УДК 004.043

Анализ данных с использованием Python

Мысив Виталина Владимировна,
бакалавр
Лаврушина Елена Геннадьевна,
преподаватель
ФБГОУ ВО «Владивостокский государственный университет»
Россия. Владивосток
E-mail: millervitalina2@gmail.com; Тел. +79841471877
ул. Гоголя, 41, г. Владивосток, Приморский край, Россия, 690014

Данная работа посвящена актуальной теме – анализу данных на Python, который сегодня востребован в науке, бизнесе и IT. Основная задача данного доклада — на практике продемонстрировать, как с помощью Python и его библиотек (Pandas, NumPy, Matplotlib) можно эффективно анализировать структурированные данные, используя в качестве примера датасет наблюдений НЛО.

Ключевые слова: *анализ данных, Python, Pandas, NumPy, Matplotlib, обработка данных, наблюдения НЛО.*

Data analysis using Python

This paper is devoted to the current topic - data analysis in Python, which is in demand today in science, business and IT. The main task of this paper is to demonstrate in practice how Python and its libraries (Pandas, NumPy, Matplotlib) can be used to efficiently analyze structured data, using a dataset of UFO observations as an example.

Keywords: *data analysis, Python, Pandas, NumPy, Matplotlib, data processing, UFO sightings.*

Введение

Анализ данных — это процесс изучения и обработки информации для поиска полезных закономерностей. Сейчас его применяют в бизнесе, медицине, науке и даже в исследовании необычных явлений, таких как наблюдения НЛО.

Python — один из лучших инструментов для анализа данных благодаря простоте и мощным библиотекам. В этой работе мы разберём набор данных о наблюдениях НЛО с сайта National UFO Reporting Center и покажем, как легко обрабатывать такие данные.

Подготовка данных.

Набор данных содержит информацию о наблюдениях:

- дата и время (datetime);
- место (city, state, country);
- длительность наблюдения (duration (seconds), duration (hours/min));
- форма объекта (shape);
- комментарии очевидцев (comments);
- координаты (latitude, longitude).

Установка библиотек.

Для работы понадобятся:

- Pandas – для обработки таблиц;
- Matplotlib – для визуализации;
- NumPy – для математических операций;
- PyCountry – для работы с названиями стран.

Посчитаем количество наблюдений по странам и построим график (Рис.1, Рис.2).

```
import pandas as pd

# Загрузка данных
df = pd.read_csv('C:/Users/User/Desktop/scrubbed.csv', low_memory=False, on_bad_lines='skip')

# Замена отсутствующих значений
df['shape'] = df['shape'].fillna('unknown')

import matplotlib.pyplot as plt
# Группировка по странам
country_counts = df['country'].value_counts().head(10)

# Построение графика
plt.figure(figsize=(10, 5))
plt.bar(country_counts.index, country_counts.values, color='skyblue')
plt.title('Топ стран по наблюдениям НЛО')
plt.xlabel('Страна')
plt.ylabel('Количество наблюдений')
plt.xticks(rotation=45)
plt.show()
```

Рис.1. Код для определения количества наблюдений по странам

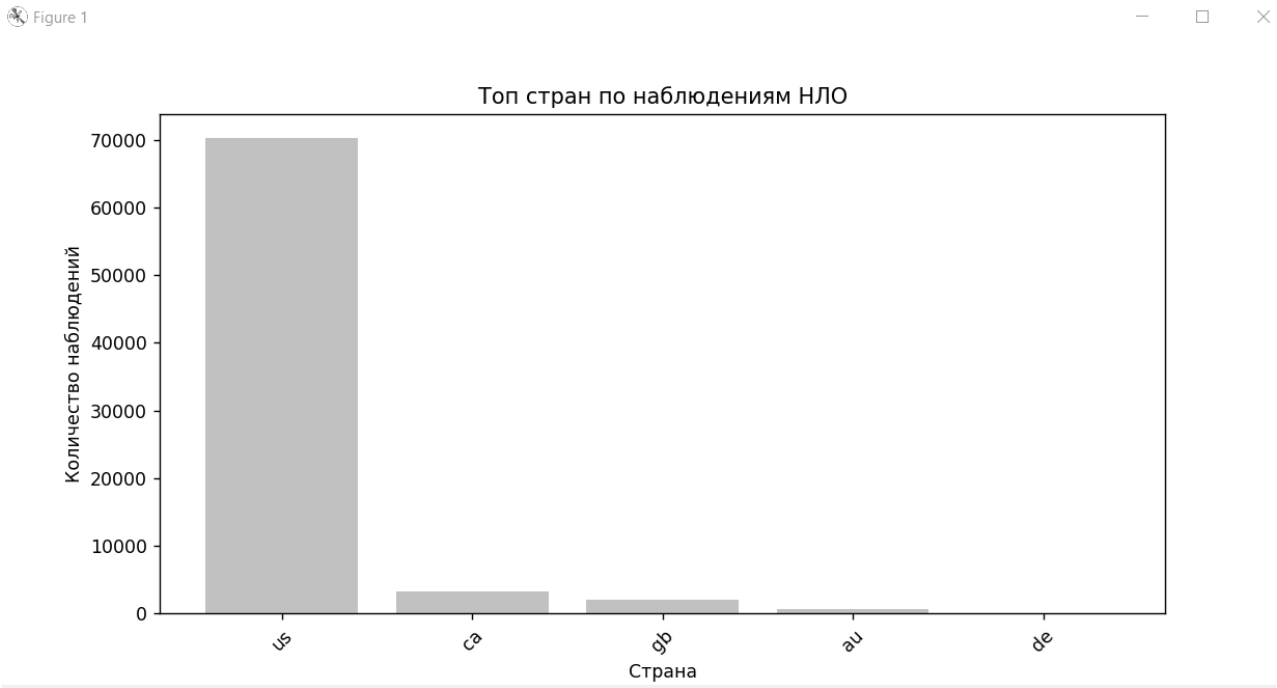

Рис.2. Страны, где больше всего наблюдений

Видно, что больше всего наблюдений в США, затем идут Канада и Великобритания.

Посмотрим какие формы объектов на небе видели и сколько раз (Рис.3, Рис.4).

```
shape_counts = df['shape'].value_counts().head(10)

plt.figure(figsize=(10, 5))
plt.barh(shape_counts.index, shape_counts.values, color='orange')
plt.title('Самые частые формы НЛО')
plt.xlabel('Количество наблюдений')
plt.show()
```

Рис.3. Код для определения форм объектов

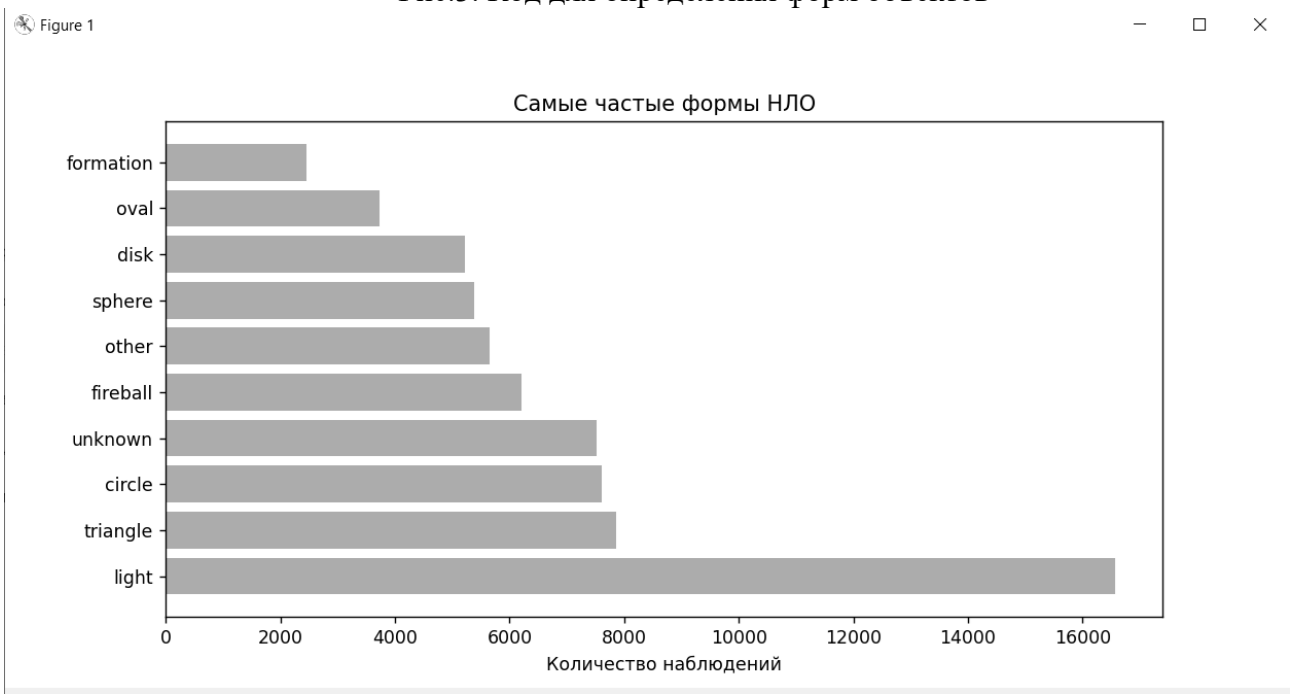

Рис.4. Типы объектов наблюдения

Световые явления - наиболее частые "НЛО". Скорее всего, это отражения прожекторов/огней, метеоры или техногенные объекты.

Можно ещё посмотреть среднее время, на которое в небе появлялся каждый из объектов (Рис.5, Рис.6).

```
import numpy as np
def clean_duration(duration):
    try:
        return float(duration)
    except:
        return np.nan

df['duration_seconds'] = df['duration (seconds)'].apply(clean_duration)

# Очистка от выбросов (удаляем наблюдения > 24 часов)
df = df[df['duration_seconds'] <= 86400]  # 86400 секунд = 24 часа

# 2. Анализ длительности по форме объекта
# Переводим секунды в часы
df['duration_hours'] = df['duration_seconds'] / 3600

# Группируем по форме объекта, считаем среднее и медиану
duration_stats = df.groupby('shape')['duration_hours'].agg(['mean', 'median'])
duration_stats = duration_stats.sort_values('median', ascending=False).head(10)

# 3. Визуализация результатов
plt.figure(figsize=(12, 6))

# График средних значений
plt.subplot(1, 2, 1)
duration_stats['mean'].plot(kind='bar', color='purple')
plt.title('Средняя длительность наблюдений')
plt.ylabel('Часы')
plt.xticks(rotation=45)
plt.grid(True)

# График медианных значений
plt.subplot(1, 2, 2)
duration_stats['median'].plot(kind='bar', color='orange')
plt.title('Медианная длительность наблюдений')
plt.ylabel('Часы')
plt.xticks(rotation=45)
plt.grid(True)

plt.tight_layout()
plt.show()
```

Рис.5. Код для анализа длительности наблюдений НЛО

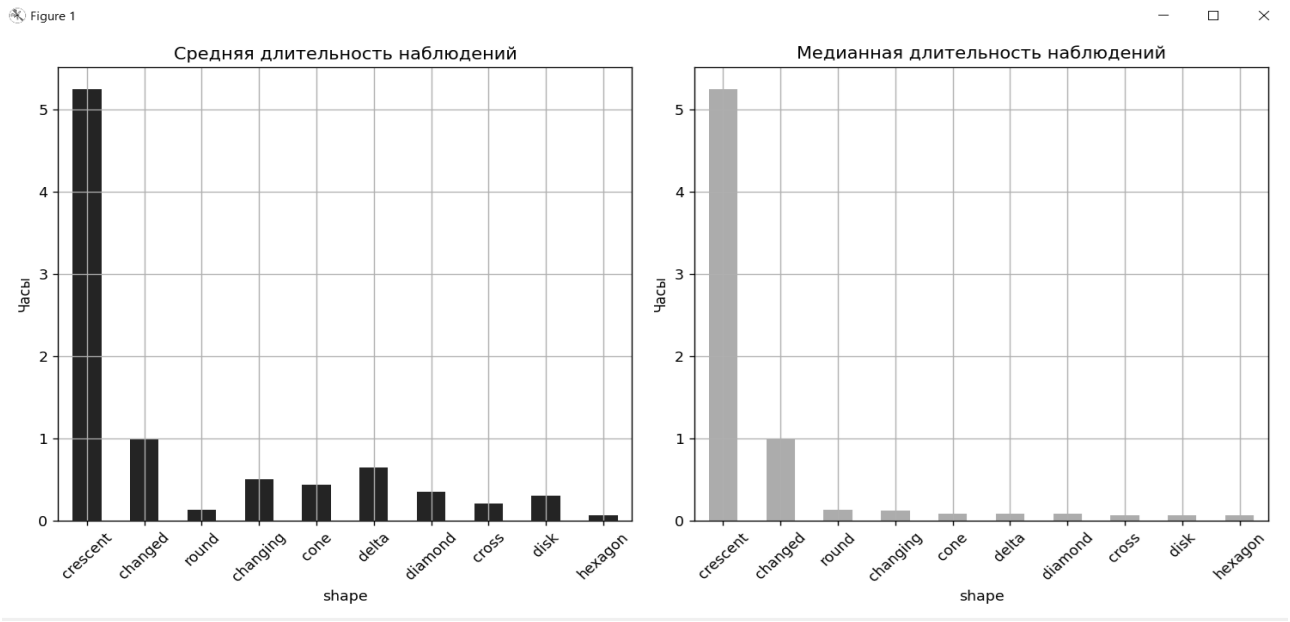

Рис.6. Среднее время появления каждого объекта

Диаграмма слева показывает, что средняя длительность искажается единичными случаями с крайне высокими значениями (например, сообщениями о "висящих" объектах в течение многих часов). Выбросы могут сильно искажать среднее значение, поэтому при больших отклонениях используют медиану. На графике справа видно, что медианная длительность дает более реалистичную оценку: для большинства форм объектов она составляет 5–15 минут, что соответствует кратковременным визуальным наблюдениям. Полумесяц наблюдали в небе чуть больше 5-ти часов, а другие объекты ненадолго появились в небе. Это уже наиболее достоверно.

Больше всего наблюдений в США, что может быть связано с высокой активностью в области наблюдений.

Чаще всего люди видят светящиеся объекты, что может быть связано с различными естественными или искусственными источниками света.

Среднее время появления объектов в небе также дало интересные результаты, и использование медианы вместо среднего значения позволяет более точно отразить данные, особенно при наличии выбросов.

Использование Python и его библиотек (Pandas, NumPy, Matplotlib, и других) делает процесс анализа данных более доступным и эффективным. Графическое представление данных (гистограммы, бар-графики) помогает лучше понять и интерпретировать результаты анализа.

Анализ данных о наблюдениях НЛО иллюстрирует, как статистические методы могут быть использованы для исследования и понимания явлений, которые могут показаться необычными или загадочными. Это подчеркивает важность критического мышления и научного подхода в интерпретации данных.