Лабораторная работа № 1. Предварительная обработка данных

Цель лабораторной работы – получение навыков работы с библиотеками анализа данных, предварительной обработки данных, средствами визуализации.

Замечания:

1. Если разместить файл с данными в папке «:\WinPython-64bit-3.5.4.0Qt5\notebooks\» при его загрузке с помощью функции `pandas.read_csv` не нужно указывать путь к данному файлу.

2. Обработку больших csv файлов можно выполнять по частям:

```
chunksize = 10 ** 6
for chunk in pd.read_csv(filename, chunksize=chunksize):
    process(chunk)
```

3. При выполнении лабораторной работы для обработки данных не нужно использовать циклы. Используйте функции библиотеки Pandas.

4. Команда для отдельной установки библиотек:

```
pip install numpy scipy matplotlib ipython jupyter pandas sympy nose spyder seaborn
```

5. Чтение первых нескольких строк из файла:
read_csv(..., nrows=999999)

6. Задание «Вторичный рынок машин» - открывать набор данных нужно открывать набор с параметром `encoding='iso-8859-1'`

Отчёт должен включать:

1. ФИО студента, номер варианта, текст задания, а также результат выполнения задания.

2. Описание значений используемых полей (признаков) в исследуемом наборе данных.

3. Имя файла с результатом выполнения лабораторной работы должен включать ФИО студента, номер лабораторной работы и номер варианта.

4. Результат выполнения следующих функций библиотеки Pandas: head, tail, info, describe, dropna, drop_duplicates, shape. Показать результат их применения к данным.

1. Titanic

1. Определите количество мужчин и женщин, которые ехали на корабле.

2. Определите какой части пассажиров удалось выжить. Посчитайте долю выживших пассажиров.

3. Какую долю пассажиры первого класса составляли среди всех пассажиров?

4. Какого возраста были пассажиры? Посчитайте среднее и медиану возраста пассажиров.

5. Коррелируют ли число братьев/сестер с числом родителей/детей? Посчитайте корреляцию Пирсона между признаками SibSp и Parch. Оцените значение p-value. Постройте плотность распределения признаков SibSp и Parch.

6. Какое самое популярное женское имя на корабле? Извлеките из полного имени пассажира (колонка Name) его личное имя (First Name). Попробуйте вручную разобрать несколько значений столбца Name и выработать правило для извлечения имен, а также разделения их на женские и мужские.

7. Коррелирует ли класс, которым ехал пассажир, с выживаемостью?

8. Визуализируйте гистограммы возраста для выживших и не выживших пассажиров. Сделайте выводы. Отобразите данные на одном и нескольких графиках

9. Визуализируйте гистограммы возраста для выживших и не выживших пассажиров по классам. Сделайте выводы.

10. Постройте столбчатую диаграмму количества людей: мужчины, женщины, дети.

2. Отмена рейсов

1. Подсчитайте количество отменённых рейсов.

2. Определите аэропорт, рейсы для которого отменяются наиболее часто.

3. Определите коэффициент корреляции Пирсона и Спирмена между отменой рейса и днём недели, месяцем, авиакомпанией, аэропортом. Оцените значение p-value. Постройте плотность распределения признаков.

4. Подсчитайте для трёх выбранных авиакомпаний: количество рейсов, количество отменённых рейсов, количество перенаправленных рейсов.

5. Определите скорость полёта для каждого рейса, скорость полёта среднюю для трёх выбранных авиакомпаний.

6. Визуализируйте тепловую карту частоты отмены рейсов. По одной оси – дни, по другой оси – рейс (для двух аэропортов).

7. Посчитайте и визуализируйте время задержки отправки и прибытия по трём аэропортам.

8. Определите для трёх выбранных аэропортов и визуализируйте задержки по каждой причине.

9. Определите авиакомпанию с максимальными задержками рейсов по отправке и прибытию.

3. Вторичный рынок машин

1. Удалите столбцы, ценность которых для оценки стоимости машины низка.

2. Удалите повторяющиеся строки, строки содержащие пропуски в данных. Выведите размер набора данных до и после удаления.

3. Удалите данные в строках, выходящие за некоторые пределы (год регистрации, цена, мощность двигателя). Выведите размер набора данных до и после удаления. Для оценки диапазонов значений признаков используйте BoxPlot.

4. Заполнить пропущенные данные в строковых полях.

5. Постройте гистограмму по маркам автомобилей, типам кузова и используемому топливу.

6. Добавьте в данные новый признак, который представляет собой длину названия автомобиля.

7. Постройте карту корреляций между признаками. Выведите также числовые значения признаков.

8. Определите коэффициент корреляции Пирсона и Спирмена между стоимостью автомобиля и типом кузова. Оцените значение p-value. Постройте плотность распределения признаков.

9. Добавьте в набор данных признак, являющийся суммой двух других признаков.

4. Дожди в Индии

1. Постройте графики количества осадков по годам в разных штатах. Сделайте выводы по построенным графикам.

2. Постройте графики количества осадков по месяцам по всем годам кумулятивно. Сделайте выводы по построенным графикам.

3. Постройте графики количества осадков по штатам, используйте boxplot. Сделайте выводы по построенным графикам.

4. Определите штаты, для которых количество осадков минимально и максимально.

5. Выведите уникальные имена штатов и территорий Индии.

6. Определите количество лет наблюдений для каждого штата.

7. Определите штат с наименьшим количеством наблюдений

8. Добавьте в набор данных признак, являющийся суммой двух других признаков.

9. Вычислите корреляцию Пирсона и Спирмена между признаками NOV и MAR. Сравните полученные величины корреляции. Постройте гистограммы для указанных признаков. Оцените значение p-value.

5. Метеоритная защита земли

1. Определите местоположение появления метеорита с наибольшим количеством высвободившейся энергии.

2. Определите метеорит с максимальной и минимальной скоростью.

3. Определите место максимальной концентрации метеоритов.

4. Определите время года, в которое вероятность появления метеоритов максимальна. Постройте график.

5. Удалите строки данных, в которых для метеоритов не указана скорость. Выведите размер набора данных до и после удаления.

6. Построить график, на котором по оси OX отложено время суток, по OY – частота появления метеоритов.

7. Заполните строки в которых для метеоритов не указана скорость.

8. Определите месяц года, для которого появление метеоритов наиболее вероятно (постройте необходимые графики).

9. Вычислите корреляцию Пирсона и Спирмена между признаками Altitude и Total Radiated Energy. Сравните полученные величины корреляции. Постройте гистограммы для указанных признаков. Оцените значение p-value.

6. Индекс счастья

1. Визуализируйте корреляции между признаками, находящимися в наборе данных (heatmap). Сделайте выводы.

2. Постройте график счастья по регионам, страны на графике должны представляться отдельными точками.

3. Оцените количество счастья по годам по регионам.

4. Определите наиболее сильно изменяющиеся параметры по разным странам в разные годы.

5. Определите страны, появляющиеся и исчезающие в рейтинге стран.

6. Удалите строки с пропущенными значениями. Выведите размер набора данных до и после удаления.

7. Вычислите корреляцию Пирсона и Спирмена между признаками Trust.Government.Corruption и Family. Сравните полученные величины корреляции. Постройте гистограммы для указанных признаков. Оцените значение p-value.

8. Добавьте в набор данных признак, являющийся суммой двух других признаков.

7. Камеры

1. Визуализируйте корреляции между признаками, находящимися в наборе данных (heatmap). Сделайте выводы.

2. Вычислите корреляцию Пирсона и Спирмена между Price и Max (Min) Resolution. Сравните полученные величины корреляции. Постройте гистограммы для указанных признаков. Оцените значение p-value.

3. Постройте график изменения средней цены на камеры по годам.

4. Определите компанию, камеры которой наиболее часто встречаются в наборе данных.

5. Добавьте в набор данных признак, являющийся произведением двух других признаков.

6. Удалите строки с нулевыми значениями в данных. Выведите размер набора данных до и после удаления.

7. Определите камеру с максимальной стоимостью, определите камеру с минимальной стоимостью.

8. Определите год, в который было выпущено максимально количество новых камер.

8. Астронавты (космонавты)

1. Определите астронавта из США, который провёл наибольшее количество времени в открытом космосе, постройте график, на котором по оси OX отображены астронавты, а по оси OY – время, которое они провели в открытом космосе.

2. Укажите университет, выпустивший наибольшее количество астронавтов, постройте график, на котором по оси OX отображены университеты, а по оси OY – количество астронавтов, которое учились в данном университете.

3. Определите количество военных и гражданских астронавтов.

4. Определите наиболее часто встречающееся среди астронавтов военное звание.

5. Определите количество женщин среди астронавтов

6. Удалите из набора данных астронавтов, для которых не указан бакалавриат. Выведите размер набора данных до и после удаления.

7. Определите количество астронавтов, родом из Техаса.

8. Определите количество миссий, в ходе выполнений которых погибли астронавты.

9. Вычислите корреляцию Пирсона и Спирмена между признаками Gender и Missions. Сравните полученные величины корреляции. Постройте гистограммы для указанных признаков. Оцените значение p-value.

9. NBA

1. Определите игрока, принявшего участие в наибольшем числе игр.

2. Вычислите корреляцию Пирсона и Спирмена между признаками Age и G. Сравните полученные величины корреляции. Постройте гистограммы для указанных признаков. Оцените значение p-value.

3. Определите год, в котором наиболее интенсивно играли в баскетбол.

4. Определите позицию (Pos), находящийся на которой игрок наиболее и наименее результативен (PTS).

5. Постройте гистограмму по признаку Year.

6. Удалите строки с нулевыми значениями в данных. Выведите размер набора данных до и после удаления.

7. Добавьте в набор данных признак, являющийся произведением двух других признаков.

8. Определите самое популярное имя среди игроков.