ЦИФРОВАЯ КУЛЬТУРА
УНИВЕРСИТЕТ ИТМО

Визуализация данных

Высшая школа цифровой культуры
Университет ИТМО
dc@itmo.ru

Содержание

1 Задачи визуализации

В этом мире можно посчитать и выразить цифрами все, что угодно. Наш мозг превосходно работает, когда речь касается абстрактного мышления, но он не способен эффективно обрабатывать тысячи цифр, и потому, имея дело с большими объёмами информации, мы прибегаем к её визуализации, то есть к наглядному представлению массивов различной информации.

Рис. 1: Пример неинформативной визуализации данных

Под визуализацией данных подразумевается представление абстрактной информации в графической форме. Визуализация данных позволяет выявлять модели, тенденции и корреляции, которые могут остаться незамеченными в традиционных отчетах и таблицах (в том числе электронных). Исследования показывают, что человеческий мозг обрабатывает визуальную информацию в 60 000 раз быстрее, чем текст. 90 процентов информации, передаваемой в мозг, составляют визуальные данные.

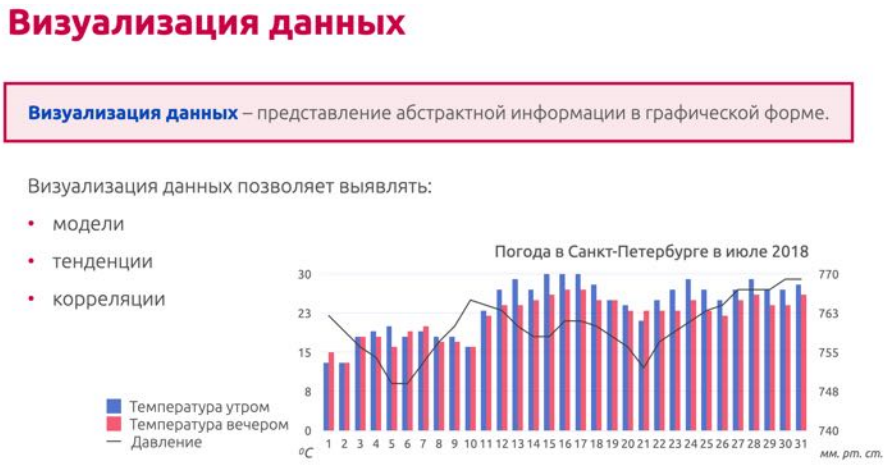


Рис. 2: Пример информативной визуализации данных

Попробуйте в ячейках таблицы быстро найти минимальное и максимальное значения. А теперь то же самое на графике. Инструкцию легче запомнить по схеме , чем читать ее в текстовом виде. 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. Почему визуализа-

ция так важна?

Задача любой визуализации – представить информацию упрощённо, позволив нам с одного взгляда составить о ней определённое мнение. При этом, идеальная визуализация является понятной сама по себе, и может сохранить свой смысл, даже лишившись всего сопровождающего текста. У людей, при-

нимающих решения, как правило, нет времени вникать в бесконечные ряды данных, поэтому им требуется материал, на основании которого они могут быстро принять качественное решение и оценить ситуацию, не углубляясь в анализ первичной информации. Именно поэтому качество визуализации – крайне важный элемент принятия решений.

Можно выделить несколько задач визуализации данных. В первую очередь, это Иллюстрация идей. Целью в этом случае являются обучение, разъяснение. Используется в качестве замены развернутому описанию. Типичные примеры: организационные схемы и схемы бизнес-процессов.

Рис. 3: Визуализация для иллюстрация идей

Визуализацию часто используют для генерации идей. Когда целью является решение проблемы, выяснение истины. Используется при мозговых штурмах. Типичным примером является ментальная карта. Как уже было

Рис. 4: Визуализация для генерация идей

сказано, визуализация используется как самостоятельный вид анализа. Такой вид анализа можно назвать визуальным исследованием.

Используется, чтобы лучше понять и проследить закономерности. Сюда относятся сложные многофакторные представления. И, наконец, рутинная визуализация, которой называют сообщение, помещенное в контекст. Используется при составлении отчетов и презентаций для руководства и партнеров. Так мы информируем нашу аудиторию о положении вещей.

Визуализация может быть очень разной. Это и обычное визуальное представление количественной информации в схематической форме. К этой группе можно отнести всем известные круговые и линейные диаграммы, гистограммы и спектрограммы, таблицы и различные точечные графики. Данные при визуализации могут быть преобразованы в форму, усиливающую восприятие и анализ этой информации. Например, карта и полярный график, временная линия и график с параллельными осями, диаграмма Эйлера.

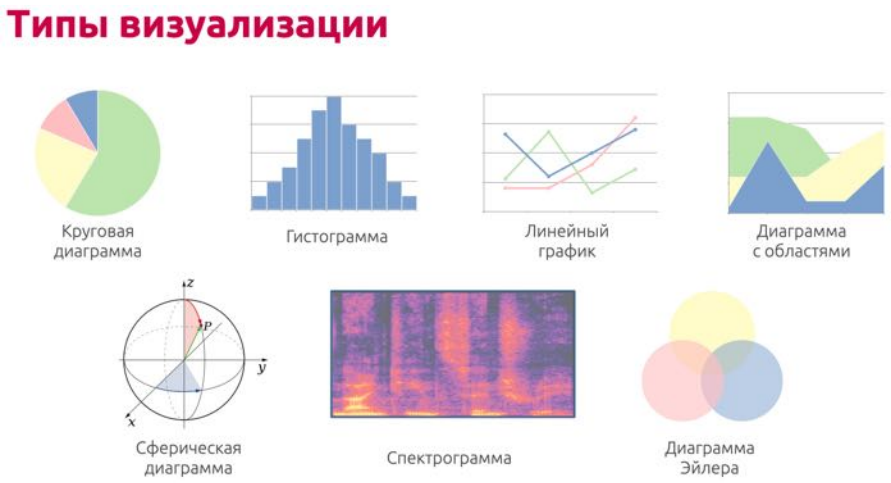


Рис. 5: Различные типы визуализации

Концептуальная визуализация позволяет разрабатывать сложные концепции, идеи и планы с помощью концептуальных карт, диаграмм Ганта, графов с минимальным путем и других подобных видов диаграмм.

Стратегическая визуализация переводит в визуальную форму различные данные об аспектах работы организаций. Это всевозможные диаграммы производительности, жизненного цикла и графики структур организаций.

Рис. 6: Стратегическая визуализация

Графически организовать структурную информацию с помощью пирамид, деревьев и карт данных поможет метафорическая визуализация, ярким примером которой является карта метро. Комбинированная визуализация позволяет объединить несколько сложных графиков в одну схему, как в карте с прогнозом погоды.

Рис. 7: Комбинированая визуализация

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.

«Цифры часто обманывают меня, особенно когда я сам их организую», – это высказывание приписывают Дизраэли. Можно также привести известное высказывание Марка Твена: «Существует три вида лжи: ложь, проклятая ложь и статистика».

Тут следует заметить, что насколько корректная визуализация помогает проанализировать данные, настолько некорректная может создать совсем неправильное, искаженное представление о данных.

Например, рассмотри данные колебания веса, приведенные в таблице. Теперь изобразим это на диаграмме. Создается ощущение, что вес стремительно увеличивался.

А теперь нарисуем другую диаграмму, используя те же данные таблицы. Теперь кажется, что вес менялся совсем незначительно.

Возраст	Вес (кг)
20	57
25	58
30	61
35	63
40	64

Причиной такого разного восприятия является изменение начальной точки вертикальной оси – в первом случае мы начинали отсчет от 50, а во втором случае – от нуля.

Этот пример наглядно иллюстрирует, как одни и те же данные могут быть визуализованы по-разному, что может привести к их различной интерпретации.

2 Методы визуализации

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

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

Цели визуализации – это реализация основной идеи информации, это то, ради чего нужно показать выбранные данные, какого эффекта нужно добиться – выявления отношений в информации, показа распределения данных, композиции или сравнения данных.

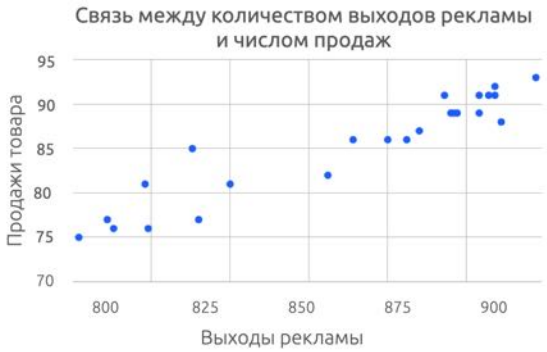


Рис. 8: Цели визуализации. Отношения в данных

Отношения в данных – это связи и зависимости между ними. С помощью отношений можно выявить наличие или отсутствие зависимостей между переменными. Если основная идея информации содержит фразы «относится к», «снижается/повышается при», то нужно стремиться показать именно отношения в данных.

Распределение отображает количество наблюдений, в которых признак принимает определенные значения в заданных интервалах. Основная идея

Рис. 9: Цели визуализации. Распределение данных

при этом будет содержать фразы «в диапазоне от x до y», «концентрация», «частотность», «распределение».

Композиция данных – объединение данных с целью анализа общей картины в целом, сравнения компонентов, составляющих процент от некоего целого. Ключевыми фразами для композиции являются «составило x%», «доля», «процент от целого».

Рис. 10: Цели визуализации. Композиция данных

Сравнение данных – объединение данных, с целью сравнения некоторых показателей, выявление того, как объекты соотносятся друг с другом. Также это сравнение компонентов, изменяющихся с течением времени. Ключевые фразы для идеи при сравнении – «больше/меньше чем», «равно», «изменяется», «повышается/понижается».

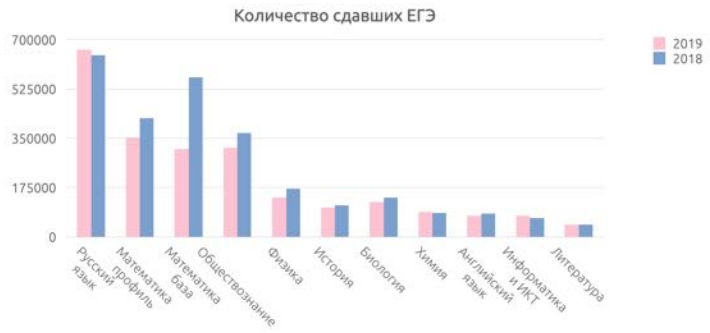


Рис. 11: Цели визуализации. Сравнение данных

После определения цели визуализации требуется определить тип данных. Они могут по своему типу и структуре быть очень разнородными, но в самом простом случае выделяют непрерывные числовые и временные данные, дискретные данные, географические и логические данные. Непрерывные числовые данные содержат в себе информацию зависимости одной числовой величины от другой, например графики функций, такой как $y = 2x$. Непрерывные временные содержат в себе данные о событиях, происходящих на каком-либо промежутке времени, как график температуры, измеряемой каждый день. Дискретные данные могут содержать в себе зависимости категорийных ве-

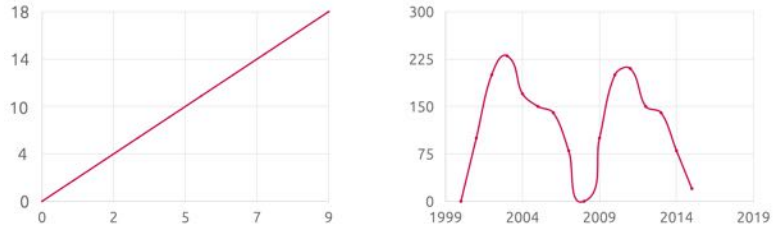


Рис. 12: Примеры непрерывных данных

личин, например график количества продаж товаров в разных магазинах. Географические данные содержат в себе различную информацию, связан-

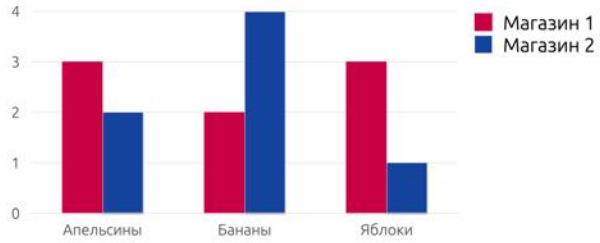


Рис. 13: Пример дискретных данных

ную с местоположением, геологией и другими географическими показателями, яркий пример – это обычная географическая карта. Логические данные показывают логическое расположение компонентов относительно друг друга, например генеалогическое древо семьи.

Рис. 14: Пример логических данных

Изображение диаграммы состоит из различных элементов – названий осей, единиц измерения, заголовка диаграммы, легенды и других элементов. Назначение этих элементов – сделать диаграмму максимально понятной. Наиболее распространенный случай диаграммы – линейный график, или линейная диаграмма. Объединяет линией набор точек, соответствующих значениям по осям.

Рис. 15: Элементы диаграммы

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 отоб-

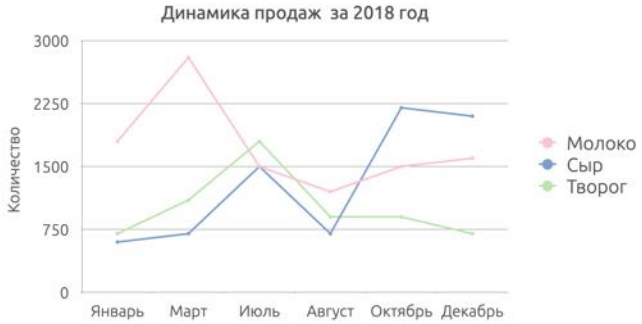


Рис. 16: Пример линейной диаграммы

ражать линии разными цветами и указывать в легенде какая линия чему соответствует.

Не нужно нагружать график большим количеством информации. Оптимальное количество разных типов данных, категорий – это не более 4-5, иначе же целесообразнее разделить такую диаграмму на несколько штук.

Диаграмма с областями основана на линейной диаграмме. Область между осью и линией обычно подчеркивается цветами, текстурами и штрихами. Обычно при помощи диаграммы с областями сравнивают два или более ряда данных. Используйте диаграмму с областями и накоплением для отображения вклада каждого значения к общему по времени или по категориям.

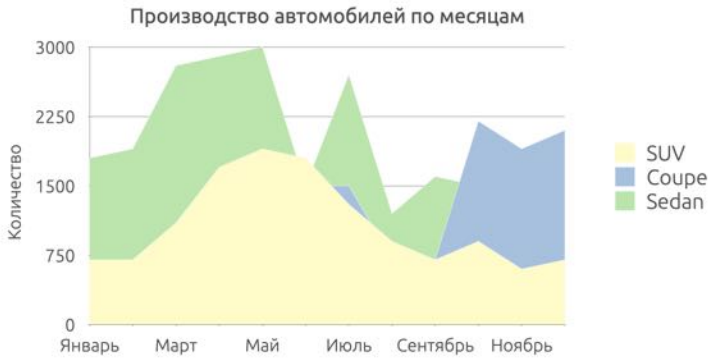


Рис. 17: Диаграмма с областями

Столбчатая диаграмма 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.

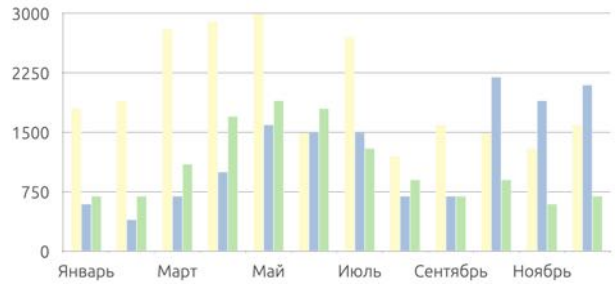


Рис. 18: Столбчатая диаграмма

На гистограммах количественные соотношения некоторого показателя изображаются в виде прямоугольников. Чаще всего для удобства восприятия ширину прямоугольников берут одинаковую, при этом их высота определяет соотношения отображаемого параметра.

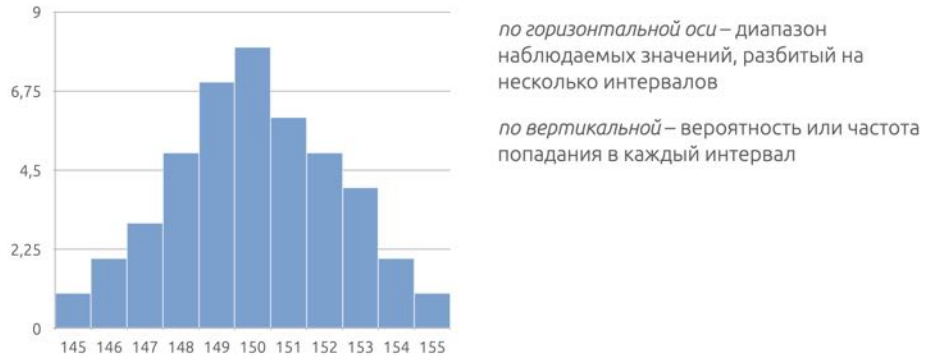


Рис. 19: Гистограмма

Диаграмма используется в статистике для графического представления распределения вероятностей значений некоторой случайной величины. По го-

ризонтальной оси гистограммы откладывается диапазон наблюдаемых значений, разбитый на несколько интервалов, а по вертикальной – вероятность или частота ее попадания в каждый из них. Тогда прямоугольник будет отражать значения этих показателей для интервала, на который он опирается.

Линейные диаграммы, графики с областями и гистограммы могут содержать в одном аргументе для одной категории несколько значений, которые к тому же дают суммарный вклад в общие итоги. Если нужно изобразить и сравнить суммарные итоги, то можно использовать диаграмму с накоплением. Это позволяет не только сравнить отдельные ряды данных, но и суммарный показатель в целом.

Нормированная диаграмма – подобна предыдущей, но здесь значения нормированы, т.е. приведены к процентам. Суммарный показатель всегда – 100%. На такой диаграмме проще оценить долевое участие каждого из параметров в совокупном результате.

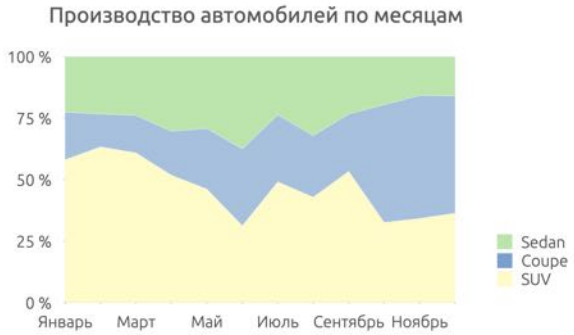


Рис. 20: Нормированная диаграмма

Облако тегов – 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.

Графики пиктограмма используют значки, чтобы дать более привлекательный общий вид небольших наборов дискретных данных. Как правило,

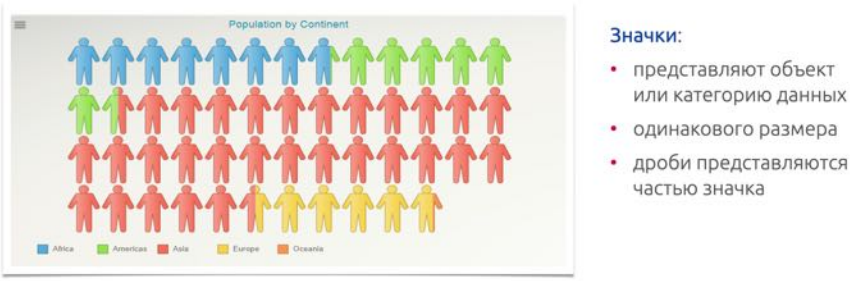


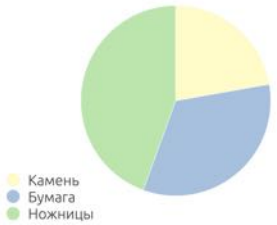
Рис. 21: Пиктограмма

значки представляют объект или категорию данных, например, данные о на-

селении будут использовать значки людей. Все значки должны быть одинакового размера, а дроби обычно представляются частью значка. Каждый значок может представлять собой единицу или любое количество единиц (например, каждый значок представляет 10).

Мы рассмотрели множество различных способов визуализации данных. Но иногда перед визуализацией нужно сделать определенные преобразования данных.

Круговые диаграммы помогают показать пропорции и процентные доли между категориями, разделяя круг на пропорциональные сегменты. Каждая длина дуги представляет собой долю соответствующей категории, а весь круг представляет собой сумму всех данных, равную 100%. Круговые диаграммы



Данные			
Камень	Бумага	Ножницы	*Итого*
2	3	4	*9*
Данные в процентах			
(2/9)*100% = 22%	(3/9)*100% = 33%	(4/9)*100% = 44%	*100%*
Градусы для каждого «Куска пирога»			
(2/9)*360° = 80°	(3/9)*360° = 120°	(4/9)*360° = 160°	*360°*

Рис. 22: Круговая диаграмма

идеально подходят для представления о пропорциональном распределении данных. Основным недостатком круговых диаграмм можно считать то, что на них не подходят для отображения больше, чем 3-5 значений, потому что по мере увеличения числа показанных значений размер каждого сегмента/-среза становится меньше. Это делает их непригодными для больших объемов данных.

Для удобства сравнения располагать сегменты следует по мере убывания длин дуг. Диаграммы рассеивания (или точечные диаграммы) используют декартовы координаты для отображения значений двух переменных в виде точек на плоскости. Такое отображение переменных по каждой оси позволяет визуально предположить, существует ли связь или корреляция между двумя переменными.

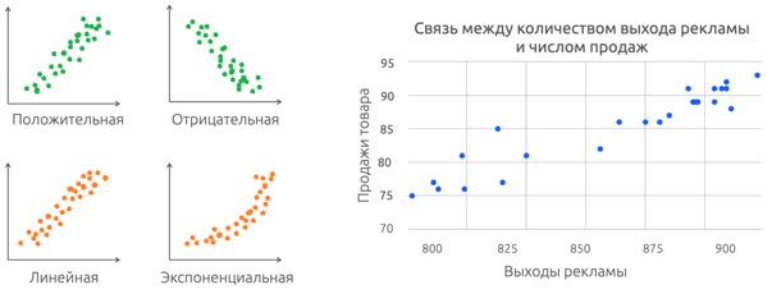


Рис. 23: Диаграмма рассеивания

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.

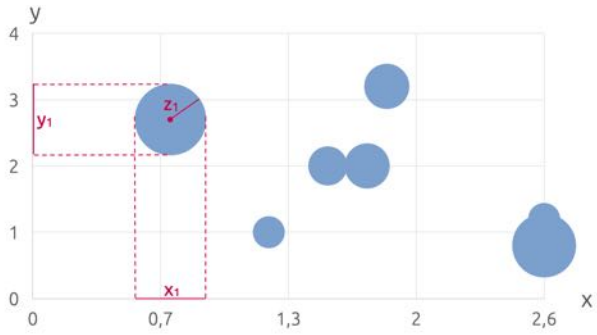


Рис. 24: Пузырьковая диаграмма

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

Продолжим разговор про виды визуализации данных. Рассмотрим диаграммы размаха, или Box Plot, которые иногда называют ящиками с усами-удобный способ наглядного отображения групп числовых данных с помощью прямоугольников, или ящиков. Границами ящика служат первый и третий квартили , линия в середине ящика – медиана. Линии, идущие параллельно от коробок, известны как"усы. Концы усов – минимальные и максимальные значения после удаления выбросов, которые используются для обозначения изменчивости вне верхней и нижней квартилей. Ящики с усами могут быть нарисованы вертикально или горизонтально.

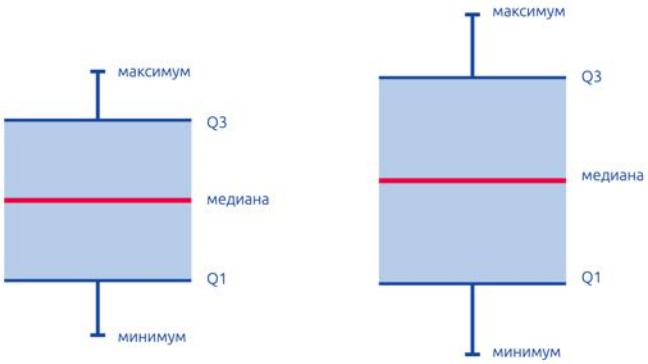


Рис. 25: Ящик с усами

Несколько таких ящиков можно нарисовать бок о бок, чтобы визуально сравнивать одно распределение с другим; их можно располагать как горизон-

тально, так и вертикально. Межквартильный размах позволяет определить степень разброса (дисперсии) и асимметрии данных.

Свечной график 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.

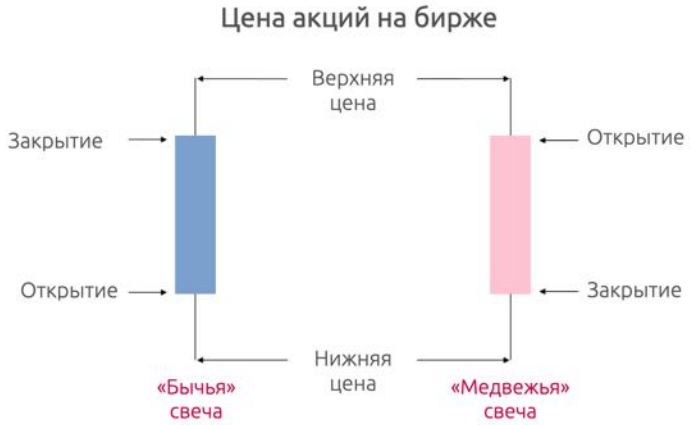


Рис. 26: Свечной график

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.

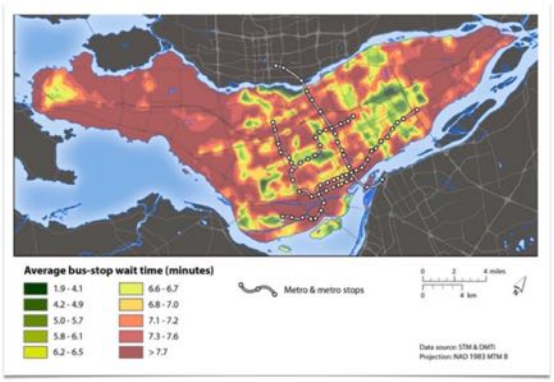


Рис. 27: Тепловая карта

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.

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.

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.

Если набор данных имеет более трех измерений, то существуют специальные методы визуализации.

Наиболее известные способы представления многомерных данных – это параллельные координаты, радарные диаграммы, лица Чернова. В параллельных координатах график представляется как объединение двумерных проекций многомерного набора данных. Параллельные проекции могут отображаться как по вертикали, так и по горизонтали.

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.

Рис. 28: Параллельные координаты

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.

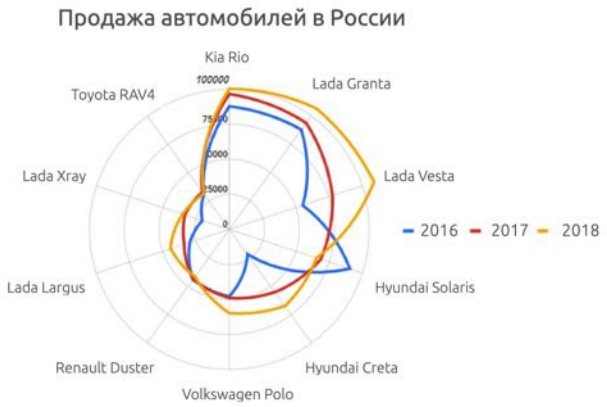


Рис. 29: Радарная диаграмма

Для каждого наблюдения рисуется свой polygon. Основная идея визуализации методом лиц Чернова – 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.

Рис. 30: Лица Чернова

Один из наиболее известных примеров называется *жизнь в Лос-Анджелесе*, где при помощи лиц Чернова изображены занятость населения, уровень дохода, пропорции белого населения и другие показатели.