**Лабораторная работа 3 «Предобработка данных»**

**Загрузка данных**

1. Загрузите данные файла ***planets.csv*** о планетах в разных планетных системах галактики (планеты одной системы идут в списке рядом). Поля данных: способ обнаружения, число планет в системе своего «солнца», период обращения вокруг своего «солнца», масса планеты, расстояние от центра планетной системы до нашего Солнца, год обнаружения.

**Описательная статистика**

1. Получите описание данных: число записей, число полей.
2. По каждому полю определите:

тип поля: категориальный, порядковый, числовой, текстовый)

число имеющихся значений,

число уникальных значений,

число отсутствующих значений.

1. По каждому числовому полю определите: мин, макс, среднее, медиану, среднеквадратическое отклонение.
2. По каждой категории: число планет и средние значения остальных признаков.
3. Определите число записей, имеющих пропуски данных и общее число пропусков.
4. Для каждого числа возможных пропусков в отдельной записи (от 0 до укажите, в скольких записях столько пропусков.
5. Для каждого года обнаружения планеты укажите среднее число пропусков данных для одной планеты.
6. Для каждой категории укажите среднее число пропусков данных для одной планеты.

**Очистка датасета**

Сделайте новый набор данных, где:

1. Удалите записи с числом пропусков более 2-х.
2. Замените пропуски на (по вариантам: 0 – на среднее значение признака по всему набору, 1 – на среднее значение признака по категории, 2 – на значение случайного объекта этой же категории, 3 – на среднее значение трех случайных объектов этой же категории, 4 – на значение ближайшего объекта этой же категории, 5 – на среднее значение трех ближайших объектов этой же категории, номер варианта равен N mod 6, где N – ваш номер по списку преподавателя).
3. Добавьте поле «код планеты», куда внесите составной код, типа В74, построенный по правилу: в качестве числа уникальный номер планетной системы, где обнаружена планета (признак принадлежности нескольких планет к одной системе: расположение в списке рядом, одинаковое число планет в системе, одинаковое расстояние до Солнца), а в качестве буквы – номер планеты от центра системы (1-А, 2-В, 3-С, 4-D, …), который определяется по возрастанию периода обращения.

**Визуализация данных.**

1. Для каждого нового набора постройте диаграммы четырех типов:

* - гистограмма каждого числового признака по всему датасету и по отдельным категориям,
* - для каждого числового признака набор ящиков с усами по всем категориям,
* - тепловую карту корреляций для числовых признаков (номер планеты в системе, число планет в системе, период обращения, масса, расстояние, год обнаружения) для всего датасета, для самой массовой и для самой редкой категории.
* - парные диаграммы рассеяния для всех пар числовых признаков, имеющих более 50 уникальных значений, по всему датасету.

**Анализ данных.**

1. Укажите, у скольких звезд обнаружены планетные системы.
2. Исследуя полученные результаты (числовые и графические) опишите все обнаруженные вами закономерности и особенности.

В качестве решения сдать отчёт (Блокнот + PDF), где по каждому пункту из всех 15-ти указать:

- текст задания (можно вырезать из скриншота задания)

- код реализации

- результат вывода (если задание – просто действие, например, «удалить …», то показать, что оно выполнено и что получилось)

- комментарий по реализации и/или результатах .

Вначале привести общие команды (в т.ч. подключение необходимых библиотек).

Каждая диаграмма должна иметь заголовок, шкалы и подписи по осям.

**Вся работа – 8 баллов**.