Тема 8. Разведочный анализ данных

Практические задания для самостоятельного выполнения

Задания выполняются по вариантам. Формулировки заданий общие для всех вариантов; набор данных, подлежащих анализу, выбирается в соответствии с номером своего варианта.

Результаты выполнения заданий необходимо представить в виде двух файлов:

1) ноутбук в формате *ipynb*, содержащий программный код, результаты его выполнения, а также все необходимые пояснения, выводы и комментарии (в текстовых ячейках);
2) файл в формате *pdf* (или *html*), полученный путем экспорта (или вывода на печать) ноутбука из п. 1).

Внимание: в названии файлов должна обязательно присутствовать фамилия автора. Безымянные работы проверяться не будут.

Максимальная оценка за выполнение задания – **3 балла**.

Внимание. Учебной целью данного задания является не механическое выполнение набора действий, а приобретение навыков анализа конкретного набора данных в контексте конкретной задачи. Поэтому при выполнении задания требуется не только написание программного кода, но и интерпретация всех результатов, полученных с помощью этого кода, обоснование всех принимаемых решений по выбору того или иного метода исследования, а также выводов, полученных по результатам этих исследований.

Все рассуждения (пояснения, обоснования, выводы) должны быть записаны в текстовых ячейках (а не с помощью print). Эти рассуждения должны содержать интерпретацию конкретных числовых значений, полученных с помощью программного кода, и конкретные выводы, относящиеся к конкретному набору данных (с обоснованиями, почему выводы именно такие – как они получились из результатов работы программного кода). Общие фразы и лозунги, позаимствованные с различных ресурсов, писать не требуется (это оцениваться не будет).

Общее оформление работы: программный код для каждого пункта и подпункта задания должен быть написан в отдельной ячейке. Ниже в текстовой ячейке должны быть написаны пояснения, рассуждения, обоснования и выводы, полученные по результатам работы этого кода. Для каждой кодовой ячейки – своя текстовая ячейка с рассуждениями и комментариями.

При несоблюдении указанных требований представленное «решение» оцениваться не будет.

Задание.

Выполнить разведочный анализ данных о продажах домов в Нью-Йорке. Описание признаков исходного набора данных.
Наборы данных по вариантам представлены в csv-файлах; они имеют одну и ту же структуру, соответствующую общему описанию, но отличаются набором записей. Имена файлов: Вариант N, где N – номер варианта.

В ходе проведения анализа выполнить следующие действия.

1. Импортировать из файла данные о проданных домах.
2. Изучить описание всех признаков, характеризующих продажи (по ссылке выше).
3. В рамках первичного знакомства с данными:
   - вывести несколько записей (для проверки корректности импорта и получения первого представления о данных);
   - изучить признаки на наличие пропущенных значений, типы данных; сопоставить типы столбцов и значения в столбцах с описанием признаков, сделать выводы о корректности имеющихся значений.
4. Выполнить исследование одномерных распределений количественных входных признаков BEDS, BATH, PROPERTYSQFT, а также прогнозируемого признака PRICE:
   - для каждого признака найти описательные статистики, асимметрию и эксцесс;
   - визуализировать распределения;
   - проанализировать степень асимметричности, «хвосты», наличие в данных групп, аномальных значений – на основе полученных статистик и визуального оценивания;
   - сформулировать предположения о нормальности/отличии от нормального распределения каждого из рассмотренных признаков;
   - для признаков, распределение которых было оценено как близкое к нормальному, выполнить визуальную оценку соответствия гистограммы и предполагаемого распределения;
   - для непрерывных признаков, распределение которых заметно отличается от нормального вследствие явной асимметрии, проверить предположение о возможной принадлежности к логнормальному распределению (на основе визуальной оценки соответствия гистограммы и теоретического распределения).
5. Выполнить исследование на наличие связей между признаками BEDS, BATH и PROPERTYSQFT, а также зависимости между каждым из этих признаков и прогнозируемым признаком PRICE:
   - для каждой пары признаков обосновать применение корреляции Пирсона либо ранговой корреляции;

- применить (в учебных целях) все три метода корреляционного анализа, проанализировать полученные результаты; сделать выводы;
- построить парные графики рассеяния (можно использовать seaborn.pairplot(), либо pandas.plotting.scatter_matrix()), соотнести результаты визуализации с результатами корреляционного анализа;
- сформулировать выводы о возможном наличии/отсутствии зависимостей между изученными признаками.

6. Выполнить исследование одномерных распределений категориальных признаков BROKERTITLE, TYPE, LOCALITY и SUBLOCALITY: построить столбцовые диаграммы и изучить распределение категорий, обращая внимание на возможное присутствие малочисленных категорий.
   Пояснение. Малочисленные категории в перспективе обучения предсказательной модели могут рассматриваться как аналоги «тяжелых хвостов»/выбросов для количественных переменных. Наличие таких особенностей может затруднять обучение модели регрессии, приводить к нестабильной работе алгоритма обучения.
7. Выполнить исследование на наличие связи между признаками BROKERTITLE и LOCALITY (из-за небольшого объема набора данных – только по наиболее крупным агентствам):
   - отобрать записи, относящиеся к агентствам недвижимости, входящим в топ-5 по общему количеству продаж (в пределах рассматриваемого набора);
   - построить таблицу сопряженности рассматриваемых признаков;
   - оценить правомерность применения критерия «хи-квадрат»; в случае значительного числа клеток с малыми частотами – выполнить объединение малочисленных категорий признака LOCALITY (например, в одну категорию «Others»);
   - применить критерий «хи-квадрат», сделать выводы о наличии связи;
   - вычислить коэффициент Крамера и оценить силу взаимосвязи между признаками (при наличии).
8. Выполнить исследование на наличие зависимости между каждым из рассмотренных качественных признаков и прогнозируемым признаком PRICE (в связи с большим разнообразием значений признаков ограничить анализ 10 наиболее крупными категориями, объединив все остальные значения в категорию «Others»):
   - выполнить визуализацию в виде диаграмм «ящик с усами»; проанализировать полученные диаграммы, сформулировать предположения о наличии/отсутствии зависимостей;
   - отдельно изучить «ящик», полученный для категории «Others» (разброс значений, наличие выбросов) и сделать вывод о правомерности выполненного объединения категорий;

- соотнести «ящики с усами» с построенными ранее столбцовыми диаграммами, проанализировать возможность снижения размерности данных путем дополнительного объединения категорий с близкими значениями прогнозируемого признака;
- в тех случаях, где это целесообразно, выполнить объединение значений признаков в более крупные категории; построить «ящики с усами» на объединенных категориях.

9. По каждому пункту исследования сделать выводы (записать в текстовых ячейках). Привести все необходимые пояснения и комментарии.
10. Рассмотреть вопрос о целесообразности использования в предсказательной модели признаков из исходного набора данных, не рассмотренных в данном исследовании. Привести аргументы «за» или «против» необходимости дополнительного изучения этих признаков.
11. Дополнительное задание (+0,5 балла – только при условии качественного выполнения всех основных заданий, включая п. 9):
    - разбить весь имеющийся набор записей на три ценовые группы (низкие, средние и высокие цены);
    - повторить проведенное ранее исследование отдельно для каждой группы; сопоставить описательные статистики, корреляции, диаграммы, полученные для каждой группы; сравнить их с показателями полного набора записей;
    - анализ результатов и полученные выводы описать в текстовых ячейках.

Замечание.
Для более тщательного изучения вопроса о наличии зависимости количественного признака от категорий можно применять дисперсионный анализ – ANOVA (за пределами данного курса). Для желающих познакомиться с методом:

Простое изложение.

Презентация с подробным разбором.

ANOVA в Python.