ФГАОУ ВО «СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
Институт космических и информационных технологий
Кафедра «Информатика»

Методы анализа данных

Практическая работа № 2
Предварительная обработка данных

Красноярск, 2024

Цель: знакомство с основными задачами предварительной обработки исходных данных, изучение основных методов предварительной обработки данных, формирование навыков выполнения предварительной обработки исходных данных с помощью языка программирования Python.

Задачи:

Выполнение практической работы предполагает решение следующий задач:

1. Визуальный анализ исходных данных
2. Поиск аномальных значений
3. Поиск и восстановление отсутствующих значений
4. Преобразование данных

Исходные данные

Файл с исходными данными в .xlsx формате.

Общая последовательность действий

1. Визуальный анализ данных

Построить визуальное представление для каждого столбца (признака) в исходном наборе данных. Провести анализ полученных диаграмм. Примечание: для качественных переменных должны быть построены столбчатая и круговая диаграммы, для количественных переменных – гистограмма, оценка плотности распределения и диаграмма «ящик с усами».

2. Провести проверку правдоподобности исходных данных

Проверка правдоподобности исходных данных должна включать проверку типов исходных данных, лишних пропусков, невозможных значений и т.п. Привести найденные значения к нужному формату

3. Поиск аномальных значений

Провести поиск значений в исходном наборе данных, резко отличающихся от других значений (выбросов). Строки с найденными выбросами удалить из исходного набора данных.

Примечание: для поиска выбросов воспользоваться методом сигм (использовать готовую реализацию **scipy.stats.sigmaclip**) и метод квартилей (реализовать самостоятельно). Провести анализ полученных результатов. Использовать результаты очистки данных, полученных с помощью метода сигм.

4. Поиск и восстановление пропущенных значений

Провести поиск пропущенных значений в исходных данных. Вывести статистику по пропускам для каждого признака. Восстановить пропущенные значения. Примечание: для первого признака для восстановления пропусков использовать метод k-ближайших соседей, для второго пропущенные значения заменить самым популярным значением, для третьего использовать среднее значение, для четвертого – медиану, для пятого – метод «k-ближайших соседей». Для восстановления пропусков методом «k-ближайших соседей» использовать одну из готовых реализаций **sklearn.impute.KNNImputer**, **impyute.imputation.cs.fast_knn**).

5. Преобразование данных

Привести числовые признаки к стандартному виду. Для категориальных признаков выполнить их кодировку. Примечание: для количественных переменных выполняем стандартизацию и нормализацию, для качественных переменных – one-hot encoding (для первого) и label encoding (для второго). Для преобразования использовать готовую реализацию **sklearn.preprocessing**.

Распределение вариантов

В таблице приведены номер варианта и список признаков для каждого варианта

№	first	second	third	fourth	fifth
0	cat16_city	cat12_blood type	num1_1	num2_1	num3_1
1	cat4_income	cat1_gender	num1_2	num2_2	num3_2
2	cat5_customerSatisfaction	cat8_motivesForTravelling	num1_3	num2_3	num3_3
3	cat11_nationality	cat5_customerSatisfaction	num1_4	num2_4	num3_4
4	cat16_city	cat11_nationality	num1_5	num2_5	num3_5
5	cat1_gender	cat7_motivesForEmployeesToWorkBetter	num1_6	num2_6	num3_6
6	cat13_productType	cat7_motivesForEmployeesToWorkBetter	num1_7	num2_7	num3_7

7	cat11_nationality	cat14_temprature	num1_8	num2_8	num3_8
8	cat15_programmingLanguage	cat4_income	num1_9	num2_9	num3_9
9	cat5_customerSatisfaction	cat1_gender	num1_10	num2_10	num3_10
10	cat16_city	cat15_programmingLanguage	num1_11	num2_11	num3_11
11	cat16_city	cat7_motivesForEmployeesToWorkBetter	num1_12	num2_12	num3_12
12	cat1_gender	cat12_blood type	num1_13	num2_13	num3_13
13	cat4_income	cat15_programmingLanguage	num1_14	num2_14	num3_14
14	cat6_brandOfSoaps	cat2_hairColor	num1_15	num2_15	num3_15
15	cat8_motivesForTravelling	cat6_brandOfSoaps	num1_16	num2_16	num3_16
16	cat7_motivesForEmployeesToWorkBetter	cat11_nationality	num1_17	num2_17	num3_17
17	cat16_city	cat4_income	num1_18	num2_18	num3_18
18	cat9_age	cat12_paymentMethod	num1_19	num2_19	num3_19
19	cat8_motivesForTravelling	cat13_productType	num1_20	num2_20	num3_20
20	cat3_education	cat16_city	num1_21	num2_21	num3_21
21	cat1_gender	cat3_education	num1_22	num2_22	num3_22
22	cat14_temprature	cat13_productType	num1_23	num2_23	num3_23
23	cat10_proficiencyLevel	cat9_age	num1_24	num2_24	num3_24
24	cat13_productType	cat14_temprature	num1_25	num2_25	num3_25

Требования к выполнению практической работы:

1. Написание программного кода и формирование результатов согласно заданию и установленному варианту.

2. Составление отчета, содержащего описание решаемых задач методов решения и полученных результатов.

Программный код и отчет должны быть выполнены в среде Jupyter notebook.