Final Project – Phase 1

EDA & Baseline Model

Supervisor: Sajjad Ramezani

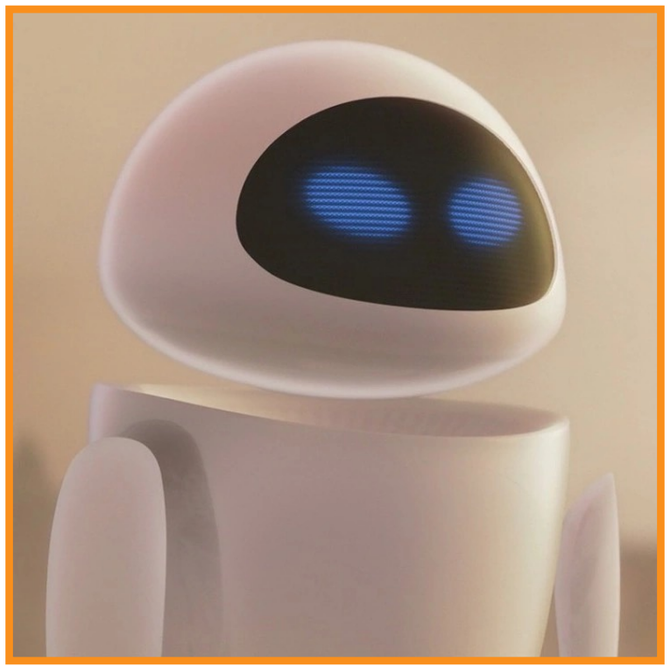
Mohammad Hashemi – Parastoo Falak Aflaki

Outline

Data Exploration

Outline

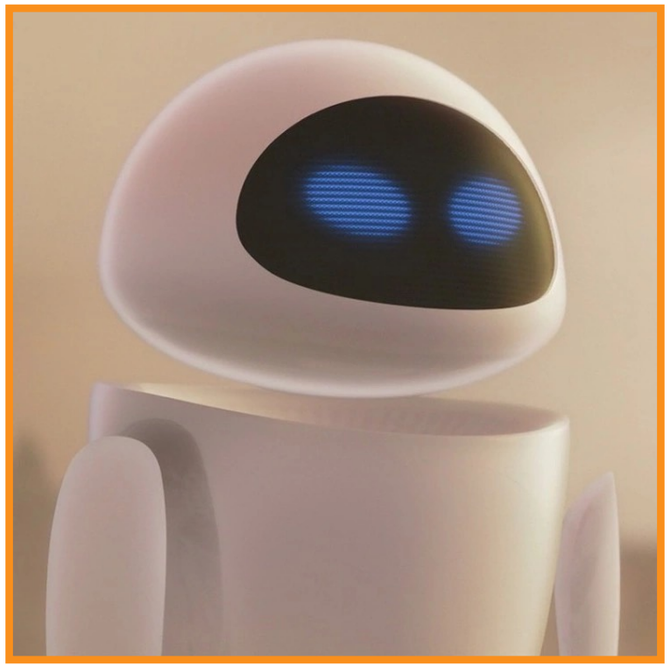
Data Exploration



Feature Engineering

Outline

Data Exploration



Feature Engineering

Baseline Models

Outline

Data Exploration

Feature Engineering

Baseline Models

Data Exploration

Data Exploration

	ip	time	method	status_code	path	response_length	user_agent	response_time
0	207.213.193.143	2021-05-12 05:06:00+04:30	Get	304	cdn/profiles/1026106239	0	Googlebot-Image/1.0	32.0
1	207.213.193.143	2021-05-12 05:06:00+04:30	Get	304	images/badge.png	0	Googlebot-Image/1.0	4.0
2	35.110.222.153	2021-05-12 05:06:00+04:30	Get	200	pages/630180847	52567	Mozilla/5.0 (Linux; Android 6.0.1; SAMSUNG SM-...	32.0
3	35.108.208.99	2021-05-12 05:06:00+04:30	Get	200	images/fav_icon2.ico	23531	Mozilla/5.0 (Linux; Android 6.0; CAM-L21) Appl...	20.0
4	35.110.222.153	2021-05-12 05:06:00+04:30	Get	200	images/sanjagh_logo_purpule5.png	4680	Mozilla/5.0 (Linux; Android 6.0.1; SAMSUNG SM-...	8.0
...	...	...	...	...	...	...	...	...
1260028	35.117.86.75	2021-05-12 15:08:59+04:30	Get	304	images/graystar_min.png	0	Mozilla/5.0 (Linux; Android 9; Redmi 7A) Apple...	4.0
1260029	35.117.86.75	2021-05-12 15:08:59+04:30	Get	304	fonts/sanjagh_icon_font_5.woff	0	Mozilla/5.0 (Linux; Android 9; Redmi 7A) Apple...	4.0
1260030	153.126.251.199	2021-05-12 15:08:59+04:30	Get	101	api/v2/connect/215865643	0	okhttp/3.12.1	60003.0
1260031	207.213.207.102	2021-05-12 15:08:59+04:30	Get	304	cdn/profiles/1289255230	0	Googlebot-Image/1.0	20.0
1260032	207.213.207.224	2021-05-12 15:09:00+04:30	Get	304	cdn/profiles/1536446365	0	Googlebot-Image/1.0	44.0

1260033 rows × 8 columns

Data Exploration

```
df.isna().sum()
```

Data Exploration

```
df.isna().sum()
```

```
ip                   18090
time                     0
method                   0
status_code              0
path                     0
response_length          0
user_agent               0
response_time        19808
dtype: int64
```

Data Exploration

```
df[df[“ip”].isna()]
```

Data Exploration

```
df[df["ip"].isna()]
```

	ip	time	method	status_code	path	response_length	user_agent	response_time
25	NaN	2021-05-12 05:06:01+04:30	Get	301		169	kube-probe/1.21	NaN
85	NaN	2021-05-12 05:06:03+04:30	Get	301		169	kube-probe/1.21	NaN
145	NaN	2021-05-12 05:06:05+04:30	Get	301		169	kube-probe/1.21	NaN
175	NaN	2021-05-12 05:06:07+04:30	Get	301		169	kube-probe/1.21	NaN
215	NaN	2021-05-12 05:06:09+04:30	Get	301		169	kube-probe/1.21	NaN
...	...	...	...	...	...	...	...	...
1259777	NaN	2021-05-12 15:08:51+04:30	Get	301		169	kube-probe/1.21	NaN
1259831	NaN	2021-05-12 15:08:53+04:30	Get	301		169	kube-probe/1.21	NaN
1259900	NaN	2021-05-12 15:08:55+04:30	Get	301		169	kube-probe/1.21	NaN
1259949	NaN	2021-05-12 15:08:57+04:30	Get	301		169	kube-probe/1.21	NaN
1260005	NaN	2021-05-12 15:08:59+04:30	Get	301		169	kube-probe/1.21	NaN

18090 rows × 8 columns

Data Exploration

```
df[df["ip"].isna()]
```

	ip	time	method	status_code	path	response_length	user_agent	response_time
25	NaN	2021-05-12 05:06:01+04:30	Get	301		169	kube-probe/1.21	NaN
85	NaN	2021-05-12 05:06:03+04:30	Get	301		169	kube-probe/1.21	NaN
145	NaN	2021-05-12 05:06:05+04:30	Get	301		169	kube-probe/1.21	NaN
175	NaN	2021-05-12 05:06:07+04:30	Get	301		169	kube-probe/1.21	NaN
215	NaN	2021-05-12 05:06:09+04:30	Get	301		169	kube-probe/1.21	NaN
...	...	...	...	...	...	...	...	...
1259777	NaN	2021-05-12 15:08:51+04:30	Get	301		169	kube-probe/1.21	NaN
1259831	NaN	2021-05-12 15:08:53+04:30	Get	301		169	kube-probe/1.21	NaN
1259900	NaN	2021-05-12 15:08:55+04:30	Get	301		169	kube-probe/1.21	NaN
1259949	NaN	2021-05-12 15:08:57+04:30	Get	301		169	kube-probe/1.21	NaN
1260005	NaN	2021-05-12 15:08:59+04:30	Get	301		169	kube-probe/1.21	NaN

18090 rows × 8 columns

Data Exploration

```
df[df[“response_time”].isna()]
```

Data Exploration

```
df[df["response_time"].isna()]
```

	ip	time	method	status_code	path	response_length	user_agent	response_time
776	20.62.177.11	2021-05-12 05:06:31+04:30	Get	200	pros/1993352776	53479	Mozilla/5.0 (compatible; SemrushBot/7~bl; +htt...	NaN
2010	20.62.177.60	2021-05-12 05:07:27+04:30	Get	200	pros/1797822247	55330	Mozilla/5.0 (compatible; SemrushBot/7~bl; +htt...	NaN
2708	20.62.177.133	2021-05-12 05:08:04+04:30	Get	200	pros/763244865	20947	Mozilla/5.0 (compatible; SemrushBot/7~bl; +htt...	NaN
2866	207.213.193.118	2021-05-12 05:08:18+04:30	Get	301	pages/1939232229	169	Mozilla/5.0 (Linux; Android 6.0.1; Nexus 5X Bu...	NaN
3468	20.62.177.4	2021-05-12 05:08:49+04:30	Get	200	pros/2084824811	37060	Mozilla/5.0 (compatible; SemrushBot/7~bl; +htt...	NaN
...	...	...	...	...	...	...	...	...
1257191	20.62.177.11	2021-05-12 15:07:34+04:30	Get	200	pros/1644096504	24540	Mozilla/5.0 (compatible; SemrushBot/7~bl; +htt...	NaN
1257984	20.62.177.11	2021-05-12 15:07:59+04:30	Get	200	pros/743056796	36129	Mozilla/5.0 (compatible; SemrushBot/7~bl; +htt...	NaN
1258077	20.62.177.161	2021-05-12 15:08:02+04:30	Get	200	pros/1177343248	51334	Mozilla/5.0 (compatible; SemrushBot/7~bl; +htt...	NaN
1258454	207.213.207.17	2021-05-12 15:08:12+04:30	Get	301	services/1404674245	169	Mozilla/5.0 (Linux; Android 6.0.1; Nexus 5X Bu...	NaN
1259451	153.126.111.1	2021-05-12 15:08:39+04:30	Get	200	pros/884190773	58833	Mozilla/5.0 (Linux; Android 4.4.4; SM-J100H) A...	NaN

1718 rows × 8 columns

Data Exploration

```
df.isna().sum()
```

Data Exploration

```
df.isna().sum()
```

```
ip                  0
time                0
method              0
status_code         0
path                0
response_length     0
user_agent          0
response_time       0
dtype: int64
```

Data Exploration

```
df[df[“path”] == “”]
```

Data Exploration

```
df[df["path"] == ""]
```

	ip	time	method	status_code	path	response_length	user_agent	response_time
155	20.117.146.75	2021-05-12 05:06:06+04:30	Get	307		0	Go-http-client/2.0	4.0
360	20.92.247.170	2021-05-12 05:06:13+04:30	Get	307		0	Go-http-client/2.0	8.0
622	76.212.164.3	2021-05-12 05:06:24+04:30	Get	307		0	Go-http-client/2.0	0.0
645	93.113.99.115	2021-05-12 05:06:25+04:30	Get	307		0	Go-http-client/2.0	12.0
828	36.67.23.210	2021-05-12 05:06:36+04:30	Get	307		0	Go-http-client/2.0	12.0
...	...	...	...	...	...	...	...	...
1258999	20.92.247.170	2021-05-12 15:08:25+04:30	Get	307		0	Go-http-client/2.0	4.0
1259139	93.113.99.115	2021-05-12 15:08:28+04:30	Get	307		0	Go-http-client/2.0	8.0
1259154	238.129.28.160	2021-05-12 15:08:29+04:30	Get	307		0	Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.3...	4.0
1259504	60.148.0.167	2021-05-12 15:08:41+04:30	Get	307		0	Go-http-client/2.0	4.0
1259973	186.236.39.213	2021-05-12 15:08:58+04:30	Head	405		0	Mozilla/5.0+(compatible; UptimeRobot/2.0; http...	16.0

5975 rows × 8 columns

Data Exploration

```
df[df["path"] == ""]
```

	ip	time	method	status_code	path	response_length	user_agent	response_time
155	20.117.146.75	2021-05-12 05:06:06+04:30	Get	307		0	Go-http-client/2.0	4.0
360	20.92.247.170	2021-05-12 05:06:13+04:30	Get	307		0	Go-http-client/2.0	8.0
622	76.212.164.3	2021-05-12 05:06:24+04:30	Get	307		0	Go-http-client/2.0	0.0
645	93.113.99.115	2021-05-12 05:06:25+04:30	Get	307		0	Go-http-client/2.0	12.0
828	36.67.23.210	2021-05-12 05:06:36+04:30	Get	307		0	Go-http-client/2.0	12.0
...	...	...	...	...	...	...	...	...
1258999	20.92.247.170	2021-05-12 15:08:25+04:30	Get	307		0	Go-http-client/2.0	4.0
1259139	93.113.99.115	2021-05-12 15:08:28+04:30	Get	307		0	Go-http-client/2.0	8.0
1259154	238.129.28.160	2021-05-12 15:08:29+04:30	Get	307		0	Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.3...	4.0
1259504	60.148.0.167	2021-05-12 15:08:41+04:30	Get	307		0	Go-http-client/2.0	4.0
1259973	186.236.39.213	2021-05-12 15:08:58+04:30	Head	405		0	Mozilla/5.0+(compatible; UptimeRobot/2.0; http...	16.0

5975 rows × 8 columns

Data Exploration

```
df[df["path"] == ""]
```

	ip	time	method	status_code	path	response_length	user_agent	response_time
155	20.117.146.75	2021-05-12 05:06:06+04:30	Get	307		0	Go-http-client/2.0	4.0
360	20.92.247.170	2021-05-12 05:06:13+04:30	Get	307		0	Go-http-client/2.0	8.0
622	76.212.164.3	2021-05-12 05:06:24+04:30	Get	307		0	Go-http-client/2.0	0.0
645	93.113.99.115	2021-05-12 05:06:25+04:30	Get	307		0	Go-http-client/2.0	12.0
828	36.67.23.210	2021-05-12 05:06:36+04:30	Get	307		0	Go-http-client/2.0	12.0
...	...	...	...	...	...	...	...	...
1258999	20.92.247.170	2021-05-12 15:08:25+04:30	Get	307		0	Go-http-client/2.0	4.0
1259139	93.113.99.115	2021-05-12 15:08:28+04:30	Get	307		0	Go-http-client/2.0	8.0
1259154	238.129.28.160	2021-05-12 15:08:29+04:30	Get	307		0	Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.3...	4.0
1259504	60.148.0.167	2021-05-12 15:08:41+04:30	Get	307		0	Go-http-client/2.0	4.0
1259973	186.236.39.213	2021-05-12 15:08:58+04:30	Head	405		0	Mozilla/5.0+(compatible; UptimeRobot/2.0; http...	16.0

5975 rows × 8 columns

Data Exploration

```
df[df[“path”] == “”].status_code.unique()
```

Data Exploration

```
df[df[“path”] == “”].status_code.unique()
```

```
array([307, 405, 301, 499])
```

Data Exploration

```
df[df[“path”] == “”].status_code.unique()
```

```
array([307, 405, 301, 499])
```

307: **Temporary Redirect** redirect status response code indicates that the resource requested has been temporarily moved to the URL given by the Location headers.

301: **Moved Permanently** redirect status response code indicates that the resource requested has been definitively moved to the URL given by the Location headers.

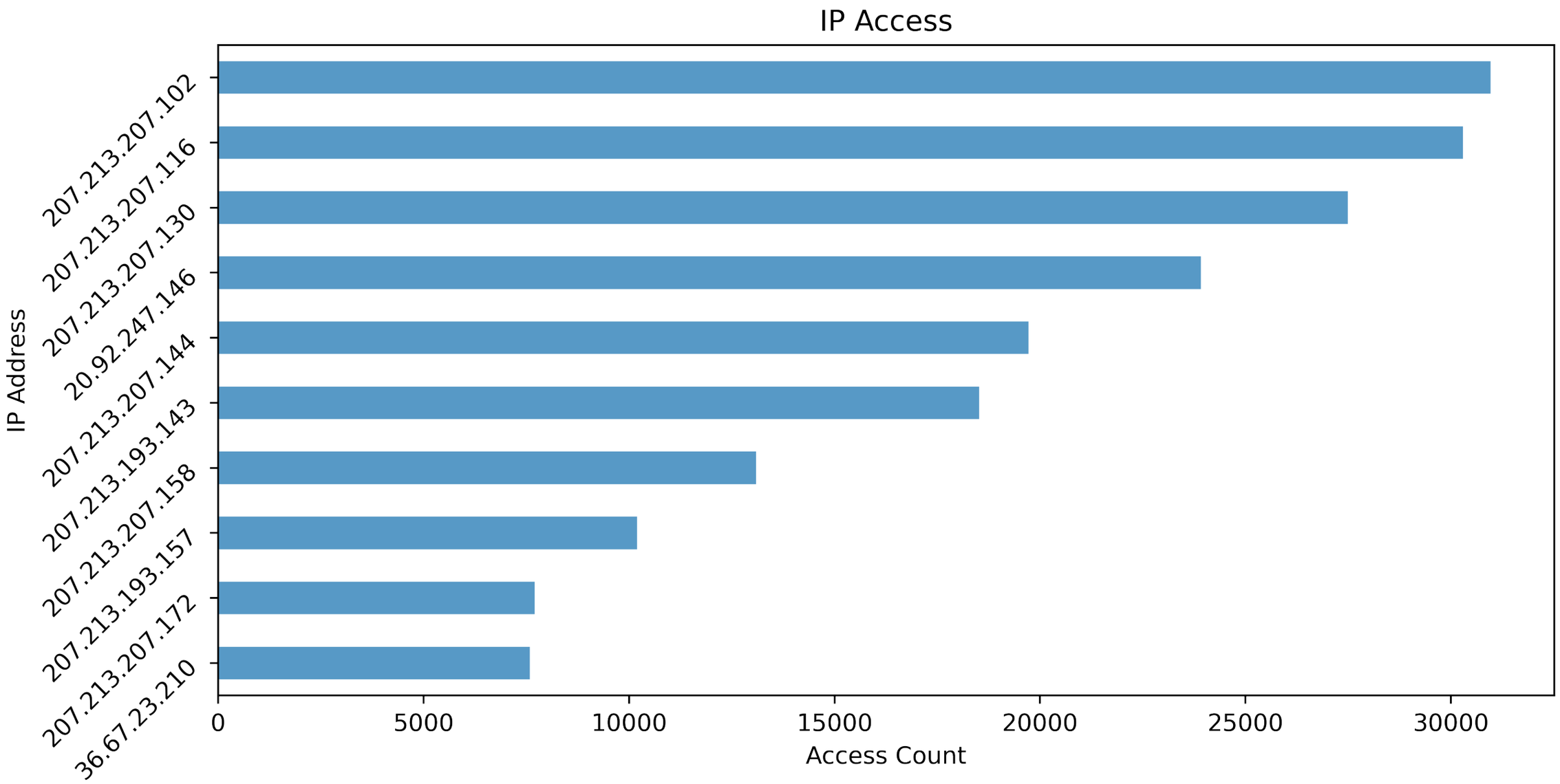
405: **Method Not Allowed** response status code indicates that the request method is known by the server but is not supported by the target resource.

499: **Error code** simply means that the client shut off in the middle of processing the request through the server.

The Most Commons

Data Exploration

1. The most visited IP addresses



Data Exploration

1. The most visited IP addresses

Data Exploration

1. The most visited IP addresses

	ip	time	method	status_code	path	response_length	user_agent	response_time
0	207.213.207.102	2021-05-12 06:25:56+04:30	Get	304	cdn/articles/1148001967	0	Googlebot-Image/1.0	16.0
30968	207.213.207.116	2021-05-12 07:40:46+04:30	Get	304	cdn/profiles/1074674108	0	Googlebot-Image/1.0	16.0
61258	207.213.207.130	2021-05-12 09:25:34+04:30	Get	200	amp/price/1313296747	125767	Mozilla/5.0 (Linux; Android 6.0.1; Nexus 5X Bu...	28.0
88748	20.92.247.146	2021-05-12 05:08:46+04:30	Get	200	js/profession.c67de06df71c34fc126d.js	107970	sentry/21.4.1 (https://sentry.io)	16.0
112660	207.213.207.144	2021-05-12 09:02:56+04:30	Get	200	amp/price/252451961	121639	Mozilla/5.0 (Linux; Android 6.0.1; Nexus 5X Bu...	28.0
132378	207.213.193.143	2021-05-12 05:06:00+04:30	Get	304	cdn/profiles/1026106239	0	Googlebot-Image/1.0	32.0
150901	207.213.207.158	2021-05-12 09:25:55+04:30	Get	304	cdn/articles/2121333045	0	Googlebot-Image/1.0	16.0
163992	207.213.193.157	2021-05-12 05:07:04+04:30	Get	200	amp/blog/article/1197238235	101087	Mozilla/5.0 (Linux; Android 6.0.1; Nexus 5X Bu...	144.0
174189	207.213.207.172	2021-05-12 09:28:50+04:30	Get	304	cdn/pro_photo_gallery/1540103160	0	Googlebot-Image/1.0	28.0
181890	36.67.23.210	2021-05-12 05:06:00+04:30	Head	200	877499224	0	Go-http-client/2.0	28.0

Data Exploration

1. The most visited IP addresses

	ip	time	method	status_code	path	response_length	user_agent	response_time
0	207.213.207.102	2021-05-12 06:25:56+04:30	Get	304	cdn/articles/1148001967	0	Googlebot-Image/1.0	16.0
30968	207.213.207.116	2021-05-12 07:40:46+04:30	Get	304	cdn/profiles/1074674108	0	Googlebot-Image/1.0	16.0
61258	207.213.207.130	2021-05-12 09:25:34+04:30	Get	200	amp/price/1313296747	125767	Mozilla/5.0 (Linux; Android 6.0.1; Nexus 5X Bu...	28.0
88748	20.92.247.146	2021-05-12 05:08:46+04:30	Get	200	js/profession.c67de06df71c34fc126d.js	107970	sentry/21.4.1 (https://sentry.io)	16.0
112660	207.213.207.144	2021-05-12 09:02:56+04:30	Get	200	amp/price/252451961	121639	Mozilla/5.0 (Linux; Android 6.0.1; Nexus 5X Bu...	28.0
132378	207.213.193.143	2021-05-12 05:06:00+04:30	Get	304	cdn/profiles/1026106239	0	Googlebot-Image/1.0	32.0
150901	207.213.207.158	2021-05-12 09:25:55+04:30	Get	304	cdn/articles/2121333045	0	Googlebot-Image/1.0	16.0
163992	207.213.193.157	2021-05-12 05:07:04+04:30	Get	200	amp/blog/article/1197238235	101087	Mozilla/5.0 (Linux; Android 6.0.1; Nexus 5X Bu...	144.0
174189	207.213.207.172	2021-05-12 09:28:50+04:30	Get	304	cdn/pro_photo_gallery/1540103160	0	Googlebot-Image/1.0	28.0
181890	36.67.23.210	2021-05-12 05:06:00+04:30	Head	200	877499224	0	Go-http-client/2.0	28.0

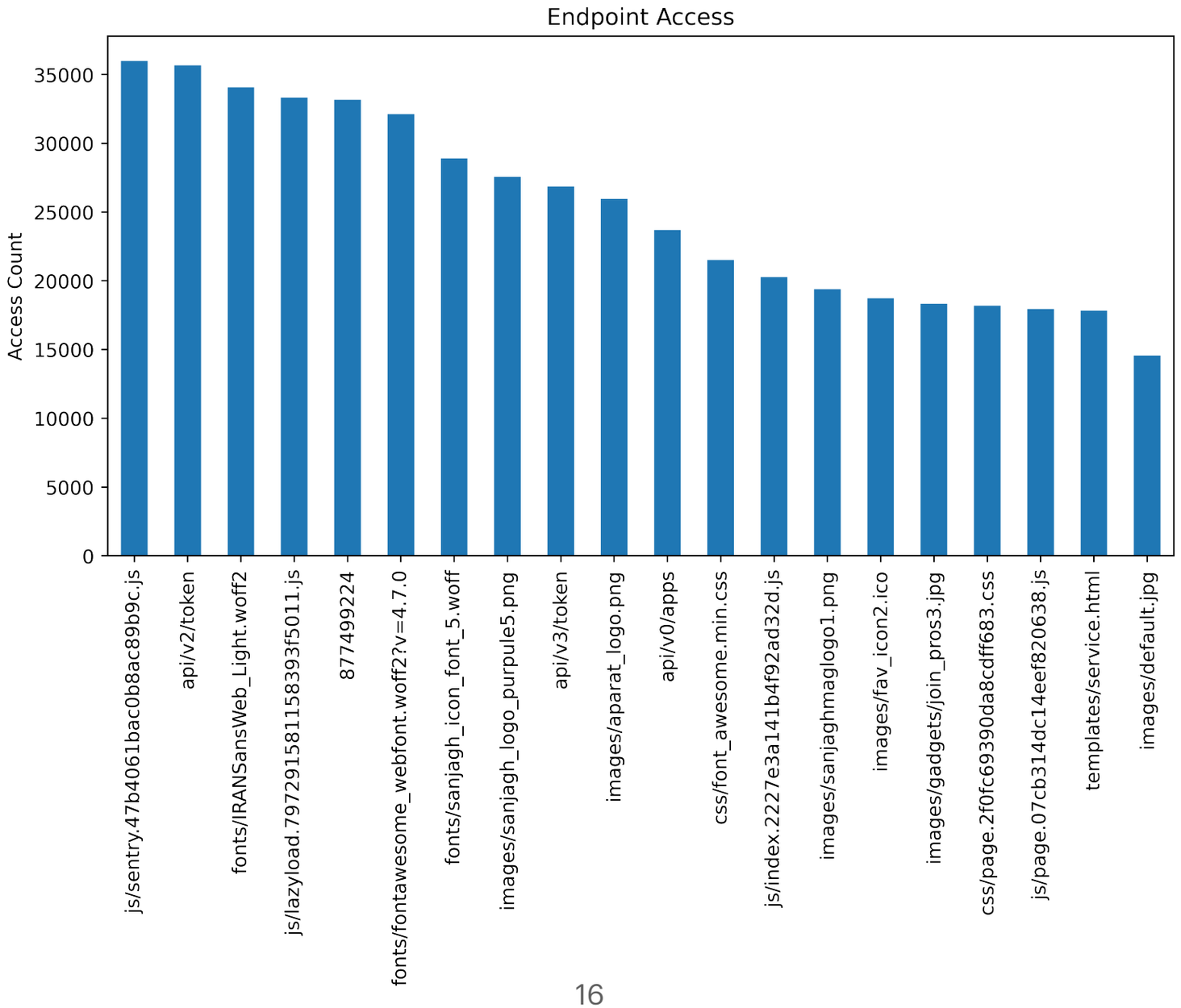
Data Exploration

```
len(df[df[“ip”] == “207.213.207.102”].user_agent.unique())
```

6

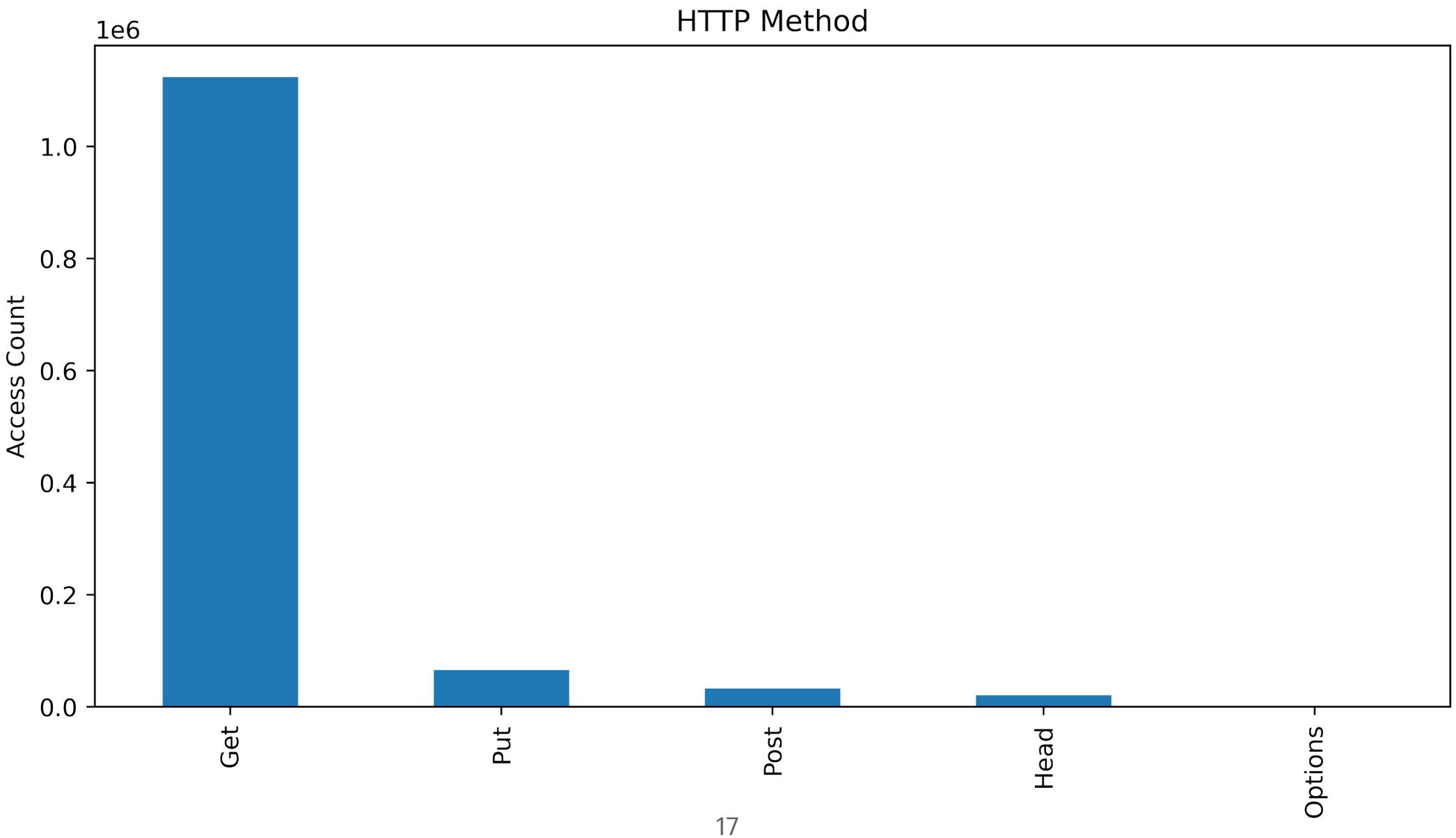
Data Exploration

2. The most requested endpoints



Data Exploration

3. The most common HTTP methods



Data Exploration

4. The most common user agents

user-agents 2.2.0

```
pip install user-agents
```

Latest version

Released: Aug 23, 2020

A library to identify devices (phones, tablets) and their capabilities by parsing browser user agent strings.

Navigation

- Project description
- Release history
- Download files

Project links

- Homepage

Statistics

GitHub statistics:

- **Stars:** 1,168
- **Forks:** 186
- **Open issues/PRs:** 36

Project description

Python User Agents

`user_agents` is a Python library that provides an easy way to identify/detect devices like mobile phones, tablets and their capabilities by parsing (browser/HTTP) user agent strings. The goal is to reliably detect whether:

- User agent is a mobile, tablet or PC based device
- User agent has touch capabilities (has touch screen)

`user_agents` relies on the excellent ua-parser to do the actual parsing of the raw user agent string.

Installation

build passing

`user-agents` is hosted on PyPI and can be installed as such:

```
pip install pyyaml ua-parser user-agents
```

Meta

License: MIT License (MIT)

Author: Selwin Ong

Maintainers

selwin

Classifiers

Development Status

- 5 - Production/Stable

Environment

- Web Environment

Intended Audience

- Developers

License

- OSI Approved :: MIT License

Operating System

- OS Independent

Programming Language

- Python
- Python :: 2
- Python :: 2.7
- Python :: 3
- Python :: 3.4
- Python :: 3.5
- Python :: 3.6
- Python :: 3.7

Usage

Various basic information that can help you identify visitors can be accessed `browser`, `device` and `os` attributes. For example:

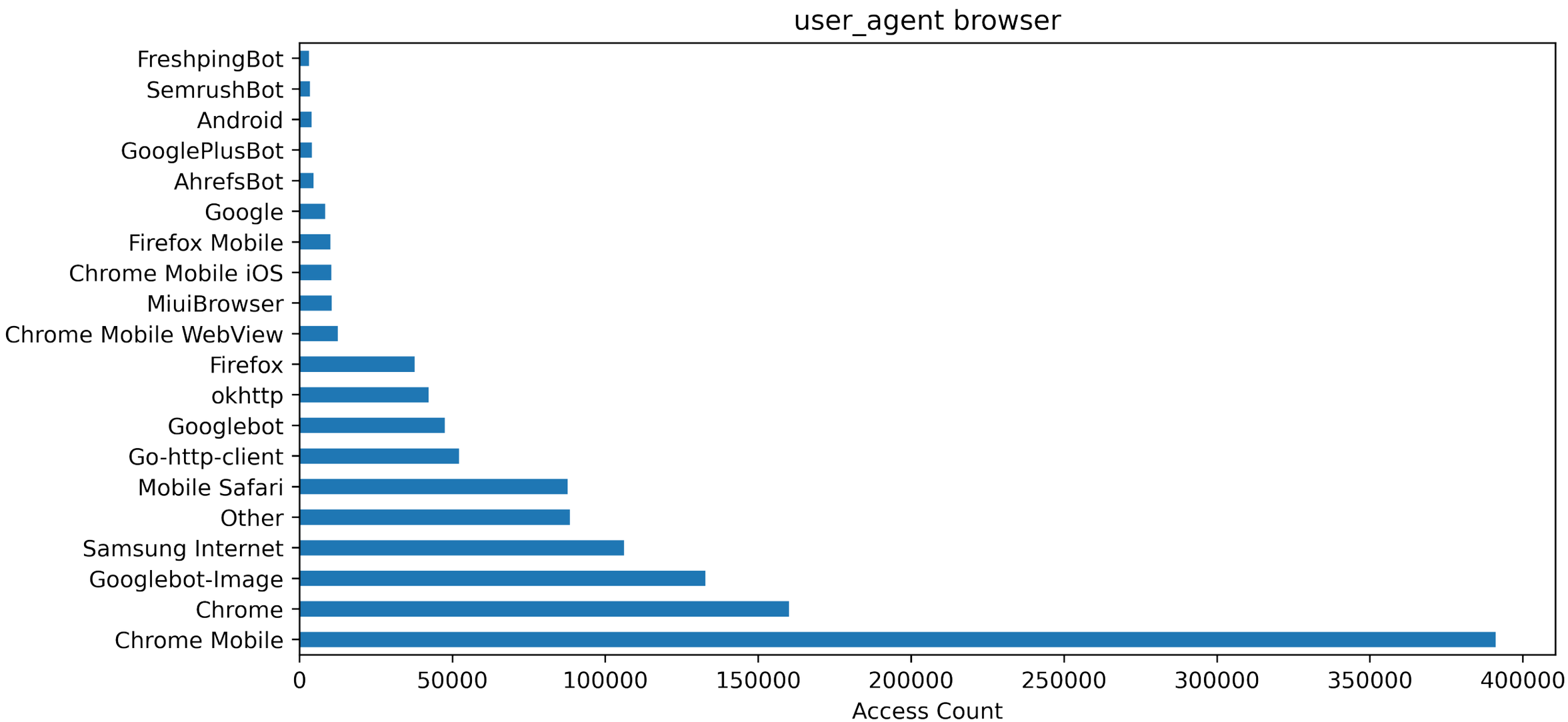
```
from user_agents import parse

# iPhone's user agent string
ua_string = 'Mozilla/5.0 (iPhone; CPU iPhone OS 5_1 like Mac OS X) AppleWebKit/534.46 (KHTM
user_agent = parse(ua_string)

# Accessing user agent's browser attributes
user_agent.browser  # returns Browser(family=u'Mobile Safari', version=(5, 1), version_stri
user_agent.browser.family  # returns 'Mobile Safari'
user_agent.browser.version  # returns (5, 1)
user_agent.browser.version_string   # returns '5.1'

# Accessing user agent's operating system properties
user_agent.os  # returns OperatingSystem(family=u'iOS', version=(5, 1), version_string='5.1
user_agent.os.family  # returns 'iOS'
user_agent.os.version  # returns (5, 1)
user_agent.os.version_string  # returns '5.1'

# Accessing user agent's device properties
user_agent.device  # returns Device(family=u'iPhone', brand=u'Apple', model=u'iPhone')
user_agent.device.family  # returns 'iPhone'
user_agent.device.brand # returns 'Apple'
user_agent.device.model # returns 'iPhone'

# Viewing a pretty string version
str(user_agent) # returns "iPhone / iOS 5.1 / Mobile Safari 5.1"
```

`user_agents` also expose a few other more "sophisticated" attributes that are derived from one or more basic attributes defined above. As for now, these attributes should correctly identify popular platforms/devices, pull requests to support smaller ones are always welcome.

Currently these attributes are supported:

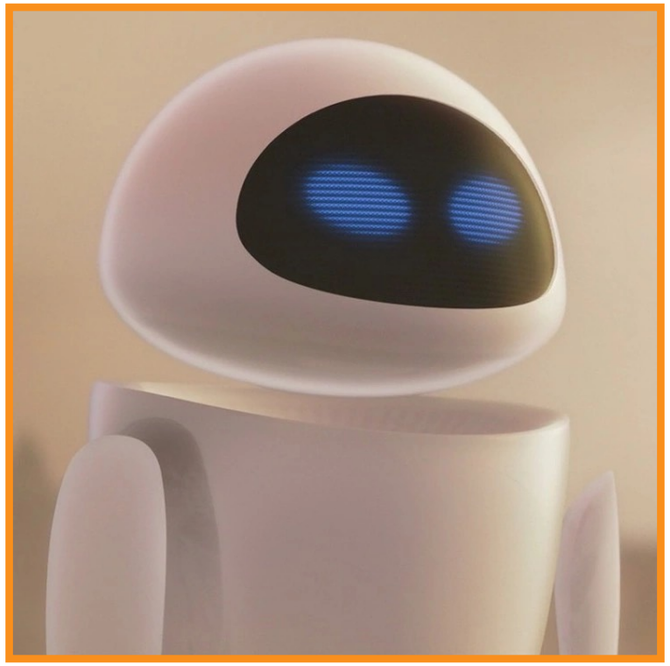
Data Exploration

4. The most common user agents



Outline

Data Exploration



Feature Engineering

Baseline Models

Feature Engineering

What We Have,

Vs.

What We Want to Do.

Feature Engineering

Session Identification:

Simply it can be performed by grouping all HTTP requests by their **IP**s and **user agents.**

Feature Engineering

Session Identification:

Simply it can be performed by grouping all HTTP requests by their **IP**s and **user agents.**

Types of Features:

Per Request

Per Session

Features Per Request

Feature Engineering

```
df.columns
```

Feature Engineering

```
df.columns
```

```
Index(['ip', 'time', 'method', 'status_code', 'path', 'response_length',
       'user_agent', 'response_time', 'browser', 'os', 'is_bot', 'is_pc',
       'path_count_normalized', 'browser_norm', 'os_norm'],
      dtype='object')
```

Features Per Session

Feature Engineering

1. Click rate – Higher click rate can only be achieved by an automated script.

```
user_df.sort_values(“request_count”, ascending=False)
```

Feature Engineering

1. Click rate – Higher click rate can only be achieved by an automated script.

```
user_df.sort_values("request_count", ascending=False)
```

ip	user_agent	requests_count
20.92.247.146	**sentry/21.4.1 (https://sentry.io)**	23912
207.213.207.102	**Googlebot-Image/1.0**	23627
207.213.207.116	**Googlebot-Image/1.0**	23380
207.213.207.130	**Googlebot-Image/1.0**	21494
207.213.207.144	**Googlebot-Image/1.0**	15363
...	**...**	...
35.202.69.199	**Mozilla/5.0 (compatible; heritrix/3.4.0-20200304 +https://zarebin.ir/)**	1
35.202.76.221	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	1
	okhttp/3.12.1	1
35.202.77.168	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	1
99.96.90.10	**Mozilla/5.0 (iPhone; CPU iPhone OS 14_6 like Mac OS X) AppleWebKit/605.1.15 (KHTML, like Gecko) Version/14.1.1 Mobile/15E148 Safari/604.1**	1

51123 rows × 1 columns

Feature Engineering

2. STD of path’s depth – deeper requests usually indicates a human user.

```
user_df[user_df['requests_count'] > 5].sort_values(‘path_length_std', ascending=True)
```

Feature Engineering

2. STD of path’s depth – deeper requests usually indicates a human user.

```
user_df[user_df['requests_count'] > 5].sort_values(‘path_length_std', ascending=True)
```

ip	user_agent	requests_count	path_length_std
1.81.122.235	**Mozilla/5.0 (iPhone; CPU iPhone OS 14_6 like Mac OS X) AppleWebKit/605.1.15 (KHTML, like Gecko) Version/14.1.1 Mobile/15E148 Safari/604.1**	7	0.000000
35.47.50.38	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	11	0.000000
35.47.49.41	**okhttp/3.12.1**	14	0.000000
35.47.49.38	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	25	0.000000
35.47.45.107	**okhttp/3.12.1**	6	0.000000
...	**...**	...	...
35.202.60.172	**Mozilla/5.0 (Linux; Android 10; SAMSUNG SM-A307FN) AppleWebKit/537.36 (KHTML, like Gecko) SamsungBrowser/12.1 Chrome/79.0.3945.136 Mobile Safari/537.36**	6	1.366260
35.109.94.99	**Mozilla/5.0 (iPhone; CPU iPhone OS 10_2 like Mac OS X) AppleWebKit/602.1.50 (KHTML, like Gecko) GSA/68.0.234683655 Mobile/14C92 Safari/602.1**	6	1.366260
35.202.142.86	**Mozilla/5.0 (Linux; Android 10; SAMSUNG SM-A115F) AppleWebKit/537.36 (KHTML, like Gecko) SamsungBrowser/13.2 Chrome/83.0.4103.106 Mobile Safari/537.36**	6	1.366260
127.227.58.62	**Mozilla/5.0 (Linux; Android 10; SAMSUNG SM-A107F) AppleWebKit/537.36 (KHTML, like Gecko) SamsungBrowser/14.0 Chrome/87.0.4280.141 Mobile Safari/537.36**	7	1.463850
14.226.145.71	**Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.101 Safari/537.36**	15	2.153624

29956 rows × 2 columns

Feature Engineering

3. Percentage of 4xx status codes – Usually higher for crawlers as there is higher chances of hitting an outdated or deleted pages.

```
user_df[user_df['requests_count'] > 5].sort_values(['4xx_percentage(%)', 'requests_count'
                                                    ,ascending=True)
```

Feature Engineering

3. Percentage of 4xx status codes – Usually higher for crawlers as there is higher chances of hitting an outdated or deleted pages.

```
user_df[user_df['requests_count'] > 5].sort_values(['4xx_percentage(%)', 'requests_count'
                                                    ,ascending=True)
```

ip	user_agent	requests_count	path_length_std	3xx_percentage(%)	4xx_percentage(%)
35.124.193.182	**Dalvik/2.1.0 (Linux; U; Android 10; SM-A115F Build/QP1A.190711.020)**	104	0.00000	0.0	100.0
153.126.209.239	**Go-http-client/1.1**	35	0.00000	0.0	100.0
35.244.120.44	**Go-http-client/1.1**	18	0.00000	0.0	100.0
35.232.97.81	**okhttp/2.5.0**	9	0.00000	0.0	100.0
35.132.136.207	**MobileSafari/604.1 CFNetwork/1240.0.4 Darwin/20.5.0**	8	0.00000	0.0	100.0
...	**...**	...	...	...	...
92.144.239.236	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	6	0.83666	0.0	0.0
92.144.77.249	**Mozilla/5.0 (X11; Linux x86_64) app_version: 735**	6	0.00000	0.0	0.0
92.239.17.78	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	6	0.83666	0.0	0.0
92.239.237.42	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	6	0.83666	0.0	0.0
92.51.185.100	**Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.77 Safari/537.36**	6	0.00000	0.0	0.0

29956 rows × 4 columns

Feature Engineering

4. Percentage of 3xx status codes.

```
user_df[user_df['requests_count'] > 5].sort_values(['3xx_percentage(%)', 'requests_count'
                                                    ,ascending=True)
```

Feature Engineering

4. Percentage of 3xx status codes.

```
user_df[user_df['requests_count'] > 5].sort_values(['3xx_percentage(%)', 'requests_count'
                                                    ,ascending=True)
```

ip	user_agent	requests_count	path_length_std	3xx_percentage(%)	4xx_percentage(%)
207.213.193.213	**Googlebot-Image/1.0**	71	0.257679	100.0	0.0
217.98.85.154	**Mozilla/5.0 (Linux; Android 10; SM-A105F) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.88 Mobile Safari/537.36**	54	0.292582	100.0	0.0
14.240.9.74	**Mozilla/5.0 (Linux; Android 8.1.0; DUB-LX1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/88.0.4324.181 Mobile Safari/537.36**	50	0.274048	100.0	0.0
35.108.36.67	**Mozilla/5.0 (Linux; Android 8.1.0; SM-G610F) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/88.0.4324.51 Mobile Safari/537.36**	39	0.269953	100.0	0.0
4.115.196.73	**Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.77 Safari/537.36**	39	0.269953	100.0	0.0
...	**...**	...	...	...	...
92.144.239.236	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	6	0.836660	0.0	0.0
92.144.77.249	**Mozilla/5.0 (X11; Linux x86_64) app_version: 735**	6	0.000000	0.0	0.0
92.239.17.78	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	6	0.836660	0.0	0.0
92.239.237.42	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	6	0.836660	0.0	0.0
92.51.185.100	**Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.77 Safari/537.36**	6	0.000000	0.0	0.0

29956 rows × 4 columns

Feature Engineering

4. Percentage of HTTP HEAD requests – most crawlers, in order to reduce the amount of data requested from a site, employ the HEAD method when requesting a page.

```
user_df[user_df['requests_count'] > 5].sort_values(['HEAD_count(%)', 'requests_count'
                                                    ,ascending=True)
```

Feature Engineering

4. Percentage of HTTP HEAD requests – most crawlers, in order to reduce the amount of data requested from a site, employ the HEAD method when requesting a page.

```
user_df[user_df['requests_count'] > 5].sort_values(['HEAD_count(%)', 'requests_count'
                                                    ,ascending=True)
```

ip	user_agent	requests_count	path_length_std	3xx_percentage(%)	4xx_percentage(%)	HEAD_count(%)
20.163.161.41	**Mozilla/5.0 (iPhone; CPU iPhone OS 7_0 like Mac OS X; en-us) AppleWebKit/537.51.1 (KHTML, like Gecko) Version/7.0 Mobile/11A465 Safari/9537.53**	37	0.538321	0.000000	0.000000	72.972973
36.67.23.210	**Go-http-client/2.0**	7582	0.475879	9.773147	1.609074	45.805856
60.148.0.167	**Go-http-client/2.0**	7351	0.479779	10.053054	1.714053	43.980411
20.92.247.170	**Go-http-client/2.0**	7273	0.482874	10.325863	0.000000	42.334662
76.212.164.3	**Go-http-client/2.0**	6549	0.487955	9.406016	1.878149	42.113300
...	**...**	...	...	...	...	...
92.144.239.236	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	6	0.836660	0.000000	0.000000	0.000000
92.144.77.249	**Mozilla/5.0 (X11; Linux x86_64) app_version: 735**	6	0.000000	0.000000	0.000000	0.000000
92.239.17.78	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	6	0.836660	0.000000	0.000000	0.000000
92.239.237.42	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	6	0.836660	0.000000	0.000000	0.000000
92.51.185.100	**Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.77 Safari/537.36**	6	0.000000	0.000000	0.000000	0.000000

29956 rows × 5 columns

Feature Engineering

5. Percentage of image requests – web crawlers usually ignore images.

```
user_df[user_df['requests_count'] > 20].sort_values(['image_count(%)', 'requests_count'
                                                    ,ascending=True).head(10)
```

Feature Engineering

5. Percentage of image requests – web crawlers usually ignore images.

```
user_df[user_df['requests_count'] > 20].sort_values(['image_count(%)', 'requests_count'
                                                    ,ascending=True).head(10)
```

ip	user_agent	requests_count	path_length_std	3xx_percentage(%)	4xx_percentage(%)	HEAD_count(%)	image_count(%)
102.29.29.19	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	21	0.000000	0.0	0.0	0.0	0.0
113.11.38.30	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	21	0.749603	0.0	0.0	0.0	0.0
113.111.195.219	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	21	0.000000	0.0	0.0	0.0	0.0
113.118.175.102	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	21	0.810643	0.0	0.0	0.0	0.0
113.118.45.106	**okhttp/3.12.1**	21	0.436436	0.0	0.0	0.0	0.0
113.118.90.28	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	21	0.810643	0.0	0.0	0.0	0.0
113.118.96.94	**okhttp/3.12.1**	21	0.358569	0.0	0.0	0.0	0.0
113.74.79.241	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	21	0.810643	0.0	0.0	0.0	0.0
113.97.91.168	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	21	0.643650	0.0	0.0	0.0	0.0
127.166.17.33	**okhttp/3.12.1**	21	0.218218	0.0	0.0	0.0	0.0

Feature Engineering

6. Average & sum of response_length & response_time –

Human users retrieve info from the web via browser, so it forces the user’s session to request additional resource automatically.

```
user_df[(user_df['requests_count'] > 5) & (user_df['3xx_percentage(%)'] < 20)] \
                              .sort_values(['mean_response_time', 'mean_response_length'],
                              ascending=True)
```

Feature Engineering

6. Average & sum of response_length & response_time –

Human users retrieve info from the web via browser, so it forces the user’s session to request additional resource automatically.

```
user_df[(user_df['requests_count'] > 5) & (user_df['3xx_percentage(%)'] < 20)] \
                                .sort_values(['mean_response_time', 'mean_response_length'],
                                ascending=True)
```

		requests_count	path_length_std	3xx_percentage(%)	4xx_percentage(%)	HEAD_count(%)	image_count(%)	total_resp
ip	user_agent							
29.240.244.96	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_1) AppleWebKit/601.2.4 (KHTML, like Gecko) Version/9.0.1 Safari/601.2.4 facebookexternalhit/1.1 Facebot Twitterbot/1.0	8	0.462910	0.00	75.000000	0.0	25.000000	
93.113.11.166	Mozilla/5.0 (iPhone; CPU iPhone OS 14_4 like Mac OS X) AppleWebKit/605.1.15 (KHTML, like Gecko) Version/14.0.3 Mobile/15E148 Safari/604.1	7	0.377964	0.00	0.000000	0.0	100.000000	
155.114.254.242	Mozilla/5.0 (iPhone; CPU iPhone OS 13_5_1 like Mac OS X) AppleWebKit/605.1.15 (KHTML, like Gecko) Version/13.1.1 Mobile/15E148 Safari/604.1	16	0.250000	6.25	0.000000	0.0	43.750000	

Feature Engineering

7. Set the user agent attributes

browser	os	is_bot	is_pc
Mobile Safari	iOS	False	False
Mobile Safari	iOS	False	False
Mobile Safari	iOS	False	False
Samsung Internet	Android	False	False

The last,

but

not the least.

Feature Engineering

8. Average of time between requests.

```
user_df.sort_values(['avg_time_diff'], ascending=True)
```

Feature Engineering

8. Average of time between requests.

```
user_df.sort_values(['avg_time_diff'], ascending=True)
```

ip	user_agent	requests_count	path_length_std	3xx_percentage(%)	4xx_percentage(%)	HEAD_count(%)	image_count(%)	total_response_
20.92.247.146	**sentry/21.4.1 (https://sentry.io)**	23912	0.015839	0.000000	0.012546	0.0	0.000000	219622
207.213.207.102	**Googlebot-Image/1.0**	23627	0.215708	99.297414	0.004232	0.0	5.091632	66
207.213.207.116	**Googlebot-Image/1.0**	23380	0.215150	99.328486	0.000000	0.0	5.038494	63
207.213.207.130	**Googlebot-Image/1.0**	21494	0.214493	99.390528	0.004652	0.0	5.038615	66
207.213.193.143	**Googlebot-Image/1.0**	13320	0.337257	98.791291	0.000000	0.0	13.250751	60
...	**...**	...	...	...	...	...	...	
148.197.248.86	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	5	0.000000	0.000000	0.000000	0.0	0.000000	
123.4.15.246	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	5	0.894427	0.000000	0.000000	0.0	0.000000	
217.49.61.35	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	5	0.894427	0.000000	0.000000	0.0	0.000000	
35.96.149.43	**Mozilla/5.0 (X11; Linux x86_64) app_version: 580**	5	0.894427	0.000000	0.000000	0.0	0.000000	
35.96.121.32	**Mozilla/5.0 (X11; Linux x86_64) app_version: 581**	5	0.894427	0.000000	0.000000	0.0	0.000000	

31541 rows × 16 columns

Outline

Data Exploration

Feature Engineering

Baseline Models

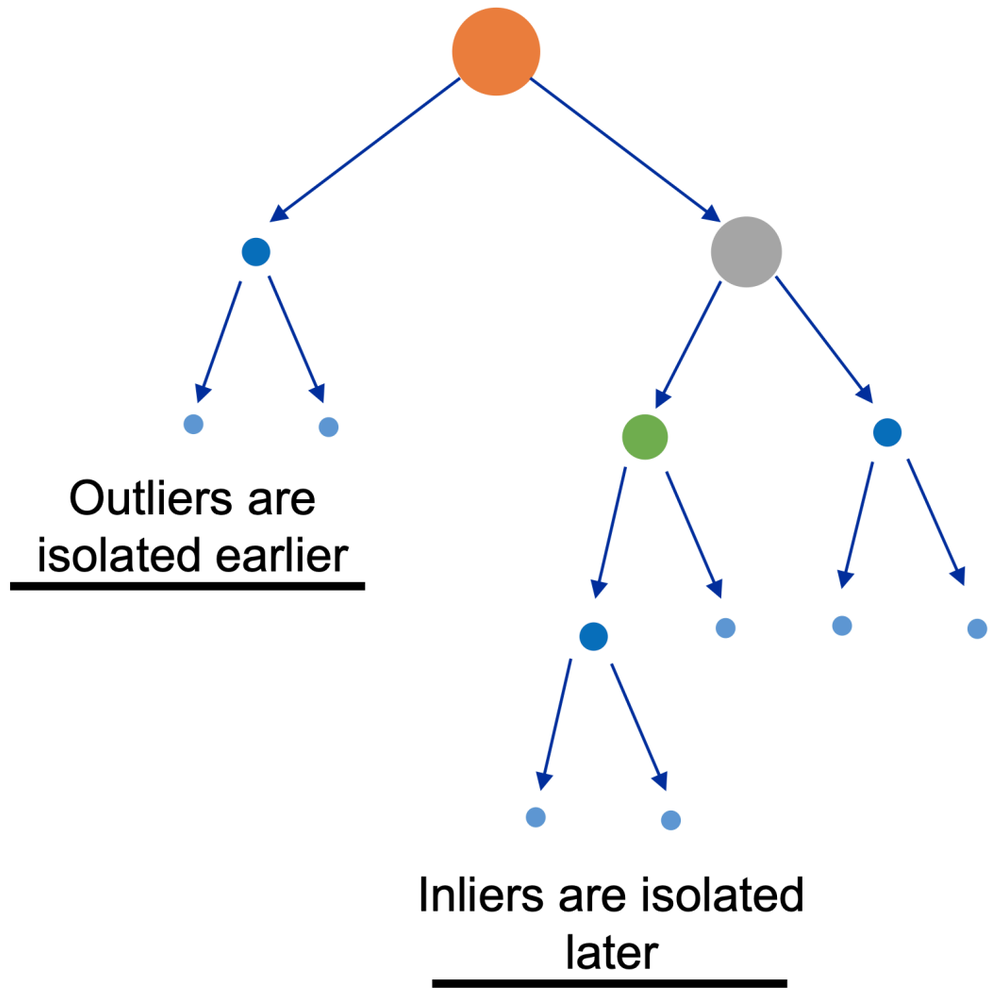
Baseline Models

Baseline Models

1. Isolation Forest

Since outliers have features X that differ significantly from most of the samples, they are isolated earlier in the hierarchy of a decision tree.

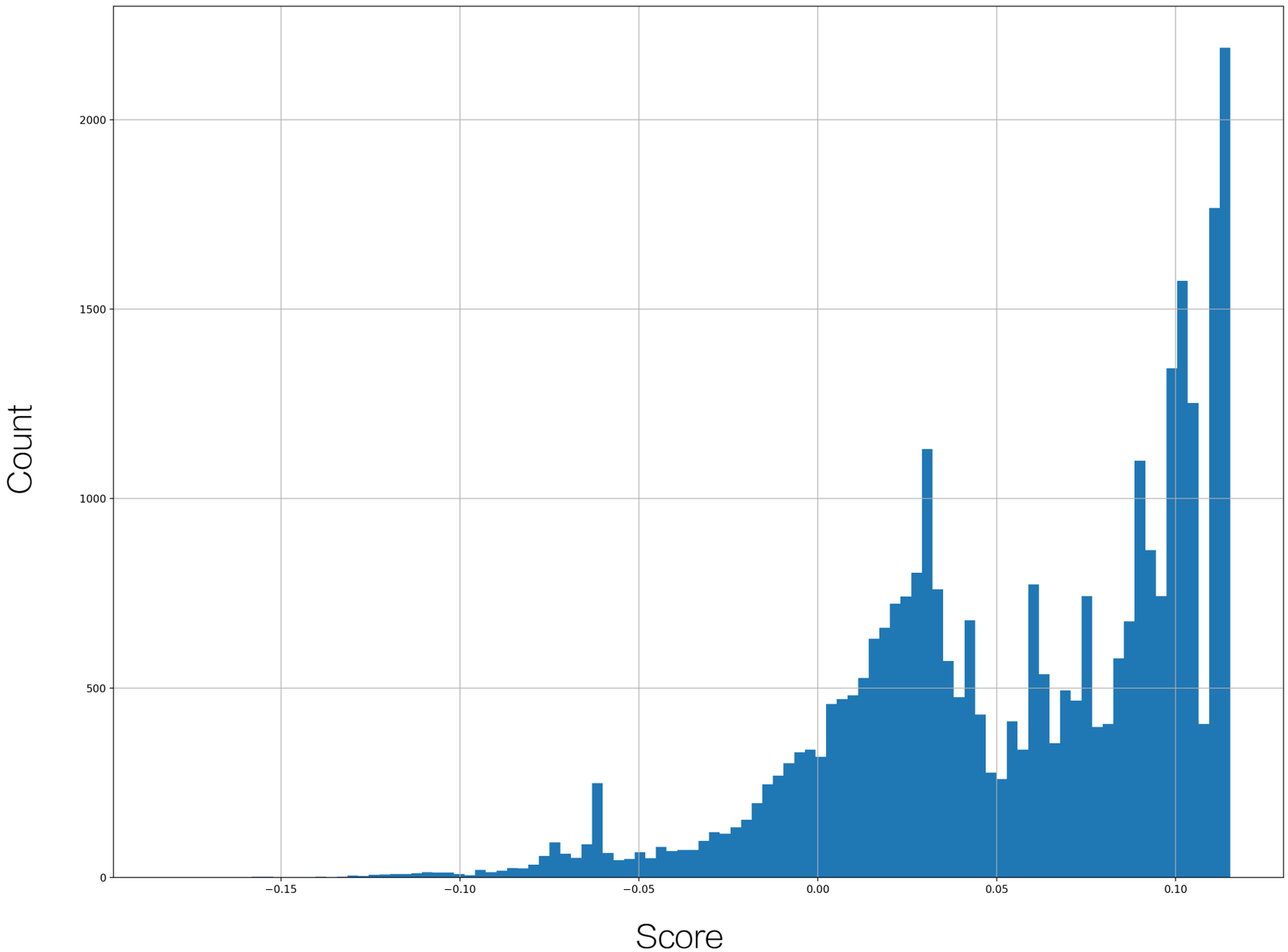
The Scikit-learn implementation provides a score for each sample that increases from -1 to 1 with the number of splits.



Baseline Models

1. Isolation Forest

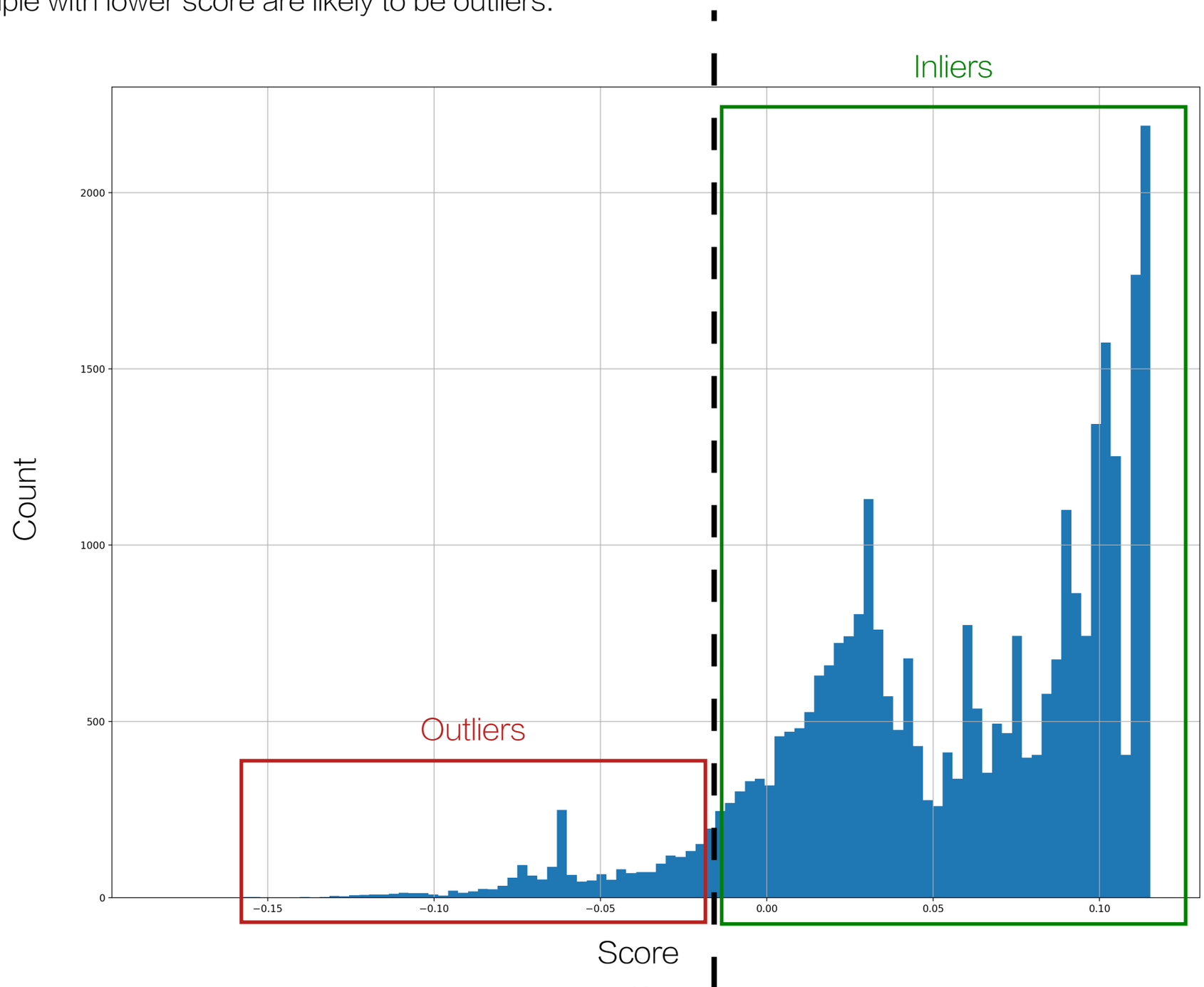
The sample with lower score are likely to be outliers.



Baseline Models

1. Isolation Forest

The sample with lower score are likely to be outliers.



Baseline Models

1. Isolation Forest

```
print(X['anomaly'].values_counts())
```

Baseline Models

1. Isolation Forest

```
print(X['anomaly'].values_counts())
```

```
 1     27756
-1      3785
Name: anomaly, dtype: int64
```

Baseline Models

1. Isolation Forest – Results

Crawler

	ip	time	method	status_code	path	response_length	user_agent	response_time	path_count_norr
296503	233.46.142.110	2021-05-12 09:32:04+04:30	Put	200	api/v2/token	289	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_13_6...	8.0	2.
297902	233.46.142.110	2021-05-12 09:32:50+04:30	Put	200	api/v2/token	289	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_13_6...	8.0	2.
302321	233.46.142.110	2021-05-12 09:35:31+04:30	Get	101	api/v2/connect/1396318207	99	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_13_6...	160053.0	0.
302369	233.46.142.110	2021-05-12 09:35:33+04:30	Put	200	api/v2/token	289	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_13_6...	12.0	2.
303761	233.46.142.110	2021-05-12 09:36:14+04:30	Get	101	api/v2/connect/1944714213	465235	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_13_6...	40604.0	0.
304029	233.46.142.110	2021-05-12 09:36:18+04:30	Put	200	api/v2/token	289	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_13_6...	8.0	2.
328160	233.46.142.110	2021-05-12 09:48:56+04:30	Put	200	api/v2/token	289	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_13_6...	12.0	2.

Baseline Models

1. Isolation Forest – Results

Crawler

	ip	time	method	status_code	path	response_length	user_agent	response_time	path_count_norma
302565	148.208.107.7	2021-05-12 09:35:40+04:30	Get	404	favicon.ico	29827	Mozilla/5.0 (Windows NT 6.3; Win64; x64) Apple...	4.0	0.03
302566	148.208.107.7	2021-05-12 09:35:40+04:30	Put	200	api/v2/token	289	Mozilla/5.0 (Windows NT 6.3; Win64; x64) Apple...	20.0	2.87
309599	148.208.107.7	2021-05-12 09:38:55+04:30	Get	200	cdn/verification_docs/1330189377	5495	Mozilla/5.0 (Windows NT 6.3; Win64; x64) Apple...	16.0	0.00
309597	148.208.107.7	2021-05-12 09:38:55+04:30	Get	200	cdn/verification_docs/483825864	3813	Mozilla/5.0 (Windows NT 6.3; Win64; x64) Apple...	16.0	0.00
309596	148.208.107.7	2021-05-12 09:38:55+04:30	Get	200	cdn/verification_docs/718839810	5258	Mozilla/5.0 (Windows NT 6.3; Win64; x64) Apple...	16.0	0.00
309595	148.208.107.7	2021-05-12 09:38:55+04:30	Get	200	cdn/verification_docs/23998397	4944	Mozilla/5.0 (Windows NT 6.3; Win64; x64) Apple...	28.0	0.00
309598	148.208.107.7	2021-05-12 09:38:55+04:30	Get	200	cdn/verification_docs/1803873472	6512	Mozilla/5.0 (Windows NT 6.3; Win64; x64) Apple...	12.0	0.00

Baseline Models

1. Isolation Forest – Results

Normal

	ip	time	method	status_code	path	response_length	user_agent	response_time	path_count_norma
631560	4.138.32.12	2021-05-12 11:58:09+04:30	Get	200	pages/630180842	50797	Mozilla/5.0 (Linux; Android 10; Redmi Note 8) ...	16.0	0.0
631958	4.138.32.12	2021-05-12 11:58:19+04:30	Get	200	css/font_awesome.min.css	30891	Mozilla/5.0 (Linux; Android 10; Redmi Note 8) ...	8.0	1.7
631959	4.138.32.12	2021-05-12 11:58:19+04:30	Get	200	css/page.2f0fc69390da8cdff683.css	50880	Mozilla/5.0 (Linux; Android 10; Redmi Note 8) ...	8.0	1.4
634810	4.138.32.12	2021-05-12 11:59:31+04:30	Get	200	js/page.07cb314dc14eef820638.js	332023	Mozilla/5.0 (Linux; Android 10; Redmi Note 8) ...	28.0	1.4
634808	4.138.32.12	2021-05-12 11:59:31+04:30	Get	200	images/gadgets/join_pros3.jpg	34053	Mozilla/5.0 (Linux; Android 10; Redmi Note 8) ...	8.0	1.4
634807	4.138.32.12	2021-05-12 11:59:31+04:30	Get	200	images/sanjagh_logo_purpule5.png	4680	Mozilla/5.0 (Linux; Android 10; Redmi Note 8) ...	4.0	2.2
634809	4.138.32.12	2021-05-12 11:59:31+04:30	Get	200	images/default.jpg	20993	Mozilla/5.0 (Linux; Android 10; Redmi Note 8) ...	8.0	1.1

Baseline Models

1. Isolation Forest – Results

Normal

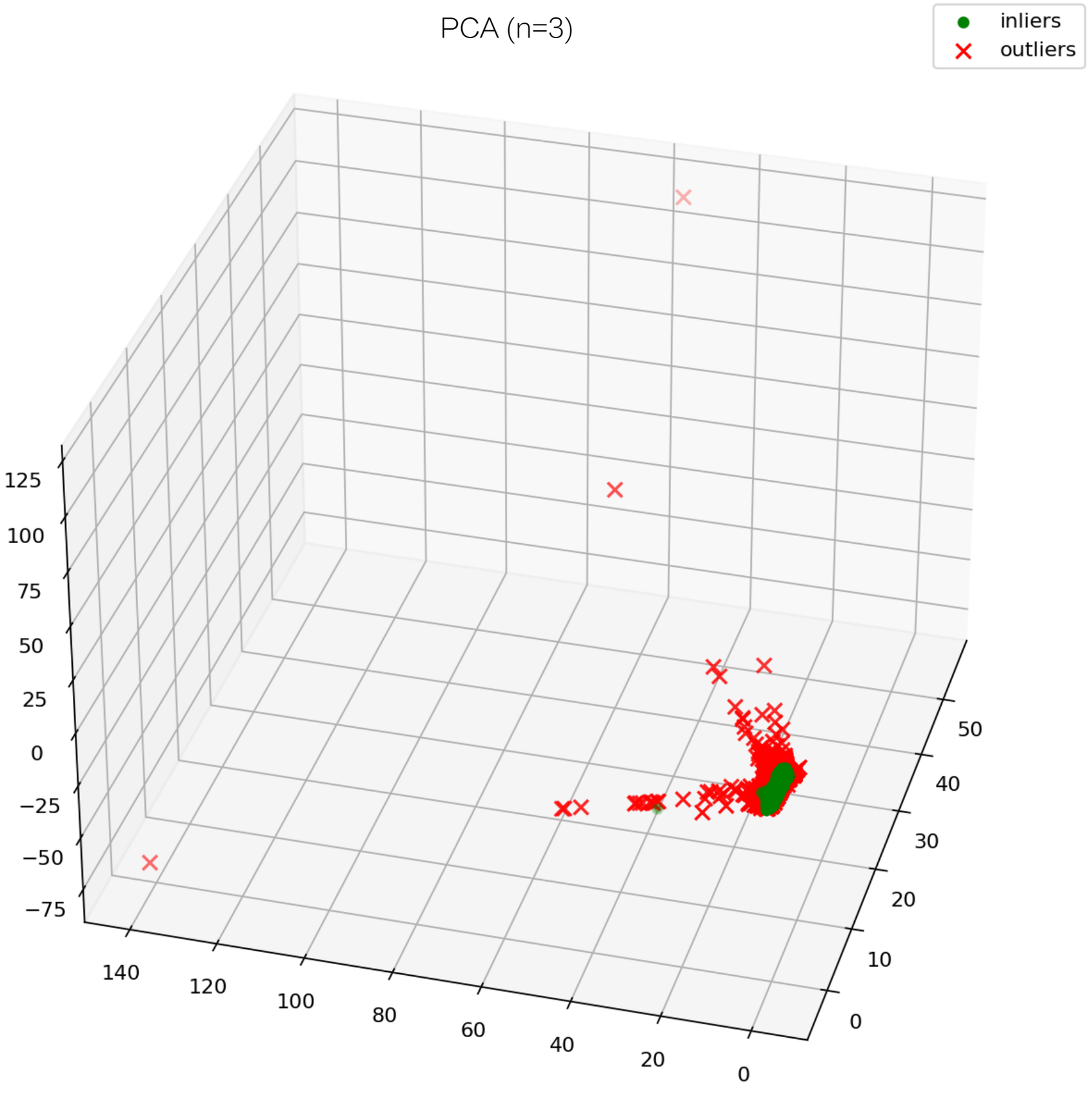
	ip	time	method	status_code	path	response_length	user_agent	response_time	path_count_nc
929570	37.199.253.251	2021-05-12 13:15:02+04:30	Get	200	pages/2098538394	52698	Mozilla/5.0 (Linux; Android 8.0.0; SM-A600FN) ...	12.0	
929597	37.199.253.251	2021-05-12 13:15:03+04:30	Get	304	css/font_awesome.min.css	0	Mozilla/5.0 (Linux; Android 8.0.0; SM-A600FN) ...	0.0	
929598	37.199.253.251	2021-05-12 13:15:03+04:30	Get	304	css/page.2f0fc69390da8cdff683.css	0	Mozilla/5.0 (Linux; Android 8.0.0; SM-A600FN) ...	0.0	
929605	37.199.253.251	2021-05-12 13:15:03+04:30	Get	200	images/sanjaghmaglogo1.png	25889	Mozilla/5.0 (Linux; Android 8.0.0; SM-A600FN) ...	8.0	
929606	37.199.253.251	2021-05-12 13:15:03+04:30	Get	200	images/gadgets/join_pros3.jpg	34053	Mozilla/5.0 (Linux; Android 8.0.0; SM-A600FN) ...	8.0	
929607	37.199.253.251	2021-05-12 13:15:03+04:30	Get	200	images/default.jpg	20993	Mozilla/5.0 (Linux; Android 8.0.0; SM-A600FN) ...	8.0	
929608	37.199.253.251	2021-05-12 13:15:03+04:30	Get	200	images/sanjagh_logo_purpule5.png	4680	Mozilla/5.0 (Linux; Android 8.0.0; SM-A600FN) ...	0.0	

Baseline Models

1. Isolation Forest – Results

```
print(len(anomalies[anomalies['is_bot']])))
print(len(non_anomalies[non_anomalies['is_bot']])))
```

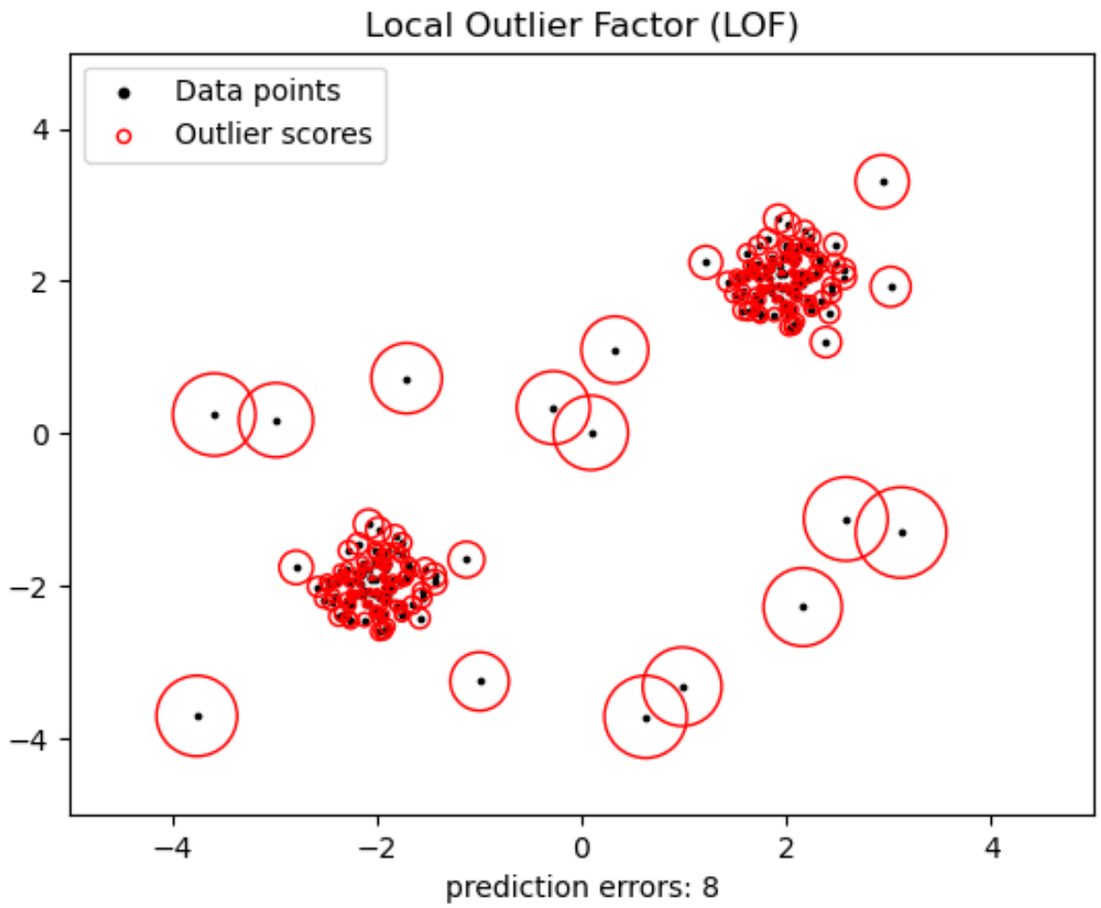
```
603
12
```


PCA (n=3)
inliers
outliers
125
100
75
50
25
0
−25
−50
−75
140
120
100
80
60
40
20
0
50
40
30
20
10
0

Baseline Models

2. Local Outlier Factor

It measures the local deviation of density of a given sample with respect to its neighbor. It is local in that the anomaly score depends on how isolated the object is with respect to the surrounding neighborhood.



Resources

[1] Detecting Web Crawlers from Web Server Access Logs with Data Mining Classifiers

Thanks.

Q/A

Mohammad Hashemi – Parastoo Falak Aflaki