**Лабораторная работа № 1. Предварительная обработка данных**

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

**Замечания:**

1. Обработку больших csv файлов можно выполнять по частям:

chunksize = 10 \*\* 6

for chunk in pd.read\_csv(filename, chunksize=chunksize):

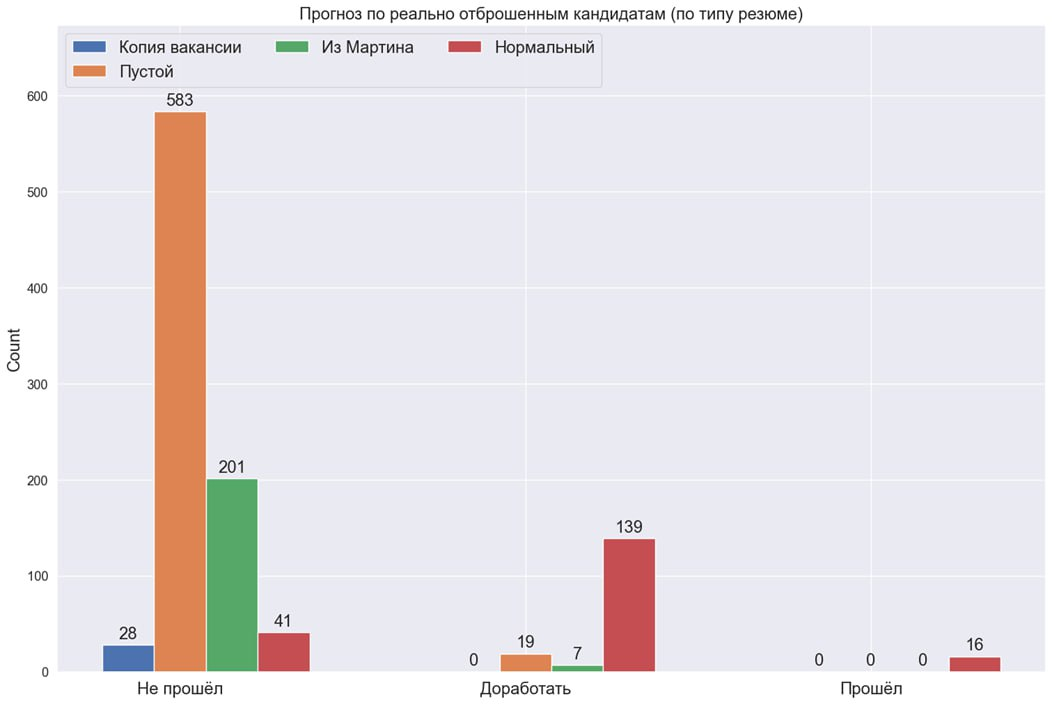
process(chunk)

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

3. Чтение первых нескольких строк из файла:

read\_csv(..., nrows=999999)

4. Если выполняется построение гистограммы, то должно быть подписано какое значение в каждом столбике (см пример ниже). Если колонок будет очень много, например по оси x возраст, то можно не указывать.



**Работа с данными**

Есть два варианта:

* Google colab - <https://colab.research.google.com/?hl=ru> Создаёте ноутбук там и работаете с ним, потом скачиваете и загружаете его на github
* Jupyter локально. Форкаете репозиторий и работаете с ноутбуками локально и потом помещаете их на github.

**Нюансы при локальной работе**.

Необходимые библиотеки:

numpy scipy matplotlib ipython jupyter pandas sympy nose spyder seaborn

**Отчёт должен включать:**

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

2. Краткое описание значений используемых полей (признаков) в исследуемом наборе данных (есть в описания файлике на английском)

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

Если данных после прочистки остаётся мало, то можно использовать [imputers](https://scikit-learn.org/stable/api/sklearn.impute.html). Они позволяют заполнить данные разными стратегиями.

**Общие задания:**

1. Результат выполнения следующих функций библиотеки Pandas: head, tail, info, describe, shape. Показать результат их применения к данным.
2. Пустые значения.
   1. Вывести количество пустых значений по колонкам. (отсортировать по убыванию)
   2. Вывести процент пустых значений по колонкам. (отсортировать по убыванию)
   3. Показать выводы функции info до и после очистки.
3. Удаление дубликатов. (drop\_duplicates) Вывести количество строк до удаления и после. Так же вывести количество найденных дубликатов.

Полезные ресурсы

Plotly-express [link](https://plotly.com/python/plotly-express/) для красивых и интерактивных графиков. Но после завершения сессии в jupyter они не будет. Поэтому экспортируем в html.

Более сложные примеры работы с go.Figure можно найти [здесь](https://github.com/lenow55/mdata_flow/blob/master/src/mdata_flow/plotly_ext/func_library.py) и в документации.

Evidently - [link](https://github.com/evidentlyai/evidently) для построения быстрых и шаблонных отчётов по данным. Можно использовать в лабах. [QuickStartDocs](https://docs.evidentlyai.com/quickstart_ml)

Так же можно его использовать для написания своих отчётов. [Пример](https://github.com/lenow55/mdata_flow/blob/master/src/mdata_flow/evidently_ext/column_count_by_category.py) метрики для вывода всех категорий по колонам с точным указанием количества и процентных отношений ([место](https://github.com/lenow55/mdata_flow/blob/32a69546009380dbcca47d042ef71518b461a1c8/src/mdata_flow/evidently_ext/visitors/count_by_category_report.py#L30-L36) где вызывается эта метрика).

**Далее идут задания для конкретных вариантов.**

**1. Titanic**

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

2. Определите какой части пассажиров удалось выжить. Посчитайте процент выживших пассажиров. (Две гистограммы, одна с количествами, другая с процентами. Или на одной, если получится отразить оба значения)

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

4. Посчитайте среднее и медиану возраста пассажиров. Отразите на диаграмме violitplot или на гистограмме распределение возрастов (см. пример [violinplot](https://seaborn.pydata.org/generated/seaborn.violinplot.html) – вам понравится😊)

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. Для построения используйте [kdeplot](https://seaborn.pydata.org/generated/seaborn.kdeplot.html) или [distplot](https://seaborn.pydata.org/generated/seaborn.displot.html).

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? Постройте histplot

8. Визуализируйте violinplot возраста для выживших и не выживших пассажиров. Сделайте выводы.

9. Визуализируйте violinplot возраста для выживших и не выживших пассажиров по классам. Сделайте выводы. (Можно ещё histplot для каждого класса бахнуть по каждому году с группировкой по выжившим и не выжившим, но разного цвета)

10. Постройте histplot количества людей: мужчины, женщины, дети; по группам выживаемости. Группы выживших/не выживших должны быть разных цветов.

**2. Отмена рейсов**

1. Histplot – кол-во отменённых и не отменённых рейсов.

2. Постройте гистограмму для аропортов по количеству отменённых рейсов. Выведите топ 10 самых злостных отменщиков. Столбики отсортировать по убыванию слева на право. Определите аэропорт, рейсы для которого отменяются наиболее часто.

3. Определите коэффициент корреляции Пирсона и Спирмена между отменой рейса и днём недели, месяцем. Оцените значение p-value. Постройте плотность распределения количества отмен по месяцам и по дням недели (используйте kdeplot или displot).

4. Подсчитайте для трёх выбранных авиакомпаний: количество рейсов, количество отменённых рейсов, количество перенаправленных рейсов. По каждому типу постройте histplot.

5. Сформируйте новый признак скорости полёта для каждого рейса. Рассчитайте среднюю скорость полёта по всем рейсам для трёх выбранных авиакомпаний и выведите её на histplot.

6. Возьмите 10 дней из общей выборки. Визуализируйте тепловую карту (heatplot) частоты отмены рейсов. По одной оси – дни, по другой оси – рейс (для двух аэропортов).

7. Сформируйте новый датасет по авиакомпаниям с признаками: среднее время задержки отправки и среднее время прибытия.

8. Постройте два histplot по задержкам отправления и прибытия для топа 5 компаний.

9. Определите авиакомпанию с максимальными задержками рейсов по отправке и прибытию (самый ярый опаздун).

**3. Вторичный рынок машин**

1. Заполнить NaN значения в строковых полях (нейтральными значениями, например «неизвестный».
2. Оцените диапазоны значений для числовых колонок (например год регистрации, цена, мощность двигателя) используйте violinplot или boxplot.
3. Удалите выбросы. Выведите размер набора данных до и после удаления.
4. Для числовых признаков и цены машины постройте матрицу корреляций (heatplot)
5. Удалите столбцы, ценность которых для оценки стоимости машины низка исходя из данных корреляций и из логики (опишите почему удалили тот или иной столбец).
6. Постройте histplot-ы по маркам автомобилей (топ 15), типам кузова и используемому топливу.
7. Добавьте в данные новый признак, который представляет собой длину названия автомобиля.
8. Определите коэффициент корреляции Пирсона и Спирмена между стоимостью автомобиля и мощностью двигателя. Оцените значение p-value. Постройте графики плотностей распределения признаков (kdeplot или displot).

**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 ОХ 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. Постройте гистограмму, на которой по оси OX отображены астронавты, а по оси OY – время, которое они провели в открытом космосе. Астронавты должны быть отсортированы по продолжительности пребывания в космосе слева на право. Определите астронавта, который пробыл в космосе наибольшее количество времени.

2. Постройте Гистограмму, на которой по оси OX отображены университеты, а по оси OY – количество астронавтов, которое учились в данном университете. Отсортируйте слева на право в порядке убывания. Укажите университет, выпустивший наибольшее количество астронавтов.

3. Определите количество военных и гражданских астронавтов и постройте гистограмму.

4. Определите наиболее часто встречающееся среди астронавтов военное звание.

5. Постройте гистограмму: Мужчины – Женщины 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.