

## Лабораторная работа №1

### Структура и описание данных

**Цель:** Обзор основных стандартов работы с данными. Знакомство с этапом понимания данных стандарта CRISP-DM.

Задача этапа – понять, что именно содержится в этих данных, какие факты они описывают, понять, что они могут дать.

#### **Задания**

Откройте файл с данными. Приведите его к формату, подходящему для обработки.

1. Выясните структуру данных. Для этого:
2. Разделите переменные на категориальные и количественные
3. Для каждой переменной определите возможные шкалы, выберите шкалу, которая даст возможность получить больше информации.

Для категориальных переменных укажите – номинальными или порядковыми они являются.

Определите, какие качественные переменные могут задавать структуру набора данных. Какие из них будет полезно использовать для фильтрации, разделения выборки и выяснения того, какие зависимости есть в выборке, что может помочь ответить на те вопросы, которые заданы в описании набора данных, выявить, какие исследования можно провести. В вашем исследовании должны быть отражены частоты появлений разных значений качественных переменных; распределения количественных переменных по выборке, распределения, полученные по разным категориям

4. Используя Excel, постройте сводные таблицы для категориальных переменных и таблицы частот для количественных
5. Постройте диаграммы, которые иллюстрируют структуру, для качественных переменных (4-6 штук, с заголовками, подписями осей и пояснениями, что они показывают)
6. Используйте значения качественных переменных как фильтры для количественных
7. Для количественных переменных постройте гистограммы и графики (кумулятивный, частотный, простой, диаграмма рассеяния - какие будут уместны) которые иллюстрируют распределения (4-6 штук, с заголовками, подписями осей и пояснениями, что они показывают)
8. Посчитайте возможные описательные характеристики переменных
- 9. Должны быть задействованы в исследовании все переменные выборки**
10. Сформулируйте, какие задачи могут решаться с применением этих данных. Приведите несколько формулировок

***Теоретические сведения.***

Измерения и шкалы. **Шкалы**

*Существует пять типов шкал измерений: номинальная, порядковая, интервальная, относительная и дихотомическая.*

**Номинальная шкала** (nominal scale) - шкала, содержащая только категории; данные в ней не могут упорядочиваться, с ними не могут быть произведены никакие арифметические действия.

**Порядковая шкала** (ordinal scale) - шкала, в которой числа присваивают объектам для обозначения относительной позиции объектов, но не величины различий между ними. равно ( $=$ ), не равно ( $\neq$ ), больше ( $>$ ), меньше ( $<$ )

**Интервальная шкала** (interval scale) - шкала, разности между значениями которой могут быть вычислены, однако их отношения не имеют смысла.

Эта шкала позволяет находить разницу между двумя величинами, обладает свойствами номинальной и порядковой шкал, а также позволяет определить количественное изменение признака.

Номинальная и порядковая шкалы являются дискретными, а интервальная шкала - непрерывной, она позволяет осуществлять точные измерения признака и производить арифметические операции сложения, вычитания, умножения, деления.

равно ( $=$ ), не равно ( $\neq$ ), больше ( $>$ ), меньше ( $<$ ), операции сложения ( $+$ ) и вычитания ( $-$ ).

**Относительная шкала** (ratio scale) - шкала, в которой есть определенная точка

отсчета и возможны отношения между значениями шкалы.

Пример такой шкалы: вес новорожденного ребенка (4 кг и 3 кг). Первый в 1,33 раза тяжелее

Относительные и интервальные шкалы являются числовыми.

( $=$ ), ( $\neq$ ), ( $>$ ), ( $<$ ), ( $+$ ), ( $-$ ), умножения ( $*$ ) и деления ( $/$ ).