МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Кафедра Информационных систем

КУРСОВОЙ ПРОЕКТ

по дисциплине «Интеллектуальный анализ данных»

на тему: «исследование взаимодействий ip-адресов и их параметров»

Студент гр. 0374 _____ Наливайко Д. Д.

Преподаватель _____ Татчина Я. А.

Санкт-Петербург

2025

Целью данного исследования является исследование взаимодействий ip-адресов и их параметров. Анализ проводится на основе набора данных, содержащего события, связанные с атаками в области кибербезопасности.

Будет применён описательный анализ данных с визуализацией и выявлением повторяющихся паттернов атак. Особое внимание уделено наиболее распространённым протоколам, типам трафика и взаимодействиям между IP-адресами.

Набор данных состоит из 40 000 записей и включает 25 признаков, таких как временные метки (Timestamp), IP-адреса источника и назначения, порты, протоколы (TCP, UDP, ICMP и др.), длина пакетов, типы трафика (HTTP, FTP, DNS и др.), оценка аномальности поведения (Anomaly Scores), уровень опасности события (Severity Level) и другие. Полный список характеристик представлен в таблице, где указаны названия столбцов, типы данных и их описание **(**рис. 1):

Название столбца	**Тип данных**	**Описание**
Timestamp	datetime64[ns]	Время события
Source IP Address	object	IP-адрес источника
Destination IP Address	object	IP-адрес назначения
Source Port	int64	Порт источника
Destination Port	int64	Порт назначения
Protocol	object	Протокол передачи (TCP, UDP, ICMP и др.)
Packet Length	int64	Длина сетевого пакета в байтах
Packet Type	object	Тип сетевого пакета (например, SYN, ACK, FIN и др.)
Traffic Type	object	Тип сетевого трафика (HTTP, FTP, DNS и др.)
Payload Data	object	Содержимое полезной нагрузки пакета
Malware Indicators	object	Признаки наличия вредоносного ПО
Anomaly Scores	float64	Оценка аномальности поведения (например, от 0 до 1)
Alerts/Warnings	object	Срабатывания систем мониторинга или предупреждения

Attack Type	object	Тип атаки (например, DDoS, Malware, Port Scan и др.)
Attack Signature	object	Сигнатура атаки, зафиксированная системой обнаружения
Action Taken	object	Мера, принятая системой (Blocked, Logged, Ignored и др.)
Severity Level	object	Уровень опасности события (Low, Medium, High, Critical)
User Information	object	Сведения о пользователе, если известны
Device Information	object	Сведения об устройстве, инициировавшем трафик
Network Segment	object	Сегмент сети, в котором произошло событие
Geo-location Data	object	Географическое положение источника/назначения
Proxy Information	object	Данные о прокси, если использовались
Firewall Logs	object	Логи межсетевого экрана
IDS/IPS Alerts	object	Срабатывания систем обнаружения и предотвращения вторжений
Log Source	object	Источник лога (Firewall, IDS, SIEM и др.)

	Timestamp	Source IP Address	Destination IP Address	Source Port	Destination Port	Protocol	Packet Length	Packet Type	Traffic Type	Payload Data	...	Action Taken	Severity Level	User Information	Device Information	Network Segment	Geo-location Data	Proxy Information	Firewall Logs	IDS/IPS Alerts	Log Source
0	2023-05-30 06:33:58	103.216.15.12	84.9.164.252	31225	17616	ICMP	503	Data	HTTP	Qui natus odio asperiores nam. Optio nobis ius...	...	Logged	Low	Reyansh Dugal	Mozilla/5.0 (compatible; MSIE 8.0; Windows NT ...	Segment A	Jamshedpur, Sikkim	150.9.97.135	Log Data	NaN	Server
1	2020-08-26 07:08:30	78.199.217.198	66.191.137.154	17245	48166	ICMP	1174	Data	HTTP	Aperiam quos modi officiis veritatis rem. Omni...	...	Blocked	Low	Sumer Rana	Mozilla/5.0 (compatible; MSIE 8.0; Windows NT ...	Segment B	Bilaspur, Nagaland	NaN	Log Data	NaN	Firewall
2	2022-11-13 08:23:25	63.79.210.48	198.219.82.17	16811	53600	UDP	306	Control	HTTP	Perferendis sapiente vitae soluta. Hic delectu...	...	Ignored	Low	Himmat Karpe	Mozilla/5.0 (compatible; MSIE 9.0; Windows NT ...	Segment C	Bokaro, Rajasthan	114.133.48.179	Log Data	Alert Data	Firewall
3	2023-07-02 10:38:46	163.42.196.10	101.228.192.255	20018	32534	UDP	385	Data	HTTP	Totam maxime beatae expedita explicabo porro l...	...	Blocked	Medium	Fateh Kibe	Mozilla/5.0 (Macintosh; PPC Mac OS X 10_11_5; ...	Segment B	Jaunpur, Rajasthan	NaN	NaN	Alert Data	Firewall
4	2023-07-16 13:11:07	71.166.185.76	189.243.174.238	6131	26646	TCP	1462	Data	DNS	Odit nesciunt dolorem nisi iste iusto. Animi v...	...	Blocked	Low	Dhanush Chad	Mozilla/5.0 (compatible; MSIE 5.0; Windows NT ...	Segment C	Anantapur, Tripura	149.6.110.119	NaN	Alert Data	Firewall
5	2022-10-28 13:14:27	198.102.5.160	147.190.155.133	17430	52805	UDP	1423	Data	HTTP	Repellat quas illum harum fugit incidunt exerc...	...	Logged	Medium	Zeeshan Viswanathan	Opera/8.58.(X11; Linux i686; nl-NL) Presto/2.9...	Segment C	Aurangabad, Meghalaya	NaN	NaN	NaN	Server

Рисунок 1 – Первые 5 записей

Первичный анализ данных показал, что типы данных соответствуют содержимому: числовые признаки представлены целыми числами (int64) или числами с плавающей точкой (float64), временные данные — в формате datetime, а категориальные — в виде строк (object) (рис.2). Большинство признаков в датасете заполнены корректно, пропуски встречаются крайне редко (рис. 3). Категориальные признаки, такие как Protocol, Traffic Type, Attack Type, Severity Level, Action Taken и Log Source, содержат ограниченное

количество уникальных значений и хорошо подходят для анализа и визуализации.

Числовые признаки, например Packet Length и Anomaly Scores, имеют широкий диапазон значений, что позволяет использовать их для статистического анализа, построения распределений и выявления выбросов. Признак Timestamp особенно важен для анализа временных паттернов — он поможет определить пиковые периоды активности атак. Целевым признаком для исследования могут служить либо Attack Type (для классификации типов атак), либо Severity Level (для оценки уровня угрозы), в зависимости от конкретной задачи (рис. 2).

Таким образом, датасет структурирован и подходит для дальнейшего анализа: визуализации, построения корреляционных связей, выявления закономерностей.

```
df.dtypes # определяем тип данных столбцов
```

```
Timestamp                datetime64[ns]
Source IP Address                object
Destination IP Address           object
Source Port                       int64
Destination Port                  int64
Protocol                         object
Packet Length                     int64
Packet Type                      object
Traffic Type                     object
Payload Data                     object
Malware Indicators               object
Anomaly Scores                  float64
Alerts/Warnings                  object
Attack Type                      object
Attack Signature                 object
Action Taken                     object
Severity Level                   object
User Information                 object
Device Information               object
Network Segment                  object
Geo-location Data                object
Proxy Information                object
Firewall Logs                    object
IDS/IPS Alerts                   object
Log Source                       object
dtype: object
```

Рисунок 2 – Типы данных

```
df.isnull().sum() # проверяем на нулевые значения
```

```
Timestamp                  0
Source IP Address          0
Destination IP Address     0
Source Port                0
Destination Port           0
Protocol                   0
Packet Length              0
Packet Type                0
Traffic Type               0
Payload Data               0
Malware Indicators     20000
Anomaly Scores             0
Alerts/Warnings        20067
Attack Type                0
Attack Signature           0
Action Taken               0
Severity Level             0
User Information           0
Device Information         0
Network Segment            0
Geo-location Data          0
Proxy Information      19851
Firewall Logs          19961
IDS/IPS Alerts         20050
Log Source                 0
dtype: int64
```

Рисунок 3 – Количество нулевых значений

```
df.nunique() # уникальные значения
```

```
Timestamp              39997
Source IP Address      40000
Destination IP Address 40000
Source Port            29761
Destination Port       29895
Protocol                   3
Packet Length           1437
Packet Type                2
Traffic Type               3
Payload Data           40000
Malware Indicators         1
Anomaly Scores          9826
Alerts/Warnings            1
Attack Type                3
Attack Signature           2
Action Taken               3
Severity Level             3
User Information       32389
Device Information     32104
Network Segment            3
Geo-location Data       8723
Proxy Information      20148
Firewall Logs              1
IDS/IPS Alerts             1
Log Source                 2
dtype: int64
```

Рисунок 4 – Количество уникальных значений

Количество запросов по каждому протоколу распределено равномерно (рис. 5). Что можно сказать и о запросах по типу трафика (рис. 6).

Анализ распределения запросов по протоколам показал, что ICMP, UDP и TCP встречаются почти одинаково часто, что может свидетельствовать о разнообразии атак или сбалансированности данных (рис. 5).

График распределения запросов по типу трафика (рис. 6) демонстрирует схожую картину: столбцы почти одинаковы, что указывает на равномерное распределение между категориями, такими как HTTP, FTP, DNS.

Кроме того, аналих взаимодействий между IP-адресами показал, что ни одна пара не повторяется, даже если учитывать обратное направление (например, A → B и B → A). Это может указывать на то, что данные охватывают широкий спектр уникальных атак, происходящих между различными источниками и целями (рис. 7).

```
Количество запросов по каждому протоколу:
ICMP: 13429
UDP: 13299
TCP: 13272
```

Рисунок 5 – Количество запросов по каждому протоколу

```
plt.xlabel("Тип трафика", fontsize=12)
plt.ylabel("Количество запросов", fontsize=12)
plt.xticks(rotation=0)
plt.grid(axis='y', linestyle='--', alpha=0.7)

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

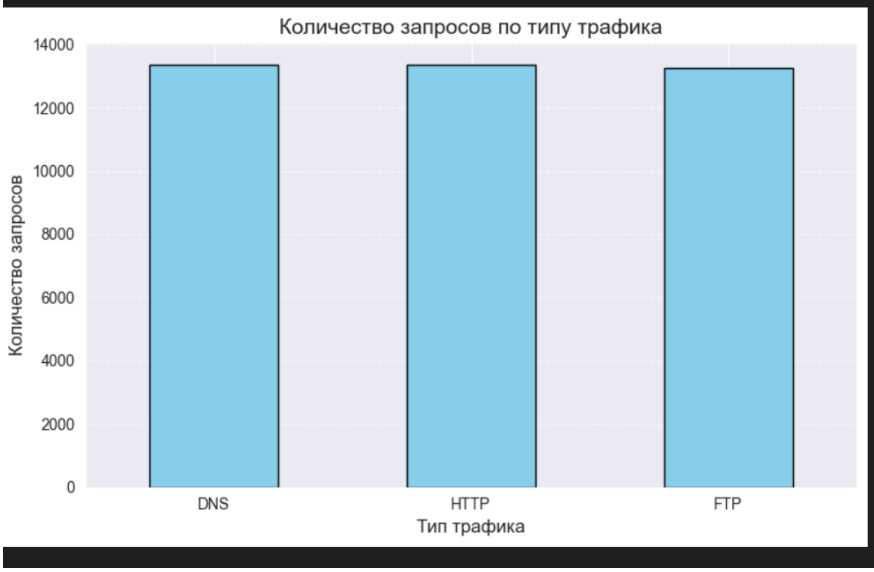

Рисунок 6 – График количества запросов по типу трафика

```
duplicate_pairs = pair_counts[pair_counts > 1]

print(f"Количество дублирующихся взаимодействий (независимо от порта и направления): {len(duplicate_pairs)}")
```

```
Количество дублирующихся взаимодействий (независимо от порта и направления): 0
```

Рисунок 7 – Количество дублирующийся взаимодействий

График показывает, что среди заблокированных запросов, которые сопровождались предупреждениями систем IDS/IPS и исходили от фаервола,

чаще всего встречается протокол ICMP , за ним следует UDP , а на последнем месте — TCP (рис. 8). Это может говорить о том, что системы безопасности чаще реагируют на определённые типы сетевой активности. Например, ICMP часто используется в атаках типа DDoS или для сканирования доступности хостов, что делает его подозрительным для систем защиты. UDP, в свою очередь, часто применяется в атаках спуфинга или DNS-флуда, так как он не требует установления соединения и легче маскируется. TCP, несмотря на его устойчивость к подобным атакам, всё же блокируется, но реже — например, при обнаружении необычных флагов в заголовках пакетов или попытках эксплуатации уязвимостей.

Разница в количестве блокировок между протоколами может быть связана с их особенностями. ICMP и UDP менее защищены от манипуляций, поэтому злоумышленники активно используют их для атак. TCP же сложнее подделать, но при этом его аномалии (например, нестандартные комбинации флагов) всё равно фиксируются системами IDS/IPS.

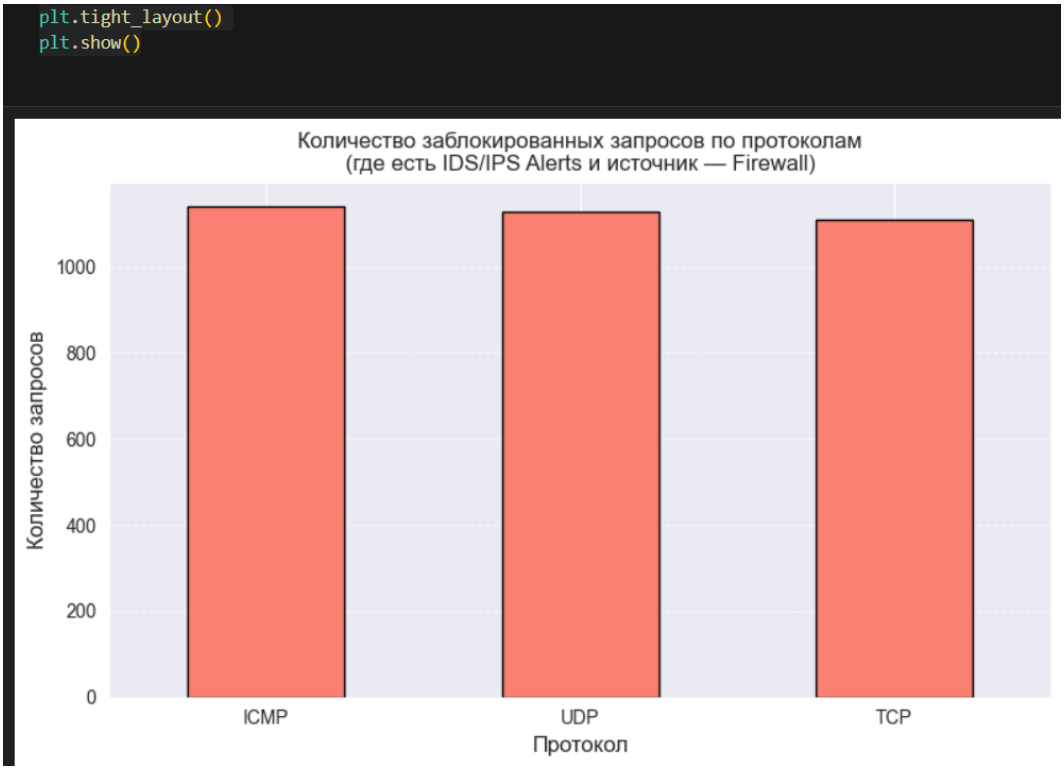

Рисунок 8 – График заблокированных запросов по протоколам

Построил графики распределения числовых признаков из набора данных о кибератаках.

Каждый график отображает:

•Гистограмму — частоту встречаемости значений.

•KDE (ядерную оценку плотности) — плавную линию, показывающую форму распределения.

•Значение skewness (асимметрии) — меру отклонения распределения от симметричного (нормального).

Результаты показали, что все четыре параметра (source port, destination port, packet length, anomaly scores) имеют почти симметричные распределения с минимальной асимметрией (значения skewness от -0.01 до 0.02) (рис. 9). Это означает, что данные по этим признакам равномерно распределены вокруг среднего значения, без явного перекоса в сторону больших или меньших значений.

Порты источника и назначения используются равномерно, что может указывать на разнообразие атак, а не на фокусировку на определённых сервисах. Длина пакетов варьируется в широком диапазоне, но без доминирования коротких или длинных пакетов, что характерно для смешанного трафика. Низкие значения аномальности говорят о том, что большинство событий в данных имеют нормальные характеристики, а явно вредоносные атаки встречаются редко.

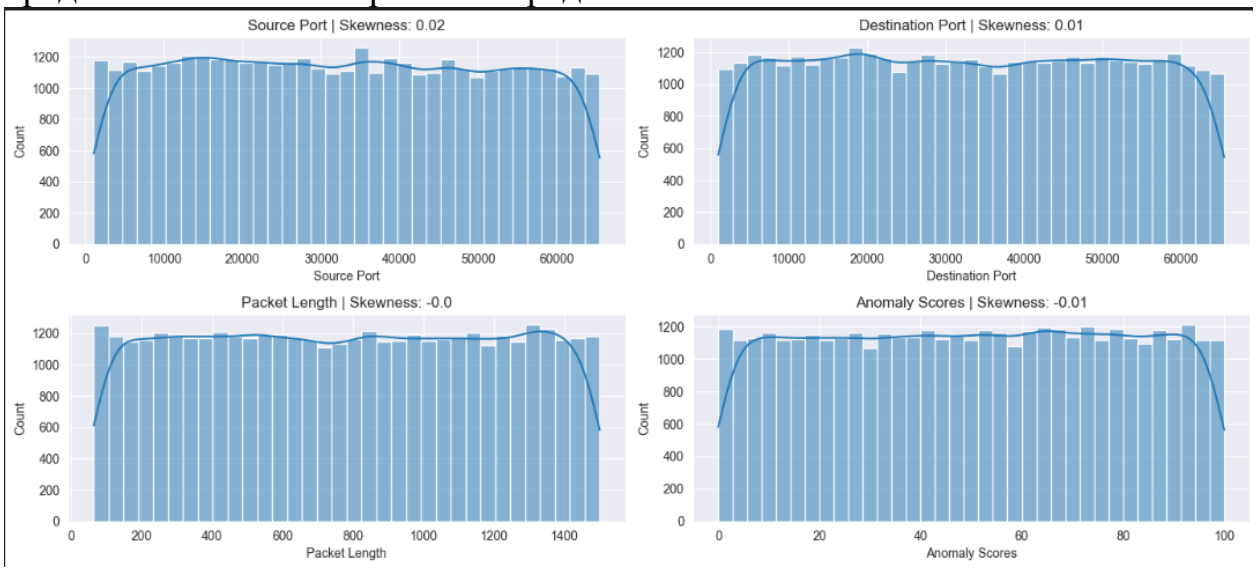

Рисунок 9 – Графики распределения числовых признаков

Также было определено количество взаимодействий (записей), которые содержат хотя бы один IP-адрес попадающий в диапазон от 114.0.0.0 до 116.255.255.255 – 1004 (рис. 10).

```
df = pd.read_csv('D:/cybersecurity_attacks.csv')
mask = df["Source IP Address"].apply(ip_in_custom_range) | df["Destination IP Address"].apply(ip_in_custom_range)
filtered_df = df[mask]

interaction_count = filtered_df.shape[0]

interaction_count
```

```
1004
```

Рисунок 10 – Количество взаимодействий от 114.0.0.0 до 116.255.255.255

График показывает, сколько раз IP-адреса из подсетей /8 (диапазоны вида X.0.0.0 — X.255.255.255, где X — число от 50 до 60) участвовали в атаках.

Каждый столбец на графике соответствует одному числу в первом октете IP-адреса, а высота столбца отражает общее количество взаимодействий (атак), связанных с этой подсетью. Некоторые диапазоны IP-адресов чаще участвуют в атаках, чем другие (рис. 11).

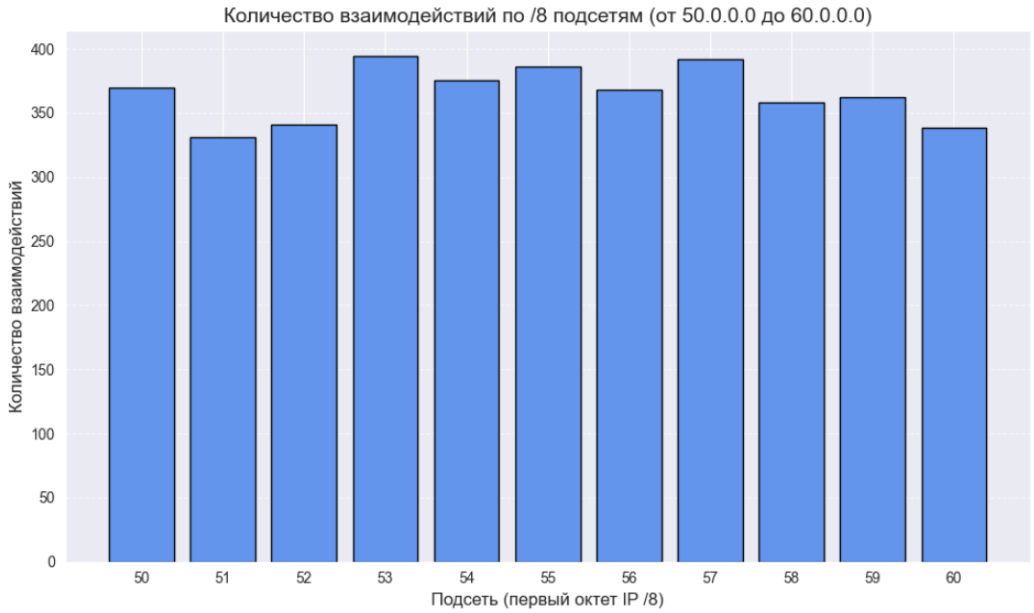

Рисунок 11 – График количества взаимодействий по /8 подсетям

Был построен график, который показывает, как часто система предпринимала те или иные действия в ответ на атаки:

- «Blocked» — атака была заблокирована,
- «Ignored» — атака проигнорирована,
- «Logged» — событие записано в логи.

К сожалению, несмотря на преимущество заблокированных атак, почти какая же часть атак была проигнорирована, а другая записана в логи (рис. 12). Соответственно можно сделать вывод, что система безопасности недостаточно эффективна.

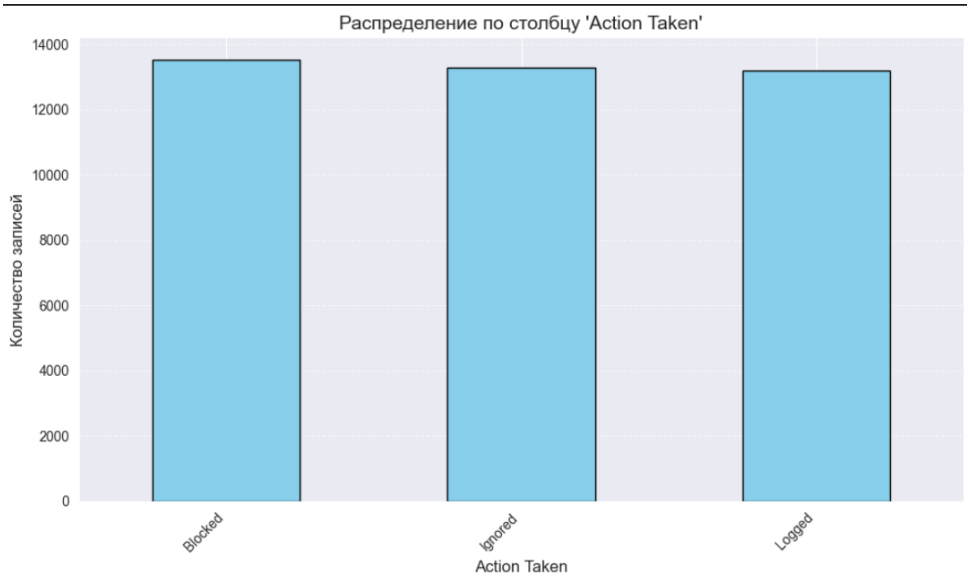

Рисунок 12 – Графики распределения по столбцу «Action Taken»

Также были получен топ 10 городов которые преобладают по TCP взаимодействию (рис. 13).

```
 Топ-10 городов по количеству TCP-взаимодействий:
City
Aurangabad         81
Ghaziabad          80
Kakinada           64
Allahabad          59
Kalyan-Dombivli    57
Medininagar        57
Bally              56
Avadi              56
Muzaffarpur        56
Vijayanagaram      55
```

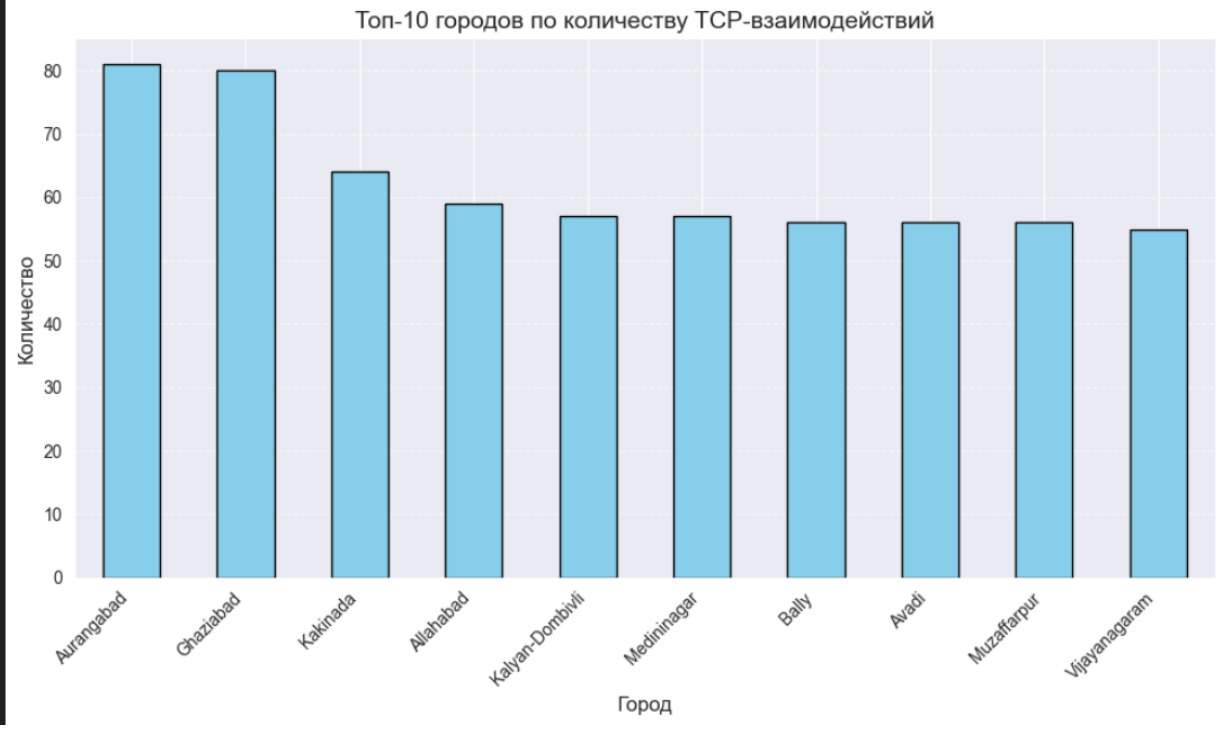

Рисунок 13 – Топ 10 городов по TCP взаимодействиям

Построил график, который показывает распределение номеров портов источника для разных типов сетевого трафика (HTTP, DNS, FTP).

Для HTTP-трафика (обозначен зелёным) медианное значение порта — около 33 000. Диапазон довольно широкий — от 200 до 65 000. Это связано с тем, что браузеры и другие клиентские приложения используют случайные высокие порты, в то время как сервер обычно работает на стандартных портах 80 или 443. Такое распределение — норма для клиент-серверного взаимодействия в HTTP.

У DNS-трафика (оранжевый) медиана чуть ниже — примерно 32 000. Это может быть связано с кэшированием, повторными запросами или даже попытками скрытой передачи данных. Такой разброс может указывать как на обычную активность, так и на аномалии.

FTP-трафик (синий) показывает медиану около 33 000. Распределение похоже на HTTP, но чуть смещено в сторону более низких значений. Это можно объяснить тем, что FTP использует как стандартный порт 21, так и динамический диапазон для передачи данных, особенно в пассивном режиме. Также возможны влияния прокси-серверов или специфических настроек.

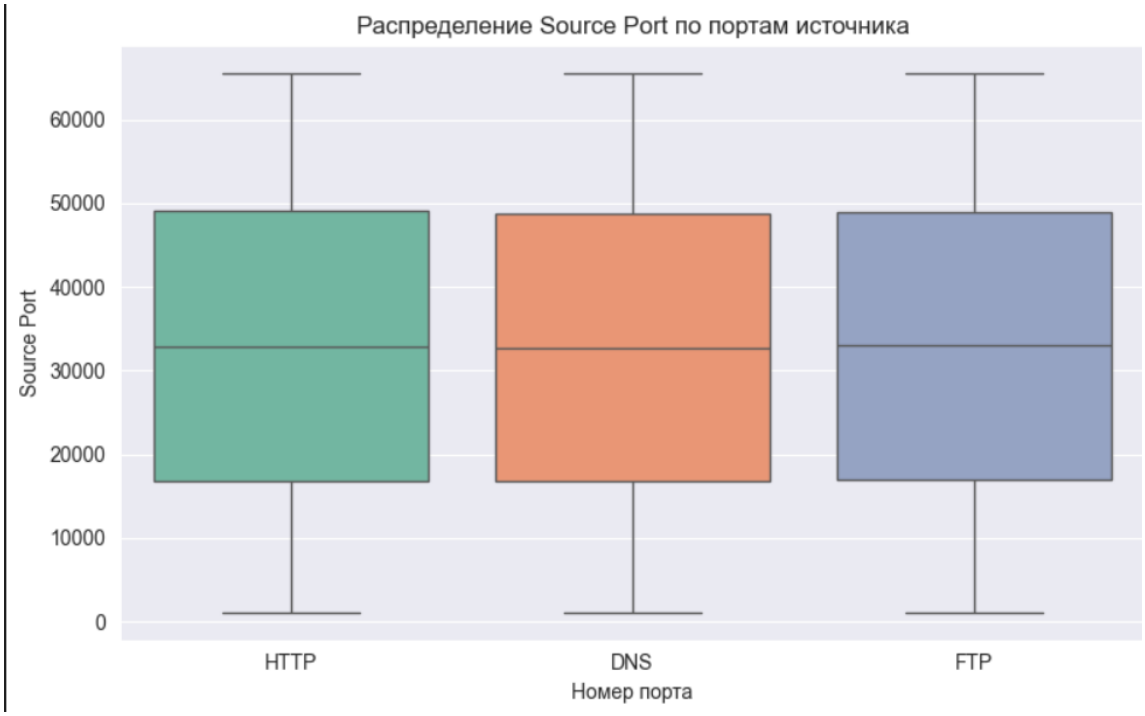

Рисунок 14 – График распределения «Source Port» по портам источника

Делаем вывод, что сетевой трафик распределён равномерно по протоколам и типам, взаимодействия между IP-адресами уникальны, а система безопасности не всегда эффективно реагирует на угрозы. Не весь трафик с высокими портами или необычными параметрами указывает на атаки.