Институт информационных технологий

Кафедра: Математическое и программное обеспечение ЭВМ

Дисциплина: Основы Data Science

ЛАБОРАТОРНАЯ РАБОТА № 1

Тема: Структура и описание данных

Выполнил:

студент гр. 1ПИб-02-2оп-23

Кринкин Олег Алексеевич

Проверил:

Юдина Ольга Вадимовна

# 1. Задание

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

1. Используя Excel, постройте сводные таблицы для категориальных переменных и таблицы частот для количественных
2. Постройте диаграммы, которые иллюстрируют структуру, для качественных переменных (4-6 штук, с заголовками, подписями осей и пояснениями, что они показывают)
3. Используйте значения качественных переменных как фильтры для количественных
4. 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)
5. Посчитайте возможные описательные характеристики переменных
6. Должны быть задействованы в исследовании все переменные выборки
7. Сформулируйте, какие задачи могут решаться с применением этих данных. Приведите несколько формулировок

# 2. Ход работы

1. Данные из csv файла импортированы в xlsx-файл путём выбора *Данные->Получить->Из текста*. В качестве разделителя выбран символ запятой. Также для числовых значений с плавающей точкой был заменён разделитель на точку.
2. Для каждой переменной была определена шкала и указано номинальная ли она или порядковая, а также качественная или количественная и имеет ли смысл использовать её для фильтра.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Тип | Шкала | Фильтр |
| Sex | Категориальные | Номинальная | Да |
| School Code | Категориальные | Номинальная | Да |
| Playing Often | Категориальные | Порядковая | Да |
| Playing Games | Категориальные | Номинальная | Да |
| Mother Education | Категориальные | Порядковая | Да |
| Father Education | Категориальные | Порядковая | Да |
| Parent Revenue | Категориальные | Порядковая | Да |
| Playing Years | Количественные | Относительная | Да |
| Playing Hours | Количественные | Относительная | Да |
| Grade | Количественные | Относительная | Нет |

1. Для общего исследования сформулирована тема: «Как влияет воспитание ребёнка на его времяпрепровождение».
2. Для исследования были выбраны следующие сводные таблицы и определены следующие вопросы о структуре данных:

* Пол/Количество – какое соотношение пола в данных?
* Образование матери, отца/Соотношение – каково распределение учеников по образованию родителей?
* Время игры/Количество – каково распределение учеников по времени игры в день?
* Достаток родителей/Соотношение – каково распределение выборки учеников по достатку родителей?
* Оценка/Количество – каков уровень образования учеников в выборке?
* Школа/Количество – какое количество учеников из какой школы?
* Игровой стаж/Количество – как много человек, вовлечённых в игры уже долгое время?

1. Получены следующие сводные таблицы:

|  |  |
| --- | --- |
| **Названия строк** | **Количество по полю Sex** |
| 0 | 386 |
| 1 | 384 |
| **Общий итог** | **770** |

|  |  |
| --- | --- |
| **Названия строк** | **Количество по полю Mother Education** |
| 0 | 0.13% |
| 1 | 3.51% |
| 2 | 21.43% |
| 3 | 25.84% |
| 4 | 31.82% |
| 5 | 13.77% |
| 6 | 3.51% |
| **Общий итог** | **100.00%** |

|  |  |
| --- | --- |
| **Названия строк** | **Количество по полю Father Education** |
| 0 | 0.52% |
| 1 | 0.78% |
| 2 | 15.97% |
| 3 | 23.12% |
| 4 | 34.03% |
| 5 | 19.87% |
| 6 | 5.71% |
| **Общий итог** | **100.00%** |

|  |  |
| --- | --- |
| **Названия строк** | **Количество по полю Playing Hours** |
| 0 | 29.48% |
| 1 | 25.06% |
| 2 | 23.51% |
| 3 | 14.68% |
| 4 | 3.64% |
| 5 | 3.64% |
| **Общий итог** | **100.00%** |

|  |  |
| --- | --- |
| **Названия строк** | **Количество по полю Parent Revenue** |
| 0 | 11.82% |
| 1 | 24.03% |
| 2 | 39.22% |
| 3 | 18.44% |
| 4 | 6.49% |
| **Общий итог** | **100.00%** |

|  |  |
| --- | --- |
| **Названия строк** | **Количество по полю School Code** |
| 0 | 29.48% |
| 1 | 14.94% |
| 2 | 9.35% |
| 3 | 14.55% |
| 4 | 11.43% |
| 5 | 20.26% |
| (пусто) | 0.00% |
| **Общий итог** | **100.00%** |

|  |  |
| --- | --- |
| **Названия строк** | **Количество по полю Playing Years** |
| 0 | 29.48% |
| 1 | 25.06% |
| 2 | 17.79% |
| 3 | 12.86% |
| 4 | 14.81% |
| (пусто) | 0.00% |
| **Общий итог** | **100.00%** |

|  |  |
| --- | --- |
| **Названия строк** | **Количество по полю Grade** |
| <6 или (пусто) | 0.00% |
| 6-16 | 0.13% |
| 26-36 | 0.13% |
| 36-46 | 3.64% |
| 46-56 | 9.74% |
| 56-66 | 13.12% |
| 66-76 | 19.74% |
| 76-86 | 17.53% |
| 86-96 | 19.09% |
| 96-106 | 16.88% |
| **Общий итог** | **100.00%** |

1. Для полученных сводных таблиц построены следующие гистограммы и диаграммы:

Рис. . Диаграмма «Распределение пола»

Диаграмма (рис. 1) показывает, что в выборке в равной степени присутствуют ученики обоих полов.

Рис. . Гистограмма «Образование матери»

Рис. . Гистограмма «Образование отца»

Обе гистограммы (рис. 2-3) показывают, что больше всего встречается уровень образования родителей по шкале 4, а затем 3 и 5. Реже всего встречаются родители с самым высшим уровнем образования и без образования совсем.

Рис. . Гистограмма «Игровое время»

Из этой гистограммы (рис. 4) видно, что большая часть учеников играет в игры не слишком долго – 1-2 часа или не играет вовсе.

Рис. . Гистограмма «Достаток родителей»

Если интерпретировать шкалу 0-4, как привычную 1-5, то по данным из гистограммы (рис. 5) можно сказать, что в выборке преобладают семьи со средним уровнем достатка. Также чуть чаще встречаются семьи с достатком ниже среднего или низким, чем выше среднего или высоким.

Рис. . Гистограмма «Код школы»

Из гистограммы (рис. 6) видно, что больше всего данных было собрано по школам 0 и 5, а меньше всего в школах 2 и 4.

Рис. . Гистограмма «Игровой стаж»

Из гистограммы (рис. 7) видно, что большая часть учеников достаточно недавно начали играть в компьютерные игры или не играла в них совсем.

Рис. . Гистограмма «Оценка»

Из гистограммы (рис. 8) видно, что большая часть учеников успешно справляется со школьной программой, когда как не справляются всего в сумме 13,51% учеников.

1. С применением полученных данных, могут решаться следующие задачи:

* **Анализ влияния развлечений на успеваемость** – влияет ли количество проводимого за развлечениями времени на оценки?
* **Анализ доступности развлечений** – поиск взаимосвязи между доходом семьи и количеством проводимого времени за играми учеником.
* **Анализ влияния ситуации в семье на успеваемость** – влияет ли образование родителя на то, как проводит свободное время ученик или влияет ли образование родителя на оценки ученика?
* **Анализ предпочтений учеников** – какой досуг больше предпочитают ученики? Выборку также можно разделить на группы: по школе или по полу.
* **Разработка уникального подхода к образованию** – вытекает из исследования выше: разработка подхода к образованию для каждой из школ на основе данных о предпочтениях учеников.