Лабораторная работа №4
Использование библиотеки pandas

При решении задач данной лабораторной работы необходимо использовать функции библиотеки pandas, если иное не указано в условии. Также можно использовать библиотеки numpy, scipy, matplotlib (либо любую другие для построения графиков).

По данной работе необходимо подготовить отчет в формате блокнота Jupyter Notebook (.ipynb) либо в текстовом виде (.pdf). В отчете должны быть:

1) исходные коды
2) результаты выполнения
3) ответы на текстовые вопросы (если ответ не содержится в явном виде в результатах выполнения).

Задание 1 (7 баллов)

1. Загрузите данные из файла «athlete_events.csv» о спортсменах – участниках олимпийских игр (ОИ).
   Данные содержат следующие признаки:
   ID – уникальный идентификатор спортсмена
   Name – имя спортсмена
   Sex – пол (M – мужской, F – женский)
   Age – возраст (полных лет, целое число)
   Height – рост в сантиметрах
   Weight – вес в килограммах
   Team – название команды (страны)
   NOC – трехбуквенное обозначение страны (по стандарту МОК)
   Games – год и вид ОИ (летние или зимние)
   Year – год проведения ОИ
   Season – вид ОИ (летние или зимние)
   City – город проведения ОИ
   Sport – вид спорта
   Event – соревнование (дисциплина)
   Medal – завоеванная медаль (Gold, Silver, Bronze или NA)
2. Определите количество значений каждого из признаков в загруженных данных. По каким значениям имеются не все данные? По какому значению отсутствующих данных больше всего? Подсказка: воспользуйтесь функцией count или info.
3. Выведите статистическую информацию (среднее значение, стандартное отклонение, минимальное и максимальное значение, значение квартилей) по полям: возраст, рост, вес. Подсказка: воспользуйтесь функцией describe.
4. Ответьте на вопросы, написав соответствующий код
   1) Сколько лет было самому молодому участнику олимпийских игр в 1992 году? Как звали этого участника и в какой дисциплине он(а) участвовал(а)?
   2) Выведите список всех видов спорта, которые когда-либо входили в программу олимпийских игр. (Каждый вид спорта должен присутствовать в списке один раз.)
   3) Каков средний рост теннисисток (пол – женский, вид спорта – большой теннис), участвовавших в играх 2000 года?
   4) Сколько золотых медалей в настольном теннисе выиграл Китай на ОИ в 2008 году?
   5) Как изменилось количество видов спорта на летних ОИ в 2004 году по сравнению с летними ОИ в 1988 году?

6) Постройте гистограмму распределения возраста мужчин-керлингистов (Sport == 'Curling'), участвовавших в олимпиаде 2014 года. Подсказка: для построения гистограммы можно использовать функцию hist() из библиотеки matplotlib с параметрами по умолчанию (либо можете использовать любую другую функцию на свое усмотрение).
7) Рассмотрим зимнюю олимпиаду 2006 года. Сгруппируйте данные по стране (используйте признак «NOC») и посчитайте для каждой страны количество завоеванных медалей и средний возраст спортсменов. Выведите только те страны, которые завоевали хотя бы одну медаль.
8) Продолжим рассматривать зимнюю олимпиаду 2006 года. Посчитайте, сколько медалей каждого достоинства завоевала каждая из стран-участниц (страны, не завоевавшие ни одной медали, можно не выводить). Для этого сгруппируйте данные по стране и по виду медали. Представьте данные в виде сводной таблицы (pivot_table). В сводной таблице не должно быть отсутствующих значений (NaN), замените их на 0.

Задание 2 (6 баллов)

Загрузите данные из файла «telecom_churn.csv» о клиентах оператора сотовой связи. Данные содержат, в числе прочих, следующие признаки:
State – обозначение территории (штата)
Area code – код города
International plan – подключена ли услуга международного роуминга
Number vmail messages – количество голосовых сообщений
Total day minutes – суммарная продолжительность дневных звонков (в минутах)
Total day calls – количество дневных звонков
Total eve minutes, Total eve calls – аналогичные показатели по вечерним звонкам
Total night minutes, Total night calls – аналогично по ночным звонкам
Customer service calls – количество звонков в службу поддержки
Churn – отток клиентов (False – клиент активен, True – клиент потерян, то есть расторг договор)
Остальные столбцы можно сразу проигнорировать при загрузке данных.

1. Выведите общую информацию о датафрейме с помощью методов info или describe. Есть ли отсутствующие данные?
2. С помощью метода value_counts определите, сколько клиентов активны, а сколько потеряно. Сколько процентов клиентов в имеющихся данных активны, а сколько потеряны?
3. Добавьте дополнительный столбец в датафрейм – средняя продолжительность одного звонка (вычислить как суммарная продолжительность всех звонков, деленная на суммарное количество всех звонков). Отсортируйте данные по этому значению по убыванию и выведите 10 первых записей.
4. Сгруппируйте данные по значению поля «Churn» и вычислите среднюю продолжительность одного звонка в каждой категории. Есть ли существенная разница в средней продолжительности одного звонка между активными и потерянными клиентами?
5. Сгруппируйте данные по значению поля «Churn» и вычислите среднее количество звонков в службу поддержки в каждой категории. Есть ли существенная разница между активными и потерянными клиентами?
6. Исследуйте подробнее связь между параметрами «Churn» и «Customer service calls», построив таблицу сопряженности (факторную таблицу) по этим признакам. Подсказка: используйте функцию crosstab. При каком количестве звонков в службу поддержки процент оттока становится существенно выше, чем в целом по датафрейму? (В качестве уточнения фразы «существенно выше» можете использовать «более 40%».)

7. Аналогично предыдущему пункту исследуйте связь между параметрами «Churn» и «International plan». Можно ли утверждать, что процент оттока среди клиентов, использующих международный роуминг, существенно выше или ниже, чем среди клиентов, не использующих его?
8. Добавьте в датафрейм столбец «Прогнозируемый отток», заполнив его на основе значений столбцов «Customer service calls» и «International plan». Сравните значение в этом столбце со значением столбца «Churn». Если мы будем пользоваться построенным прогнозом, то какой процент ошибок первого и второго рода (ложноположительных и ложноотрицательных) мы получим?

Задание 3 (5 баллов)

1. Загрузите данные о чемпионах автогонок «Формула 1» с сайта https://en.wikipedia.org/wiki/List_of_Formula_One_World_Drivers%27_Champions (используйте таблицу с указанием всех чемпионов по годам)
2. Посмотрите результат загрузки, при необходимости удалите лишние строки (подсказка: «подвал» таблицы, то есть последние несколько строк, не нужны).
3. Проверьте тип данных. Преобразуйте числовые данные (возраст, число поулов, побед, подиумов, быстрейших кругов, очки и т.д.) к числовому типу.
4. Ответьте на вопросы, написав соответствующий код.
    1) Средний, минимальный и максимальный возраст гонщиков, становившихся чемпионами.
    2) Гонщики какой команды выиграли больше всего личных титулов? (Под командой в этом вопросе подразумевается конструктор шасси.)
    3) Имена гонщиков, наибольшее число раз становившихся чемпионами.
    4) Кто из гонщиков смог стать чемпионом, выиграв менее 30% гонок в сезоне? (Подсказка: число гонок в сезоне можно определить по столбцу, где указано в какой гонке из скольки был завоеван титул.)
    5) Рассмотрим гонщиков, которые становились чемпионами хотя бы дважды. Рассмотрим промежутки времени между их «соседними» чемпионствами. Каков максимальный перерыв между двумя последовательными чемпионствами одного и того же гонщика? Например, Нельсон Пике становился чемпионом в 1981, 1983 и 1987 годах. Для него такие перерывы составляют 2 года (с 1981 до 1983) и 4 года (с 1983 до 1987), максимальный из них – 4 года. Вам нужно найти максимум среди всех гонщиков. (В этом задании можно не ограничиваться только функциями библиотек, но и написать свой код на Python.)