Видео 0

Д. Всем привет. Это - Дмитрий Покровский и
В. Иван Станкевич, и мы приветствуем вас на втором блоке видео к курсу Анализ данных
Д. Ваня, в предыдущих видео мы обсудили то, как выбирать тему для исследования, и как составлять анкету. А о чем будет эта серия видео?
В: Мы поговорим о типах данных, статистических показателях и мерах связи между признаками
Д: Будут формулы или примеры?
В: Будем обсуждать содержательные моменты, а в качестве примеров будем рассматривать олимпиадные задачки
А какие книжки наши участники могут посмотреть дополнительно к нашим видео?
Д. Манга Занимательная Статистика и Статистика (очень краткое введение),
В. Итак, нас ждут 5 выпусков, и первый выпуск будет посвящен типам переменных и шкалам.

Видео 1

Д: Тема нашего видео: типы показателей и шкалы. Что это вообще такое “тип показателя”?
В: Типы переменных, просьба к Д - привести примеры
Д: Срок хранения молока (днях) , - качеств, т.к про качество молока
В: Нет потому что это количественная переменная - дни
Д: а какая переменная качественная?
В: цельное\нецельное молоко
Д: В задачке про хомяков - масса - это колич непрерывная переменная, верно?
В: точно, а активность?
Д: ранговая, потому что часы бывают 1,2,3, 4,.. И т.д.
В: нет, потому что…, а вот образование - как раз ранговая, потому что…..
Д: где ноль , и почему масса ограничена 40?/
В: …. Кстати, ответь на вопрос а)

Д: Связь линейная, а не стоит делать так делать, т.к. мало категорий, лучше - срок образования в месяцах.
В: Частично правильно, но это не критично а главные недостатки в другом….
Д: Здорово, а я напоминаю, что больше информации можно прочесть в книжке….

Видео 2

показывать картинки с ввп, типа:
Просто ВВП: https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?most_recent_value_desc=true&view=map
ВВП на душу: https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?most_recent_value_desc=true&view=map
А потом ВВП по ППС: https://data.worldbank.org/indicator/NY.GDP.MKTP.PP.CD?most_recent_value_desc=true&view=map
И ВВП по ППС на душу: https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.CD?most_recent_value_desc=true&view=map
Covid 19 https://www.worldometers.info/coronavirus/#countries

Д. Нужно ли что-то делать с собранными данными до начала анализа?
В. Бывает полезным, потому что есть абсолютные и относ показатели…
Д. Можно какие-нибудь примеры? И желательно экономические
В. Про ВВП и ВВП на душу
Д. Это похоже на то, как если бы я спрашивал у своих друзей насколько большая у них квартира, а на самом деле меня должно интересовать сколько кв метров приходится на одного члена семьи
В. Да, верно...
Д. Правильно ли я понимаю, что чем выше ВВП на душу, тем лучше живут люди?
В. Не всегда, иногда нужно сравнивать с нормой, нормировка, например цены разных лет нужно корректировать на инфляцию…...
Если говорить про Исландию и Нидерланды: где бы хотел жить?
Д. В Исландии, так как там ВВП на душу выше, но как далеко от всего…
В. На самом деле ВВП по ППС в Нидерландах выше, жить лучше там
Д. Если бы у меня были данные о з/п отдельных людей, то чтобы судить о том насколько они богаты я должен был бы еще знать средний уровень цен
В. Да, а лучше - стоимость потребительской корзины.
Д. А как применить эту технику относительных показателей к пандемии?

В. Можно говорить о смертности общей (как ВВП) и смертности на 10000 (стандпоказатель) или о смертности к числу заболевших + картинки (таблички) + вопрос: где ситуация хуже?

Видео 3

Д: мы все интуитивно понимаем что такое “среднее”, когда говорим о среднем росте, средней цене или среднем времени ожидания. А зачем вообще нужна такая характеристика как среднее?
В: Чтобы описать центр распределения
Д: А какие еще бывают показатели центра ?
В: Разные, в том числе ср.-взвешенное и медиана
Д: А как мы можем графически показать средние?
В: гистограмма, ящичковые диаграммы
Д: А можем мы на одном графике показывать разные средние, и вообще их сравнивать?
В: Иногда да, иногда нет…. Все сильно зависит от того, как собираются данные, от выборки… Вот например задача из ВП 2016. Как бы ты собирал данные

Д: Ну вариант № 3 очень странный, потому что (*смещенная выборка*)
В: Да, а еще потому что… И что же ты выберешь из оставшихся?
Д: я бы выбрал первый вариант: просто по алфавиту обзвонил бы
В: На самом деле лучше второй
Д: Спасибо, а где бы наши слушатели могли больше узнать про средние и про выборки?
В: Манга, Статистика и котики, Статистика в комиксах, Статистика (очень краткое введение)

Видео 4

Доля богатства у 10% самых богатых:
https://data.worldbank.org/indicator/SI.DST.10TH.10?most_recent_value_desc=true&view=map&year=2016

GINI https://data.worldbank.org/indicator/SI.POV.GINI?most_recent_value_desc=true&view=map&year=2016

Д. В прошлом видео говорили про средние , достаточно ли это чтобы описывать совокупность в целом?
В. Нет, потому что надо учитывать вариативность….
Д. Какие есть показатели неоднородности?
В. Размах, СКО, коэффициент вариации
Д. И когда какой надо применять?
В. Пример про богаство на основе 10%
Д. А как же 10% снизу? Нет ли чего более общего?
В. Джини…
Д. Это интересно, но все про неравенство по доходу, а есть ли универсальные показатели?
В. Есть СКО
Д. А что делать, если в двух выборкаъ разные средние и вариация, как сравнивать?
В. Коэффициент вариации

Видео 5

Д. Как понять связаны ли и как сильно значения 2-х показателей
В. Вспомним 1 видео и картинку - скатерплот, который показывает линейную связь
Д. И как понять насколько связь сильная?
В. Чем более узкое и вытянутое облако вдоль линии - тем лучше
Д. Хотелось бы число…

В. Коэффициент корреляции.
Д. Где его взять?
В. Есть формула, а в Excel есть даже функция
Д. То есть корр можно считать для любых переменных?
В. Не совсем так, например… для колич можно, а для категориальных есть другое, вот например для бинарных можно пользоваться таблицей сопряженности

Д. Есть ли подводные камни?
В. Есть ложные корреляции.
https://www.tylervigen.com/spurious-correlations