Введение в EDA

Варианты самостоятельных заданий v.4.0

Кафедра БИСУП – 2023

МИСиС
Национальный исследовательский технологический университет

Рекомендации по выполнению заданий.

Язык программирования: Python 3.9 - 3.1X: www.python.org

Используемые библиотеки Python

Pandas: pip install pandas - установка

import pandas as pd -использование

Matplotlib: pip install matplotlib - установка

import matplotlib.pyplot as plt

Seaborn: pip install seaborn

import seaborn as sns

Среда выполнения Jupyter Lab: www.jupyter.org

Установка:

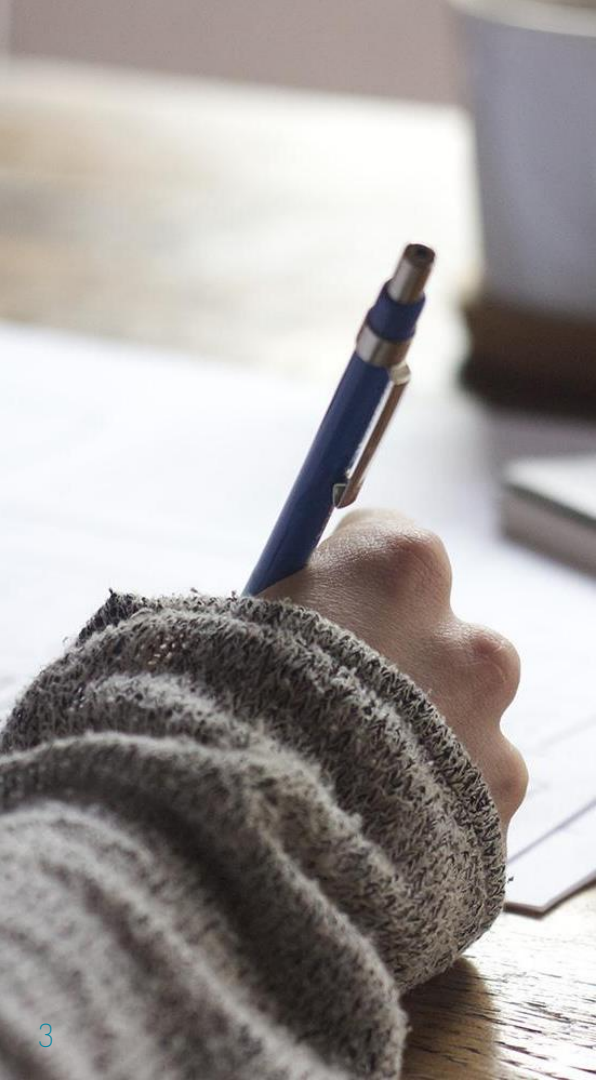
pip install jupyterlab

Запуск блокнота из командной строки:

Jupyter lab

Рекомендация: Сначала необходимо загрузить и установить последнюю версию Python с официального сайта, затем установить Jupyter Lab и далее необходимые пакеты

Примечание: При работе в ОС Linux, возможно потребуется отдельная установка pip: для дистрибутивов на основе Ubuntu командой : apt install python3-pip; Kaggle.com



Задание № 1:Наиболее важные факторы, связанные с увольнением и эффективностью сотрудников!

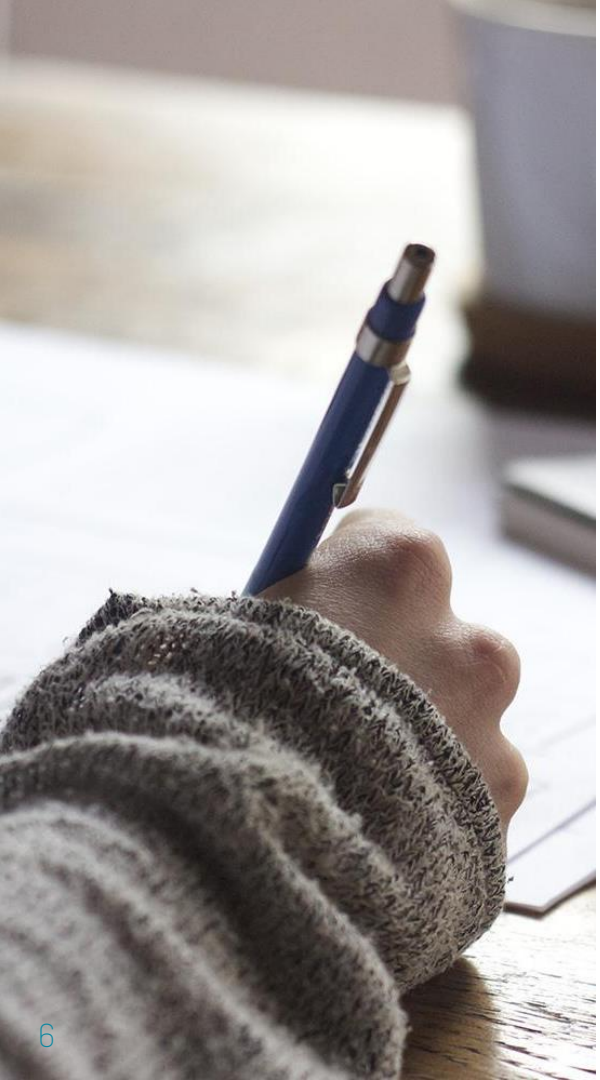
employe
e.zip

IBM создала синтетический набор данных, который можно использовать, чтобы понять, как различные факторы влияют на увольнение и удовлетворенность сотрудников. Некоторые из переменных включают образование, вовлеченность в работу, рейтинг производительности и баланс между работой и личной жизнью.

Изучите этот набор данных и посмотрите, есть ли какие-либо существенные переменные, которые действительно влияют на удовлетворенность сотрудников. Сделайте еще один шаг и посмотрите, сможете ли вы расположить переменные от наиболее важных до наименее важных.

Исходные данные

- Eduction: 1- среднее, 2-студент, 3- бакалавр, 4-магистр, 5-кандидат наук
- EnvironmentSatisfaction: 1 - низкое, 2-среднее, 3-высокое, 4- очень высокое
- JobInvolvyment: 1 - низкое, 2-среднее, 3-высокое, 4- очень высокое
- JobSatisfaction: 1 - низкое, 2-среднее, 3-высокое, 4- очень высокое
- PerformanceRating: 1 - низкий, 2-хороший, 3-отличный, 4- выдающийся
- WorkLifeBalance: 1- плохой, 2-хороший, 3-лучше, 4- самый лучший



Задание № 2: Международный рейтинг университетов

universit
y.zip

Как вы думаете, в нашей стране лучшие университет в мире? Что для начала значит быть «лучшим» университетом?

Этот набор данных содержит три глобальных рейтинга университетов. Используя эти данные, посмотрите, сможете ли вы ответить на следующие вопросы:

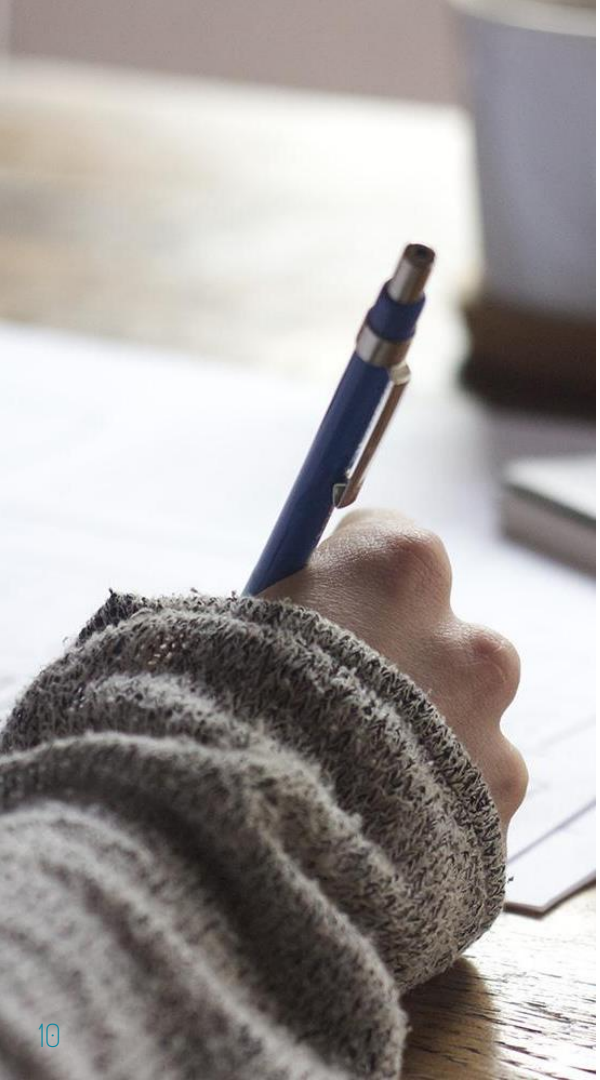
- В каких странах лучшие университеты?
- Каковы основные факторы, определяющие мировой рейтинг?
- Как различные рейтинги соотносятся друг к другу?
- Как наш университет соотносится с позициями в рейтингах?

Исходные данные:

- Международные рейтинги:
  - Times Higher Educational World University – считается одним из самых влиятельных и широко цитируемых рейтингов с 2010
  - Academic Ranking of World Universities - также известный как Шанхайский рейтинг, не менее влиятельный рейтинг. Он был основан в Китае в 2003 году
  - Center for World University Rankings - это менее известный рейтинг из Саудовской Аравии, он был основан в 2012 году.

ВОЗ создала набор данных о состоянии здоровья во всех странах с течением времени и включает статистические данные о продолжительности жизни, смертности взрослых и многом другом.

- Используя этот набор данных, исследуйте отношения между различными переменными. Что больше всего влияет на продолжительность жизни
- Данный набор данных создавался для ответа на следующие вопросы:
  - Действительно ли различные прогностические факторы, выбранные изначально, влияют на продолжительность жизни? Какие прогнозные переменные действительно влияют на ожидаемую продолжительность жизни?



Задание № 3: Исследование факторов, влияющих на продолжительность жизни

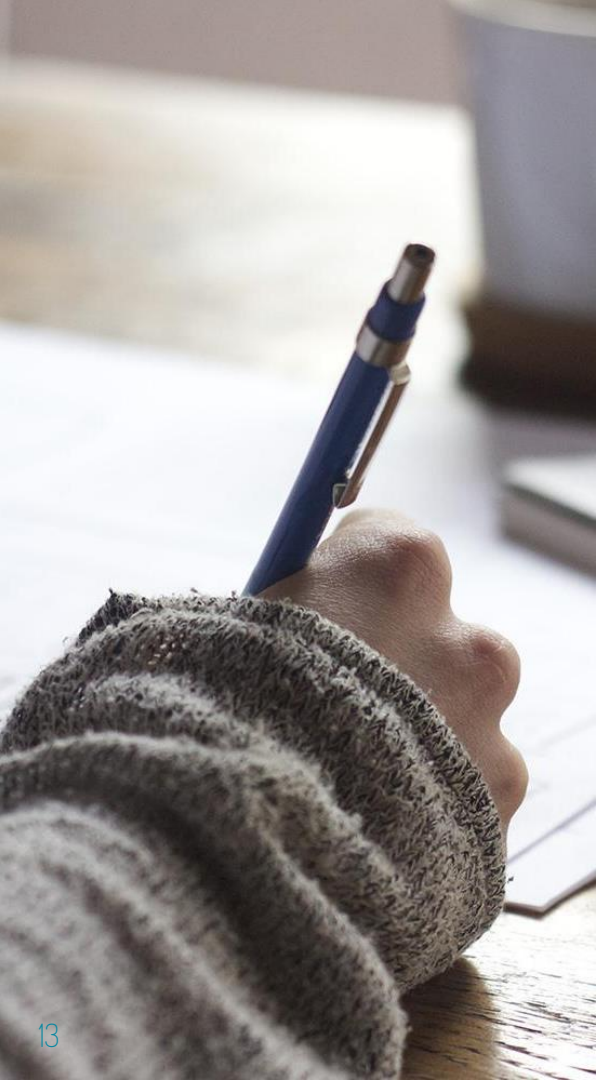
life_expe
ct.zip

Факторы продолжительности жизни

- Следует ли странам с низкой ожидаемой продолжительностью жизни (<65) увеличивать свои расходы на здравоохранение, чтобы улучшить среднюю продолжительность жизни
- Как показатели детской и взрослой смертности влияют на продолжительность жизни?
- Имеет ли ожидаемая продолжительность жизни положительную или отрицательную корреляцию с пищевыми привычками, образом жизни, физическими упражнениями, курением, употреблением алкоголя и т. д.

Факторы продолжительности жизни

- Как образование влияет на продолжительность жизни людей?
- Имеет ли продолжительность жизни положительное или отрицательное отношение к употреблению алкоголя?
- У густонаселенных стран обычно более низкая продолжительность жизни?



Задание № 4:
Анализ факторов, влияющих на прекращение абонентом контракта с оператором телефонной связи

WA_Fn-U
seC_-Telc

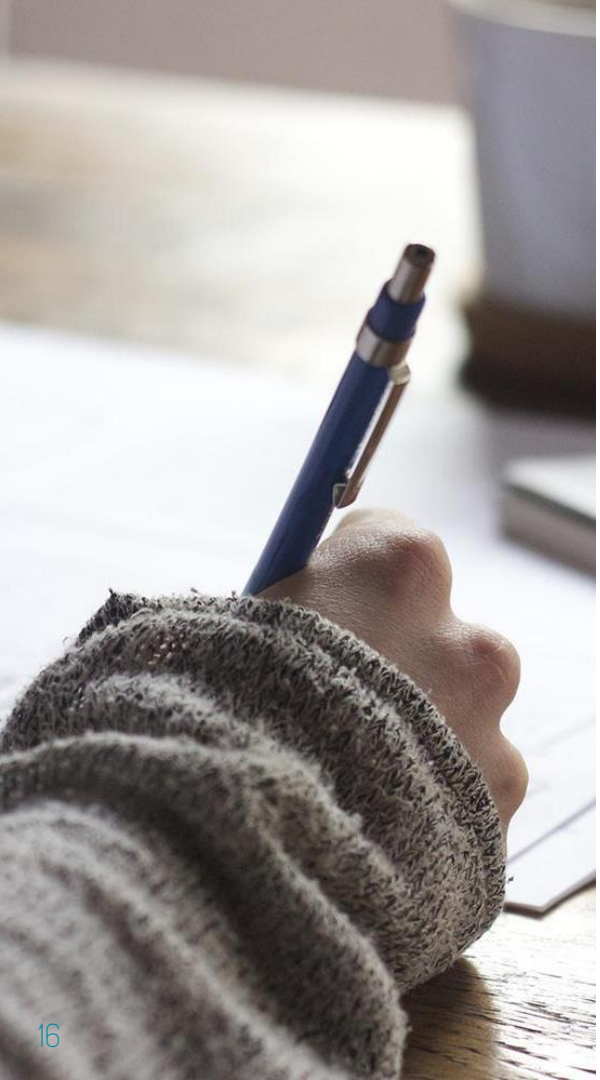
Исходные данные

- CustomerID: **идентификатор пользователя**
- Gender: **пол** (female, male)
- SeniorCitizen: **пенсионер** (1,0)
- Partner: **Наличие партнера** (yes,no)
- Dependants: **наличие иждевенцев** (yes,no)
- PhoneService: **подключена услуга тф связи** (yes,no)
- MultiLines: **подключена услуга нескольких линий** (yes,no)
- Internetservice: **тип услуг интернета** (DSL, Optic, no)
- OnlineSecurity: **подключена услуга** online **безопасности** (yes,no,no internet)
- OnlineBackup: **есть ли услуга создания резервной копии** (yes,no, no internet)
- DeviceProtection: **подключена ли услуга защиты устройства** (yes, no, no internet)
- TechSupport: **предоставлена услуга тех. поддержки** (yes,no, no internet)

Исходные данные

- StreamingTV: Подключена услуга ТВ (yes,no, no internet)
- StreamingMovies: Подключены потоковые кино сервисы (yes,no, no internet)
- Tenure: Кол-во месяцев пользования услугами
- Contract: условия контракта, по-месячно, ежегодный, двухлетний
- PaperlessBilling: подключена услуга получения элек. счетов (yes,no)
- Payment method: метод оплаты
- MonthlyCharges: сумма, снимаемая клиента ежемесячно
- TotalCharges: Общая сумма, перечисленная клиентом
- Churn - прекратил клиент действие контракта или нет

Постановка задачи: *Провести предварительный анализ данных (EDA). Получить представление о структуре данных, оценить качество данных. Провести предварительную обработку данных, очистку, удаление аномалий. Оценить какие параметры в наибольшей степени влияют на решение клиента о прекращении пользования услугами связи. На каждом этапе анализа делать промежуточные выводы*



Задание № 5:
Анализ доходов ведущих спортсменов мира

sport_for
bes.zip

Исходные данные

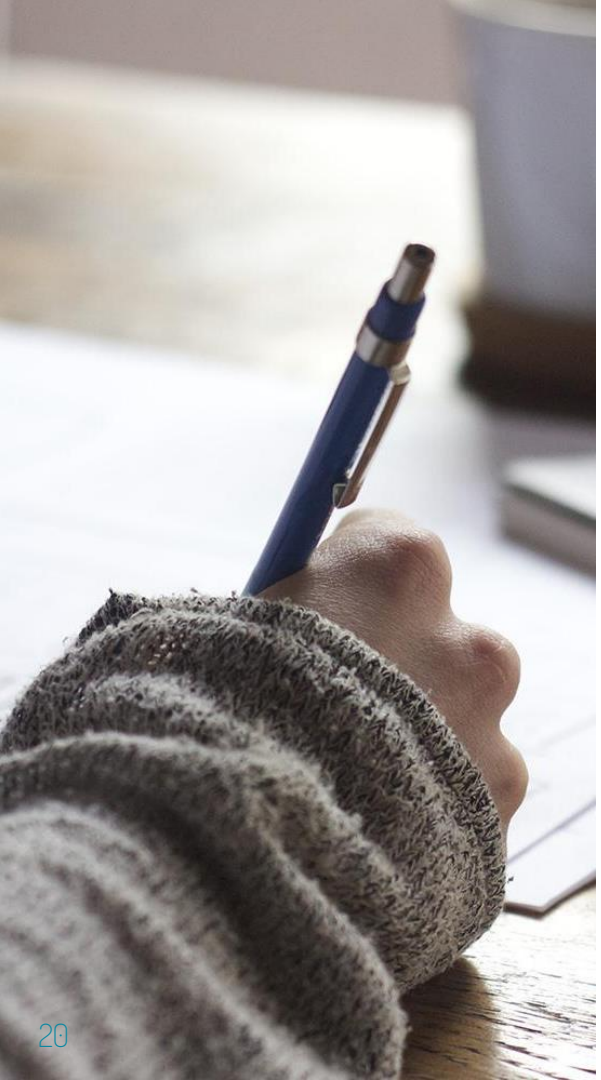
- S.NO: Идентификатор
- Name: Имя
- Nationality: Страна, за которую выступает спортсмен
- Current Rank: текущий рейтинг
- Previous Year Rank: рейтинг за прошлый год
- Sport: вид спорта
- Year: год рейтинга
- Earnings: доходы за год рейтинга

Задача и этапы выполнения

- Загрузить данные в dataframe, используя библиотеку pandas
- Преобразовать параметр Year в тип datatime
- Преобразовать текстовые значения Sport либо к нижнему, либо к верхнему регистру
- Получить представление о структуре данных и качестве данных
- Использовать графические библиотеки seaborn или matplotlib (допускается использование любых других библиотек, например plotly или bokeh) для визуализации данных и получения предварительных выводов о зависимости дохода спортсмена от прочих параметров
- Представить выводы

Задача и этапы выполнения

- Следует проанализировать следующее:
  - Спортсмены дольше всего находившиеся в рейтингах
  - Страны, в которых спортсмены заработали больше всего
  - Сколько зарабатывает самый высокооплачиваемый спортсмен за каждый год
  - Вид спорта, который доминирует по заработкам
  - Страна, которая превосходит всех по заработкам спортсменов
  - Максимальное время повления спортсменов в топ рейтинге
  - Показать соотношение спортсменов и спортсменок в данном рейтинге
  - Анализ доходов трех ведущих спортсменов



Задание № 6:
Популярность музыкальных композиций

Исходные данные

- song duration_ms: длительность трека
- acousticness: акустика
- danceability: танцевальность
- energy: текущий рейтинг
- instrumentalness: инструментальность
- key: ключ
- liveness: живость
- loudness: громкость
- Audio_mode: режим аудио
- Speechiness: речь
- Tempo: темп
- Time signature: число ритмических единиц в такте
- Audio valence: мера описывающая позитивное воздействие на слушателя
- Song popularity: популярность трека

Задача и этапы выполнения

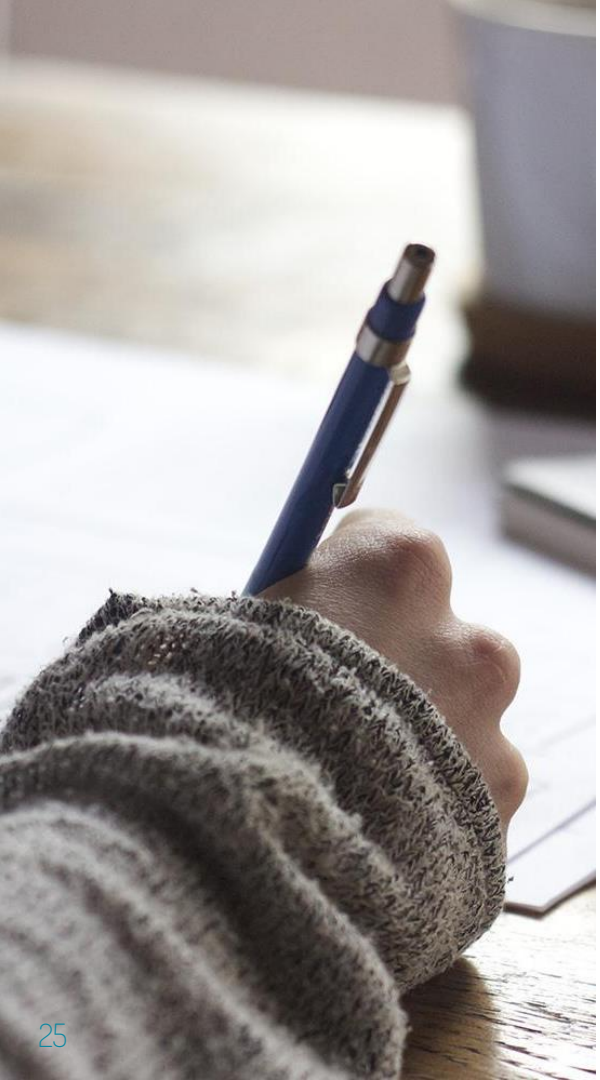
- Загрузить данные в dataframe, используя библиотеку pandas
- Получить представление о структуре данных и качестве данных
- Использовать графические библиотеки seaborn или matplotlib (допускается использование любых других библиотек, например plotly или bokeh) для визуализации данных и получения предварительных выводов о зависимости популярности трека от прочих параметров
- Представить выводы

Исходные данные

- school – имя школы (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
- sex: пол учащегося (двоичный код: «Ж» - женский или «М» - мужской)
- age: возраст студента (числовое: от 15 до 22)
- address: тип домашнего адреса учащегося (двоичный: «U» - городской или «R» - сельский)
- famsize: размер семьи (двоичный: «LE3» — меньше или равно 3 или «GT3» — больше 3)
- Pstatus: статус совместного проживания родителей (бинарное: «T» - совместное проживание или «A» - раздельно)
- Medu: образование матери (числовое: 0 - нет, 1 - начальное образование (4 класс), 2 - 5-9 класс, 3 - среднее или 4 - высшее)
- Fedu: образование отца (числовое: 0 - нет, 1 - начальное образование (4 класс), 2 - 5-9 классы, 3 - среднее или 4 - высшее)
- Mjob: работа матери (номинальное: «учитель», «здравоохранение», гос.служба (например, административная или полицейская), «домашний» или «другой»)

Исходные данные

- Fjob: работа отца (номинальное: «учитель», «медицина», связанная с уходом за здоровьем, гос. служба (например, административная или полицейская), «самозанятый» или «другое»)
- Reason: причина выбора этой школы (номинальное: близость к «дому», «репутация школы», предпочтительная «программа» или «другое»)
- guardian: опекун ученика (имя: «мать», «отец» или «другой»)
- traveltime: время в пути от дома до школы (числовое значение: 1 – <15 мин., 2 – от 15 до 30 мин., 3 – от 30 мин. до 1 часа или 4 – >1 часа)
- stdudytime: еженедельное время обучения (числовое значение: 1 – <2 часов, 2 – от 2 до 5 часов, 3 – от 5 до 10 часов или 4 – >10 часов)
- failures: количество прошлых неуспешных предметов (числовое: n, если 1<=n<3, иначе 4)



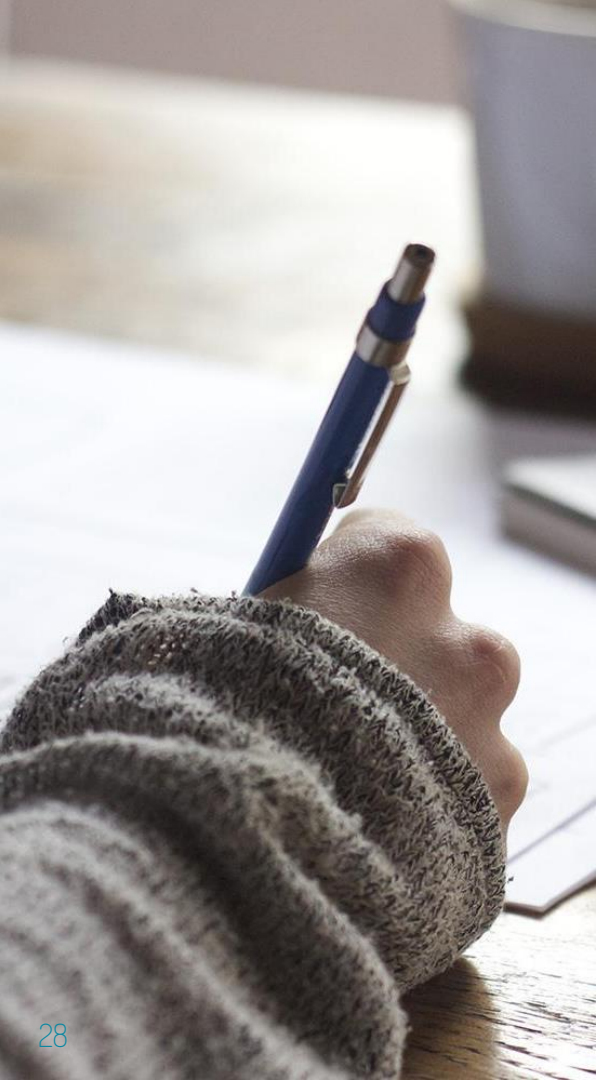
Задание № 7:
Успеваемость студентов

Исходные данные

- Schoolsup: образовательная поддержка (двоичный код: да или нет)
- famsup: семейная образовательная поддержка (двоичный код: да или нет)
- paid: дополнительные оплачиваемые занятия в рамках предмета курса (математика или португальский язык) (двоичный код: да или нет)activities: внеклассные мероприятия (бинарные: да или нет)
- nursery: посещал детский сад (двоичный код: да или нет)
- higher: хочет получить высшее образование (бинарное: да или нет)
- internet: Доступ в Интернет дома (бинарный: да или нет)
- romantic: с романтическими отношениями (бинарный: да или нет)
- famrel: качество семейных отношений (числовое: от 1 - очень плохо до 5 - отлично)
- freetime: свободное время после школы (числовое: от 1 - очень низко до 5 - очень высоко)
- goout: время с друзьями (числовое значение: от 1 — очень низкий до 5 — очень высокий)
- Dalc: потребление алкоголя в рабочие дни (числовое значение: от 1 — очень низкое до 5 — очень высокое)
- Walc: потребление алкоголя в выходные дни (числовое значение: от 1 — очень низкое до 5 — очень высокое)
- health: текущий статус здоровья (1 – очень плохо, 5 – отлично)
- Absences: количество пропусков занятий (числовое: от 0 до 93)

Задача и этапы выполнения

- Загрузить данные в dataframe, используя библиотеку pandas
- Получить представление о структуре данных и качестве данных
- Использовать графические библиотеки seaborn или matplotlib (допускается использование любых других библиотек, например plotly или bokeh) для визуализации данных и получения предварительных выводов о зависимости успеваемости студента от прочих параметров. Выделить параметры влияющие на успеваемость в наибольшей степени
- Представить выводы



Задание № 8: Изменение климата – температура поверхности Земли

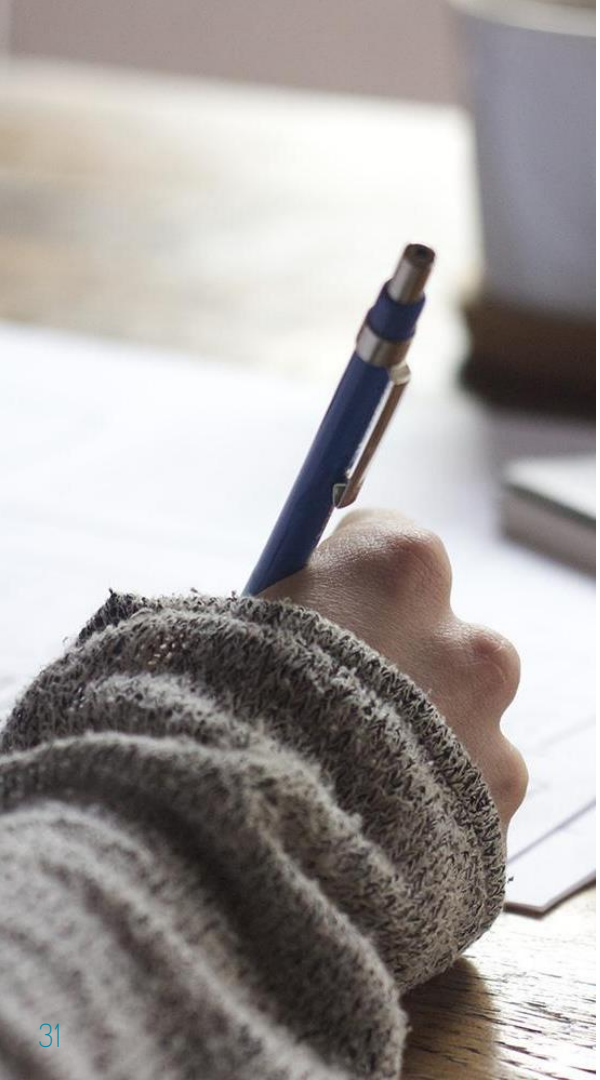
Исходные данные

- Dt – дата наблюдения
- LandAverageTemperature:
- LandAverageTemperatureUncertainity:
- LandMaxTemperature: текущий рейтинг
- LandMaxTemperatureUncertainity:
- LandMinTemperature :
- LandMinTemperatureUncertainity :
- LandAndOceanAverageTemparature
- LandAndOceanAverageTemparatureUncertainity

Задача и этапы выполнения

- Загрузить данные в dataframe, используя библиотеку pandas
- Получить представление о структуре данных и качестве данных
- Использовать графические библиотеки seaborn или matplotlib (допускается использование любых других библиотек, например plotly или bokeh) для визуализации данных и получения предварительных выводов о зависимости изменений температуры поверхности от прочих параметров, а также от температуры поверхности океана. Отдельно проанализировать изменения температуры по различным странам

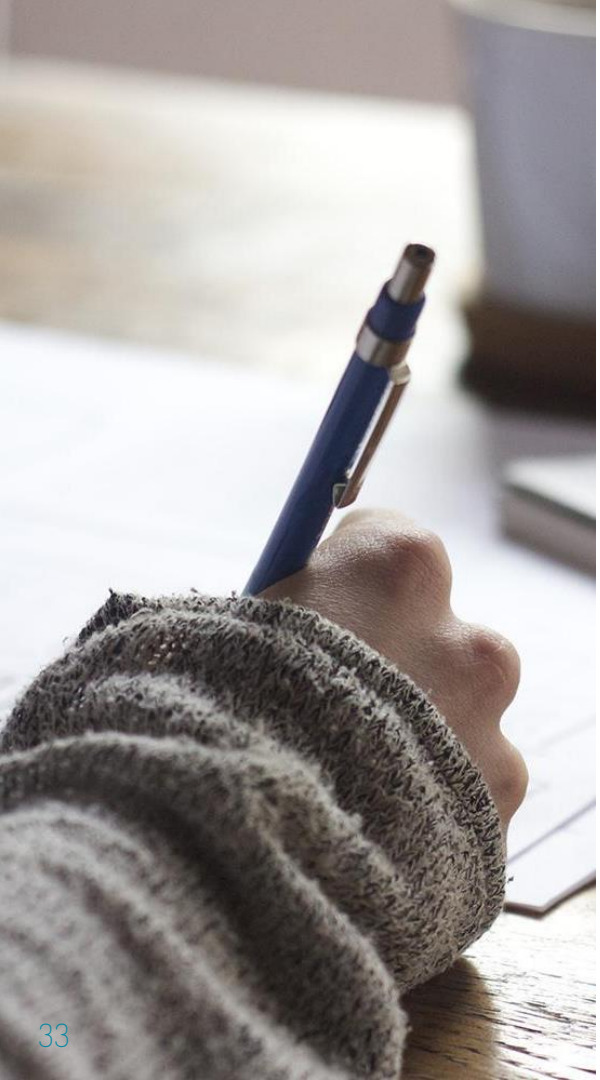
- Представить выводы



Задание № 9*:
Анализ факторов, влияющих на стоимость недвижимости

Исходные данные

- Исходные данные объяснены в файле data description.txt



Задание № 10:
Анализ и визуализация данных о возникающих цунами

Исходные данные

- База данных о цунами может также содержать ошибки, уникальные для этой базы данных. Одним из наиболее важных измерений, связанных с событием цунами, является максимальная высота заплеска или высота воды над уровнем моря в метрах. К сожалению, не всегда ясно, какой опорный уровень использовался. База данных о цунами также включает места, где наблюдалось цунами, называемые местами заплеска. Та же проблема, которая возникает при определении эпицентров землетрясений, может возникнуть при назначении мест заплеска, когда названия местностей были неправильно расшифрованы или когда некоторые места имели одинаковые или очень похожие названия. Кроме того, названия мест могут меняться со временем, что увеличивает вероятность ошибок. Если время прихода цунами и время прохождения доступны для конкретных мест заплеска, они включаются в базу данных. Эти данные важны для проверки моделей времени распространения цунами. Определение, используемое в этой базе данных, — это время прихода или прохождения первой волны, достигшей места заплеска. Первая волна, возможно, не была самой большой волной, поэтому время прохождения, указанное в первоисточнике, могло быть второй или третьей волной.

Исходные данные

- Код причины цунами
  - 0 Неизвестна
  - 1 Землетрясение
  - 2 Возможное землетрясение
  - 3 Землетрясение и оползень
  - 4 Извержение вулкана и землетрясение
  - 5 Извержение вулкана, землетрясение и оползень
  - 6 Извержение вулкана
  - 7 Извержение вулкана и оползень
  - 8 Оползень
  - 9 Метеорологический фактор
  - 10 Взрыв

Исходные данные

- Код подтверждения цунами
  - -1 ложное событие
  - 0 событие, вызвавшее только сейшу или волнение на внутренней реке Hr
  - 1 крайне сомнительное цунами
  - 2 маловероятное цунами
  - 3 вероятно цунами
  - 4 определенно цунами

Исходные данные

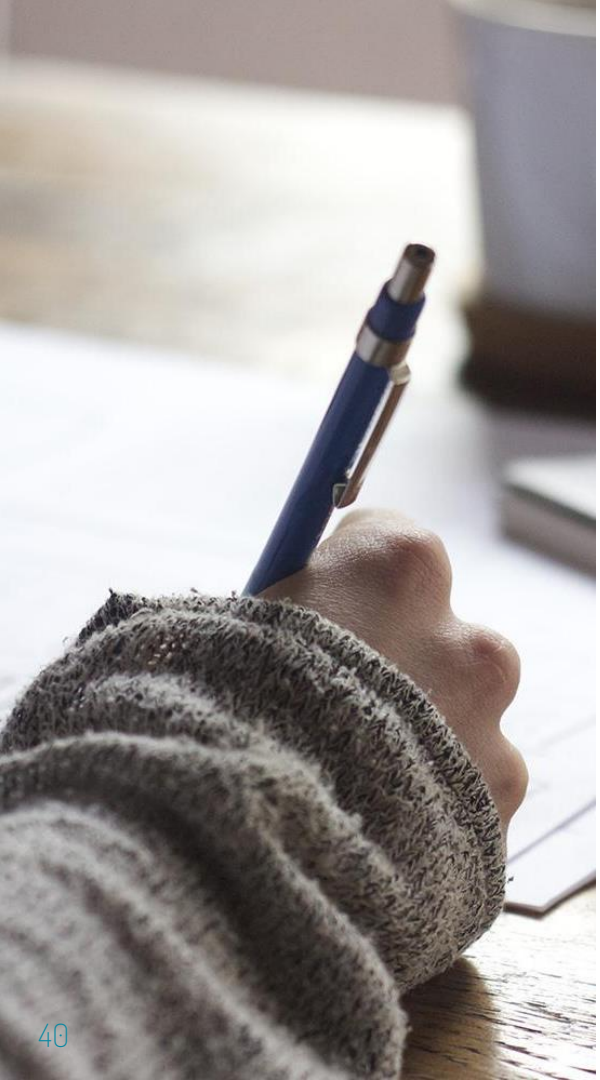
- year – дата наблюдения
- mo месяц
- dy: день
- mn: минуты
- sec: секунды
- Tsunami event validity : Действительность события цунами
- Tsunami cause code : Код причины цунами
- Earthquake Magnitude: Магнитуда землетрясения
- Deposits: Депозиты
- Country: Страна
- Location name: Название места
- Latitude: Широта
- Longitude: Долгота
- Maximum Water Height: Максимальная высота волны
- Number of Runups: Кол-во запусков

Исходные данные

- Tsunami Intensity - интенсивность
- Total death: общее число погибших
- Total missing: общее число пропавших
- Total missing description
- Total Injuries: общее число раненых
- Total Damage (Mil): суммарный ущерб (миллионы)
- Total Damage description:
- Total Houses destroyed: Кол-во разрушенных домов
- Total Houses damaged: Кол-во поврежденных домов
- Country: Страна

Задача и этапы выполнения

- Загрузить данные в dataframe, используя библиотеку pandas
- Получить представление о структуре данных и качестве данных
- Использовать графические библиотеки seaborn или matplotlib (допускается использование любых других библиотек, например plotly или bokeh) для визуализации данных и получения предварительных выводов о зависимости появление цунами от прочих параметров,
- Представить выводы



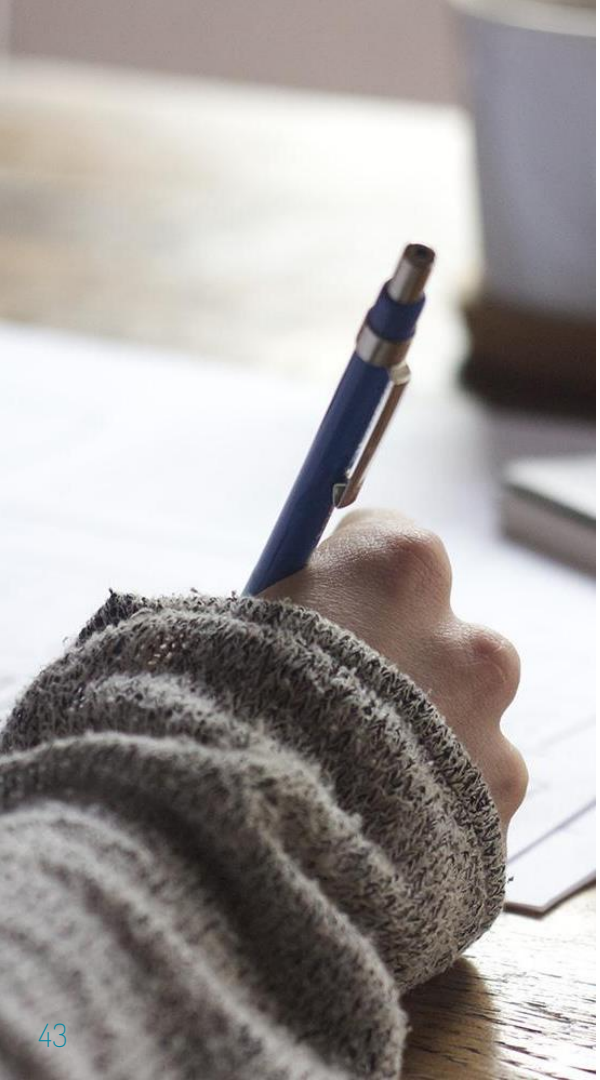
Задание № 11:
Анализ и визуализация данных рынка Airbnb NYC

Задача и этапы выполнения

- Импорт данных из трех файлов
- Очистка данных в колонке цена (price)
- Расчет средней цены
- Сравнение затрат на рынке аренды
- Очистка данных в колонке room type
- С какими сроками вы работает?
- Объедините датафреймы
- Проанализируйте цены по различным районам
- Сделайте ранжирование цен по районам

Исходные данные

- Id – идентификатор позиции
- Name: заголовок предложения по аренде
- Host_id: ид арендодателя
- Host_name: Имя арендодателя
- Neighbourhood_group: городской округ
- Neighbourhood: район в округе
- Latitide: широта
- Longitude: долгота
- Room_type: тип квартиры
- Price: цена
- Minimum_nights: минимальное время пребывания (ночей)
- Number_of_Reviews: Количество отзывов
- Last_review: Дата последнего отзыва
- Reviews_per_month: кол-во отзывов в месяц
- Availability_365: доступность в течении года



Задание № 12:
Анализ и визуализация данных об Android приложениях в Google Play

Задача и этапы выполнения

- Импорт данных
- Очистка данных
- Коррекция типов данных
- Анализ категорий приложений в магазине
- Анализ распределения рейтинга приложений
- Визуализация цены и размера приложения
- Фильтрация «мусорных» приложений
- Анализ популярности платных и бесплатных приложений
- Анализ тональности пользовательских отзывов

Исходные данные

- App: Название приложения
- Category: Категория приложения (Game, Family и т.д.)
- Rating: Пользовательский рейтинг приложения
- Reviews: Кол-во пользовательских отзывов о приложении
- Size: Размер приложения (Мб)
- Installs: Количество скачиваний приложения пользователями
- Type: Свободно распространяемое или бесплатное
- Price: Цена
- Content rating: Целевая возрастная группа пользователей (дети, взрослые..)
- Genres: Приложение может относится к нескольким жанрам
- Last Updated: дата последнего обновления
- Current ver: Текущая версия приложения
- Android ver: Текущая версия ОС Android

Спасибо!

Вопросы?

Пишите:

kotelenkosa@misis.ru