Data Science UA. Лекция 3. Exploratory Data Analysis

Январь 2017

Пару слов о домашнем задании

1. Где оно? :)
2. Не затягивайте до последнего дня.
3. Держите форму вопроса!
4. Качество данных - важно. Качество тех из них, по которым дается ответ на поставленный вопрос - критично.
5. Не выводите полные списки/DataFrame - пользуйтесь .head() и лимитами в списках.
6. Отделяйте исследование данных от непосредственно домашней работы.
7. Пользуйтесь функционалом языка.
8. Приводите финальный результат.

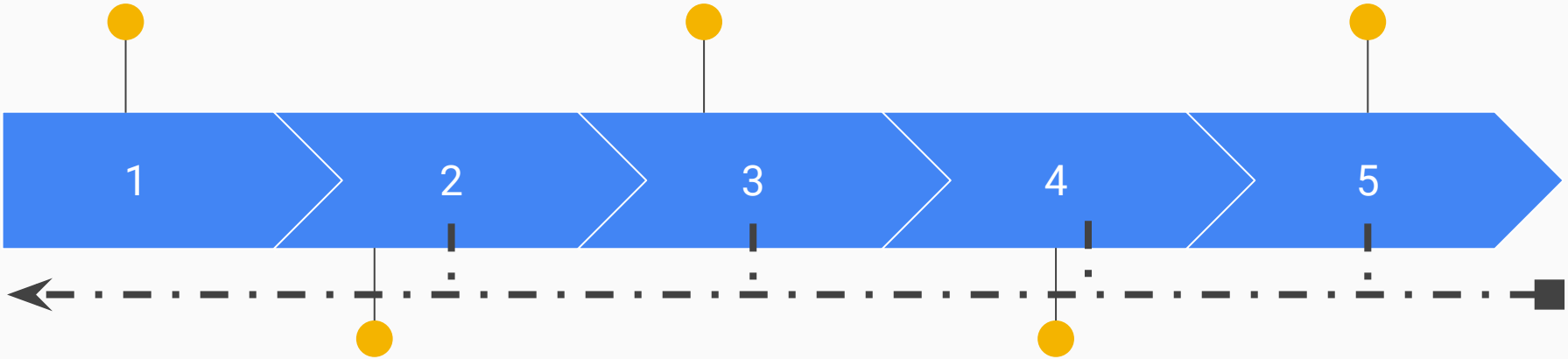
Исследование данных

Не всегда понятно что мы ищем… пока мы это не найдем.

Формулировка задачи

Анализ данных

Презентация, использование и контроль модели



Сбор и очистка данных

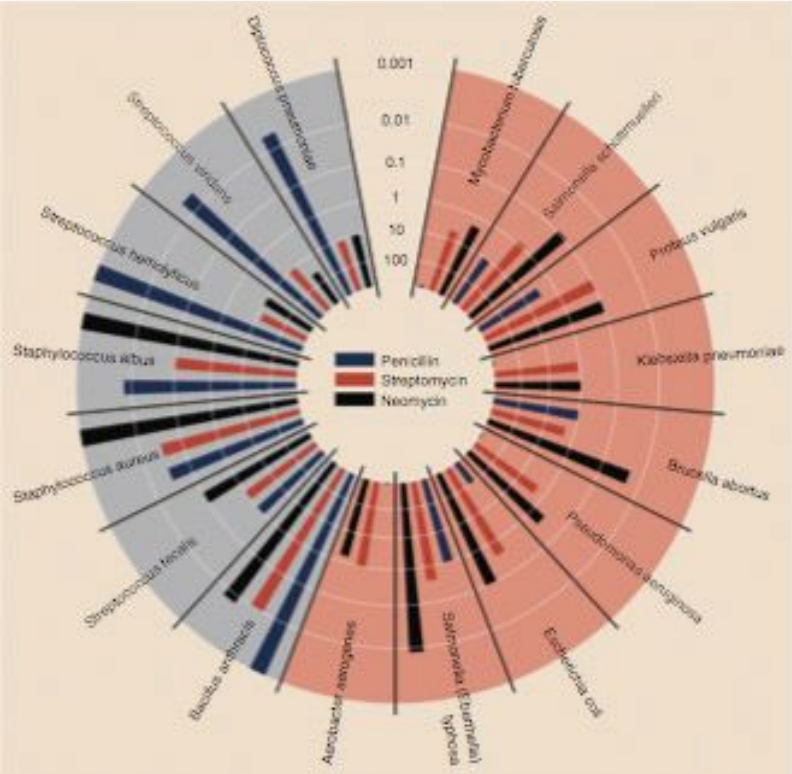
Построение модели данных

Антибиотики

Уилл Буртин, исследование 1951 года, незаданные вопросы

http://www.americanscientist.org/issues/pub/thats-funny

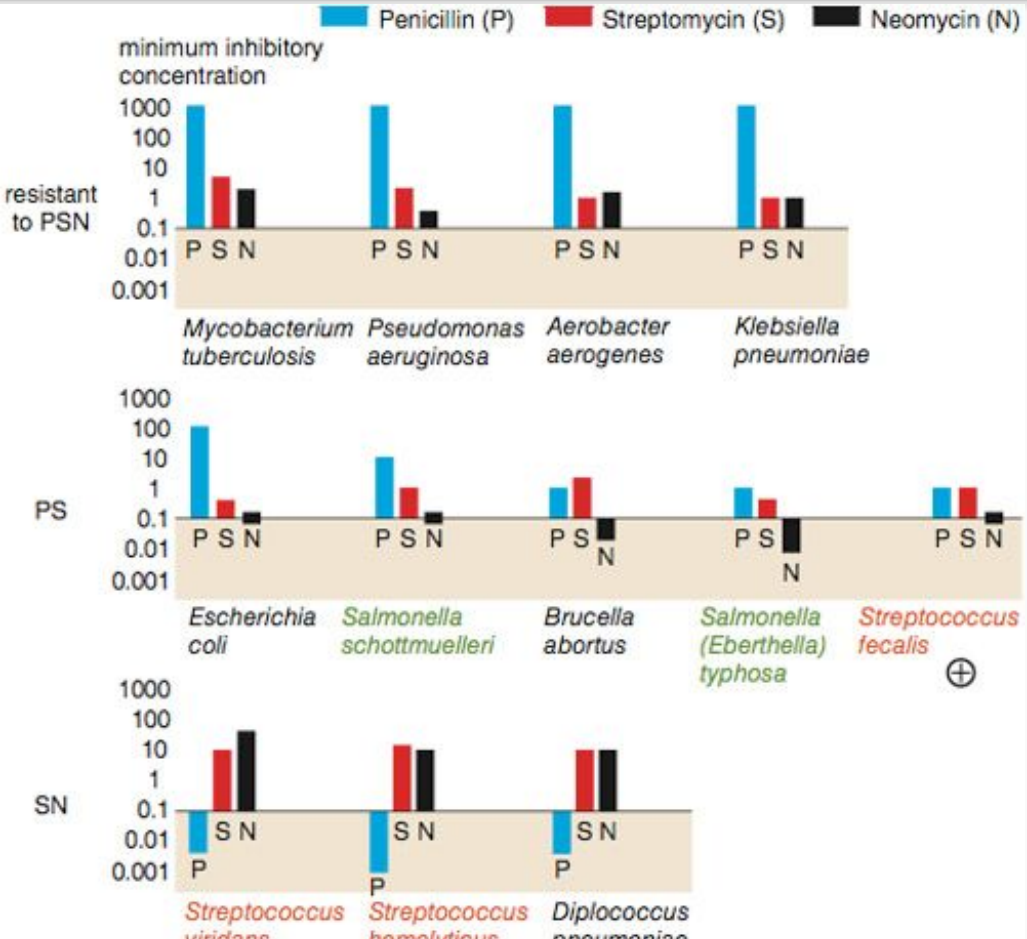
Эффективность антибиотиков



Bacteria	Penicillin	Antibiotic Streptomycin	Neomycin	Gram stain
Aerobacter aerogenes	870	1	1.6	–
Brucella abortus	1	2	0.02	–
Bacillus anthracis	0.001	0.01	0.007	+
Diplococcus pneumoniae	0.005	11	10	+
Escherichia coli	100	0.4	0.1	–
Klebsiella pneumoniae	850	1.2	1	–
Mycobacterium tuberculosis	800	5	2	–
Proteus vulgaris	3	0.1	0.1	–
Pseudomonas aeruginosa	850	2	0.4	–
Salmonella (Eberthella) typhosa	1	0.4	0.008	–
Salmonella schottmuelleri	10	0.8	0.09	–
Staphylococcus albus	0.007	0.1	0.001	+
Staphylococcus aureus	0.03	0.03	0.001	+
Streptococcus fecalis	1	1	0.1	+
Streptococcus hemolyticus	0.001	14	10	+
Streptococcus viridans	0.005	10	40	+

https://gist.github.com/borgar/cd32f1d804951034b224#file-bacteria-csv

That’s funny © Айзек Азимов



Величайшая ценность изображений в том, что они заставляют видеть то, чего вы не собирались увидеть.

Джон Таки

Цели визуализации

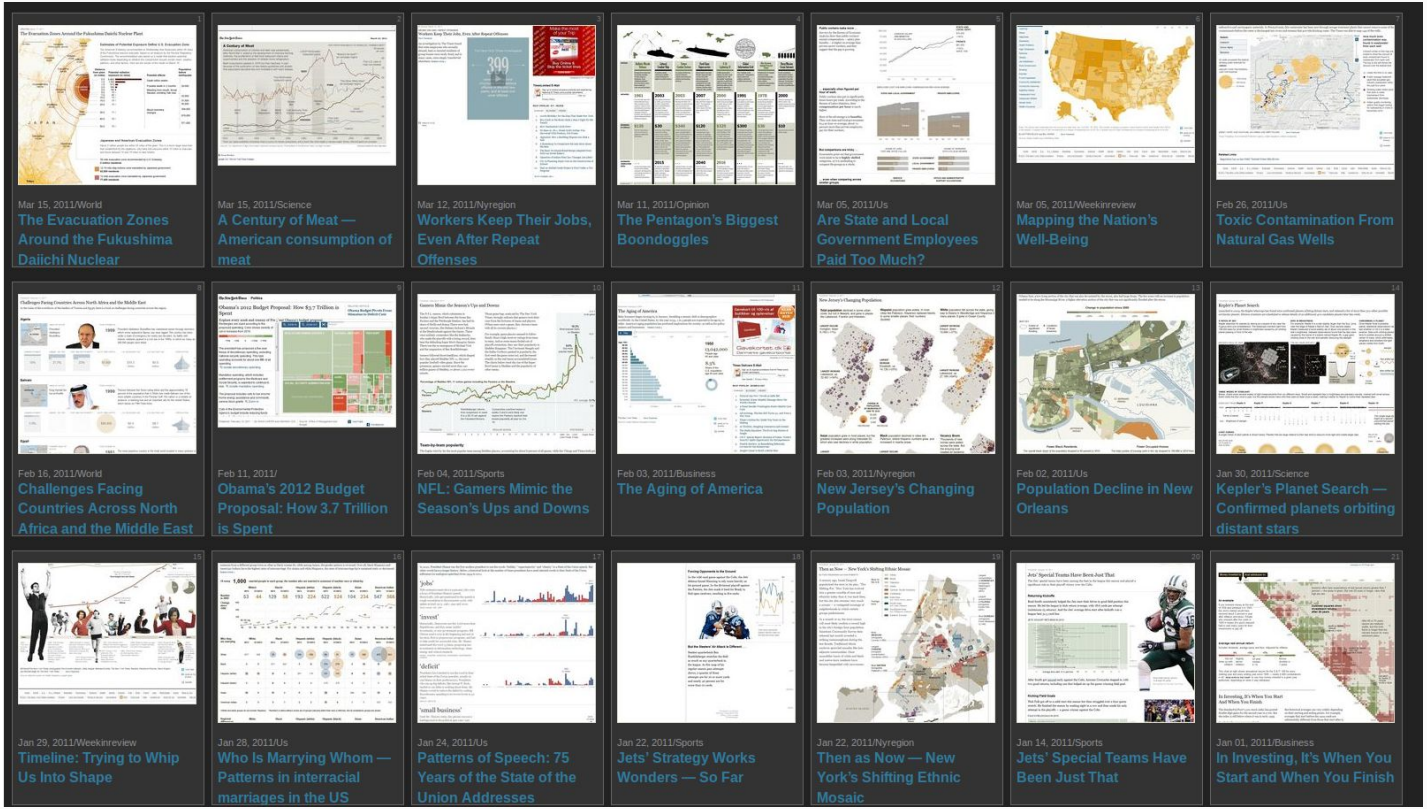
Коммуникативная (explanatory)

1. Представить данные и идеи
2. Донести и проинформировать
3. Поддержать и аргументировать
4. Повлиять и убедить

Исследовательская (exploratory)

1. Исследовать данные
2. Проанализировать ситуацию
3. Определить следующие шаги
4. Вынести решение по вопросу

New York Times



Mar 15, 2011/World
The Evacuation Zones Around the Fukushima Daiichi Nuclear

Mar 15, 2011/Science
A Century of Meat — American consumption of meat

Mar 12, 2011/Nyregion
Workers Keep Their Jobs, Even After Repeat Offenses

Mar 11, 2011/Opinion
The Pentagon's Biggest Boondoggles

Mar 05, 2011/Us
Are State and Local Government Employees Paid Too Much?

Mar 05, 2011/Weekinreview
Mapping the Nation's Well-Being

Feb 26, 2011/Us
Toxic Contamination From Natural Gas Wells

Feb 16, 2011/World
Challenges Facing Countries Across North Africa and the Middle East

Feb 11, 2011/
Obama's 2012 Budget Proposal: How $3.7 Trillion is Spent

Feb 04, 2011/Sports
NFL: Gamers Mimic the Season's Ups and Downs

Feb 03, 2011/Business
The Aging of America

Feb 03, 2011/Nyregion
New Jersey's Changing Population

Feb 02, 2011/Us
Population Decline in New Orleans

Jan 30, 2011/Science
Kepler's Planet Search — Confirmed planets orbiting distant stars

Jan 29, 2011/Weekinreview
Timeline: Trying to Whip Us Into Shape

Jan 28, 2011/Us
Who Is Marrying Whom — Patterns in interracial marriages in the US

Jan 24, 2011/Us
Patterns of Speech: 75 Years of the State of the Union Addresses

Jan 22, 2011/Sports
Jets' Strategy Works Wonders — So Far

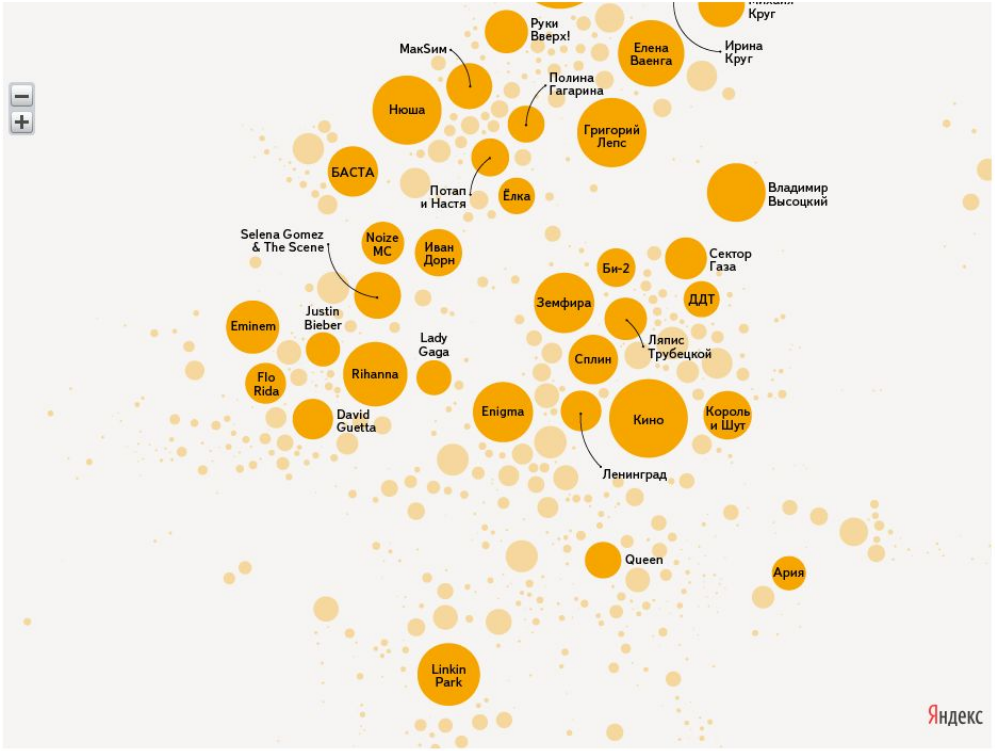
Jan 22, 2011/Nyregion
Then as Now — New York's Shifting Ethnic Mosaic

Jan 14, 2011/Sports
Jets' Special Teams Have Been Just That

Jan 01, 2011/Business
In Investing, It's When You Start and When You Finish

http://www.smallmeans.com/new-york-times-infographics/

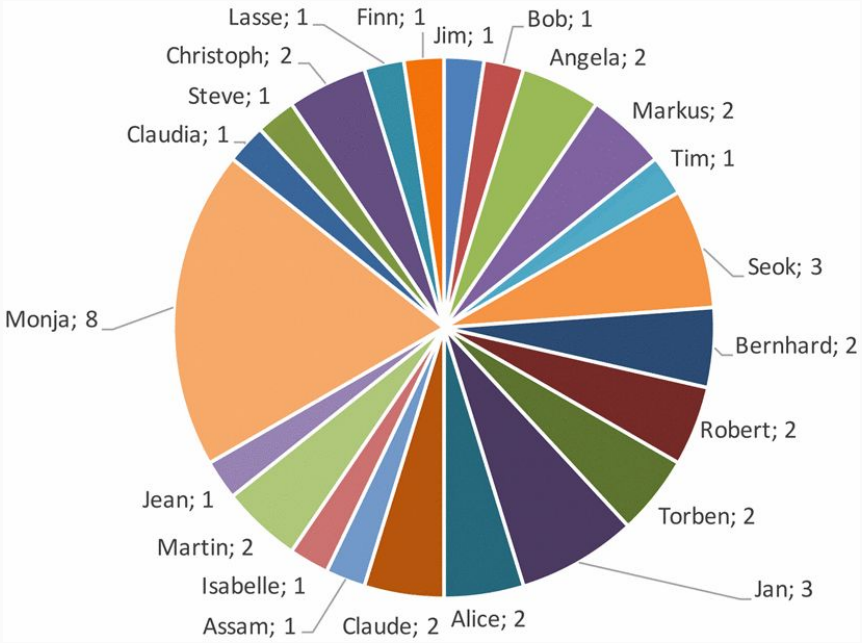
Яндекс

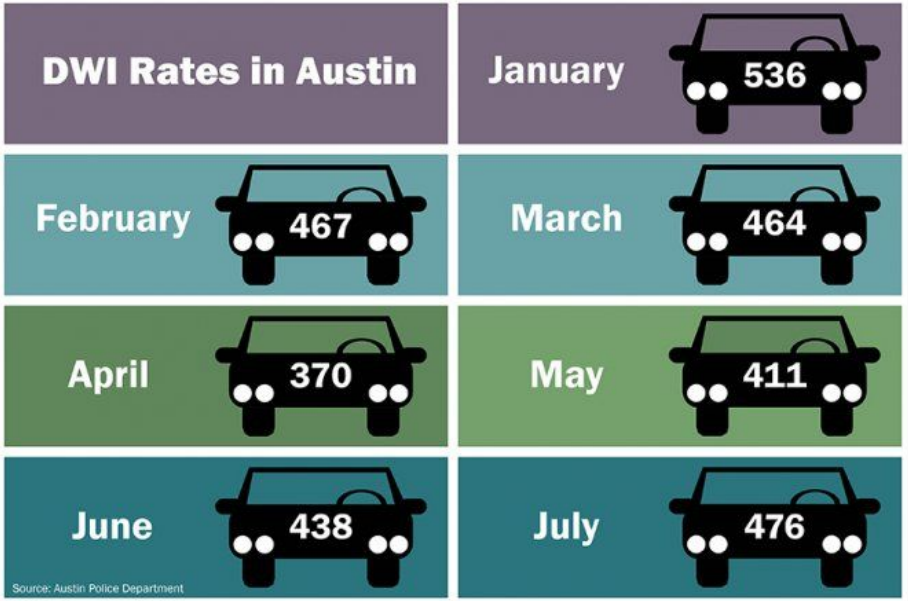


Посмотреть на карте в полный экран

https://yandex.ru/company/researches/2013/ya_music_13

Визуализируйте (не) правильно!





http://viz.wtf/

Проблемы плохой визуализации

1. Неправильный выбор типа визуализации.
2. Перегруженность материалом.
3. 3D эффекты.
4. Просто ненужные эффекты.
5. Неправильные/неэффективные цвета.
6. Слишком мало информации.
7. Информация, которую предполагается сравнивать - разнесена в пространстве.
8. Тысячи их!

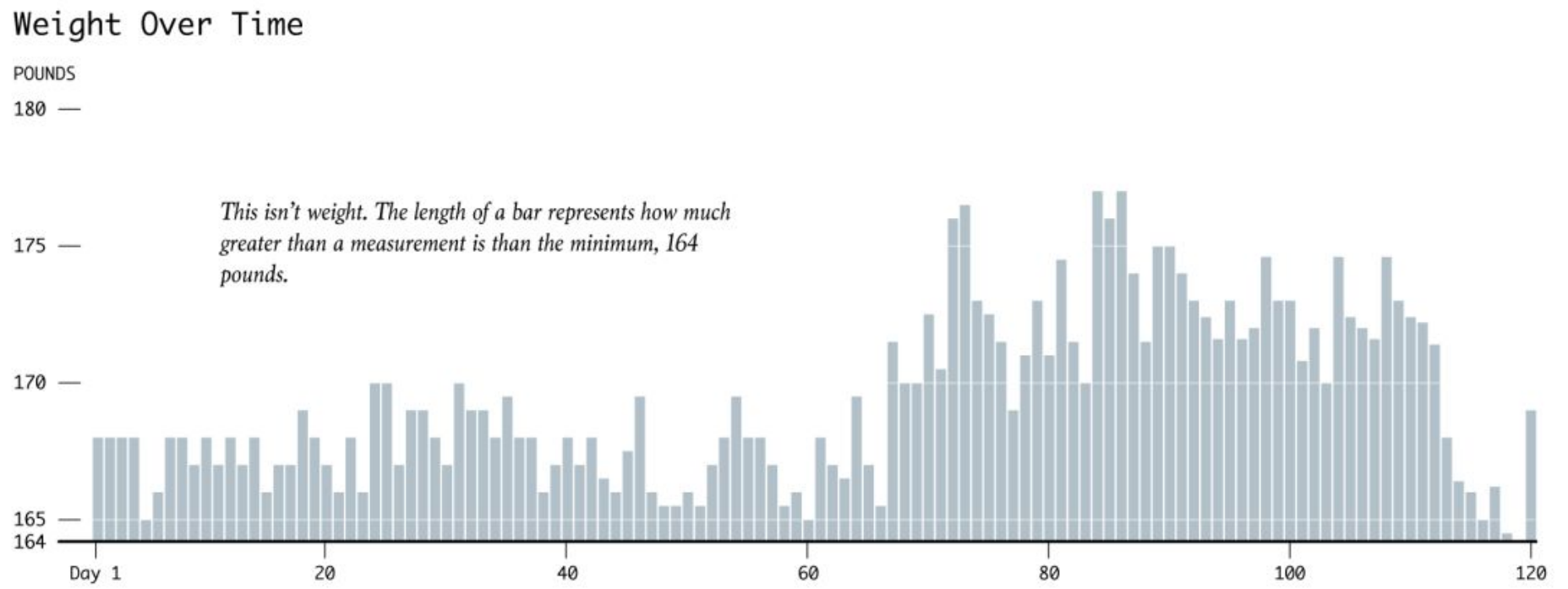
Если после просмотра графика/диаграммы у вас возникает желание посмотреть исходные данные - визуализация провалилась в своей задаче.

Визуализируйте правильно!

Ключевые моменты:

1. Графическая целостность.
2. Простота.
3. Правильная форма.
4. Правильное использование цвета.
5. Целостность повествования.

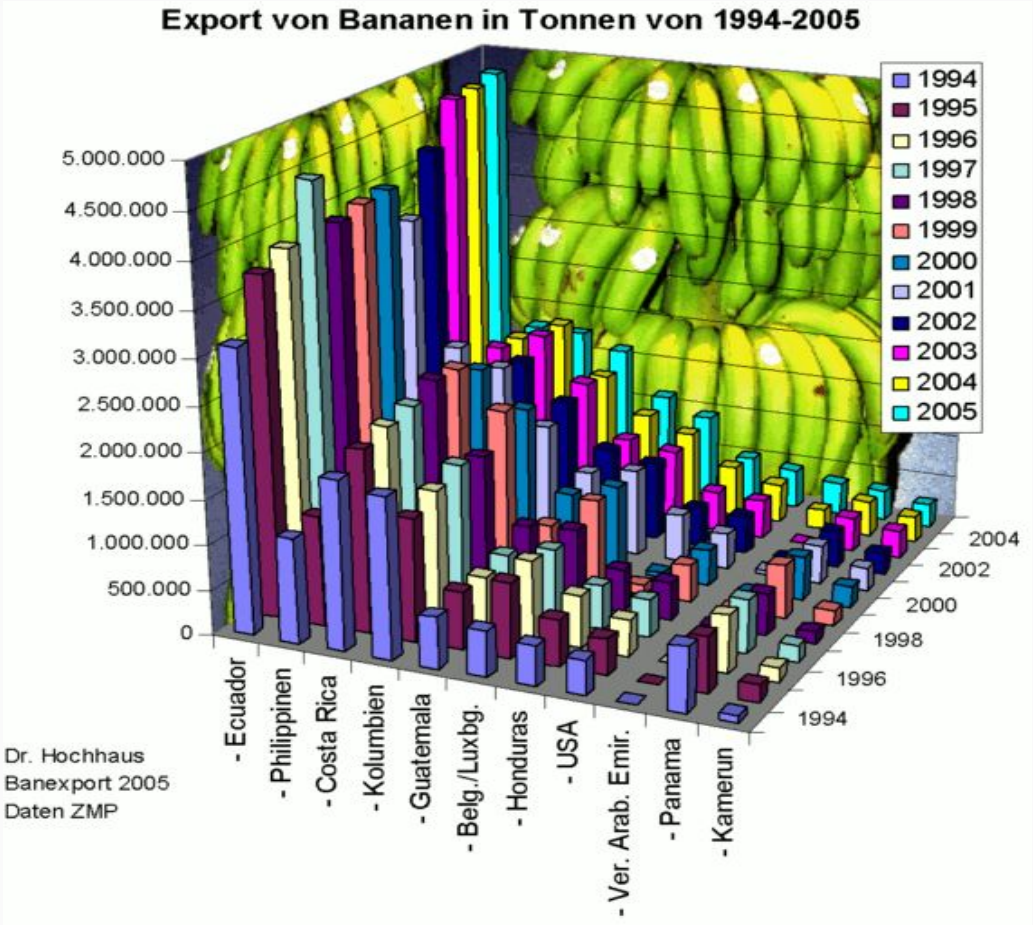
Графическая целостность



Шкалы - честная координатная сетка, честное представление данных. Равные расстояния на визуализации должны означать равную разницу в данных.

https://flowingdata.com/2015/08/31/bar-chart-baselines-start-at-zero/

Простота

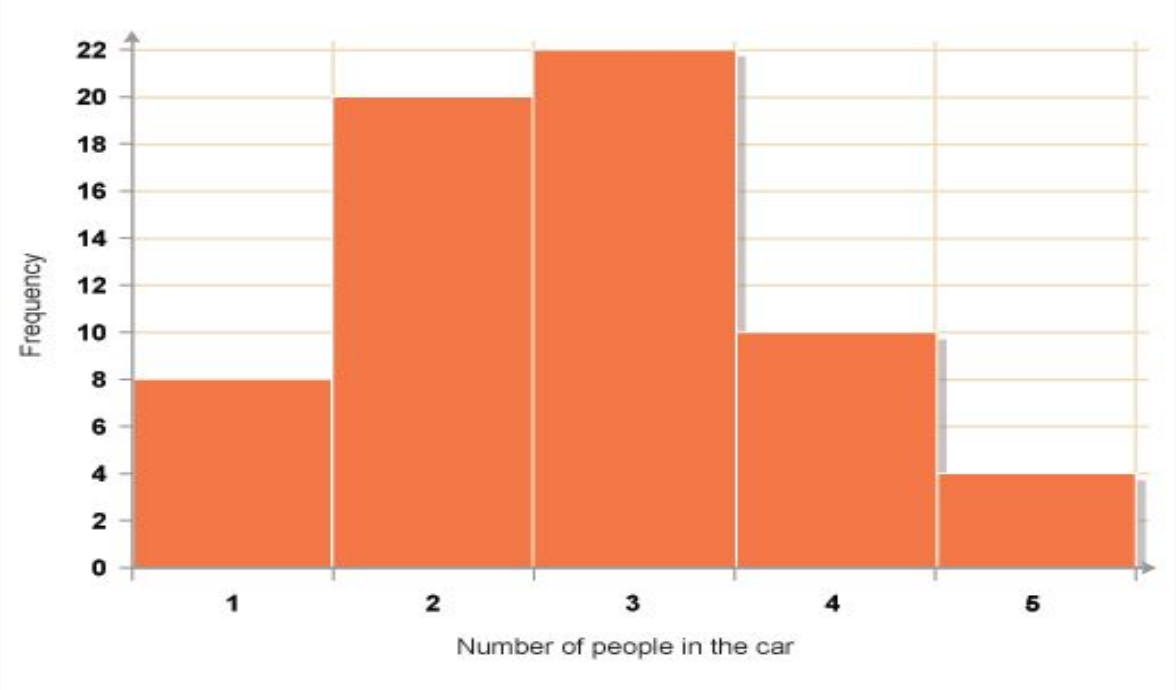


Отношение данные / чернила должно быть как можно больше. (c) Эдвард Тафти

https://en.wikipedia.org/wiki/Chartjunk

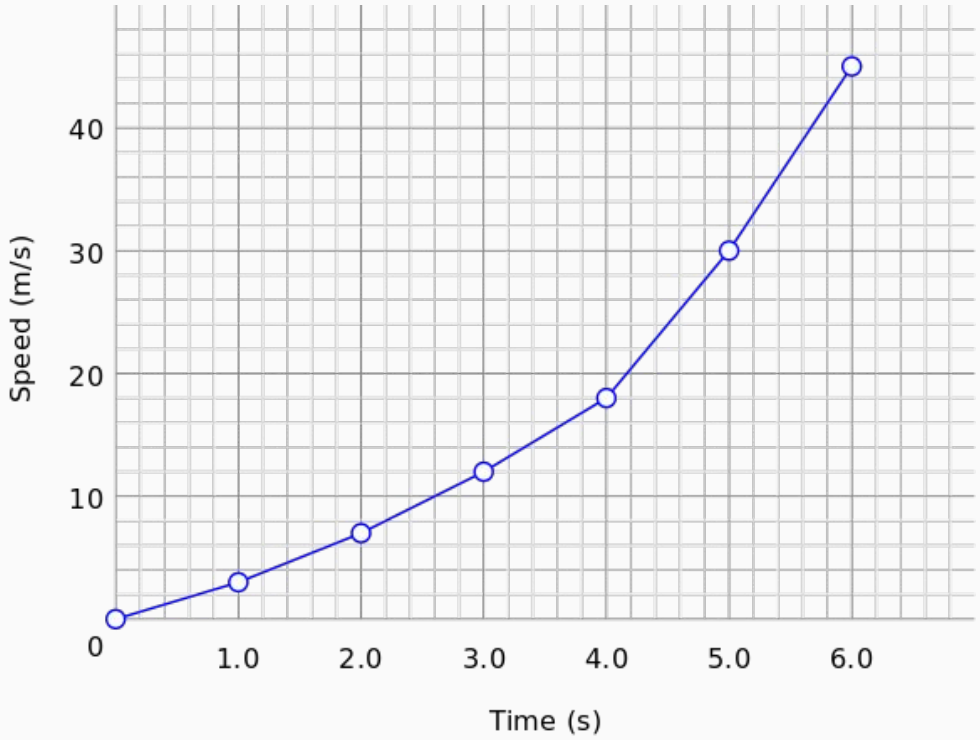
Правильная форма

Сравнение - (столбчатая диаграмма) bar charts. Линейные графики подразумевают непрерывность



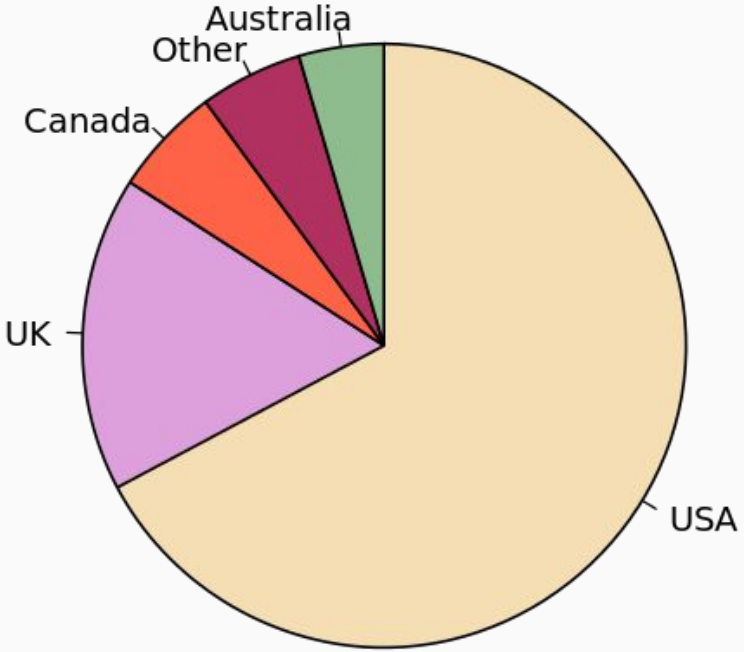
Правильная форма

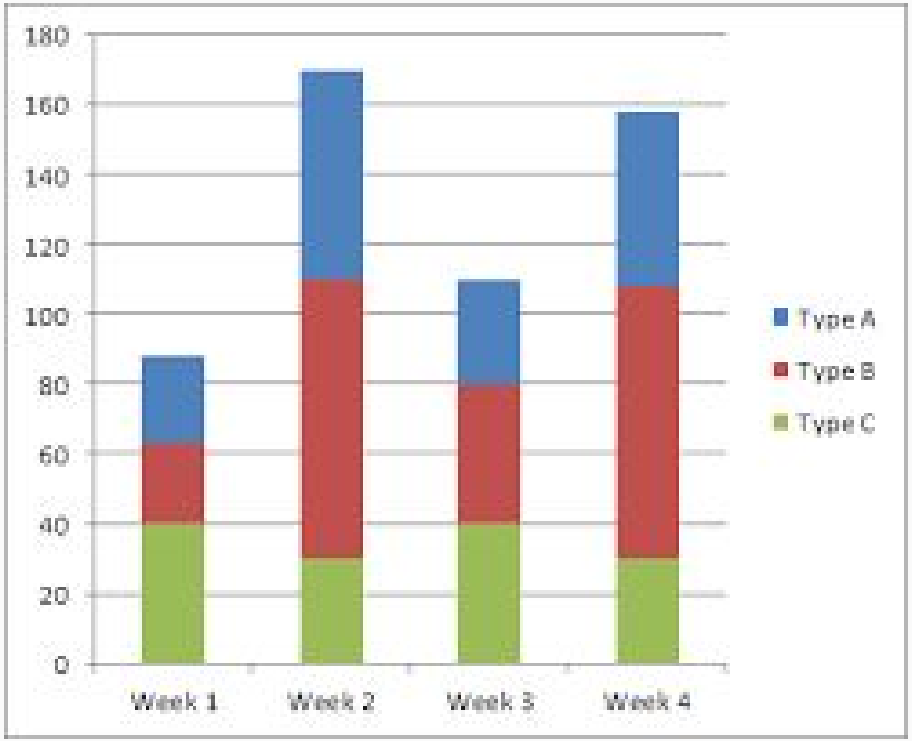
Тренды во времени, непрерывные данные - линейные графики.



Правильная форма

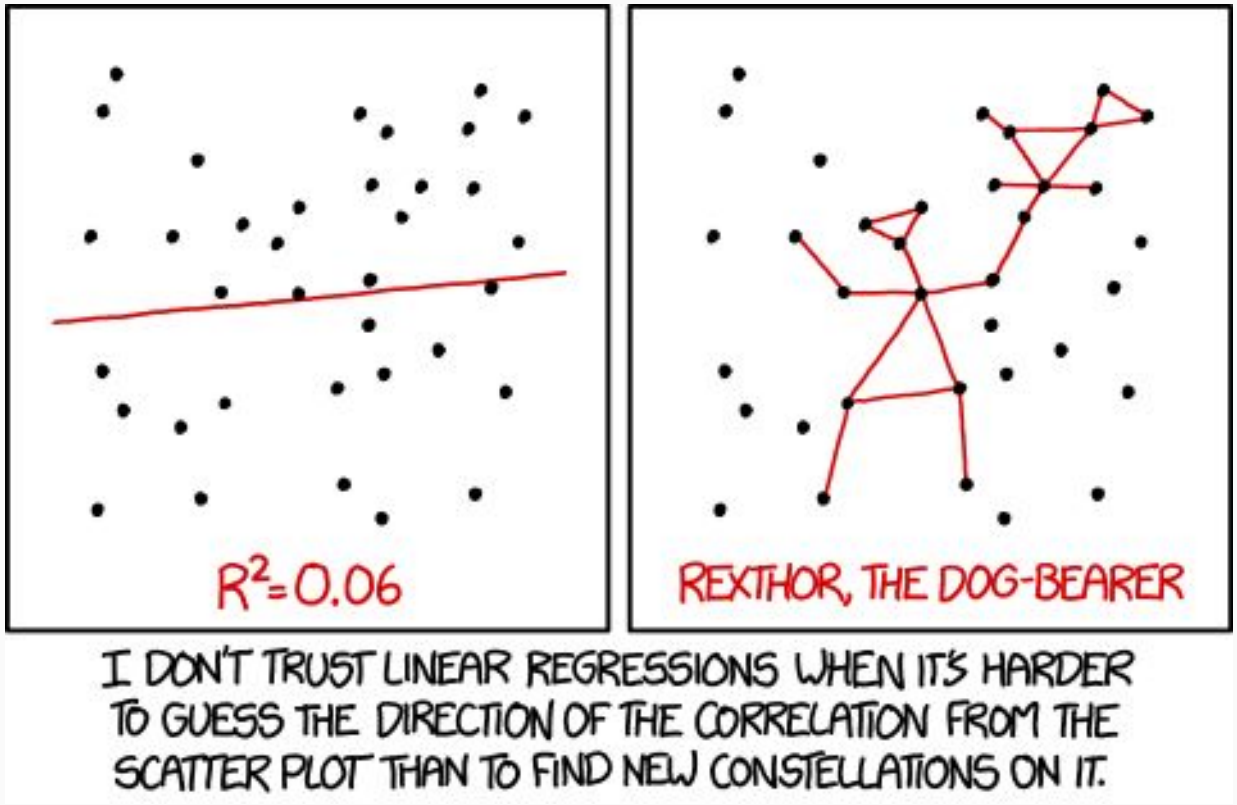
Пропорции - пироги :). Круговая диаграмма. Колонки как вариант.





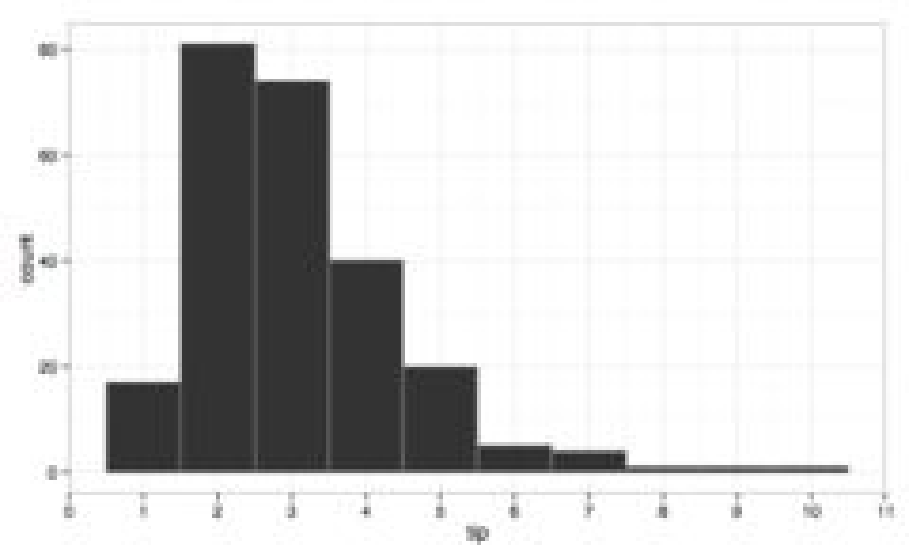
Правильная форма

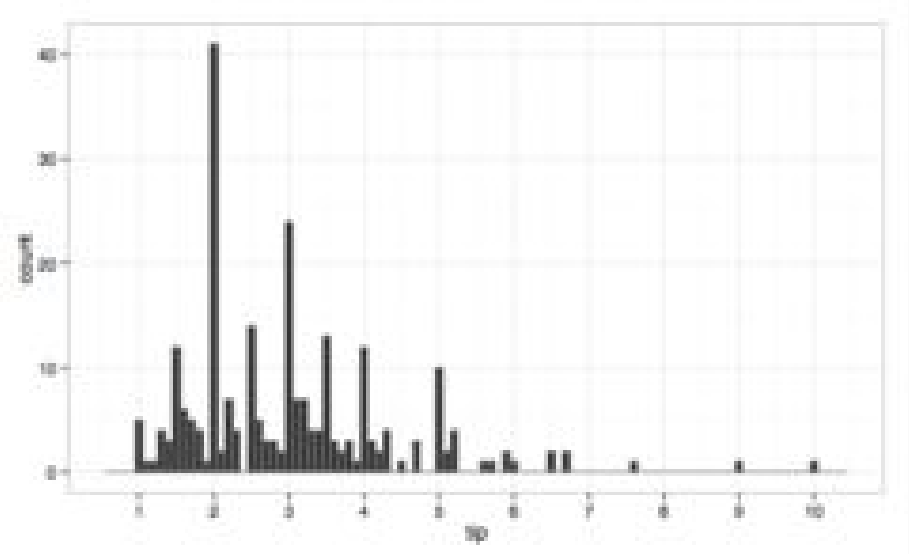
Корреляции - scatter plot (диаграмма рассеяния).



Правильная форма

Распределение - гистограмма.

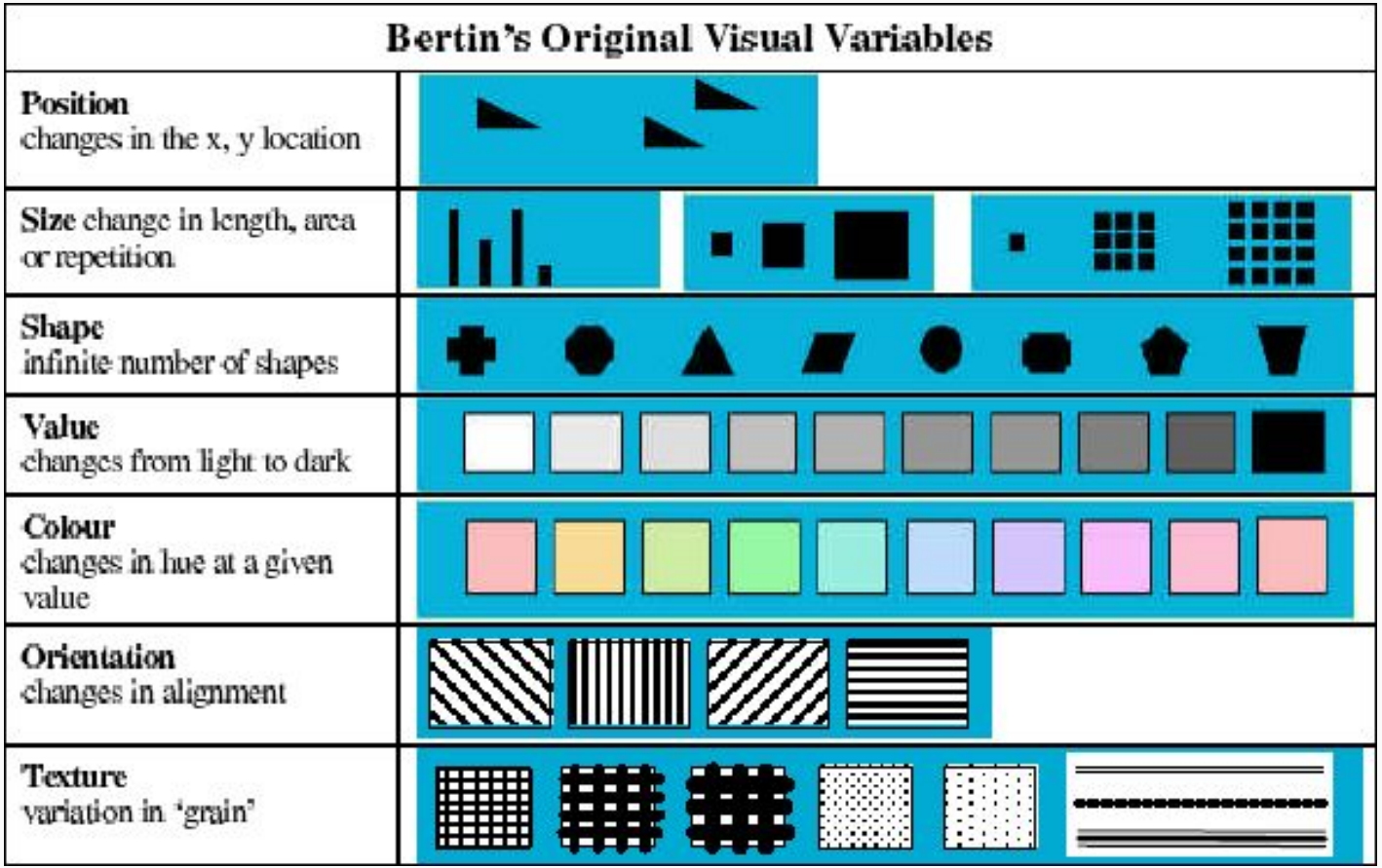




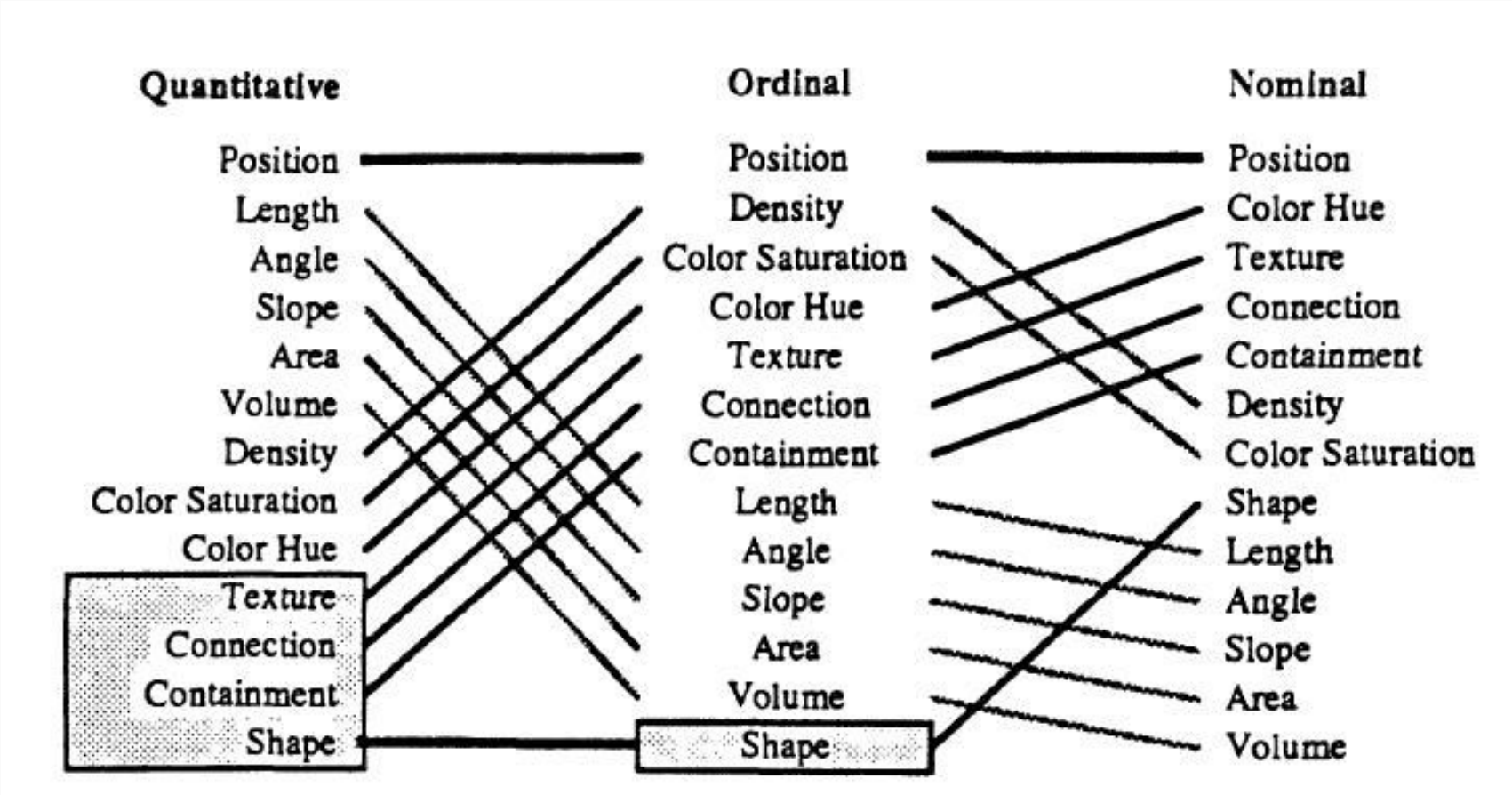
Тренировка

	1.before	2.after	3.in_1_week
0	10	15	10
1	20	10	15
2	30	35	40
3	40	40	35

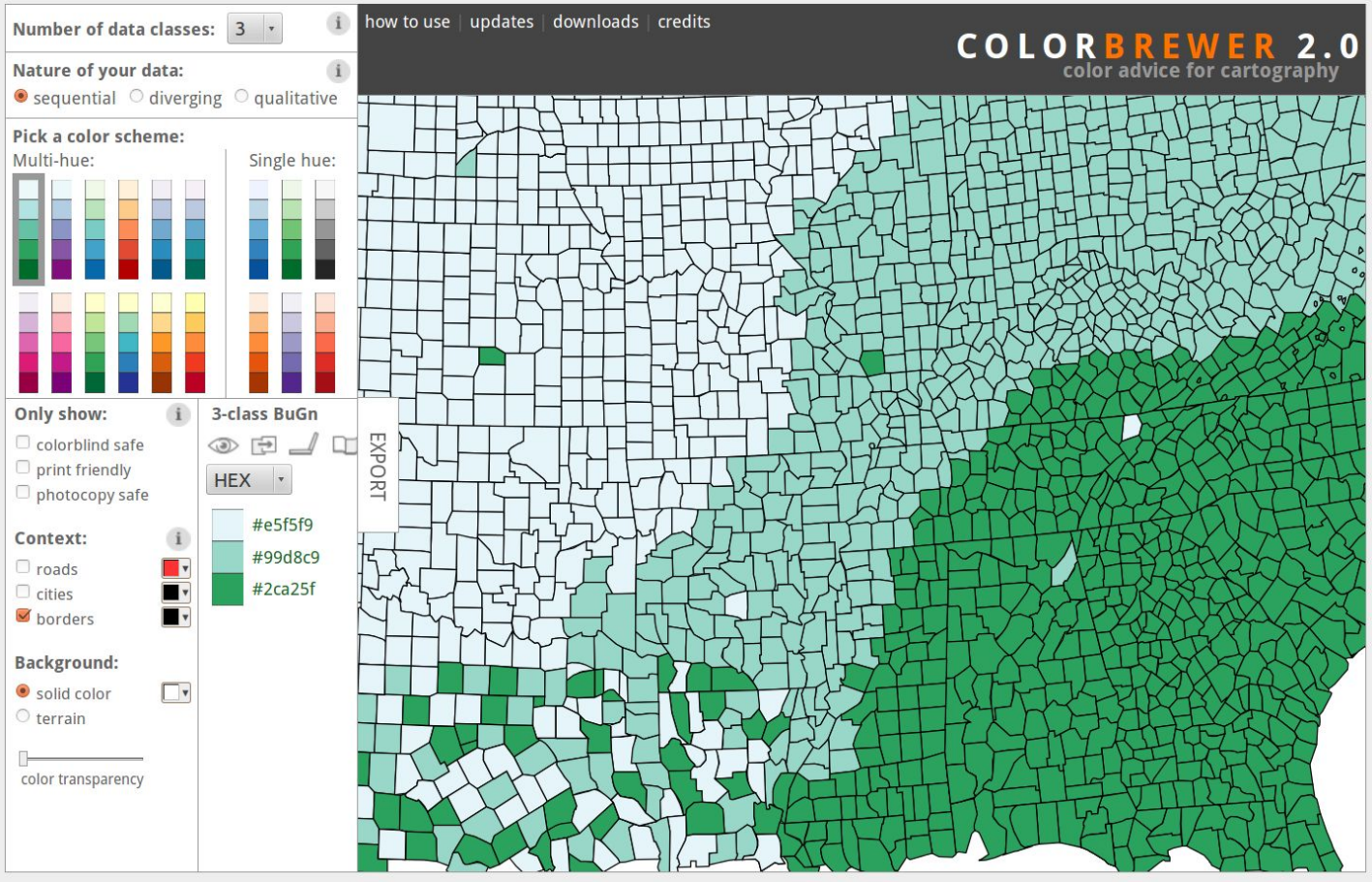
Эффективность восприятия



Эффективность восприятия



Правильное использование цвета



Рекомендуемые ресурсы

1. Сайты
    a. https://flowingdata.com/
    b. http://colorbrewer2.org/#type=sequential&scheme=BuGn&n=3
    c. http://www.datavis.ca/milestones/index.php?group=1950%2B
    d. http://www.infovis-wiki.net/index.php?title=Visual_Variables

2. Книги
    a. Эдвард Тафти
        i. Visual Display of quantitative information
        ii. Envisioning information
        iii. Visual explanations
    b. Стивен Фью
        i. Show me the numbers
        ii. Now you see it

Вопросы?

Домашнее задание

Seaborn - мощная библиотека для визуализации данных. Изучите ее базовый функционал и выведите график линейной регрессии по данным собранным в домашней работе №2. Подсказку можно найти в проверочном скрипте окружения.

Подумайте над следующими вопросами:

- О чем говорит наклон графика линейной модели?
- Как он соотносится с результатами полученными вами в предыдущем домашнем задании?
- Хорошо ли линейная аппроксимация подходит для ваших данных или видны систематические отклонения?

Copyrights

При оформлении слайдов использованы изображения с следующих страниц.

- https://flowingdata.com/
- http://colorbrewer2.org/#type=sequential&scheme=BuGn&n=3
- http://www.datavis.ca/milestones/index.php?group=1950%2B
- http://www.infovis-wiki.net/index.php?title=Visual_Variables
- https://www.explainxkcd.com/wiki/index.php/1725:_Linear_Regression
- https://en.wikipedia.org/wiki/Pie_chart
- https://en.wikipedia.org/wiki/Histogram
- https://en.wikipedia.org/wiki/Line_chart
- http://www.bbc.co.uk/bitesize/ks3/maths/handling_data/representing_data/revision/2/
- http://www.excelcharts.com/blog/change-bad-charts-in-the-wikipedia/
- http://viz.wtf/
- https://www.123rf.com/photo_25442337_hand-lens-that-magnifies-a-needle-in-a-haystack.html