.



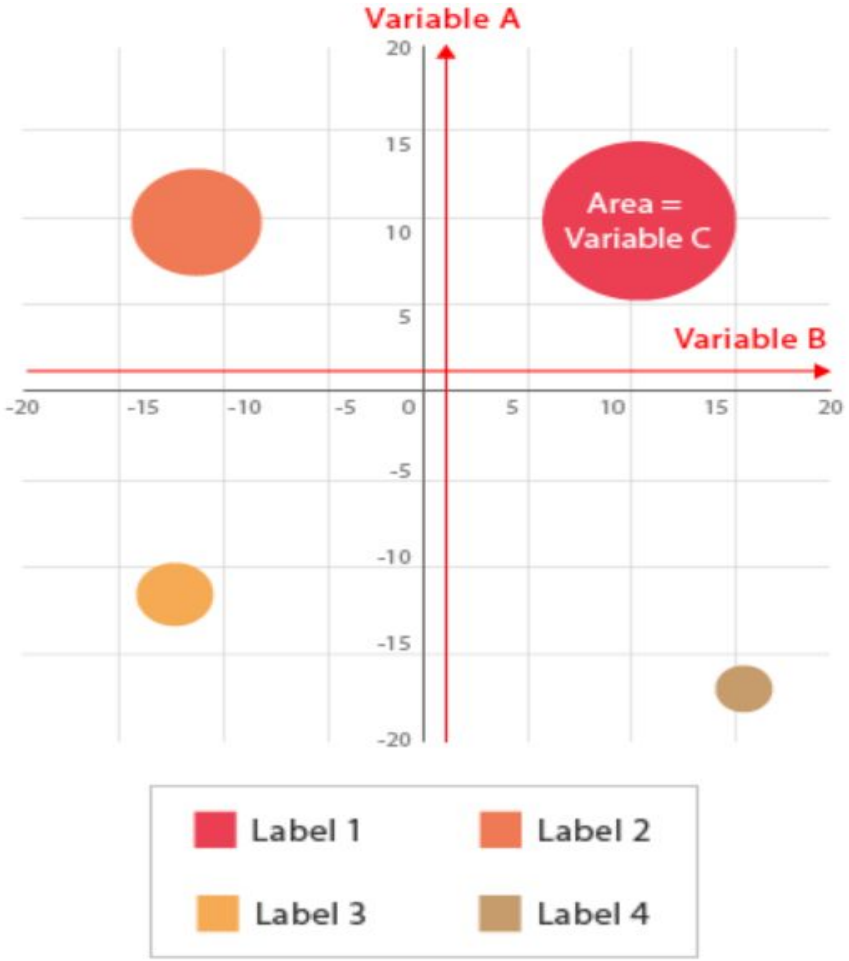
Bubble Chart



Bubble Chart

Пузырьковые диаграммы очень похожи на диаграммы рассеивания, так как каждая позиция пузыря определяется двумя координатами. Кроме того, размер окружности в каждой точке отражает дополнительное измерение. Из-за этого пузырьковые диаграммы позволяют проводить сравнение трех переменных, что позволяет легко визуализировать сложные взаимозависимости, которые не видны в диаграммах для двух переменных .

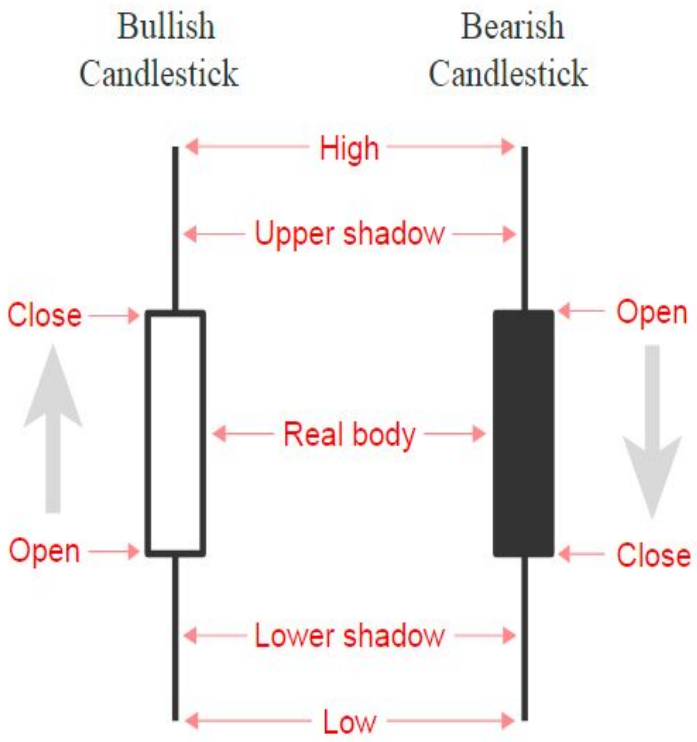
Цвета также могут использоваться для различения категорий или для представления дополнительной переменной.



Candlestick Chart

Candlestick Chart

Этот тип диаграммы используется в качестве инструмента для визуализации и анализа движения цены для ценных бумаг, производных, валюты, акций, облигаций и т. д. Диаграммы состоят из свечей, представляющих торговую деятельность за фиксированный период времени, и отображаю цену открытия, цену закрытия, минимальную и максимальную цену за этот период. Окраска используется для того, чтобы различать свечи, у которых цена открытия была больше цены закрытия и наоборот.

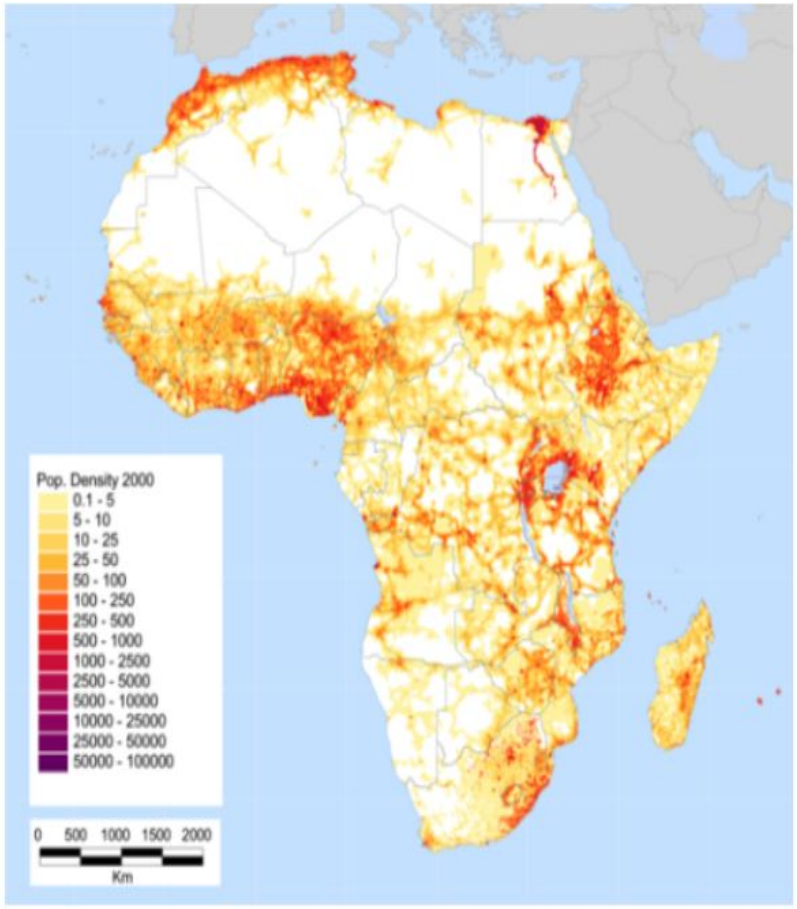


Тепловые карты

- Термин «**тепловая карта**» ввел разработчик программного обеспечения Кормак Кинни в 1991 г. в описании 2D дисплея, который показывал информацию финансового рынка в режиме реального времени.
- **Тепловые карты** – это тип визуализации, в которой цвет выступает в качестве дополнительного измерения. Тепловые карты позволяют увидеть важные переменные в цвете как функцию двух других переменных.

Тепловые карты

Плотность населения. Простейший пример цветовой карты, знакомый нам с детства – карта региона, на которой цветом показана плотность населения. Можно составить рейтинг регионов Африки по плотности населения, а можно визуализировать те же данные при помощи тепловой карты, которая наглядно покажет эту информацию.



Тепловые карты

Тепловая карта на службе таксистов. Это уже корпоративное использование тепловых карт – крупная служба такси **Uber** с помощью тепловых карт помогает своим водителям определить, где сейчас находится больше всего потенциальных клиентов. На карте города красным подсвечиваются зоны с наибольшим количеством заказов такси за последний час.

Источник: http://blog.uber.com/

Тепловые карты

Тепловые карты в таблице. Тепловые карты облегчают процесс восприятия больших массивов данных и необязательно связаны с отображением информации на географической карте. Ниже Вы видите, как выигрывает простая плоская таблица от добавления тепловой карты, и насколько облегчается первоначальное восприятие данных

Страна	Город	Январь	Февраль	Март	Апрель	Май	Июнь	Июль	Август	Сентябрь	Октябрь	Ноябрь	Декабрь
Австралия	Мельбурн	20,1	20,2	18,6	15,6	12,7	10,5	9,7	10,9	12,6	14,6	16,6	18,6
	Сидней	22,3	22,3	21,2	18,6	15,5	13,1	12,2	13,4	15,6	17,9	19,6	21,4
Австрия	Вена	0,3	1,7	5,8	10,8	15,6	18,6	20,8	20,3	15,5	10,4	5,2	1,1
	Инсбрук	-5,2	-3,7	0,2	3,4	7,8	10,8	12,8	12,7	9,3	4,8	-0,5	-4,2
	Клагенфурт	-7,2	-5,4	-1,3	2,8	7,8	11,1	12,9	12,7	9,0	4,3	-1,0	-5,2

Страна	Город	Январь	Февраль	Март	Апрель	Май	Июнь	Июль	Август	Сентябрь	Октябрь	Ноябрь	Декабрь
Австралия	Мельбурн	20,1	20,2	**18,6**	**15,6**	**12,7**	**10,5**	9,7	**10,9**	**12,6**	**14,6**	**16,6**	**18,6**
	Сидней	22,3	22,3	21,2	**18,6**	**15,5**	**13,1**	**12,2**	**13,4**	**15,6**	**17,9**	**19,6**	21,4
Австрия	Вена	0,3	1,7	5,8	**10,8**	**15,6**	**18,6**	20,8	20,3	**15,5**	**10,4**	5,2	1,1
	Инсбрук	-5,2	-3,7	0,2	3,4	7,8	**10,8**	**12,8**	**12,7**	9,3	4,8	-0,5	-4,2
	Клагенфурт	-7,2	-5,4	-1,3	2,8	7,8	**11,1**	**12,9**	**12,7**	9,0	4,3	-1,0	-5,2

Что делать с данными, имеющими более трех измерений?

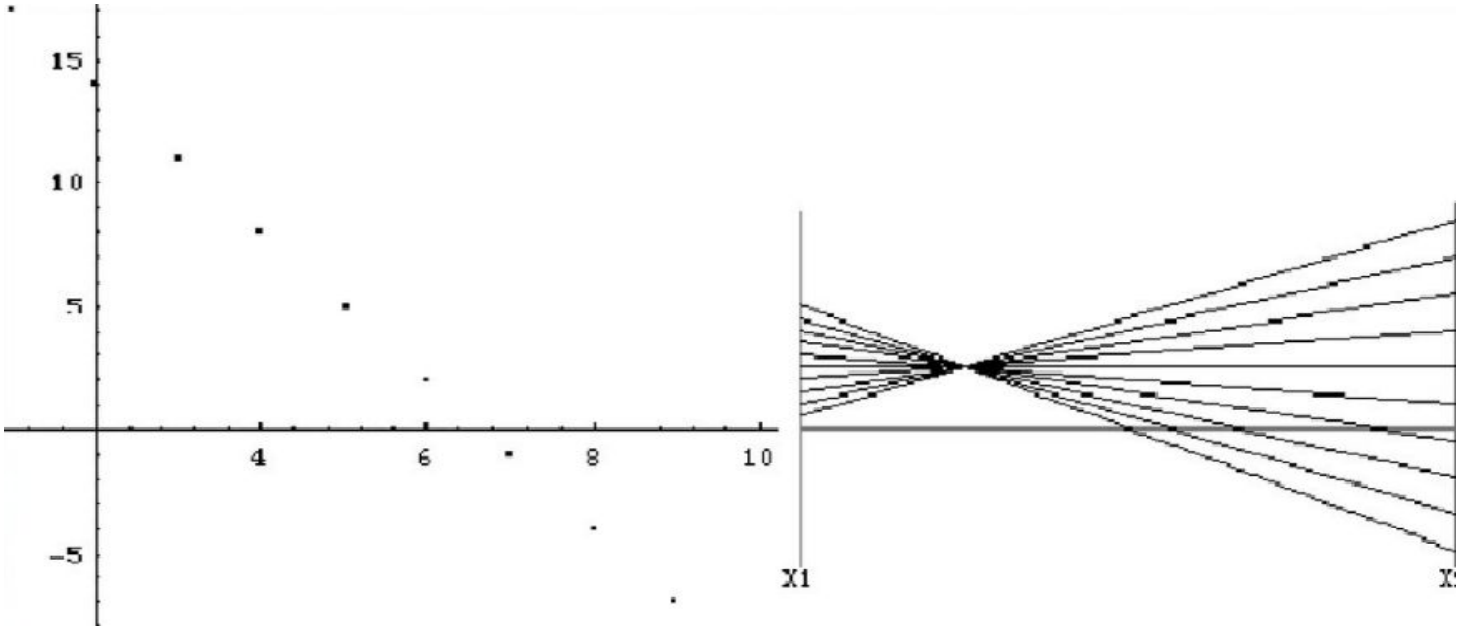
Если набор данных имеет более трех измерений, то существуют специальные методы визуализации или методы, понижающие размерность до 2 или 3 измерений. Такие методы существуют, в частности, факторный анализ. Рассмотрим некоторые из методов визуализации (факторный анализ сейчас рассматривать не будем).

Наиболее известные способы представления многомерных данных

- Параллельные координаты
- Радарные диаграммы
- Лица Чернова

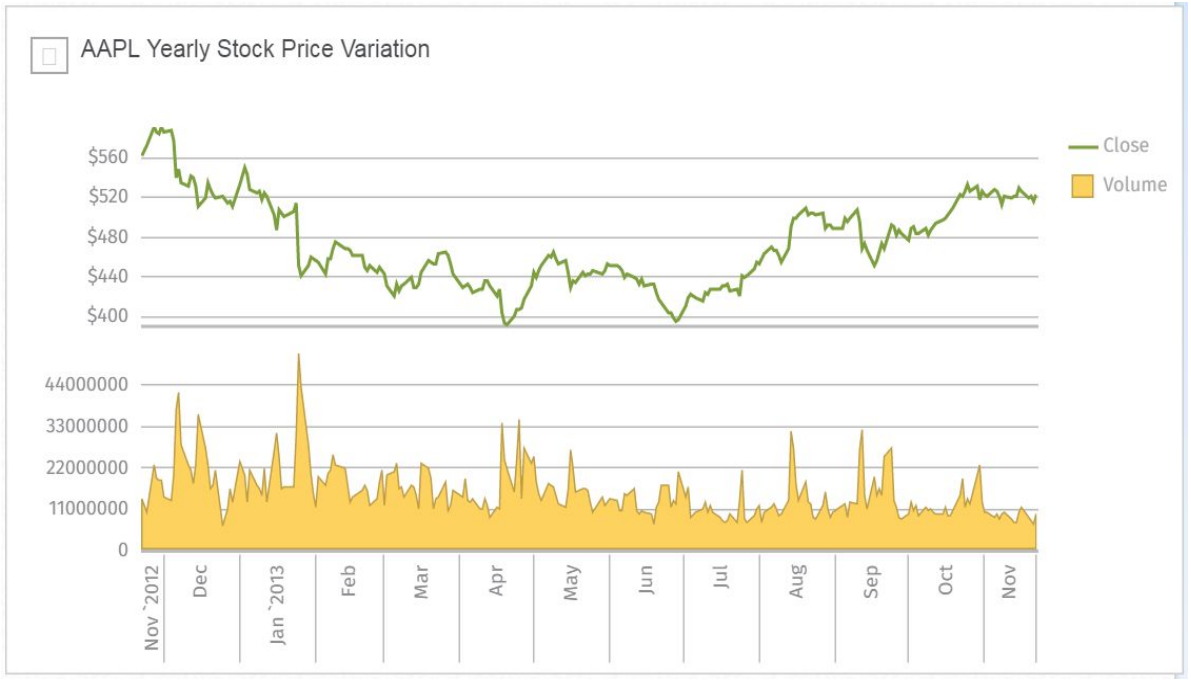
Параллельные координаты

В параллельных координатах график представляется как объединение двумерных проекций многомерного набора данных. Параллельные проекции могут отображаться как по вертикали, так и по горизонтали.

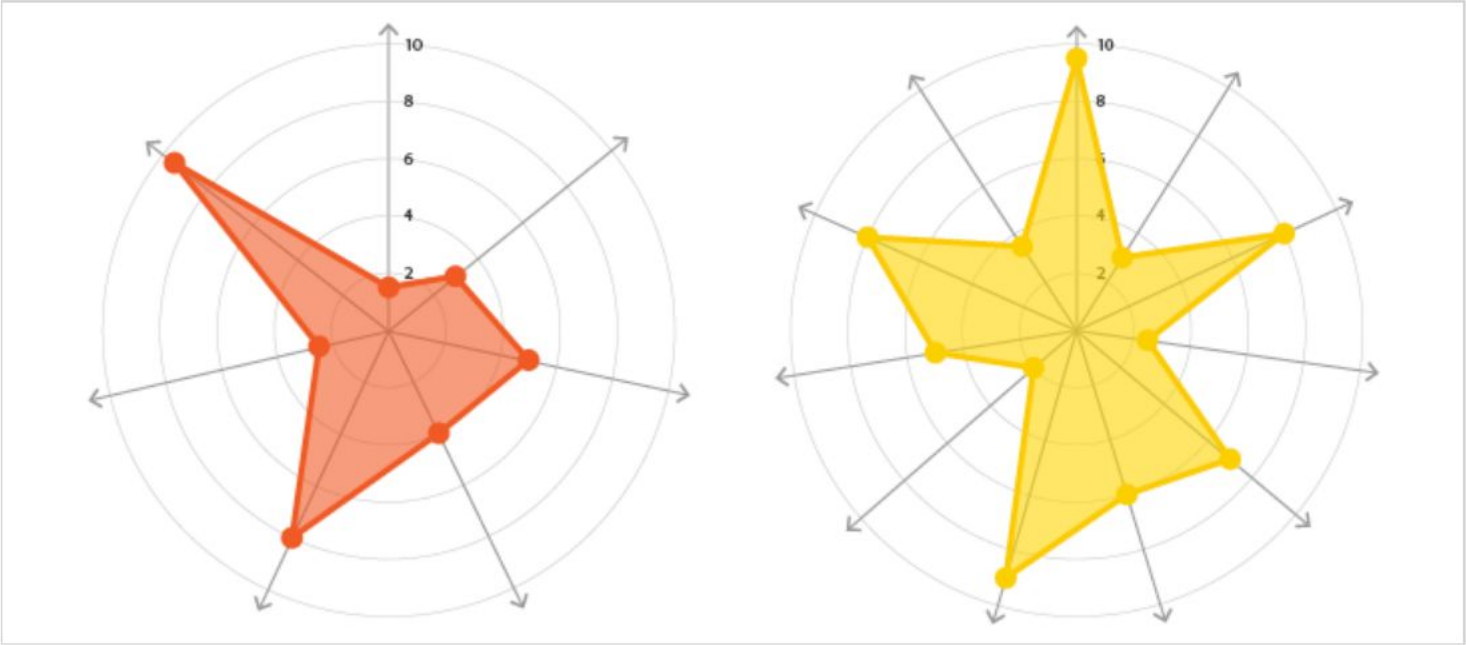


Параллельные координаты

Широко распространенный способ представления биржевых данных в виде составного графика (или графика с параллельными координатами). На одной проекции – время и цена сделки, на второй – время и объем. График можно было бы расширить еще двумя проекциями – время и количество поданных заявок на покупку и время и количество поданных заявок на продажу.



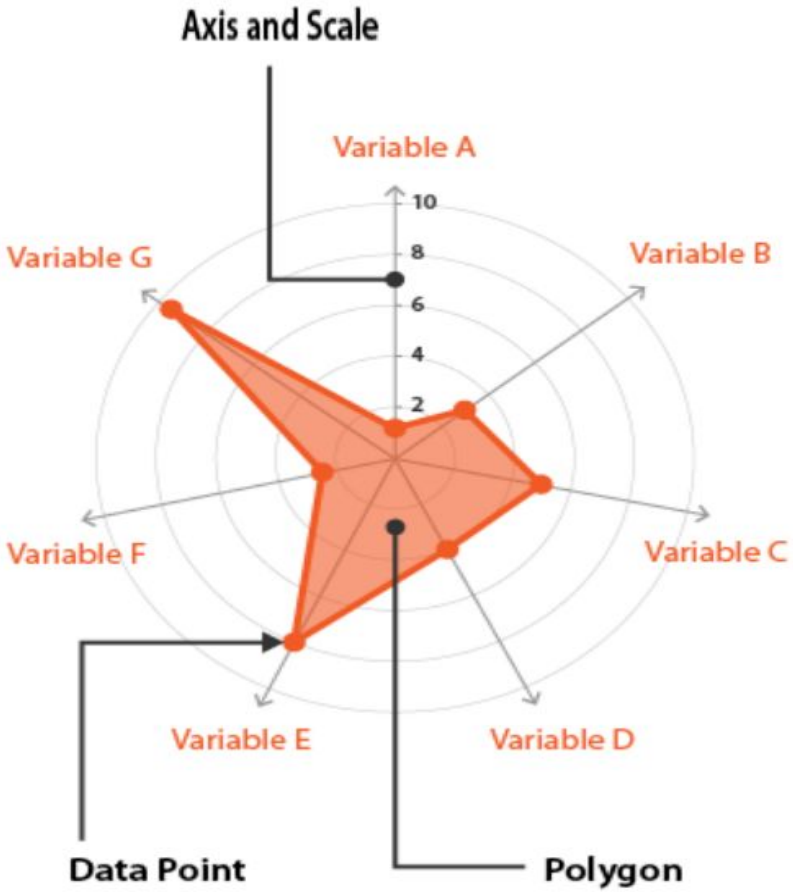
Radar Chart



Radar Chart

Радарные диаграммы-это способ сравнения значений нескольких количественных переменных (если они соизмеримы). Каждой переменной предоставляется ось, начинающаяся с центра. Все оси расположены радиально, с одинаковыми расстояниями между собой. В качестве направляющей часто используются линии сетки, соединяющиеся между осями. Каждое значение переменной прорисовывается вдоль своей отдельной оси. Все отложенные значения соединяются вместе, чтобы сформировать полигон.

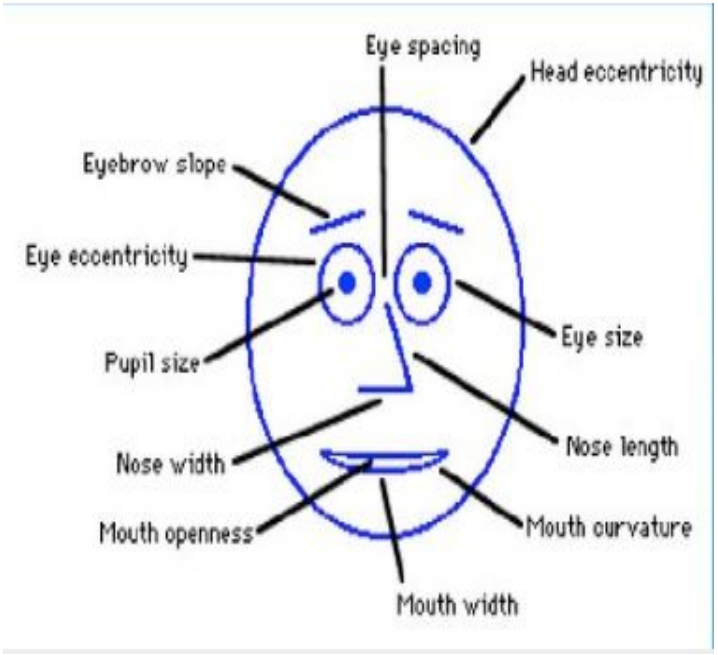
Для каждого наблюдения рисуется свой polygon.

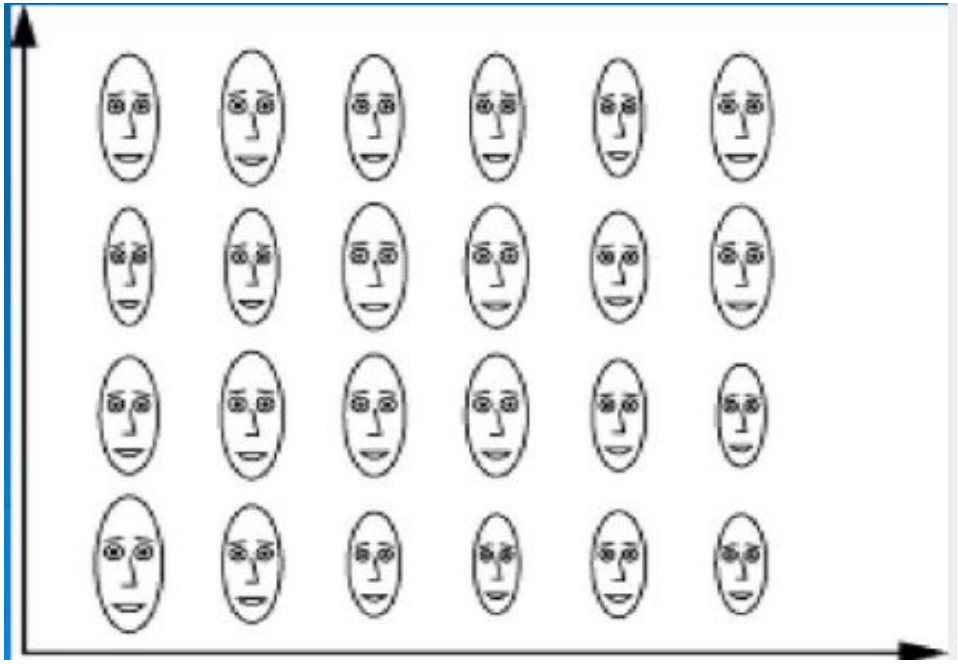


Лица Чернова

Основная идея – кодирование значений переменных в чертах человеческого лица. Для каждого наблюдения рисуется отдельное лицо. На каждом лице относительные значения переменных отображаются как размеры отдельных черт лица (например, длина и ширина носа, размер глаз, угол между бровями и т.п.). Такой анализ основан на способности человека интуитивно находить сходства и различия в чертах лица.

Пример (лица Чернова)



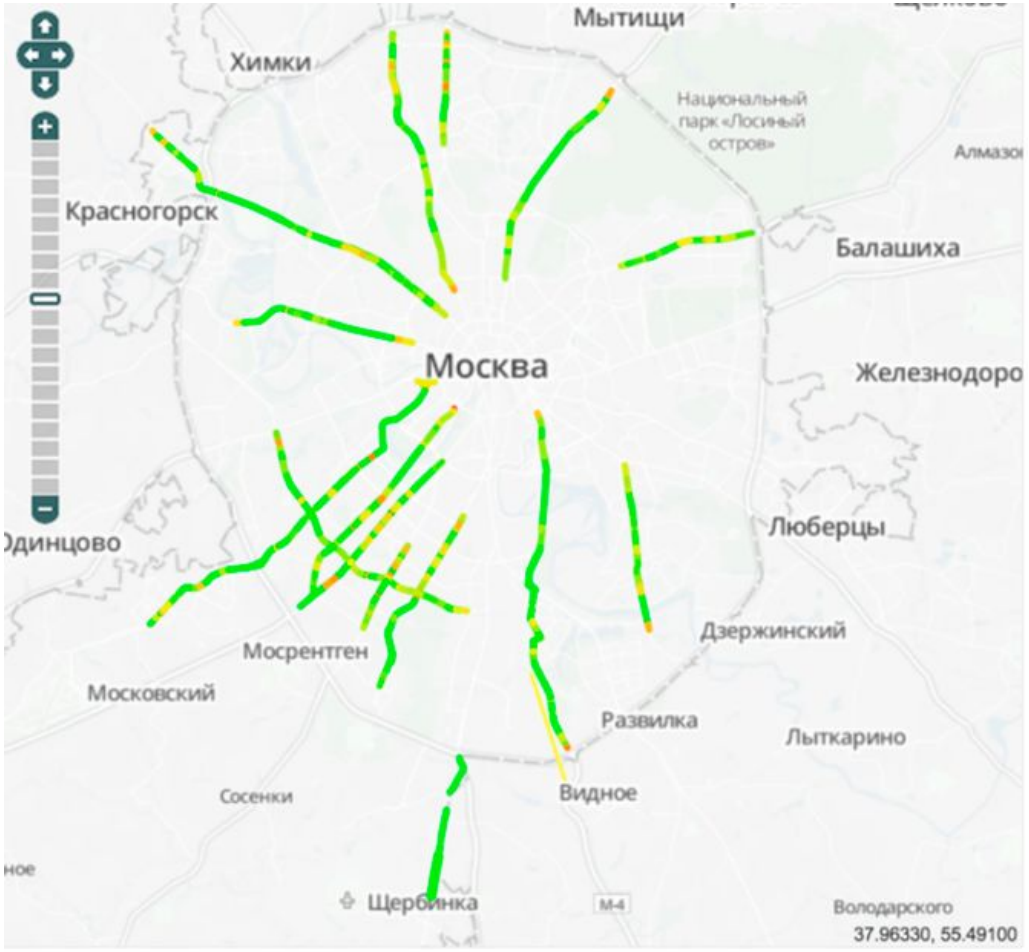


Пример использования (booking.com)

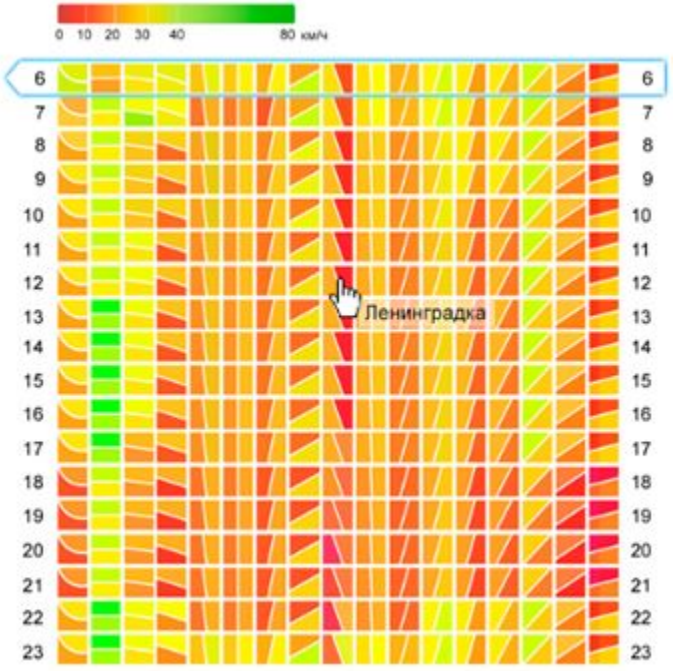
Другие способы визуализации

Нет никаких ограничений в способах представления информации. Существующие шаблоны в виде диаграмм и графиков – всего лишь начальные идеи. Главное – донести информацию в как можно более выразительном виде. Рассмотрим несколько примеров.

Стандартное представление транспортных потоков



Нестандартное представление транспортных потоков



Как голосует Америка

- https://www.nytimes.com/interactive/2016/06/10/upshot/voting-habits-turnout-partisanship.html

Как Трамп перекроил избирательную карту от побережья до побережья

- https://www.washingtonpost.com/graphics/politics/2016-election/election-results-from-coast-to-coast/

Ханс Рослинг: Самая лучшая статистика

https://ideanomics.ru/lectures/14772

20 лучших инструментов для визуализации

- https://freelance.today/poleznoe/20-luchshih-instrumentov-dlya-vizualizacii-dannyh.html

Задание 5

Визуализируйте какой-нибудь свой dataset в интерактивно-анимационной манере (примеры можно подсмотреть в GOOGLE CHART).

Примечание: Срок сдачи: 2 недели с момента выдачи. Задание отправлять по адресу: N.Grafeeva@spbu.ru.

Topic: DataMining_2018_job5