**Лабораторная работа № 1. Предварительная обработка данных (EDA)**

**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):

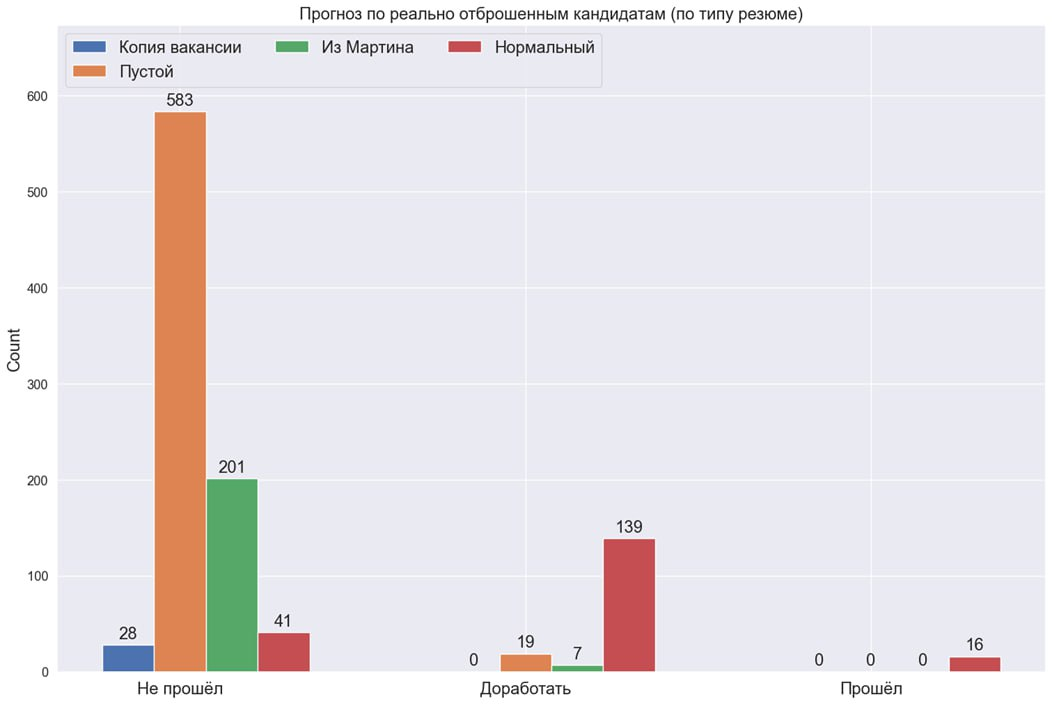
process(chunk)

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.



5. 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.

**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 описания в файлике на английском)

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

**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. 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).

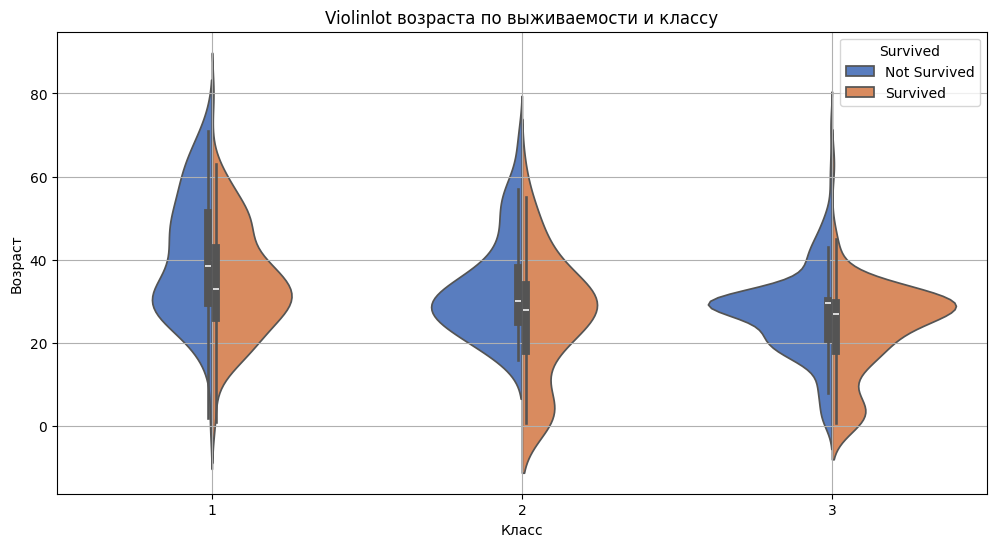
6. Какое самое популярное женское имя на корабле? Извлеките из полного имени пассажира (колонка Name) его личное имя (First Name). Попробуйте вручную разобрать несколько значений столбца Name и выработать правило для извлечения имен, а также разделения их на женские и мужские.

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.

Пример:



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

**2. Отмена рейсов**

Подготовка: Логически объясните по каким причинам в некоторых колонках очень много пустых значений (прям текстом в ноутбуке); Если можно, заполните пустые значения нейтральными, например время задержки по погоде можно выставить в 0 вместо nan; Дропните остальные пустые значения; Выведите количество записей до и после удаления пустых значений.

1. Вывести количество отменённых и не отменённых рейсов. Рассчитать процент отменённых и не отменённых рейсов ко всем рейсам. Отобразить гистограмму **для процентов** отменённых и не отменённых рейсов.

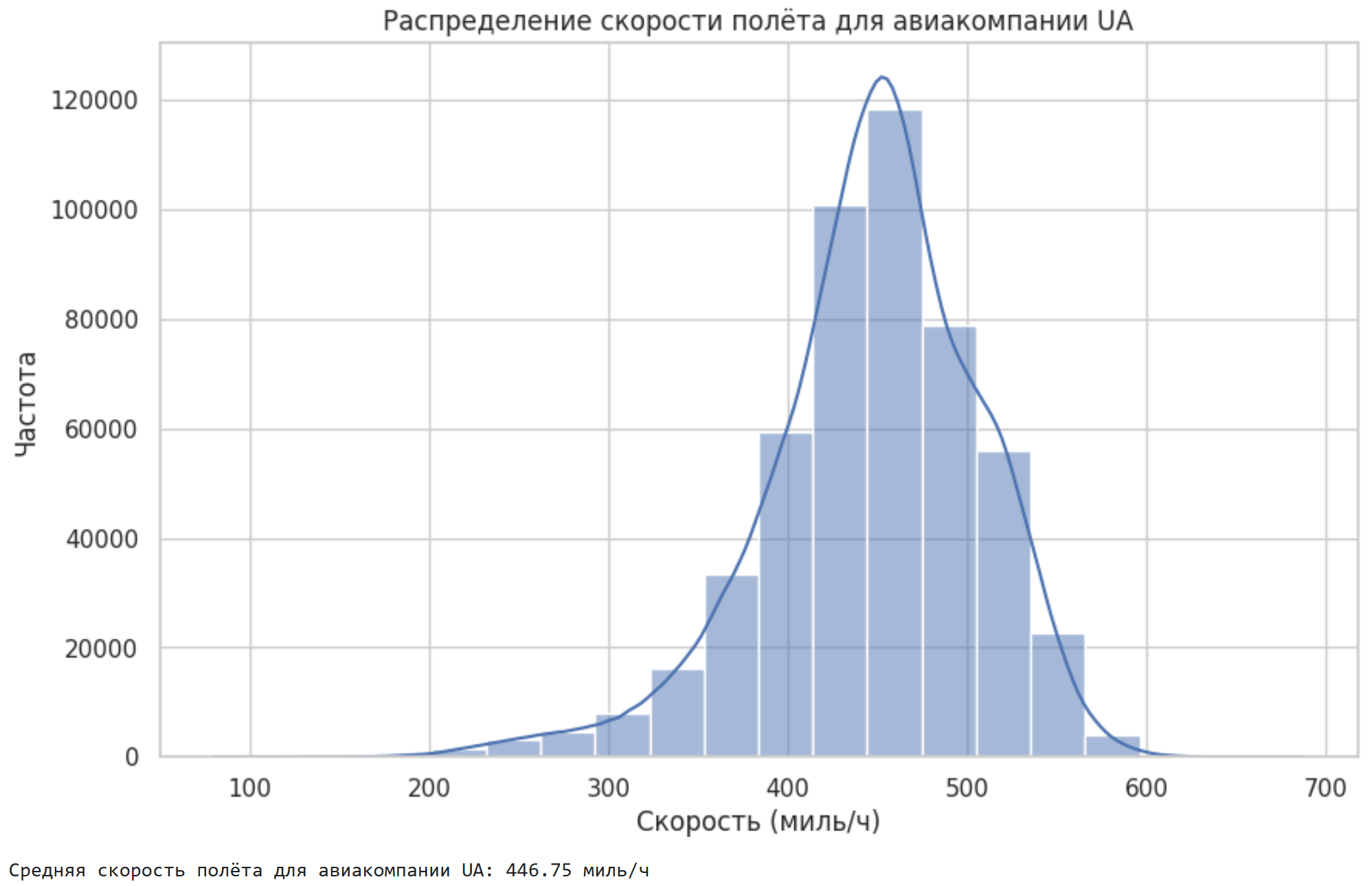
2. Рассчитать количество отменённых рейсов по каждому аэропорту. Отсортировать результаты по убыванию. Вывести топ 10 (значит 10 у кого больше всего отмен) самых злостных отменщиков. Определите аэропорт, рейсы для которого отменяются наиболее часто. Построить гистограмму: на ней те 10 аэропортов у кого больше всего отмен, отсортировать столбики слева на право по убыванию.

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-гистограммы).

4. Выберите три авиакомпании. Подсчитайте по каждой компании: количество рейсов, количество отменённых рейсов, количество перенаправленных рейсов. Отобразите гистограммы по каждому типу рейсов (отменённые, перенаправленные и нормальные) **столбики сортировать по убыванию слева на право.**

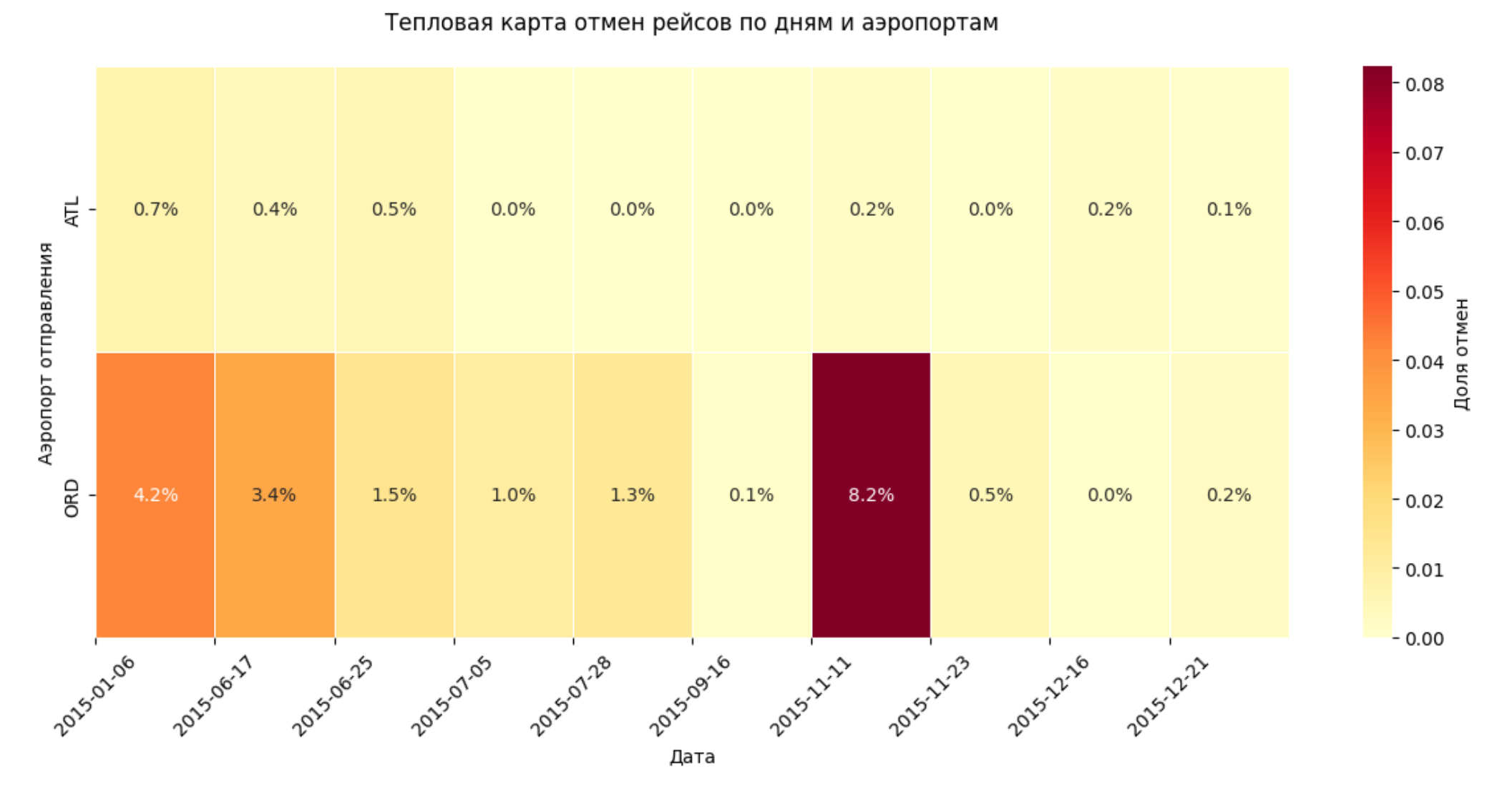
5. Сформируйте новый признак скорости полёта для каждого рейса. Рассчитайте среднюю скорость полёта по всем рейсам для трёх выбранных авиакомпаний и выведите её на histplot.

Пример:



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).

Пример:



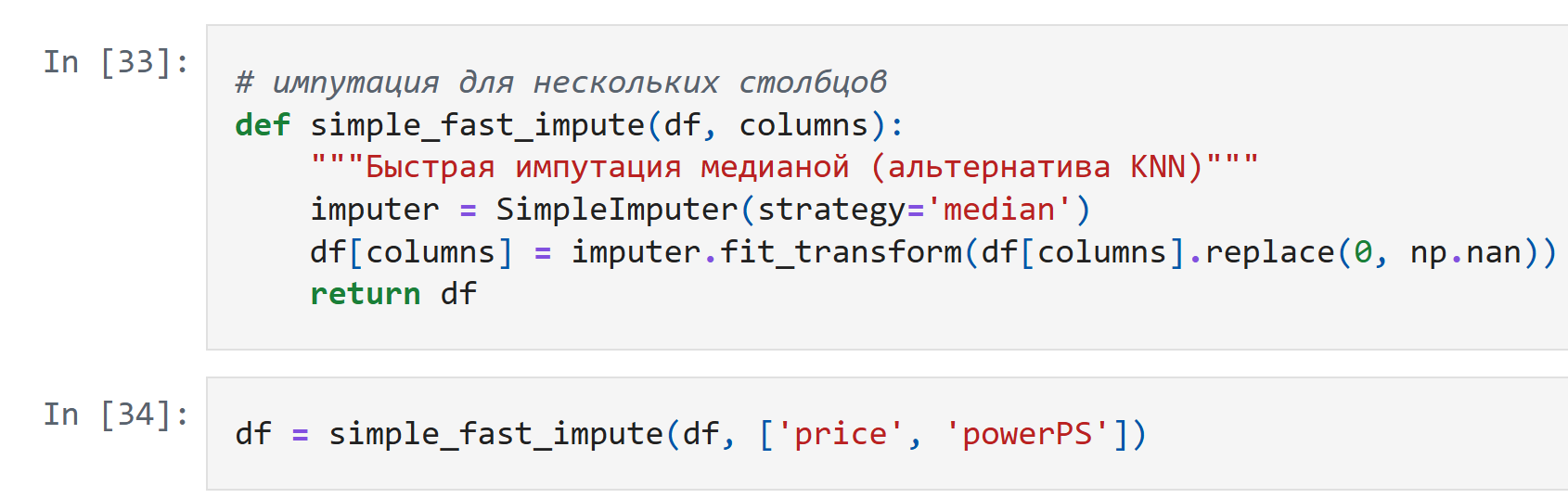
7. Сформируйте новый датасет по авиакомпаниям с признаками: среднее время задержки отправки и среднее время прибытия.

8. **Используйте новый датасет со средними временами задержки и прибытия**. Определите топ 5 компаний с наибольшими задержками прибытия, постройте гистограмму по задержкам прибытия для этих компаний. Определите топ 5 компаний с наибольшими задержками отправления, постройте гистограмму по задержкам отправления для этих компаний. **Сортировать по убыванию слева на право. На гистограмме отобразить среднюю задержку в цифрах над каждым столбиком.**

9. **Используйте новый датасет со средними временами задержки и прибытия**. Определите авиакомпанию с максимальными средними задержками рейсов по отправке и прибытию (самый ярый опаздун). Достаточно вывести название и средние задержки прибытия и отправки для этой компании.

**3. Вторичный рынок машин**

1. Заполнить NaN значения в **строковых полях** (нейтральными значениями, например «неизвестный».
2. Проведите импутацию данных в **числовых полях** с большим количеством пустых значений. Используйте knnimputer или simpleImputer.



пример заполнения. (пример копировать нельзя)

1. Оцените диапазоны значений **для числовых колонок** (например год регистрации, цена, мощность двигателя) используйте violinplot или boxplot.
2. Удалите выбросы. Выведите размер набора данных до и после удаления. В качестве алгоритма очистки можно использовать критерий 1.5\*интерквантильный размах.
3. Для **числовых признаков и цены машины** постройте матрицу корреляций (heatplot)
4. Удалите столбцы, ценность которых для оценки стоимости машины низка исходя из данных корреляций и из логики (опишите почему удалили тот или иной столбец).
5. Посчитайте количество автомобилей по их маркам. Отсортируйте по убыванию и выведите топ 15 самых распространённых марок. Постройте histplot по полученным 15 маркам.
6. Посчитайте количества автомобилей по типам кузова и используемому топливу. Отсортируйте по убыванию и выведите топ 15 самых распространённых кузовов и все типы топлива. Постройте histplot по полученным данным.
7. Добавьте в данные новый признак, который представляет собой длину названия автомобиля. Проанализируйте параметры данного признака: df.describe – медиана, стд-отклонение и тд.
8. Определите коэффициент корреляции Пирсона и Спирмена **между стоимостью автомобиля и мощностью двигателя**. Оцените значение p-value. Для двух этих признаков постройте графики плотностей распределения признаков (kdeplot или displot).

**4. Дожди в Индии**

1. Выведите уникальные имена штатов и территорий Индии.
2. Определите количество лет наблюдений **для каждого штата**.
3. Определите штат с наименьшим количеством наблюдений.
4. Определите штаты, для которых количество осадков минимально и максимально.
5. Найдите количество пропущенных значений по столбцам, выведите процентное отношение пропущенных значений. Выведите все строки со всеми пропущенными значениями и сделайте выводы.
6. Используйте KnnImputer для заполнения показаний по пропущенным осадкам. **Для каждого штата надо обучить свой Imputer и использовать его для заполнения.** (объясните почему для каждого штата свой imputer). Заполните пропущенные суммы значений осадков в строках с пропущенными данными.
7. Выведите процентное отношение пропущенных значений
8. По каждому штату постройте **график плотности осадков** (kdeplot, displot) – все графики должны быть отражены **на одном холсте** (строить нужно не по годовой оценке, а **по каждому месяцу**). Проанализируйте полученное семейство плотностей.
9. Используя предыдущий график, распределите штаты по группам осадков, например: Штат1, Штат2 – среднее значение осадков такое, Штат3,Штат4 – среднее значение другое. Каждую группу выведите отдельным списком. **Разделение по группам должно быть не ручным!**
10. С агрегируйте количество осадков за каждый месяц в каждом году по всей индии. Постройте графики количества осадков по месяцам по всем годам кумулятивно. Сделайте выводы по построенным графикам.
11. Рассчитайте **среднее значение количества осадков по годам** по каждому штату. Отсортируйте штаты по среднему количеству осадков по убыванию.
12. Выберите из предыдущего списка 10 штатов с наибольшим средним количеством осадков. Для выбранных штатов постройте графики boxplot, по каждому штату (строить не по годам, а по месяцам)
13. Выберите из предыдущего списка 10 штатов с наименьшим средним количеством осадков. Для выбранных штатов постройте графики boxplot, по каждому штату (строить не по годам, а по месяцам)
14. Добавьте в набор данных признак, являющийся суммой двух других признаков.
15. Вычислите корреляцию Пирсона и Спирмена между признаками NOV и MAR. Сравните полученные величины корреляции. Постройте гистограммы для указанных признаков. Оцените значение p-value.

(мало графиков в задании)

**5. Метеоритная защита земли**

1. Удалите строки данных, в которых для метеоритов не указана скорость. Выведите размер набора данных до и после удаления. **Дальше полученный датасет не использовать!**
2. Определите местоположение появления метеорита с наибольшим количеством высвободившейся энергии.
3. Определите место максимальной концентрации метеоритов. В качестве места используйте радиус в 10градусов по широте и долготе. Вставьте картинку с точкой на карте.
4. Заполните строки, в которых для метеоритов не указана скорость, используйте SimpleImputer или KnnImputer.
5. Определите метеориты с максимальной и минимальной скоростью.
6. Определите время года (зима, весна и т.д.), в которое вероятность появления метеоритов максимальна. Постройте гистограмму количества метеоритов по временам года.
7. Построить график (гистограмму), на котором по оси ОХ отложено время суток в часах, по OY – частота появления метеоритов.
8. Определите месяц года, для которого появление метеоритов наиболее вероятно (постройте гистограмму появления метеоритов по месяцам).
9. Вычислите корреляцию Пирсона и Спирмена между признаками Altitude и Total Radiated Energy. Сравните полученные величины корреляции. Оцените значение p-value.
10. Постройте гистограммы для признаков Altitude и Total Radiated Energy. Подберите количество бинов (bins) для нормального отображения. Если столбики слишком разные по размерам, используйте логарифмическую шкалу.