Презентация:
`https://www.overleaf.com/9372798468hxfmdkyhjmnw`

Видео 0

В. Всем привет. Это - Виталия Елисеева и

Н. Анастасия Небольсина,

В . Мы приветствуем вас на последнем блоке видео к курсу Анализа данных, на блоке про красивую и эффективную визуализацию полученных данных

Н: (или как избежать вырвиглазного эффекта)

В. О чём будет этот блок?

Н. Мы будем обсуждать
-- как выбрать тип графика в зависимости от того, какие у вас данные
-- как оформлять легенду, выбирать масштаб осей
-- какие основные ошибки допускают при составлении графиков
-- Как сделать так, чтобы график не искажал информацию и не обманывал зрителя

В: А почему мы вообще заинтересованы в оформлении графиков? Разве автоматически создаваемые в Экселе графики могут быть ужасными?

Н: Давай тогда посмотрим на реальные примеры графиков из интернета, в которых ВСЁ пошло не так. (*слайд 2)* Например, что произошло на этом графике?

В: О, этот график -- типичный продукт Экселя, который автоматически присвоил цвета переменным. Но для человеческого глаза несоответствие цветов на круговой диаграмме и цветов-названий переменных вызывает когнитивный диссонанс.

(*слайд 3)* А что не так с этим графиком?

Н: Здесь изображена операционная прибыль какого-то предприятия. На первый взгляд кажется, что квартальная прибыль выросла, но на самом деле она упала потому, что на графике 2017 и 2018 годы поменяны местами. А ещё расположение цифр внутри/над столбцами также намеренно создаёт иллюзию роста прибыли. Такой график намеренно искажает информацию для зрителя.

(*слайд 4)* А что не так со следующим графиком?

В: Здесь решили создать уродливого мутанта из пай-чарта и облака тегов. На обычную круговую диаграмму сверху наложили частично нечитаемое облако тегов.

Н: Согласна, выглядит просто ужасно.

В. Итак, нас ждут 2 выпуска и первый выпуск будет посвящен основным типам графиков и правилам их оформления

Н: Не переключайтесь!

Видео 1

В: В этом видео мы поговорим об основных типах графиков и об ошибках, которые можно допустить при их создании. Чем обоснован наш выбор типа графика?

Н: (*слайд 6)* Чаще всего типом переменной. Мы всегда хотим выбрать такой тип графика, который подчеркнёт ключевые моменты наших данных.
-- Мы можем использовать столбчатую диаграмму или гистограмму, когда одна из переменных -- количественный признак, а вторая качественный или количественный
-- Гистограмма с накоплением вместо простой гистограммы используется, когда важно показать долю/соотношение, а не абсолютное количество

В:
-- Пай-чарт и боксплот мы используем, когда одна из переменных количественная, а вторая -- категориальная (или искусственно созданная категориальная из количественной) и категорий <5.
-- А карты-инфографики мы используем, когда одна из переменных количественная/качественная, а другая -- регион/географический объект

Н: (*слайд 7)* Здесь мы используем столбчатую диаграмму для демонстрации данных о весе чемодана, с которым участники ЛЭШ приезжали в прошлом году, когда школа была в офлайне. Для столбчатой диаграммы нам нужны категории для создания столбиков. Каждый столбик -- это среднее значение внутри категории. Справа наши категории существуют так как переменная была категориальной. Слева мы создали категории самостоятельно из количественной переменной.

В: А почему одна из диаграмм горизонтальная, а другая -- вертикальная?

Н: На самом деле, здесь это исключительно вопрос эстетики. Названия федеральных округов не помещались в ширину столбиков в вертикальном варианте, поэтому мы “перевернули” диаграмму.

Что нам нужно понимать при создании категорий из количественной переменной?

В: (*слайд 8)* Одна из главных ошибок -- когда вместо категорий на ось наносят индивидуальные значения. Тогда каждый столбик вместо среднего значения в

какой-либо группе индивидов показывает индивидуальное значение. Такой график бессмысленный.

Также столбчатую диаграмму часто путают с гистограммой. Почему так происходит?

Н: (*слайд 9)* Визуально они выглядят практически одинаковыми, но в гистограмме вместо средних значений используются частоты. То есть количество респондентов, принадлежащих к той или иной категории.

В: (*слайд 10)* Что в столбчатой диаграмме, что в гисторамме важно выбрать оптимальную длину интервалов, если вы создаёте их из количественной переменной. Лучше всего вручную попробовать создать графики с различной длиной интервалов.

Какая длина интервала лучше всего подходит для этого графика?

Н: Мы видим, что на гистограмме (a) слишком много шума и тяжело заметить тренд в данных. На гистограмме (d) исчезло резкое падение количества индивидов, проживающих на расстоянии 300 метров от ЛЭШ. Поэтому нам подходят гистограммы (b) и (c), на которых видны тренды в данных.

В: (*слайд 11)* Отдельный тип гистограммы -- гистограмма с накоплением. Мы можем захотеть использовать её вместо нескольких гистограмм для различных групп данных. К примеру, здесь мы объединили гистограмму количества мужчин среди различных групп населения ЛЭШ и количества женщин среди различных групп населения ЛЭШ.

Н: Но если просто поставить один столбик на другой, то высота столбцов про женщин не может быть легко сопоставлена одна с другой потому, что они начинаются и заканчиваются на разной высоте. Один из способов решения этой проблемы -- перевести количество мужчин/женщин из абсолютной переменной в относительную. Тогда мы можем сопоставлять долю женщин и мужчин в каждой категории. Теперь легко сопоставлять доли женщин так, как столбцы начинаются на одинаковой высоте.

В: (*слайд 12)* Или другой способ решения проблемы (особенно если столбцов много и мы хотим сравнить форму распределения данных мужчин и женщин) -- *age pyramid*. Такое название дано так как чаще всего он используется для описания половозрастной структуры данных.

Раз мы начали говорить о распределениях данных -- какой следующий тип графика, который позволит нам узнать больше статистических сведений о нашей выборке?

Н: (*слайд 13)* Это боксплот, который показывает распределение какого-то либо количественного признака. Как работает боксплот?

The line in the middle of the boxplot represents the median, and the box encloses the middle 50% of the data. The top and bottom whiskers extend either to the maximum and minimum of the data or to the maximum or minimum that falls within 1.5 times the height of the box, whichever yields the shorter

whisker. The distances of 1.5 times the height of the box in either direction are called the upper and the lower fences. Individual data points that fall beyond the fences are referred to as outliers and are usually showns as individual dots.

В: Каждый индивидуальный боксплот отвечает распределению количественного признака внутри какой-либо группы. Но можно разместить несколько боксплотов рядом друг с другом, чтобы сравнить их распределения.

(*слайд 14)* In that figure, we can now see that temperature is highly skewed in December (most days are moderately cold and a few are extremely cold) and not very skewed at all in some other months, for example in July.

Н: (*слайд 15)* И последний тип графика, который можно использовать для одного количественного признака -- пай-чарт или круговая диаграмма. На нём можно показать только пропорции/доли, а за целое мы берём площадь круга, то есть 100%. Одним из его плюсов является то, что он позволяет легко выделить простые доли, например, 1/2, 1/3, 1/4.

В: Один из главных минусов пай чарта -- невозможность визуально сопоставлять значения с соседних графиков. When we visualize this dataset with pie charts, it is difficult to see what exactly is going on. It appears that the market share of company A is growing and the one of company E is shrinking, but beyond this one observation we can't tell what's going on. In particular, it is unclear how exactly the market shares of the different companies compare within each year.

Круговая диаграмма -- это, наверное, самый распространённый тип графика, который можно увидеть в реальной жизни, и при этом тот, в котором допускают больше всего ошибок. Каких, например?

Н: (*слайд 16)* Самая распространённая -- когда доли не суммируются в 100%. Это происходит, к примеру, когда в один пай чарт суммируют несколько количественных переменных.

В: (*слайд 17)* Другая ошибка -- когда размера доли не соответствует площади, занимаемой на графике.

Н: До этого момента мы обсуждали только то, как можно нарисовать один количественный признак. Если мы хотели изобразить одновременно 2 количественных признака, то один из них мы превращали в категориальный вручную. Единственный способ изобразить два количественных признака одновременно -- скаттерплот.

В: (*слайд 18)* Каждая точка на скаттерплоте представляет собой единицу наблюдения с двумя признаками, отложенными на осях x и y.

A line of best fit (or “trend” line) is a straight line that best represents the data on a Scatter Plot. This line may pass through some of the points, none of the points, or all of the points.

Н: (*слайд 19)* The blue jay dataset contains both males and females, and we may want to know whether the overall relationship between heigth and body mass holds up separately for each sex. To address this question, we can color the points in the scatter plot by the sex. This figure reveals that the overall trend is at least in part driven by the sex of the birds. At the same time, females tend to be lighter than males on average.

В: (*слайд 20)* И боксплот, и скаттерплот удобны тем, что они позволяют увидеть выбросы на графиках. Чаще всего, в вашем случае выбросы будут появляться от некорректно собранных данных (к примеру, рост в тысячах сантиметров). Один из способов предотвратить это -- устанавливать ограничения на вводимые значения в гугл-формах.

Н: (*слайд 21)* Другая распространённая проблема -- когда много индивидов имеют одни и те же значения. В таком случае точки на скаттерплоте будут накладываться друг на друга. Решение в таком случае -- сделать полупрозрачные точки, shade of the points reflects the density of points in that location of the graph.

Какие типы визуализации данных у нас остались неохваченными?

В: (*слайд 22)* Иногда данные лучше презентовать в виде таблицы, чем графика. В чём проблема этой таблицы?
-- Текстовые ячейки лучше выравнивать по левому краю;
-- Числовые ячейки лучше выравнивать по правому краю, чтобы было легче отсчитывать разряды числительных и сопоставлять;

Н: *(слайд 23)* О, а теперь к имеющимся проблемам прибавилась расцветка, в которой шрифт меняет цвет. Вообще говоря, таблицы лучше бы делать однотонными.

В: *(слайд 24)* Так, стало гораздо лучше потому, что теперь ячейки правильно выровняны. Также в таблице (c) ушли лишние вертикальные и горизонтальные линии, которые мешали воспринимать данные.

Н: Итак, это был последний тип визуализации, который мы хотели обсудить. В следующем видео мы поговорим об основных ошибках, которые можно допустить в любом типе графика.

Видео 2

Н: В этом видео мы обсудим то, как сделать график удобным и эстетичным, а также как избежать основных ошибок, которые искажают информацию для зрителя.

Какие основные ошибки существуют?

В: *(слайд 26)* У каждого графика всегда должно быть название, легенда и подписанные оси. График должен быть виден и понятен без дополнительных комментариев.

Н: *(слайд 27)* А у этого графика оси не нормированы -- значения по оси у невозможно сравнивать друг с другом.

В: *(слайд 28)* У этого графика выровняна ось у, но теперь проблема с излешней красочностью -- использовать разные цвета нужно только в том случае, если это необходимо.

Н: *(слайд 29)* А вот с этим графиком всё в порядке.

(слайд 30) Продолжая разговор о таких обязательных частях графика, как легенда -- легенда не должна пытаться обмануть зрителя. Это пример из официальной коронавирусной статистики штата Джорджия, графики появились на сайте с разницей в неделю. Если не вчитываться в легенду, то создаётся представление, что ситуация в штате практически не изменилась. Но при этом легенда показывает увеличение количества выявленных случаев коронавируса в 1,5 раза.

Какие ещё ошибки существуют в графиках?

Н: *(слайд 31)* К примеру, проблемы с масштабом -- размещение слишком большого количества графиков на одном слайде. На этом графике изображены скаттерплоты за каждый год данных. Но невозможно разглядеть что-либо.

В:*(слайд 33) проблема с осями -- их пропорциональность.*

Median income in the five counties of the state of Hawaii. This figure is misleading, because the y axis scale starts at $50,000 instead of $0. As a result, the bar heights are not proportional to the values shown, and the income differential between the county of Hawaii and the other four counties appears

much bigger than it actually is. An appropriate visualization of these data makes for a less exciting story. While there are differences in median income between the counties, they are nowhere near as big as the first figure suggested. Overall, the median incomes in the different counties are somewhat comparable.

Н.: Когда мы говорим о барчатах и не только, возникает вопрос о том, как расположить столбцы, есть ли принципиальная разница?

В.: *(слайд 34)* For example, Figure 6.5 shows the median annual income in the U.S. by age groups. In this case, the bars should be arranged in order of increasing age. Sorting by bar height while shuffling the age groups makes no sense (Figure 6.6).

Pay attention to the bar order. If the bars represent unordered categories, order them by ascending or descending data values.

(слайд 35) Перемешаны даты. Штаты на легенде в алфавитном порядке, но на графике отсортированы так, чтобы показывать уменьшение кейсов короны.

Н.: *(слайд 36-37)* А теперь мы обратимся к графической стороне визуализации и рассмотрим элементы, которые ухудшают внешний вид и читаемость графика

Сравнивая эти слайды, мы можем заметить, что наличие сетки на фоне избыточно и усложняет восприятие информации, отображаемой на графике, поэтому стоит избегать избыточных графических деталей, делая Вашу визуализацию четкой и изящной.

В.: Также не стоит усложнять внешний вид графика, используя 3D и тени

Н.:
(слайд 38) As an example, let's take a simple pie chart with two slices, one representing 25% of the data and one 75%, and rotate this pie in space (Figure 26.1). As we change the angle at which we're looking at the pie, the size of the slices seems to change as well. In particular, the 25% slice, which is located in the front of the pie, looks much bigger than 25% when we look at the pie from a flat angle (Figure 26.1a).

В.:(слайд 39)
Продолжая тему излишней графики, можем посмотреть на этот bar chart.
This figure is labeled as "bad" because the 3D perspective makes the plot difficult to read. It is difficult to judge exactly how tall the individual bars are, and it is also difficult to make direct comparisons. For example, was the mortality rate of urban females in the 65–69 age group higher or lower than that of urban males in the 60–64 age group?

Н: (слайд 40) Теперь посмотрим на инфографику. Обычно её используют при изображении geospatial data -- information linked to locations in the physical world.

(слайд 41) Эта инфографика плоха всем:
-- на ней изображены только 2 страны, что делает её бесполезной
-- шкала military spending максимально размытая
-- тяжело ассоциировать цвет со шкалы с цветом страны

В: (слайд 42) Эта инфографика уже лучше:
-- доход разбит на 5 чётких категорий
-- легко ассоциировать цвета с карты с цветами на легенде

Но всё ещё есть минус -- Аляска очень крупная, а небольшие штаты на восточном побережье видны плохо

В: (слайд 43-44) *Подвести итог, огласив тезисно пройденный материал, указать, откуда брались графики и где школьники могут поподробнее узнать материал*