Министерство цифрового развития, связи и массовых коммуникаций

Федеральное государственное бюджетное образовательное учреждения

высшего образования

Поволжский государственный университет телекоммуникаций и информатики

Кафедра «Информационные системы и технологии»

**Основы сбора и обработки больших данных**

**Лабораторная работа № \_1\_**

ВыполнилстудентгруппыМОИС-91**:**  \_\_\_\_\_\_\_\_\_\_

Принял к.т.н., доцент кафедры ИСТ: Козлов В.В.

Самара, 2021

**Цель работы:** получить базовые навыки предобработки данных в текстовом редакторе и редакторе таблиц.

**Задачи:**

1. Найдите и скачайте в сети Интернет файл формата .txt или .csv, содержащий данные, разделённые запятой, знаком табуляции, пробелом или запятой. Необходимо согласовать свой выбор с преподавателем!

2. Сделайте копию указанного файла. Откройте её в текстовом редакторе (Блокнот). При помощи глобальной замены (Ctrl+H) заменить символ-разделитель значений какой-то другой (например, запятую на точку с запятой). Сделайте скриншот фрагмента получившегося файла и сохраните его. Также сохраните изменённый файл.

3. Открой изменённый файл в редакторе таблиц (например, MS Excel).

4. Выполните, если необходимо разбиение по столбцам.

5. Если в файле отсутствуют названия столбцов, добавьте их.

6. Сделайте скриншот получившегося файла и сохраните его.

7. По согласованию с преподавателем, определитесь с целевой переменной (столбцом).

8. Включите фильтр столбцов. Выполните фильтрацию по двум значениям целевой переменной и по одному значению двух любых независимых переменных. Сделайте скриншоты результатов и сохраните их.

9. Создайте три новый столбца разными способами. Сделайте скриншот получившихся столбцов и сохраните его.

10. Создайте новый столбец на основе целевого столбца. В новом столбце число значений должно быть равно трём. Если в оригинальном целевом столбце число возможных значений равно двум, то необходимо увеличить число принимаемых значений до четырёх (в новом столбце).

11. Раскрасьте ячейки получившегося нового целевого столбца исходя из их значений (двумя способами).

12. Постройте три вида гистограмм.

13. Выполните умножение на константу для произвольного столбца из файла.

14. Выполните сложение двух столбцов.

15. Найдите максимальное, минимальное и средние значения для определённого столбца.

16. Сравните значения двух столбцов и посчитайте процент совпадений.

17. Сравните значения двух столбцов и посчитайте процент несовпадений.

18. Выполните поиск дубликатов средствами текстового редактора.

19. Выполните поиск выбросов в данных.

20. Выполните предобработку файла, выданного преподавателем.

**Ход работы:**

Для начала нужно скачать набор данных с официального сайта data.gov.ru. Был выбран файл с тематикой спортивных дисциплин.

Создаётся резервная копия скачанного файла. Далее основной файл открывается в текстовом редакторе. В моём случае это Visual Studio Code. С помощью функции «Глобальная замена» все точки с запятой в файле заменяются на запятую.

Изображение выглядит как текст

Автоматически созданное описание

Полученный файл импортируется в Excel.

Изображение выглядит как стол

Автоматически созданное описание

Изображение выглядит как текст, внутренний, компьютер, монитор

Автоматически созданное описание

Включил фильтр столбцов. Выполнил фильтрацию по двум значениям целевой переменной и по одному значению двух любых независимых переменных. Целевой переменной стал столбец section. Двумя независимыми — subsection и title.



Создал три новых столбца разными способами. Столбец time был создан при помощи формулы, где из даты окончания вычитается дата начала и предоставляет информацию о том, сколько дней проводилось мероприятие. «Столбец1» и «Столбец2» были получены в результате разделения столбца address по запятой.

Изображение выглядит как стол

Автоматически созданное описание

Создал новый столбец на основе целевого столбца. В новом столбце число значений должно быть равно трём. Если в оригинальном целевом столбце число возможных значений равно двум, то необходимо увеличить число принимаемых значений до четырёх (в новом столбце).

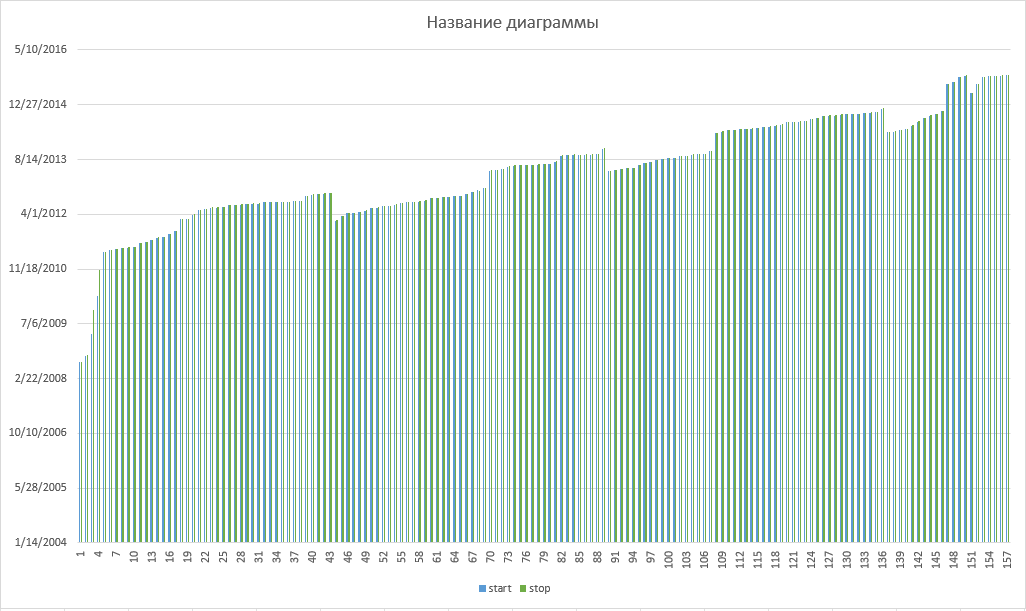
Раскрасил ячейки получившегося нового целевого столбца исходя из их значений (двумя способами).

Столбец в зависимости от времени года, в которое проводилось соревнование, и вида спорта, какой тип был у события.

Изображение выглядит как стол

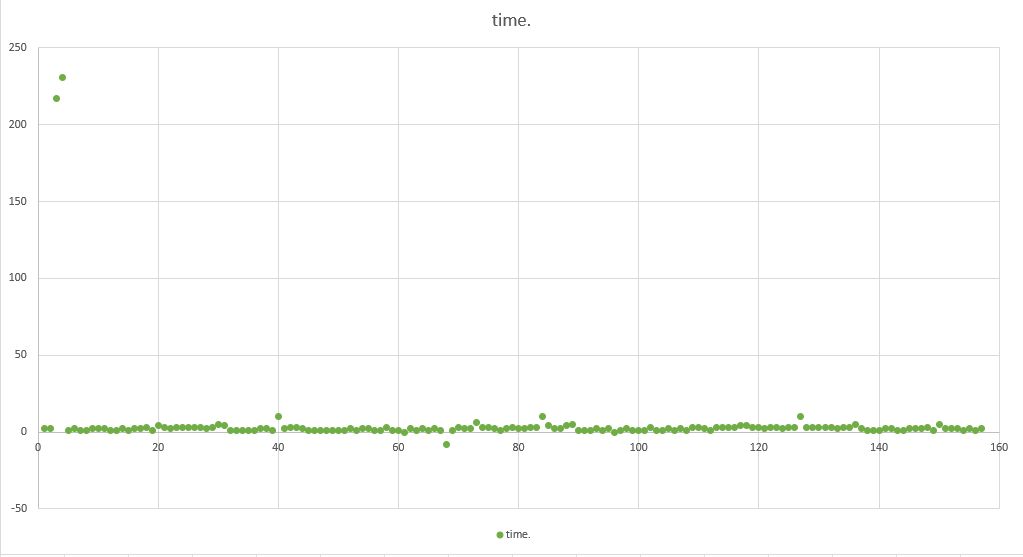
Автоматически созданное описание

Построил три вида гистограмм. Первая отображает количество мероприятий за какой-либо год, вторая — соревнования с определённым количеством участников. Третья — сколько времени проводилось событие.



Изображение выглядит как текст

Автоматически созданное описание



Выполнил умножение на константу для произвольного столбца из файла. Был выбран столбец, отображающий количество участников. Умножение производилось на число 7.

Изображение выглядит как текст, снимок экрана, электроника, дисплей

Автоматически созданное описание

Выполнил сложение двух столбцов. Складывал столбец, отображающий кол-во дней, и столбец с количеством участников соревнования.

**Изображение выглядит как текст, снимок экрана, внутренний

Автоматически созданное описание**Нашёл максимальное, минимальное и средние значения для определённого столбца.

Изображение выглядит как стол

Автоматически созданное описание

Сравнил значения двух столбцов и посчитал процент совпадений и несовпадений.



Выполнил поиск дубликатов средствами текстового редактора VS Code.

Изображение выглядит как текст, табличка

Автоматически созданное описание

Выполнил поиск выбросов (отличные значения) в данных при помощи пользовательского автофильтра.

Изображение выглядит как стол

Автоматически созданное описание

Изображение выглядит как стол

Автоматически созданное описание

**Вывод:** в ходе работы получил базовые навыки предобработки данных в текстовом редакторе и редакторе таблиц.

**Полученные результаты:**

