Машинное обучение

НИЯУ МИФИ, Кафедра финансового мониторинга

Лабораторный практикум

В.Ю. Радыгин

Семестр 1. Лабораторная работа 1

Лабораторная работа 1 рассчитана на два занятия и работу дома. Её целью является изучение основ формирования дата-фреймов (DataFrame) в библиотеке Pandas.

При выполнении данной работы студентам будет необходимо самостоятельно познакомиться с рядом методов и функций.

В качестве задания лабораторной работы студентам предлагается найти в интернете набор реальных статистических данных. Затем сформировать из него Pandas DataFrame и выполнить с ним некоторые манипуляции, такие, как сортировка, отсеивание лишнего и т.д. После чего провести простейшее статистическое исследование.

Лабораторная работа оценивается в 15 баллов. Результатом работы является файл-программы на языке Python.

Вариант 1

Зарегистрируйтесь на сайте https://www.kaggle.com/datasets и загрузите с него набор статистических данных, посвящённый обращениям в службу спасения 911 https://www.kaggle.com/mchirico/montcoalert.

Блок 1

1. На основе загруженного CSV-файла создайте Pandas DataFrame, гарантировав правильные типы данных, переменных (например, правильную загрузку дат). Назовите переменные lat, ing, desc, zip, title, accident_time, town, address, e соответственно.
2. Измените полученный в задании 1 Pandas DataFrame, исключив из исходного набора переменные desc, zip, address, e.
3. Измените полученный в задании 1 Pandas DataFrame, отсортировав набор, полученный в задании 2 в следующем порядке: town lat ing accident_time title.
4. Выполните простейший количественный анализ по переменной town Pandas DataFrame, полученного в задании 3, отсортировав при этом результаты в порядке возрастания количества появлений городов среди наблюдений. Сохраните результаты в новый Pandas DataFrame.
5. Сформируйте новый Pandas DataFrame, в котором останутся только 4 названия городов: два наиболее часто встречающихся и два наименее часто встречающихся среди наблюдений исходного массива.
6. Сформируйте новый Pandas DataFrame, исключив из Pandas DataFrame, полученного в задании 3 все наблюдения, относящиеся к 4-м городам, полученным в задании 5, а также все наблюдения, где город не указан. Кроме того, добавьте в полученный набор новую переменную hour, содержащую время суток в часах, в которое произошёл инцидент.
7. Выполните простейший количественный анализ по переменной hour Pandas DataFrame, полученного в задании 6, отсортировав при этом результаты в порядке убывания количества появлений данного часа среди наблюдений. Сохраните результаты в новый Pandas DataFrame.
8. Выполните нормализацию набора данных, полученного в пункте 7, по переменной count.
9. Постройте гистограмму и кривую распределения для переменной count. Сравните кривую распределения графически с кривой нормального распределения. Сделайте выводы.
10. Постройте график линейной регрессии для зависимости переменной count от переменной hour. Сделайте выводы об их взаимосвязи.

Вариант 2

Зарегистрируйтесь на сайте https://www.kaggle.com/datasets и загрузите с него набор статистических данных, посвящённый глобальным террористическим актам https://www.kaggle.com/START-UMD/gtd.

Блок 1

1. На основе загруженного CSV-файла создайте Pandas DataFrame, гарантировав правильные типы данных, переменных (например, правильную загрузку дат). Назовите переменные так же, как названы колонки в исходном наборе данных.
2. Измените полученный в задании 1 Pandas DataFrame, оставив из исходного Pandas DataFrame переменные iyear, imonth, iday, country_txt, region_txt, latitude, longitude. Кроме того, добавьте в Pandas DataFrame новую переменную accident_date, собрав её из значений переменных iyear, imonth, iday.
3. Измените полученный в задании 2 Pandas DataFrame, отсортировав его в следующем порядке: country_txt, region_txt, iyear, imonth, iday, latitude, longitude.
4. Выполните простейший количественный анализ по переменной country_txt Pandas DataFrame, полученного в задании 3, отсортировав при этом результаты в порядке убывания количества появлений стран среди наблюдений. Сохраните результаты в новый Pandas DataFrame.
5. Выполните простейший статистический анализ Pandas DataFrame, полученного в задании 4, рассчитав квартили. Результаты работы процедуры сохраните в новый Pandas DataFrame.
6. На основе Pandas DataFrame x, полученного в задании 3, сформируйте новый Pandas DataFrame, в котором останутся только инциденты, произошедшие в странах, попавших во второй и третий квартили по количеству инцидентов, произошедших в них.
7. Выполните простейший количественный анализ по переменной imonth Pandas DataFrame, полученного в задании 6, отсортировав при этом результаты в порядке убывания количества появлений данного месяца среди наблюдений. Сохраните результаты в новый Pandas DataFrame.
8. Выполните нормализацию Pandas DataFrame, полученного в пункте 7, по переменным imonth и count.
9. Постройте гистограммы и кривые распределения для переменных imonth и count. Сравните их кривые распределения графически с кривыми нормального распределения. Сделайте выводы.
10. Постройте график линейной регрессии для переменных imonth и count. Сделайте выводы о их взаимосвязи.

Вариант 3

Зарегистрируйтесь на сайте https://www.kaggle.com/datasets и загрузите с него набор статистических данных, посвящённый каталогу фильмов IMDB https://www.kaggle.com/oxanozaep/imdb-eda/data (файл movie_metadata.csv).

Блок 1

1. На основе загруженного CSV-файла создайте Pandas DataFrame, гарантировав правильные типы данных, переменных (например, правильную загрузку дат). Назовите переменные так же, как названы колонки в исходном наборе данных.
2. Измените полученный в задании 1 Pandas DataFrame, исключив из него все переменные, кроме director_name, budget, imdb_score, title_year.
3. Измените полученный в задании 2 Pandas DataFrame, отсортировав его в следующем порядке: director_name title_year budget.
4. Выполните простейший количественный анализ по переменной director_name набора Pandas DataFrame, полученного в задании 3, отсортировав при этом результаты в порядке возрастания количества появлений режиссёров среди наблюдений. Сохраните результаты в новый Pandas DataFrame.
5. Выполните простейший статистический анализ Pandas DataFrame, полученного в задании 4, в частности, выполните расчёт квартилей.
6. На основе данных, полученных в задании 3, сформируйте новый Pandas DataFrame, в котором останутся только фильмы, снятые режиссёрами, попавшими во второй и третий квартили по количеству снятых ими фильмов.
7. Выполните простейший количественный анализ по переменной year Pandas DataFrame, полученного в задании 6, отсортировав при этом результаты в порядке убывания количества появлений данного года среди наблюдений. Сохраните результаты в новый Pandas DataFrame.
8. Выполните нормализацию набора данных, полученного в пункте 3, по переменным budget и imdb_score.
9. Постройте гистограммы и кривые распределения для переменных budget и imdb_score. Сравните их кривые распределения графически с кривыми нормального распределения. Сделайте выводы.
10. Постройте график линейной регрессии для переменных budget и imdb_score. Сделайте выводы о их взаимосвязи.

Вариант 4

Зарегистрируйтесь на сайте https://www.kaggle.com/datasets и загрузите с него набор статистических данных, посвящённый продажам домов https://www.kaggle.com/harlfoxem/housesalesprediction

Блок 1

1. На основе загруженного CSV-файла создайте Pandas DataFrame, гарантировав правильные типы данных, переменных (например, правильную загрузку дат). Назовите переменные также, как названы колонки в исходном наборе данных.
2. Измените полученный в задании 1 Pandas DataFrame, оставив из исходного Pandas DataFrame переменные date price yr_built yr_renovated sqft_living condition. Кроме того, добавьте в Pandas DataFrame новую переменную real_year, собрав её как максимальное значение из значений переменных yr_built yr_renovated.
3. Измените полученный в задании 2 Pandas DataFrame, отсортировав его в следующем порядке real_year, sqft_living condition.
4. Выполните простейший количественный анализ по переменной real_year Pandas DataFrame, полученного в задании 3, отсортировав при этом результаты в порядке возрастания количества продаж среди наблюдений. Сохраните результаты в новый Pandas DataFrame.
5. Сформируйте новый Pandas DataFrame, в котором останутся только 4 значения real_year: два наиболее часто встречающихся и два наименее часто встречающихся среди наблюдений исходного массива.
6. Сформируйте новый Pandas DataFrame, исключив из Pandas DataFrame, полученного в задании 3 все наблюдения, относящиеся к 4-м годам, полученным в задании 5.
7. Выполните простейший количественный анализ по переменной condition Pandas DataFrame, полученного в задании 6, отсортировав при этом результаты в порядке убывания количества появлений данного состояния среди наблюдений. Сохраните результаты в новый Pandas DataFrame.
8. Выполните нормализацию Pandas DataFrame, полученного в пункте 3, по переменным real_year и condition.
9. Постройте гистограммы и кривые распределения для переменных real_year и condition. Сравните их кривые распределения графически с кривыми нормального распределения. Сделайте выводы.
10. Постройте график линейной регрессии для переменных real_year и condition. Сделайте выводы о их взаимосвязи.