Лабораторная работа 2

Лабораторная работа 2 рассчитана на два занятия. На первом занятии работа выполняется. На втором занятии проводится защита выполненных лабораторных работ. Её целью является изучение основ работы по преобразованию данных и их хранению в БД средствами языка Python.

Лабораторная работа предусматривает выполнение в малых группах (3-4 человека на один вариант) и является первым шагом выполнения итоговой курсовой работы.

Вариант 1

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/datasets/mchirico/montcoalert.
2. Напишите программу на языке Python выполняющую следующие действия:
    a. Создаёт в СУБД SQLite или Oracle набор таблиц для хранения данных рассматриваемого набора. При этом исключите поля desc, zip, address, e.
    b. Загружает данные набора в созданные таблицы.
    c. Выполняет очистку данных, удаляя все строки с пустыми или нереальными данными.
    d. Выполняет частотный анализ данных по переменной town.
    e. Удаляет из загруженных данных все строки, содержащие города с экстремальным числом обращений в службу спасения. Экстремальным числом будем считать такую величину x, которая стоит в первых или последних 10% по порядку наблюдениях. Причём есть такая величина y (может быть равная x), стоящая ближе к центру, чем x и разрыв между которой и соседней с ней по порядку следования к центру составляет более 10% от общего разброса. Также должны быть удалены города с числом обращений, меньшим 5.
    f. Строит график зависимости общего числа обращений в службу 911 от часа суток.
    g. На основе полученных данных проведите простейшее исследование наличия факта корреляции между общим числом обращений в службу 911 и временем суток (часом).

Вариант 2

1. Зарегистрируйтесь на сайте https://www.kaggle.com/datasets и загрузите с него набор статистических данных, посвящённый глобальным террористическим актам, https://www.kaggle.com/datasets/START-UMD/gtd.
2. Напишите программу на языке Python выполняющую следующие действия:
    a. Создаёт в СУБД SQLite или Oracle набор таблиц для хранения данных рассматриваемого набора. При этом оставив из исходного набора переменные iyear, imonth, iday, country_txt, region_txt, latitude, longitude. Кроме того добавляет в набор данных новую переменную accident_date, собрав её из значений переменных iyear, imonth, iday.
    b. Загружает данные набора в созданные таблицы.
    c. Выполняет очистку данных, удаляя все строки с пустыми или нереальными данными.
    d. Выполняет частотный анализ данных по переменной country_txt.
    e. Удаляет из загруженных данных все строки, содержащие страны с экстремальным числом террористических актов. Экстремальным числом будем считать такую величину x, которая стоит в первых или последних 10% по порядку наблюдениях. Причём есть такая величина y (может быть равная x), стоящая ближе к центру, чем x и разрыв между которой и соседней с ней по порядку следования к центру составляет более 10% от общего разброса. Также должны быть удалены страны с числом террористических актов, меньшим 5.
    f. Строит график зависимости общего числа терактов от календарного месяца.
    g. На основе полученных данных проведите простейшее исследование наличия факта корреляции между общим числом терактов и календарным месяцем.

Вариант 3

1. Зарегистрируйтесь на сайте https://www.kaggle.com/datasets и загрузите с него набор статистических данных, посвящённый каталогу фильмов IMDB, https://www.kaggle.com/datasets/suchitgupta60/imdb-data (файл movie_metadata.csv).
2. Напишите программу на языке Python выполняющую следующие действия:
    a. Создаёт в СУБД SQLite или Oracle набор таблиц для хранения данных рассматриваемого набора. При этом исключите из исходного набора все переменные, кроме director_name, budget, imdb_score, title_year.
    b. Загружает данные набора в созданные таблицы.
    c. Выполняет очистку данных, удаляя все строки с пустыми или нереальными данными.
    d. Выполняет частотный анализ данных по переменной director_name.
    e. Удаляет из загруженных данных все строки, содержащие директоров с экстремальным числом фильмов. Экстремальным числом будем считать такую величину x, которая стоит в первых или последних 10% по порядку наблюдениях. Причём есть такая величина y (может быть равная x), стоящая ближе к центру, чем x и разрыв между которой и соседней с ней по порядку следования к центру составляет более 10% от общего разброса. Также должны быть удалены директора с числом фильмов, меньшим 3.
    f. Строит график зависимости imdb_score от бюджета фильма.
    g. На основе полученных данных проведите простейшее исследование наличия факта корреляции между бюджетом фильма и его рейтингом.

Вариант 4

1. Зарегистрируйтесь на сайте https://www.kaggle.com/datasets и загрузите с него набор статистических данных, посвящённый продажам домов, https://www.kaggle.com/datasets/harlfoxem/housesalesprediction
2. Напишите программу на языке Python выполняющую следующие действия:
    a. Создаёт в СУБД SQLite или Oracle набор таблиц для хранения данных рассматриваемого набора. При этом оставьте только поля date, price, yr_built, yr_renovated, sqft_living, condition. Кроме того добавьте в набор данных новую переменную real_year, собрав её как максимальное значение из значений переменных yr_built yr_renovated.
    b. Загружает данные набора в созданные таблицы.
    c. Выполняет очистку данных, удаляя все строки с пустыми или нереальными данными.
    d. Выполняет частотный анализ данных по переменной real_year.
    e. Удаляет из загруженных данных строки со сзначением real_year, которому соответствует экстремальное число продаж. Экстремальным числом будем считать такую величину x, которая стоит в первых или последних 10% по порядку наблюдениях. Причём есть такая величина y (может быть равная x), стоящая ближе к центру, чем x и разрыв между которой и соседней с ней по порядку следования к центру составляет более 10% от общего разброса. Также должны быть удалены года с числом продаж, меньшим 5. Кроме того нужно удалить строки с экстремальными ценами продаж.
    f. Строит график зависимости состояния дома от величины real_year.
    g. На основе полученных данных проведите простейшее исследование наличия факта корреляции между состоянием дома и величиной real_year.

Вариант 5

1. Зарегистрируйтесь на сайте https://www.kaggle.com/datasets и загрузите с него набор статистических данных, посвящённый опросам людей, https://www.kaggle.com/freecodecamp/2016-new-coder-survey-/version/1
2. Напишите программу на языке Python выполняющую следующие действия:
    a. Создаёт в СУБД SQLite или Oracle набор таблиц для хранения данных рассматриваемого набора. При этом оставьте только поля EmploymentField, EmploymentStatus, Gender, JobPref, JobWherePref, MaritalStatus, Income.
    b. Загружает данные набора в созданные таблицы.
    c. Выполняет очистку данных, удаляя все строки с пустыми или нереальными данными.
    d. Выполняет частотный анализ данных по переменным EmploymentField, JobPref, Gender.
    e. Удаляет из загруженных данных строки со значениями полей JobPref, Gender, которым соответствует менее 2% наблюдений.
    f. Строит гистограмму распределения дохода Income в зависимости от семеного положения MaritalStatus. Можете использовать методы на основе хи-квадрат статистики или дисперсионный анализ.
    g. На основе полученных данных проведите простейшее исследование наличия зависимости между семейным статусом человека и его предпочтениями с точки зрения присутствия в офисе (JobWherePref).

Вариант 6

1. Зарегистрируйтесь на сайте https://www.kaggle.com/datasets и загрузите с него набор статистических данных, посвящённый вопросам суицидов в мире, https://www.kaggle.com/datasets/russellyates88/suicide-rates-overview-1985-to-2016
2. Напишите программу на языке Python выполняющую следующие действия:
    a. Создаёт в СУБД SQLite или Oracle набор таблиц для хранения данных рассматриваемого набора. При этом оставьте только поля country, year, sex, age, suicides_no, population, suicides/100k pop.
    b. Загружает данные набора в созданные таблицы.
    c. Выполняет очистку данных, удаляя все строки с пустыми или нереальными данными.
    d. Выполняет анализ суммарного числа инцидентов в мире по годам.
    e. Удаляет из загруженных данных строки со странами, в которых присутствует нулевой число случаев суйцида в какой-либо возрастной группе (с учётом пола) в рассматриваемый период времени.
    f. Строит график зависимости суммарного числа суицидов в России от года.
    g. На основе полученных данных проведите простейшее исследование наличия факта корреляции между числов суицидов и годом наблюдения.