

ЦИФРОВАЯ  
КУЛЬТУРА  
УНИВЕРСИТЕТ ИТМО

Лекция «Инструменты обработки данных и визуализация»

Михайлова Елена Георгиевна  
Графеева Наталья Генриховна

Санкт-Петербург  
2019

## **Содержание**

<b>1</b>	<b>Инструменты обработки данных</b>	<b>2</b>
<b>2</b>	<b>Задачи визуализации</b>	<b>7</b>
<b>3</b>	<b>Методы визуализации</b>	<b>13</b>

# 1 Инструменты обработки данных

Когда мы работаем с наборами данных, то нужно простое и удобное средство для его хранения и обработки, так как это довольно трудно делать в уме.

Для обработки данных нужно подобрать подходящее ПО – это будет зависеть от объема данных и от задач по обработке данных, которые мы решаем. Если данных совсем немного, то их можно записать в файле самого простого формата, точнее, совсем без форматирования, создав документ, например, при помощи программы Блокнот. Такие файлы называются файлами ASCII, и сохраняют только символы, введенные туда без каких-либо указаний по оформлению.

Если данные представляют собой неструктурированный текст, то для их обработки созданы специальные текстовые процессоры, например Word. Если же данные состоят из однородных элементов, имеющих явную структуру, то лучшим средством хранения будут электронные таблицы. Часто для

## CSV

*Comma-Separated Values – «значения, разделённые запятыми»*

Re3data.org, <http://www.re3data.org/>  
DataBib, <http://databib.org/>  
DataCite, <http://www.datacite.org/>  
Dryad, <http://datadryad.org/>  
DataPortals, <http://dataportals.org/>  
Open Access Directory, [http://oad.simmons.edu/oadwiki/Data\\_repositories](http://oad.simmons.edu/oadwiki/Data_repositories)  
Gapminder, <http://www.gapminder.org/data>  
Google Public Data Explorer, <http://www.google.com/publicdata/directory>  
IBM Many Eyes, <http://www.manyeyes.com/software/analytics/maneyes/datasets>  
Knoema, <http://www.knoema.com/atlas/>

	A	B
1	Re3data.org	<a href="http://www.re3data.org/">http://www.re3data.org/</a>
2	DataBib	<a href="http://databib.org/">http://databib.org/</a>
3	DataCite	<a href="http://www.datacite.org/">http://www.datacite.org/</a>
4	Dryad	<a href="http://datadryad.org/">http://datadryad.org/</a>
5	DataPortals	<a href="http://dataportals.org/">http://dataportals.org/</a>
6	Open Access Directory	<a href="http://oad.simmons.edu/oadwiki/Data_repositories">http://oad.simmons.edu/oadwiki/Data_repositories</a>
7	Gapminder	<a href="http://www.gapminder.org/data">http://www.gapminder.org/data</a>
8	Google Public Data Explorer	<a href="http://www.google.com/publicdata/directory">http://www.google.com/publicdata/directory</a>
9	IBM Many Eyes	<a href="http://www.manyeyes.com/software/analytics/maneyes/datasets">http://www.manyeyes.com/software/analytics/maneyes/datasets</a>
10	Knoema	<a href="http://www.knoema.com/atlas/">http://www.knoema.com/atlas/</a>

Рис. 1: Comma-Separated Values

сохранения табличных данных в виде текста используется формат CSV – Comma-Separated Values – значения, разделённые запятыми). Стока таблицы соответствуют строке текста, которая содержит одно или несколько полей, разделенных запятыми или другим разделителем, например, точкой с запятой или табуляцией. Многие приложения, которые работают с форматом CSV позволяют выбирать символ разделителя.

Для хранения таких наборов данных были придуманы электронные таблицы. Они помогают нам вычислять значения, упорядочивать и фильтровать

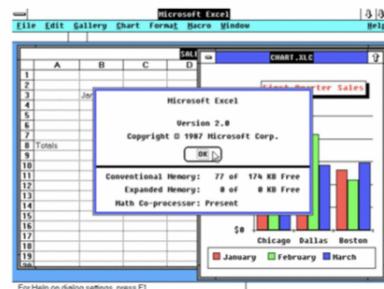
данные, делать разные преобразования, группировать, анализировать и графически представлять различные виды данных.

Первая цифровая электронная таблица – VisiCalc была выпущена в 1979 году. С течением времени цифровые таблицы стали одним из самых популярных видов использования компьютеров. И самым популярным программным обеспечением для работы с электронными таблицами за последние 30 лет является Microsoft Excel. Первая версия Excel была разработана компанией Microsoft в 1985 году. MS Excel является частью пакета MS Office, в который входят текстовый редактор Word, PPT, Outlook, Access и прочие полезные для жизни программы. Таблица Excel могут состоять из нескольких листов.

*Первая цифровая электронная таблица – VisiCalc – 1979 год*

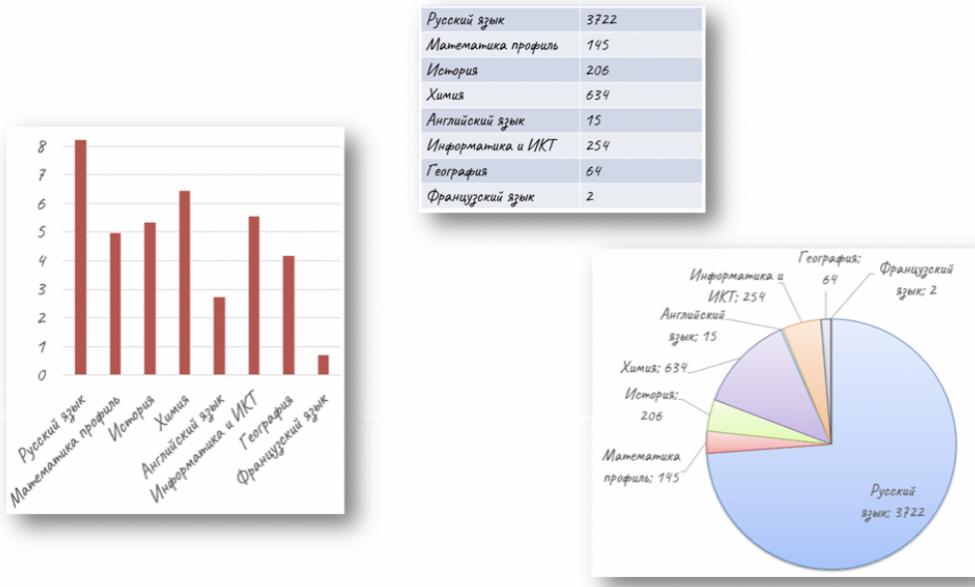
A1	B1	C1	D1
A	Result	Function	INI
1	Value		
2	0	LN	0
3	2.000000	LOG10	0
4	0.000000	SQR	1
5	2.0000005092	X^2	1
6	2.0000005092	SQRT	1
7	2.0000005092	EXP	1
8	3.141593	Pi	3.14
9	-2.71	ABS	2
10	.02	SUM	MAX
11	- .02	MIN	MIN
12	.05	COUNT	AVERAGE
13	0	4	LOOKUP
14			4
15			
16		Mode	
17	Value	Function	Mode
18	45	SIN	.7071068
19	45	COS	.6509035
20	45	TAN	1.1619725
21	.7071068	ASIN	45.00000 .7853982
22	1	ACOS	45.00000 .7853982
23	1	ATAN	45.00000 .7853982

*Первая версия Microsoft Excel – 1985 год*



Общее количество строк и столбцов на листе в версии 2016 года ограничено 1 048 576 строками и 16 384 столбцами. На первый взгляд, это количество кажется очень большим, но, во-первых, количество данных может быть больше, а во-вторых, с таким большим количеством данных на листе работать очень неудобно.

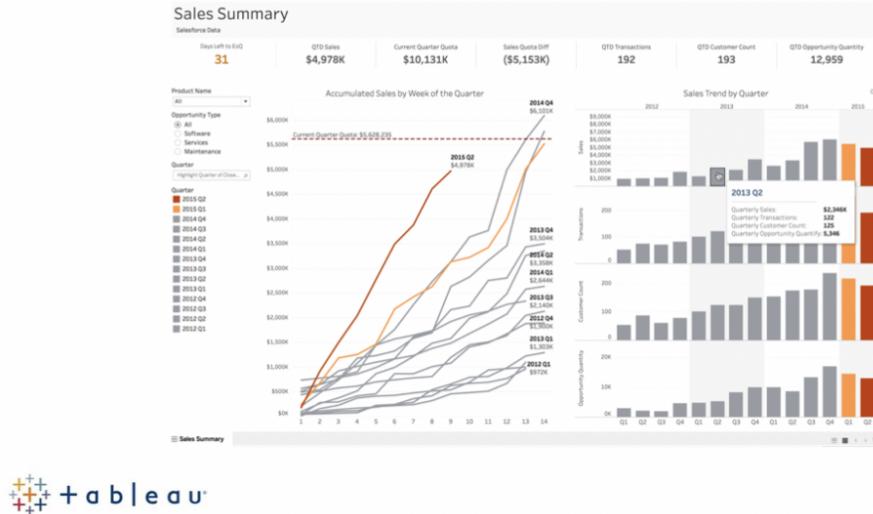
Если данных становится очень много, они имеют сложную структуру, данные должны надежно сохраняться, и при этом работать с ними нужно множеству пользователей, то для их хранения и обработки понадобятся совсем другие средства – например, системы управления базами данных. В зависимости от наличия или отсутствия четкой структуры в данных, подойдут либо реляционные СУБД, например, Oracle или PostgreSQL, которые замечательно обрабатывают структурированные данные, либо NoSQL хранилища, которые специально разрабатывались работы с неструктуризованными или слабоструктуризованными данными. Важный шаг в понимании данных дает визуализация. Наглядное представление информации использовать часто



намного удобнее, чем просто ряды цифр. Достаточно большой спектр возможностей для визуализации данных дает Excel, Google-таблицы и другие электронные таблицы.

Если для данных нужна аналитика по неизвестным заранее параметрам или сложная визуализация, например, пространственных или многомерных данных, то для этого есть специализированные пакеты. В качестве примера можно привести Tableau. Tableau специализируется на анализе данных через их визуализацию. Tableau можно назвать системой интерактивной аналитики, позволяющая в быстро проводить глубокий и разносторонний анализ больших массивов информации, используя данные из самых разных источников – это могут быть CSV- или просто текстовые файлы, PDF, файлы баз данных, расположенные на внутренних устройствах хранения или в облачных хранилищах. У Tableau есть несколько платных продуктов. Для работы с открытыми данными можно порекомендовать Tableau Online – это облачная платформа с веб-интерфейсом, которую можно использовать бесплатно, но при условии, что все решения будут храниться на общем сервере и будут опубликованы в открытом доступе. Следующий уровень исследования данных – data mining, или интеллектуальный анализ данных. Такой анализ данных представляет собой систематический и последовательный процесс выявления и обнаружения скрытых закономерностей в больших наборах данных. Для этого разработаны различные алгоритмы машинного обучения.

Чтобы применять их, вовсе не обязательно быть программистом. Можно просто воспользоваться специальным программным обеспечением, среди которых Microsoft Azure, Rapid Miner, Weka. Эти программы позволяют производить очистку и подготовку данных, обнаруживать закономерности и аномалии в данных, строить прогнозы и анализировать тексты. Все эти программ-



ные средства обладают удобным графическим интерфейсом. Можно просто загрузить туда свой набор данных, выбрать нужный алгоритм обработки, применить его – и задача решена.

Но вот какие задачи можно ставить в анализе данных, и какие алгоритмы нужно применять для их решения, и как интерпретировать результат – это большой вопрос. Чтобы в нем разобраться, нужно иметь много знаний. Для этого и предназначен наш курс. Но путь предстоит долгий.

Изучив алгоритмы анализа и обработки данных, может появиться желание применить их при помощи средств программирования. Наилучшим выбором тогда будет язык Python, высокоуровневый язык программирования общего назначения, который довольно легко освоить. Python широко применяется в образовательной сфере, для научных вычислений, больших данных и машинного обучения, в веб- и интернет-разработке, графике, GUI, играх и других направлениях.

Для программирования на Python разработано большое число библиотек, что позволяет легко собрать прикладную программу, подключая функции нужных библиотек.

Но наша задача сейчас – первичная обработка данных, поэтому в качестве инструменты мы сосредоточимся на электронных таблицах.

Обычные электронные таблицы привязаны к одному компьютеру, что затрудняет обмен данными. Кроме того, если ваш файл был случайно удален или потерян из-за сбоя компьютера, восстановить информацию было практически невозможно. Сейчас появилось множество облачных хранилищ данных, к которым можно обращаться с различных устройств и хранить там все необходимые файлы, предоставляя доступ к ним при необходимости другим людям.

В 2006 году Google вывела электронные таблицы в Интернет с помощью пакета Google Docs. Теперь в Google Sheets, или Google таблицах можно

создавать электронные таблицы, работать с ними сразу нескольким пользователям в режиме онлайн и обрабатывать данные с любого подключенного к Интернету устройства.

Google Sheets выглядит и функционирует так же, как и любой другой инструмент для работы с электронными таблицами, но поскольку это онлайн-приложение, он предлагает гораздо больше, чем большинство инструментов для работы с электронными таблицами. Например:

- веб-таблицы можно использовать где угодно, их невозможно забыть дома, как обычный файл на компьютере;
- он работает с любого устройства, с мобильными приложениями для iOS и Android вместе с основным веб-приложением;
- Google Sheets бесплатен и включает в себя Google Drive, Документы Word и Слайды PowerPoint для совместной работы и обмена файлами, документами и презентациями в Интернете;
- он включает в себя почти все те же функции электронных таблиц – если вы знаете, как использовать Excel, вы легко сможете справиться и с Google Sheets;
- вы можете загружать дополнения, создавать свои собственные и писать собственный код;
- он находится в сети, поэтому вы можете автоматически собирать данные с помощью вашей электронной таблицы и делать практически все, что захотите, даже если ваша таблица не открыта.

Справедливости ради надо сказать, что есть еще средства для удаленной работы с электронными таблицами – например, Excel Online. Для этого Вам понадобится учетная запись Microsoft. (Ссылку можно найти в полезных ссылках в ИСУ). Также надо зарегистрироваться в outlook.live.com, чтобы иметь доступ к Microsoft Online.

С Excel Online при использовании веб-браузера для создания, просмотра и редактирования книг хранения в OneDrive или Dropbox. Если ваша организация или колледжа подписка Office 365, начните использование Excel Online, Создание и Сохранение книги в библиотеках на вашем сайте.

Office Online сочетает в себе самые популярные функции Office и возможности совместного редактирования в реальном времени, чтобы вы могли и на учебе, и дома работать в команде над общими документами, презентациями и таблицами.

Кроме того, Office Online взаимодействует с установленными на компьютере приложениями Office – вы можете выбирать удобный способ работы. Используйте Office Online для динамического сотрудничества и совместного редактирования в режиме реального времени. А если у вас уже есть Office, продолжайте работать с полнофункциональными приложениями Word, PowerPoint и Excel, установленными на вашем компьютере Mac или под управлением Windows.

Мы рекомендуем Вам в рамках этого курса использовать Google Sheets, и поэтому рассмотрим их функционал более подробно, но Вы также можете выполнять упражнения в обычном MS Excel или Excel Online, если они Вам больше нравятся.

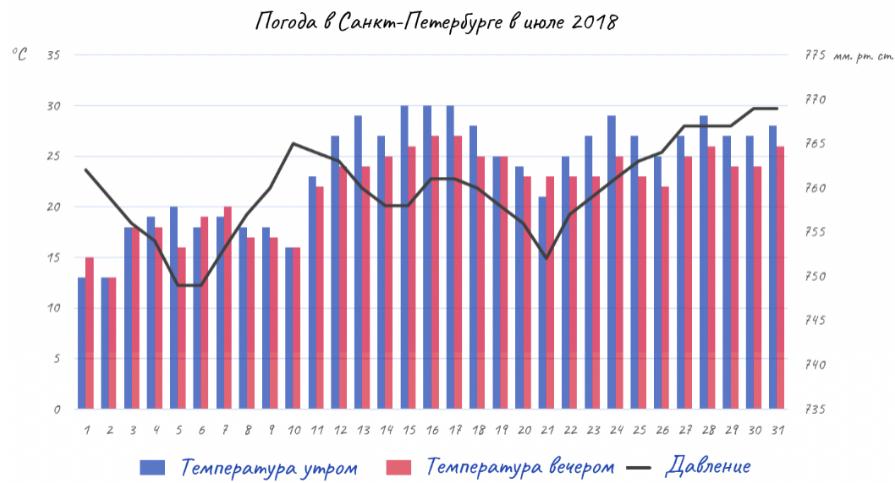
## 2 Задачи визуализации

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. Под визуализацией данных подразумевается представление абстрактной ин-

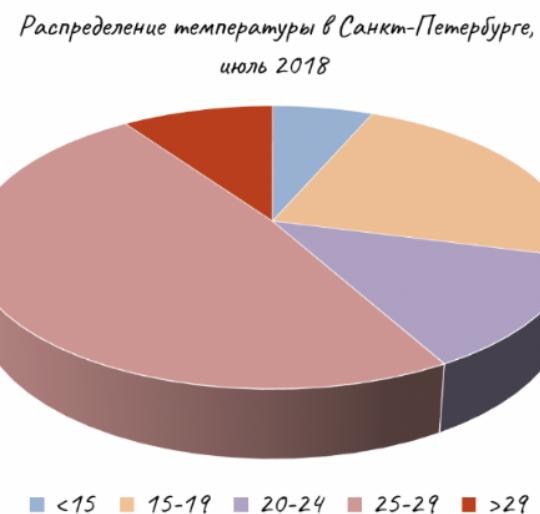
Дата	Температура утром	Температура вечером	Давление
2 июля	13	13	759
3 июля	18	18	756
4 июля	19	18	754
5 июля	20	16	749
6 июля	18	19	749
7 июля	19	20	753
8 июля	18	17	757
9 июля	18	17	760
10 июля	16	16	765
11 июля	23	22	764
12 июля	27	24	763
13 июля	29	24	760
14 июля	27	25	758
15 июля	30	26	758
16 июля	30	27	761
17 июля	30	27	761
18 июля	28	25	760
19 июля	25	25	758
20 июля	24	23	756
21 июля	21	23	752
22 июля	25	23	757
23 июля	27	23	759

формации в графической форме. Визуализация данных позволяет выявлять

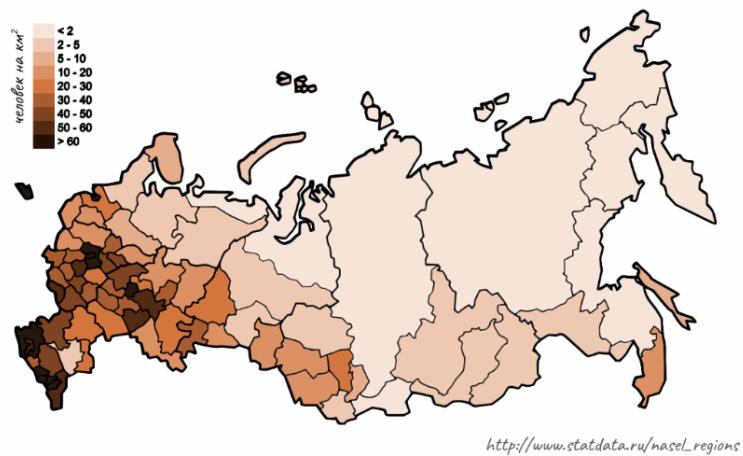
модели, тенденции и корреляции, которые могут остаться незамеченными в традиционных отчетах и таблицах (в том числе электронных). Исследования показывают, что человеческий мозг обрабатывает визуальную информацию в 60 000 раз быстрее, чем текст. 90 процентов информации, передаваемой в мозг, составляют визуальные данные. Попробуйте в ячейках таблицы быст-



ро 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 визуализация так важна?

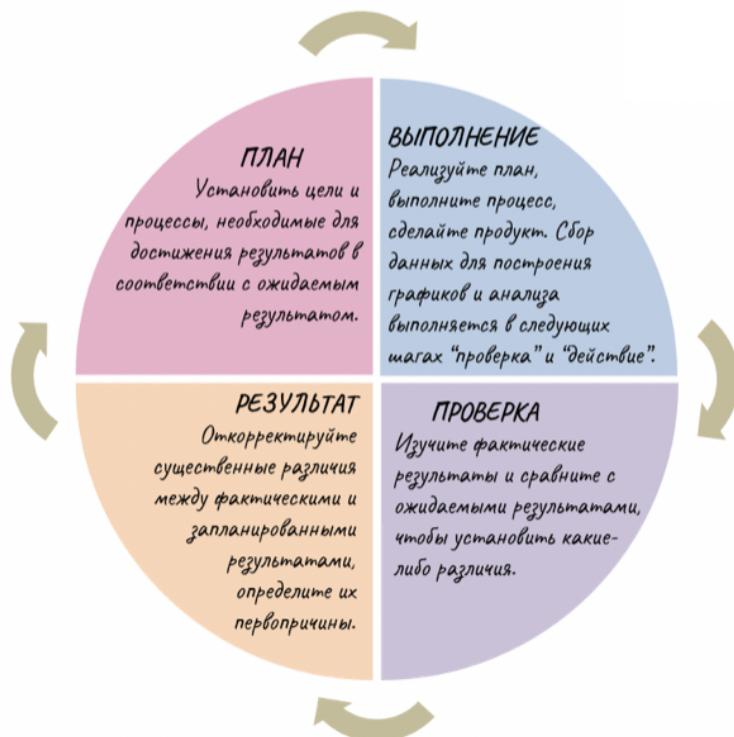


Задача любой визуализации – представить информацию упрощённо, позволив нам с одного взгляда составить о ней определённое мнение. При этом, идеальная визуализация является понятной сама по себе, и может сохранить



свой смысл, даже лишившись всего сопровождающего текста. У людей, принимающих 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. Визуализацию

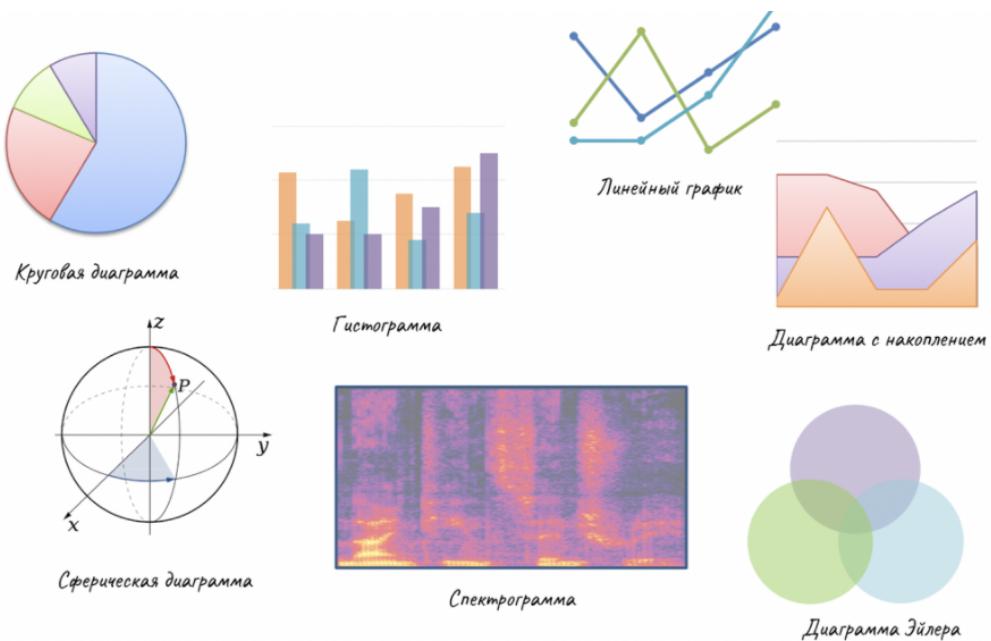


часто используют для генерации идей. Когда целью является решение про-

блемы, выяснение истины. Используется при мозговых штурмах. Типичным представлением является ментальная карта. Как уже было сказано, визуализация используется как самостоятельный вид анализа. Такой вид анализа можно назвать визуальным исследованием. Используется, чтобы лучше по-



нять 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 груп-

не можно отнести всем известные круговые и линейные диаграммы, гистограммы и спектrogramмы, 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.



*Диаграмма производительности*



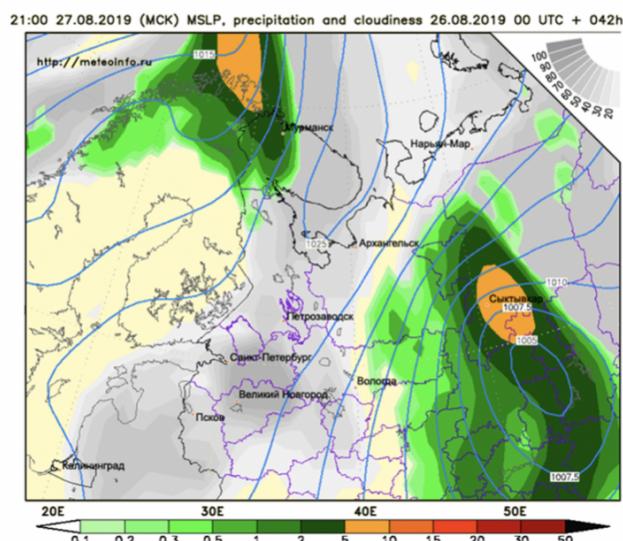
*Диаграмма жизненного цикла*



*Структура организации*

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.



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.

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.

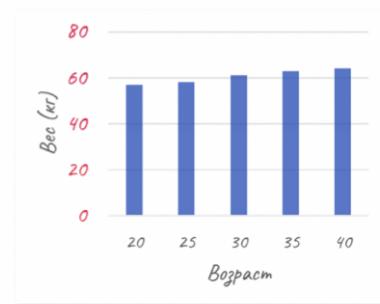
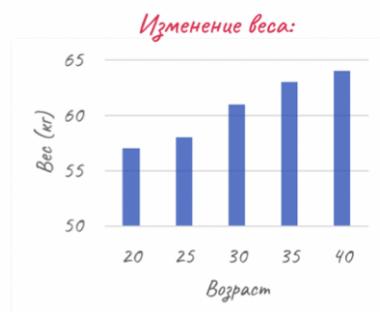
«Цифры часто обманывают меня, особенно когда я сам их организую; и в этом случае замечание, приписываемое Дизраэли, часто применялось бы с правосудием и силой: «Существует три вида лжи: ложь, проклятая ложь и статистика». – Марк Твен

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.



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



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

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.

### 3 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.

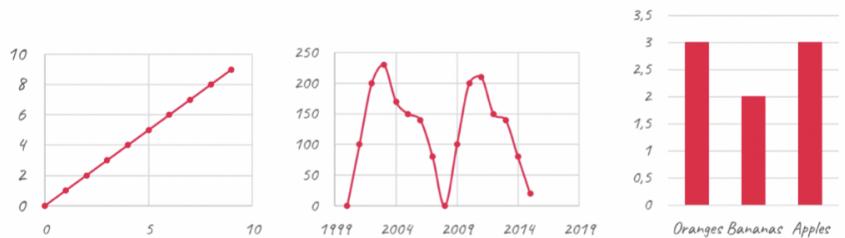
Отношения в данных – это то, как они зависят друг от друга, связь между ними. С помощью отношений можно выявить наличие или отсутствие зависимостей между переменными. Если основная идея информации содержит фразы «относится к», «снижается/повышается при», то нужно стремиться показать именно отношения в данных.

Распределение данных – то, как они располагаются относительно чего-либо, сколько объектов попадает в определенные последовательные области числовых значений. Основная идея при этом будет содержать фразы «в диапазоне от x до у», «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».

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».

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 величин, например график количества продаж товаров в разных магазинах. Географические данные содержат в себе различную информацию, связанную 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. Изображение диаграммы состоит

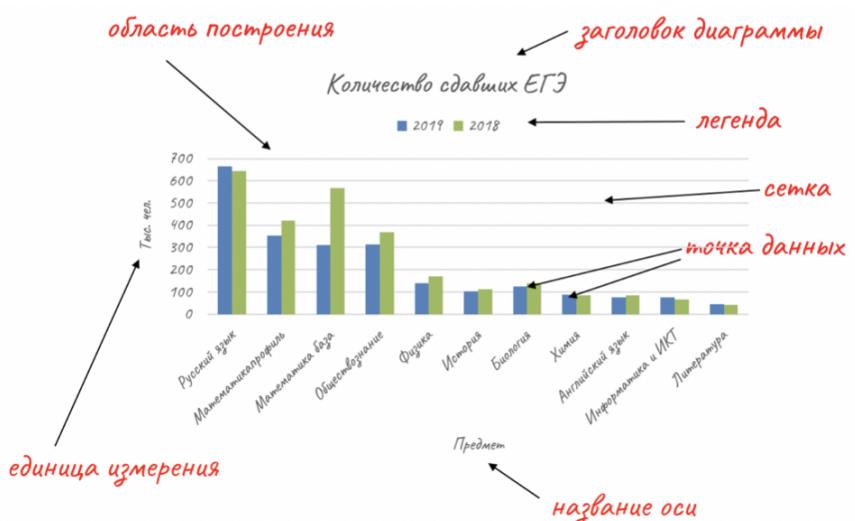


Рис. 2: Элементы диаграммы

из различных элементов – названий осей, единиц измерения, заголовка диа-

граммы, 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.

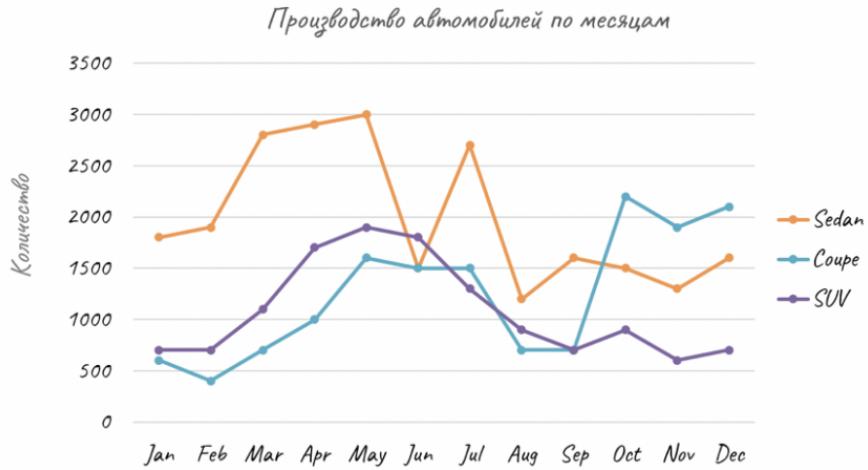


Рис. 3: Линейная диаграмма

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.

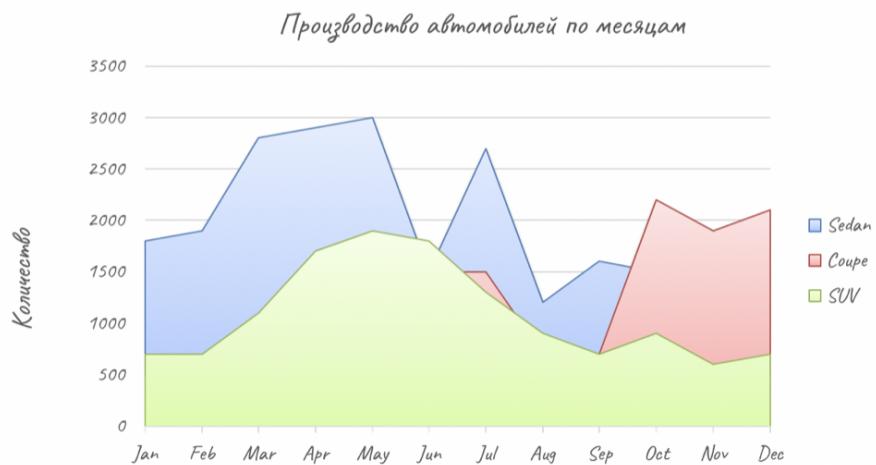


Рис. 4: Диаграмма с областями

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. 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. 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. Диаграмма используется в статистике для

*По горизонтальной оси откладывается диапазон наблюдаемых значений, разбитый на несколько интервалов, а по вертикальной – вероятность или частота попадания в каждый из них.*

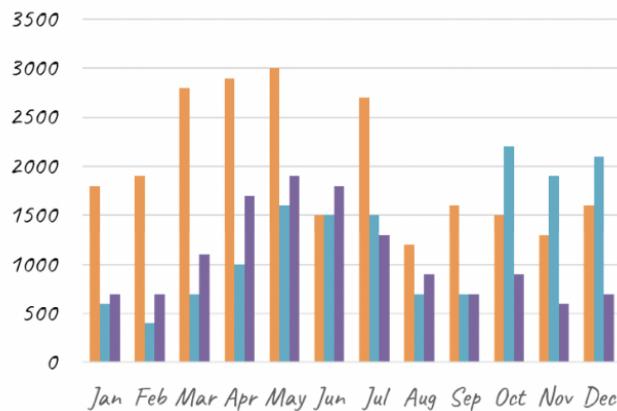


Рис. 5: Гистограмма

графического представления распределения вероятностей значений некоторой случайной величины. По горизонтальной 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,

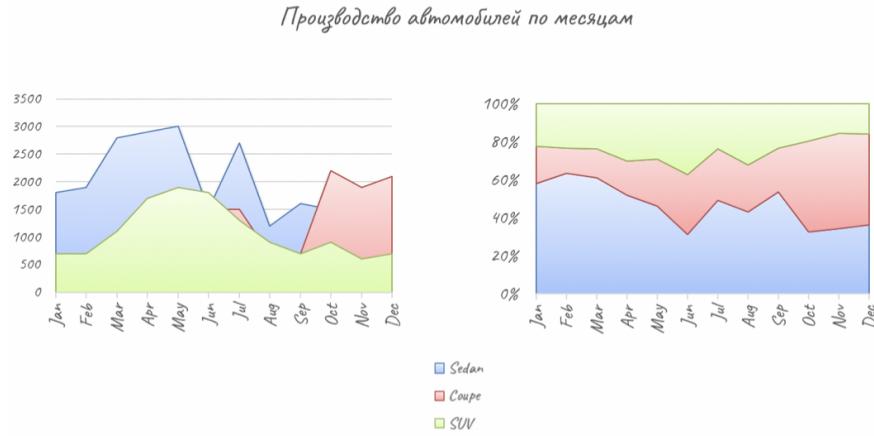


Рис. 6: Нормированная диаграмма

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.

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 Pictogram используют значки, чтобы дать более привлекательный общий вид небольших наборов дискретных данных. Как правило, значки представляют объект или категорию данных, например, данные о населении 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 идеально подходят для представления о пропорциональном распре-

делении 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/реза становится меньше. Это делает их непригодными для больших объемов данных. Для удобства сравнения располагать сегменты следует по

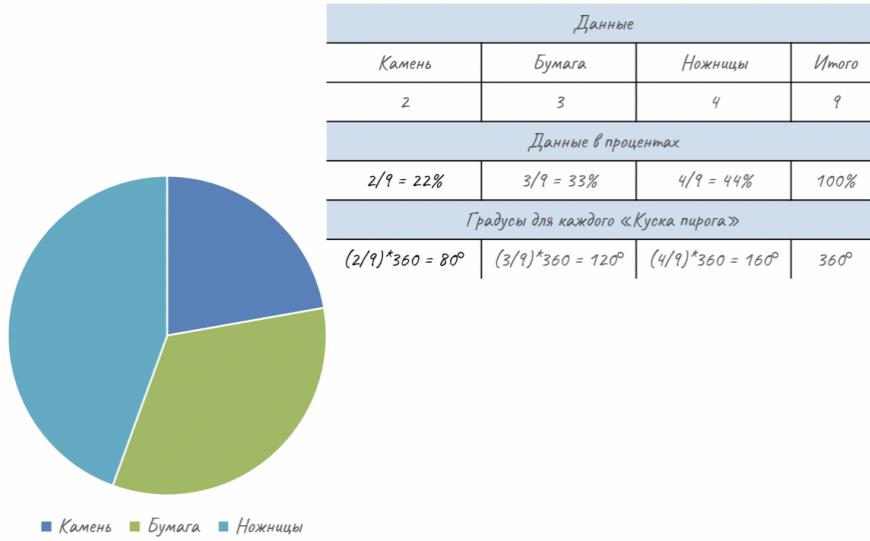


Рис. 7: Круговая диаграмма

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. Пузырьковые диаграммы очень похожи на

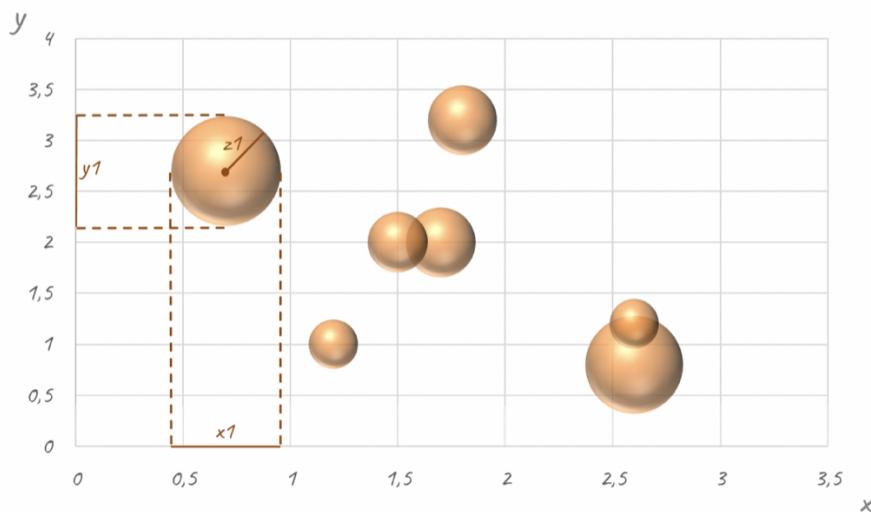


Рис. 8: Пузырьковая диаграмма

диаграммы рассеивания, так как каждая позиция пузыря определяется дву-

мя 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 различия 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.

Несколько таких ящиков можно нарисовать бок о бок, чтобы визуально сравнивать одно распределение с другим; их можно располагать как горизонтально, так и вертикально. Межквартильный размах позволяет определить степень разброса (дисперсии) и асимметрии данных. Свечной график использует

Стоимость бивалютной корзины	
Дата	Стоимость RUB
01.06.2018	66.7716
02.06.2018	66.9363
05.06.2018	66.6582
06.06.2018	66.7211
07.06.2018	66.8922
08.06.2018	66.8777
09.06.2018	67.6736
10.06.2018	67.3031
14.06.2018	68.0754
15.06.2018	67.3411
16.06.2018	67.1166
19.06.2018	67.9832
20.06.2018	68.6207
21.06.2018	68.1149
22.06.2018	68.2422
23.06.2018	67.9579
26.06.2018	67.6124
27.06.2018	67.5689
28.06.2018	67.8465
29.06.2018	67.6998
30.06.2018	67.3625
01.06.2018	66.7716
02.06.2018	66.9363
05.06.2018	66.6582
06.06.2018	66.7211
07.06.2018	66.8922
08.06.2018	66.8777
09.06.2018	67.6736
10.06.2018	67.3031
14.06.2018	68.0754
15.06.2018	67.3411
16.06.2018	67.1166
19.06.2018	67.9832
20.06.2018	68.6207
21.06.2018	68.1149
22.06.2018	68.2422
23.06.2018	67.9579
26.06.2018	67.6124
27.06.2018	67.5689
28.06.2018	67.8465
29.06.2018	67.6998
30.06.2018	67.3625

Рис. 9: Тепловая таблица

зуется 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 фиксиро-

ванный период времени, и отображают 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 – это тип визуализации, в которой цвет выступает в качестве дополнительного измерения. Тепловые карты позволяют увидеть важные переменные в цвете как функцию двух других переменных.

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

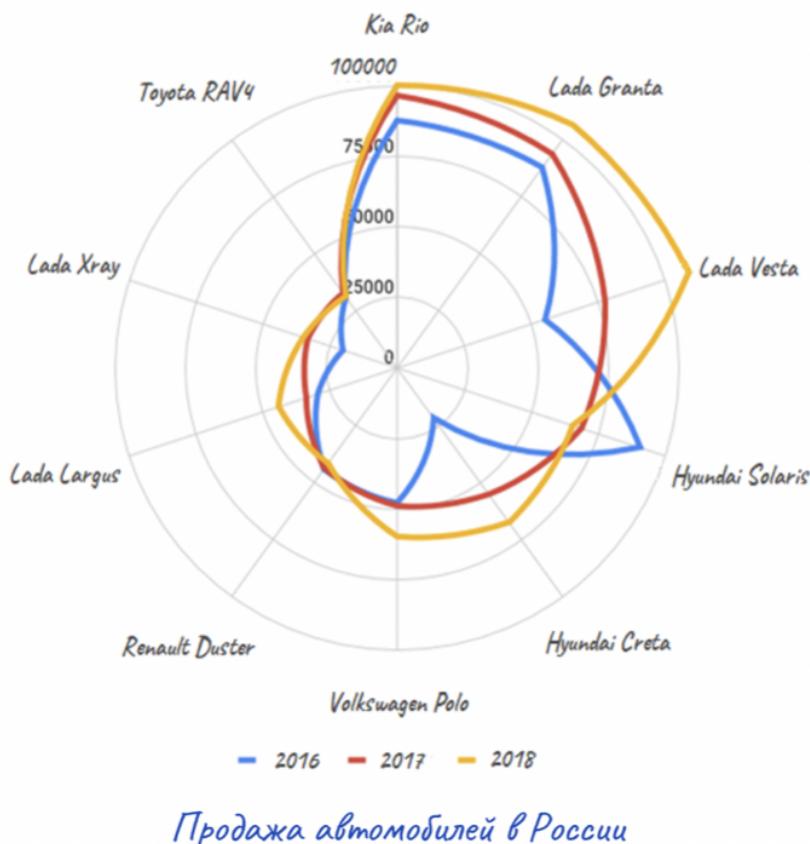


Рис. 10: Радарная диаграмма

в таблице. Тепловые карты облегчают процесс восприятия больших массивов данных и необязательно связаны с отображением информации на географи-

ческой 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.

Радарные диаграммы – это способ сравнения значений нескольких количественных переменных (если они соизмеримы). Каждой переменной представляется 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 на-

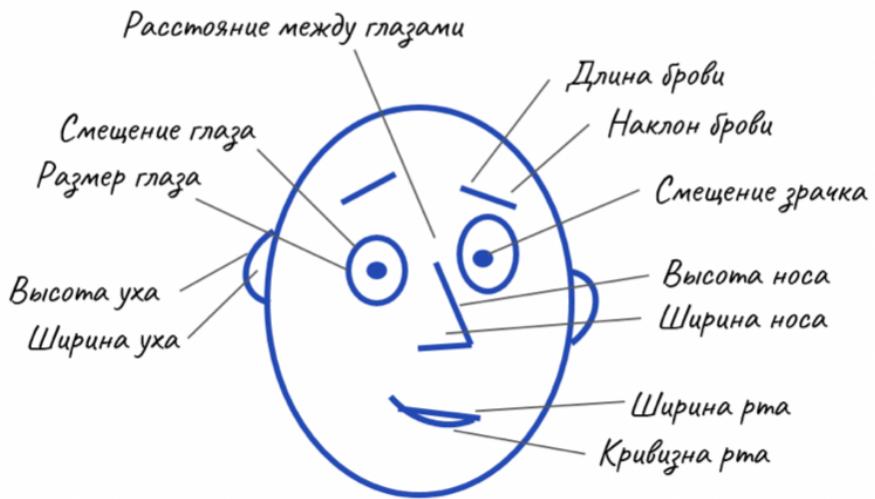


Рис. 11: Лица Чернова

блудения 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.

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