



**POLITECNICO**  
**MILANO 1863**

SCUOLA DI INGEGNERIA INDUSTRIALE  
E DELL'INFORMAZIONE

# Systems and Methods for Big and Unstructured Data Project

Author(s): **Tommaso Aiello**

Group Number: **\*\*\*\*\***

Academic Year: 2023-2024



# Contents

<b>Contents</b>	<b>i</b>
<b>1 Chapter One</b>	<b>1</b>
1.1 Introduction . . . . .	1
1.2 Dataset . . . . .	1
1.2.1 Non Relational data schema . . . . .	2
1.2.2 Attributes Description . . . . .	3
<b>2 Chapter two</b>	<b>5</b>
2.1 Queries . . . . .	5
2.1.1 Query One . . . . .	5
2.1.2 Query Two . . . . .	6
2.1.3 Query Three . . . . .	9
2.1.4 Query Four . . . . .	10
2.1.5 Query Five . . . . .	11
2.1.6 Query Six . . . . .	12
2.1.7 Query Seven . . . . .	15
2.1.8 Query Eight . . . . .	18
2.1.9 Query Nine . . . . .	21
2.1.10 Query Ten . . . . .	23
<b>3 Chapter 3</b>	<b>29</b>
3.1 Kibana . . . . .	29
3.1.1 Kibana introduction . . . . .	29
3.1.2 Trump's tweets analysis dashboard . . . . .	29
<b>List of Figures</b>	<b>33</b>



# 1 | Chapter One

## 1.1. Introduction

This project is part of a course named 'Systems and Methods for Big and Unstructured Data'. The goal of the project is to choose a dataset of at least 20000 data points and a DB technology among Neo4j, MongoDB and ElasticSearch. Perform various queries of different complexities with a good complexity variety.

I chose Trump Twitter Archive dataset that contains more than 56000 tweets and I opted for ElasticSearch as DB technology.

## 1.2. Dataset

The dataset used for analysis is the "trump\_tweets\_archive" downloaded from Kaggle (<https://www.kaggle.com/datasets/headsortails/trump-twitter-archive>) containing tweets attributed to the former U.S. President Donald Trump. The dataset was made public by Brendan of <https://www.thetrumparchive.com/>.

The dataset consists of more than 56k Trump's tweets with various attributes such as tweet id, text, is\_retweet, is\_deleted, device, favorites, retweets, datetime, is\_flagged, date. The goal of the project is to perform meaningful queries using ElasticSearch and visualize the results in Kibana.

I chose this dataset because I thought it could be interesting to explore how Trump's communication focus was shifting over the time, and how preoccupied he was with certain topics; In particular I wanted to see his behavior related to some very important events like 2016 and 2020 presidential elections and Covid-19.

I opted for ElasticSearch because of the full-text search feature and because in combination with Kibana it gave me the possibility of creating an interactive dashboard which makes visualization much more easy and comprehensive.

The dataset was upload on ElasticSearch as a csv and no prior modification was done.

Even though it may be interesting in the future to extract some more informations from the dataset. It could be interesting to perform sentiment analysis for each tweet in order to explore how those values changed over time and over topics.

### 1.2.1. Non Relational data schema

In this Elasticsearch mapping, we define the structure and characteristics of the data that will be indexed in Elasticsearch. The mapping provides a blueprint for how different fields in the data should be interpreted and stored.

```
{
  "mappings": {
    "_meta": {
      "created_by": "file-data-visualizer"
    },
    "properties": {
      "@timestamp": {
        "type": "date"
      },
      "date": {
        "type": "date",
        "format": "iso8601"
      },
      "datetime": {
        "type": "date",
        "format": "iso8601"
      },
      "device": {
        "type": "keyword"
      },
      "favorites": {
        "type": "long"
      },
      "id": {
        "type": "double"
      },
      "is_deleted": {
        "type": "keyword"
      },
      "is_flagged": {
        "type": "keyword"
      },
      "is_retweet": {
        "type": "keyword"
      }
    }
  }
}
```

```
    },  
    "retweets": {  
      "type": "long"  
    },  
    "text": {  
      "type": "text"  
    }  
  }  
}  
}
```

### 1.2.2. Attributes Description

1. **@timestamp, date and datetime:** These fields are defined as date types, indicating they store timestamp information. The format for date and datetime: "iso8601" ensures consistent interpretation of date strings in ISO 8601 format.
2. **device:** tells which device was used to tweet. I mapped it as a keyword because the options are limited.
3. **favorites:** it refers to the number of favorites of the tweet and it is mapped as a long.
4. **id:** it refers to the tweet id assigned by X (ex Twitter). It is mapped as a double.
5. **is\_deleted:** it is a flag to see if the tweet was deleted or not. It is mapped as a keyword and it assumes only two values 'TRUE' or 'FALSE'. It could have been mapped as a boolean as well.
6. **is\_flagged:** it is a flag to see if the tweet was flagged by X or not. It is important because those are the tweets that were considered clearly false or harmful. It is mapped as a keyword and it assumes only two values 'TRUE' or 'FALSE'. It could have been mapped as a boolean as well.
7. **is\_retweet:** it is a flag to see if the tweet is a retweet or not. Retweets are not always endorsement so sometimes it is good to have a field to distinguish what is a tweet and what is a retweet. It is mapped as a keyword and it assumes only two values 'TRUE' or 'FALSE'. It could have been mapped as a boolean as well.
8. **retweets:** it refers to the number of retweets and it is mapped as a long.
9. **text:** this field contains the actual content of the tweet. It is mapped as text to give the possibility of performing full-text search.





# 2 | Chapter two

## 2.1. Queries

The project involves a series of 10 queries with increasing complexity, exploring different aspects of the dataset. These queries cover a range of topics, from basic filtering and aggregation to more advanced queries involving date ranges, keyword searches, and analysis of flagged tweets.

### 2.1.1. Query One

The first query goal is to get all the tweets that contain at least the word 'Biden' or 'Sleepy Joe' and that also include at least one word between 'election', 'fraud', 'rigged', 'vote', 'fake'.

```
GET /trump_tweets_analysis/_search
{
  "query": {
    "bool": {
      "must": [
        {
          "bool": {
            "should": [
              { "match": { "text": "Biden" } }},
              { "match": { "text": "Sleepy Joe" } }
            ]
          }
        }
      ],
      "should": [
        { "match": { "text": "election" } }},
        { "match": { "text": "fraud" } }},
        { "match": { "text": "rigged" } }},
        { "match": { "text": "fake" } }},
        { "match": { "text": "vote" } }
      ]
    }
  }
}
```



'rigged', 'vote', 'fake'. Count the ones that are flagged, which means that X (aka Twitter) decided to flag because they were considered harmful or against the policies.

```
GET /trump_tweets_analysis/_search
{
  "query": {
    "bool": {
      "must": [
        {
          "bool": {
            "should": [
              { "match": { "text": "Biden" }},
              { "match": { "text": "Sleepy Joe" }}
            ]
          }
        },
        {
          "bool": {
            "should": [
              { "match": { "text": "election" }},
              { "match": { "text": "fraud" }},
              { "match": { "text": "rigged" }},
              { "match": { "text": "fake" }},
              { "match": { "text": "vote" }}
            ]
          }
        }
      ]
    }
  },
  "aggs": {
    "flagged_count": {
      "filter": {
        "term": { "is_flagged": "TRUE" }
      }
    }
  }
}
```

In Figure 2.2 we can see the partial outcome of the second query. The first part is equal to the first one but in this one there is also an aggregation that is done to count the number of tweets about Biden and the elections and what happened after, that X decided to flag because potentially harmful or false.

```
182     "_score": 14.755133,  
183     "_source": {  
184       "favorites": 216232,  
185       "date": "2020-09-15",  
186       "datetime": "2020-09-15T02:24:09Z",  
187       "is_deleted": "FALSE",  
188       "@timestamp": "2020-09-15T02:24:09.000Z",  
189       "id": 1305693861067407400,  
190       "text": "Did you see where Joe Biden – as Weak, Tired, and Sleepy as he is, went to a  
        Polling Place today in Delaware (of course!) to VOTE!? If Biden can do it, any  
        American can do it!",  
191       "retweets": 52763,  
192       "device": "Twitter for iPhone",  
193       "is_retweet": "FALSE",  
194       "is_flagged": "FALSE"  
195     }  
196   }  
197 ]  
198 },  
199 "aggregations": {  
200   "flagged_count": {  
201     "doc_count": 12  
202   }  
203 }  
204 }
```

Figure 2.2: Query 2 partial outcome

In Figure 2.3 it is shown a visualization showing that there are 12 tweets out of the 201 about the elections that are flagged.

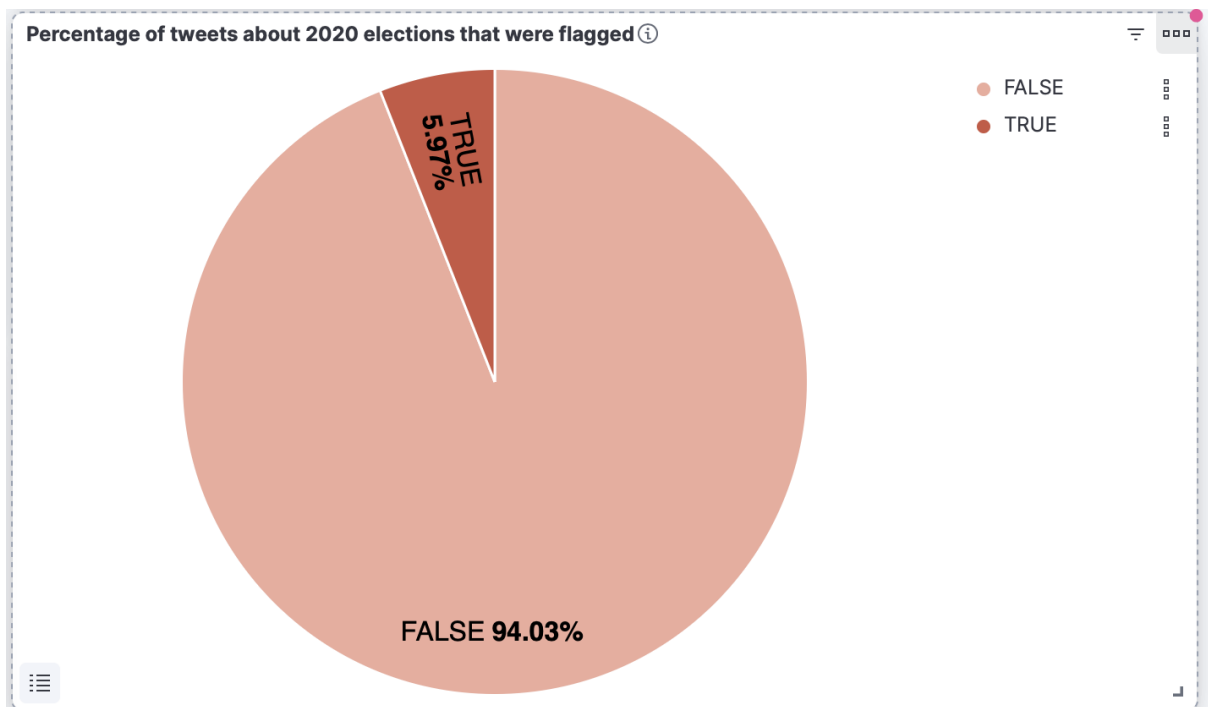


Figure 2.3: Query 2 dashboard visualization

### 2.1.3. Query Three

The third query goal is to get the average retweets and favorites per tweet containing words that may be related to elections and what happened during the protests after the presidential elections of 2020.

```
GET /trump_tweets_analysis/_search
{
  "size": 0,
  "query": {
    "bool": {
      "must": [
        { "bool": {
          "should": [
            { "match": { "text": "election" }},
            { "match": { "text": "fraud" }},
            { "match": { "text": "protest" }},
            { "match": { "text": "protests" }},
            { "match": { "text": "rigged" }},
            { "match": { "text": "vote" }}
          ]
        }
      ],
      "range": { "datetime": { "gte": "2020-11-01", "lte": "2021-12-31" }}}
    }
  },
  "aggs": {
    "avg_retweets": {
      "avg": {
        "field": "retweets"
      }
    },
    "avg_favorites": {
      "avg": {
        "field": "favorites"
      }
    }
  }
}
```

In Figure 2.4 we can see that the average retweets about elections and protests is around 39464 while the average favorite is around 154730. This computation was done over the

421 tweets that were matched by the query.



```

1- {
2-   "took": 31,
3-   "timed_out": false,
4-   "_shards": {
5-     "total": 1,
6-     "successful": 1,
7-     "skipped": 0,
8-     "failed": 0
9-   },
10-   "hits": {
11-     "total": {
12-       "value": 421,
13-       "relation": "eq"
14-     },
15-     "max_score": null,
16-     "hits": []
17-   },
18-   "aggregations": {
19-     "avg_retweets": {
20-       "value": 39464.23277909739
21-     },
22-     "avg_favorites": {
23-       "value": 154730.81235154395
24-     }
25-   }
26- }

```

Figure 2.4: Query 3 outcome

### 2.1.4. Query Four

The fourth query goal is to get the monthly tweet count.

```

GET /trump_tweets_analysis/_search
{
  "size": 0,
  "aggs": {
    "monthly_tweets": {
      "date_histogram": {
        "field": "datetime",
        "calendar_interval": "month",
        "format": "yyyy-MM",
        "order": { "_key": "asc" }
      }
    }
  }
}

```

In Figure 2.5 we can see the partial output of query 4. For each month 'doc\_count' represents the number of tweets published by Donald Trump account.

```

17-  },
18-  "aggregations": {
19-    "monthly_tweets": {
20-      "buckets": [
21-        {
22-          "key_as_string": "2009-05",
23-          "key": 1241136000000,
24-          "doc_count": 21
25-        },
26-        {
27-          "key_as_string": "2009-06",
28-          "key": 1243814400000,
29-          "doc_count": 11
30-        },
31-        {
32-          "key_as_string": "2009-07",
33-          "key": 1246406400000,
34-          "doc_count": 5
35-        },
36-        {
37-          "key_as_string": "2009-08",
38-          "key": 1249084800000,
39-          "doc_count": 7
40-        },
41-        {
42-          "key_as_string": "2009-09",
43-          "key": 1251763200000,
44-          "doc_count": 3
45-        },
46-        {
47-          "key_as_string": "2009-10",
48-          "key": 1254355200000,
49-          "doc_count": 4
50-        },

```

Figure 2.5: Query 4 partial outcome

### 2.1.5. Query Five

The fifth query goal is to get the tweets that contain the word economy or similar using fuzziness. And sort them in ascending order. Not perfect matches have lower scores.

```

{
  "query": {
    "fuzzy": {
      "text": {
        "value": "economy",
        "fuzziness": 2
      }
    }
  },
  "sort": [
    {
      "_score": {
        "order": "asc"
      }
    }
  ]
}

```

```

    }
  }
]
}

```

In Figure 2.6 we can see the partial output of query 5. The fuzziness parameter in Elasticsearch is used to implement "fuzzy" search capabilities. This means it allows for the return of results that are similar, but not exactly the same, as the search term. It's particularly useful when you want to accommodate for misspellings, typos, and variations in the data you're searching through. In this case the tweet in the photo doesn't contain the word "economy" but the word "economic" and it is matched anyway but with a lower score.

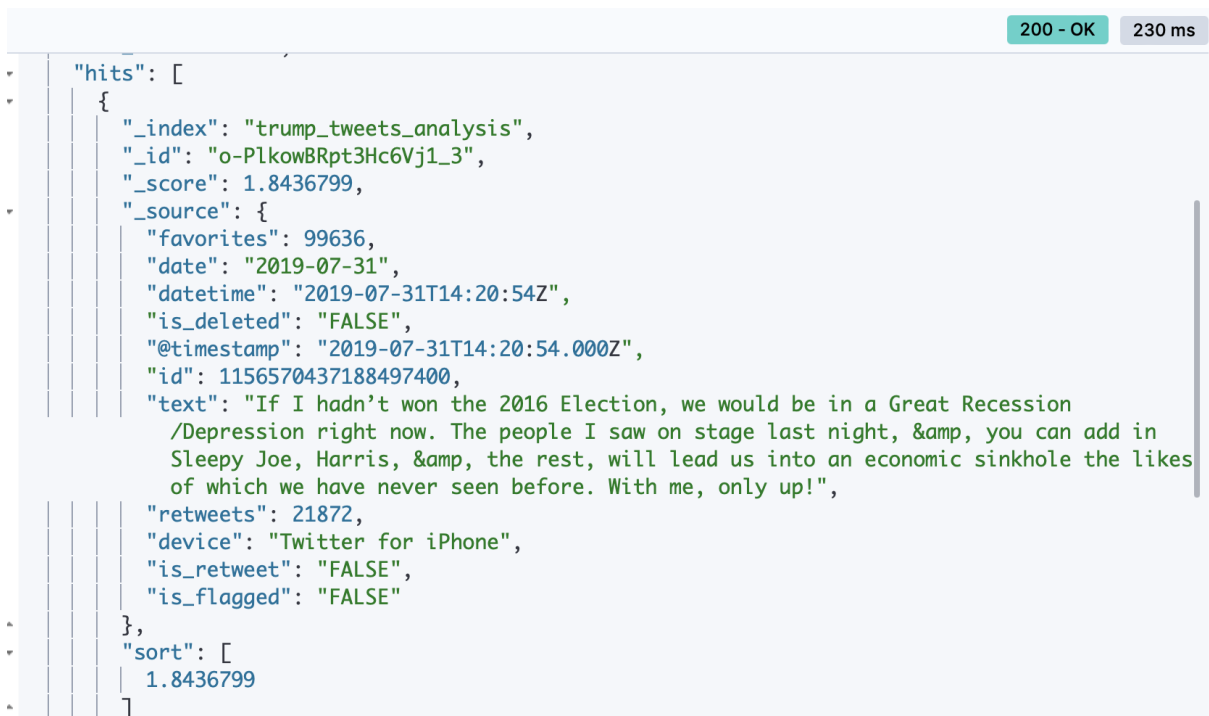


Figure 2.6: Query 5 partial outcome

### 2.1.6. Query Six

The sixth query goal is to retrieve tweet and datetime of the tweet with the most retweets per year.

```

GET /trump_tweets_analysis/_search
{
  "size": 0,
  "aggs": {

```



```
"tweets_by_year": {  
  "date_histogram": {  
    "field": "datetime",  
    "calendar_interval": "year",  
    "format": "yyyy",  
    "order": { "_key": "asc" }  
  },  
  "aggs": {  
    "top_retweets": {  
      "terms": {  
        "field": "id",  
        "order": { "max_retweets.value": "desc" },  
        "size": 1  
      },  
      "aggs": {  
        "max_retweets": {  
          "max": {  
            "field": "retweets"  
          }  
        },  
        "tweet_details": {  
          "top_hits": {  
            "size": 1,  
            "_source": {  
              "includes": ["text", "datetime"]  
            }  
          }  
        }  
      }  
    }  
  }  
}
```

In Figure 2.7 we can see the most retweeted tweet from 2019 which is about the released of ASAP Rocky, who is a very famous rapper and that was in jail in Sweden because he was accused of assault. The tweet was retweeted 226235 times.

```

1  {
2    "key_as_string": "2019",
3    "key": 1546300800000,
4    "doc_count": 7818,
5    "top_retweets": {
6      "doc_count_error_upper_bound": -1,
7      "sum_other_doc_count": 7817,
8      "buckets": [
9        {
10         "key": 1157345692517634000,
11         "doc_count": 1,
12         "tweet_details": {
13           "hits": {
14             "total": {
15               "value": 1,
16               "relation": "eq"
17             },
18             "max_score": 1,
19             "hits": [
20               {
21                 "_index": "trump_tweets_analysis",
22                 "_id": "d-PlkowBRpt3Hc6Vj1_3",
23                 "_score": 1,
24                 "_source": {
25                   "datetime": "2019-08-02T17:41:30Z",
26                   "text": "A$AP Rocky released from prison and on his way home to the United
27                     States from Sweden. It was a Rocky Week, get home ASAP! A$AP!"
28                 }
29               }
30             ]
31           }
32         },
33         "max_retweets": {
34           "value": 226235
35         }
36       ]
37     }
38   }

```

Figure 2.7: Query 6 partial outcome

In Figure 2.8 we can see the visualization that is available in the kibana dashboard where it is possible to see the content of the tweet with most retweets for every year with also the day it was sent and the number of retweets it got.

In 2021 the most retweeted tweet was about the Capitol Hill attack while in 2020 it was the announcement of POTUS and FLOTUS being tested positive for COVID-19.

Top Tweets ordered by retweet for each year ⓘ

↓ Year	▼ Tweet Content	▼ Date	▼ retweets	▼
2021	I am asking for everyone at the U.S. Capitol to remain peaceful. No violence! Remember, WE are the Party of Law & Order – respect the Law and our great men and women in Blue. Thank you!	Jan 6, 2021 @ 01:00:00.000	156,100	
2020	Tonight, @FLOTUS and I tested positive for COVID-19. We will begin our quarantine and recovery process immediately. We will get through this	Oct 2, 2020 @ 02:00:00.000	408,866	

Rows per page: 13 ▼

< 1 >

Figure 2.8: Dashboard query 6

### 2.1.7. Query Seven

The seventh query goal is to retrieve tweets having at least one word that is 'covid', 'virus', 'vaccine'. That are not a retweet and that are made in 2020. Get the aggregate count by month and get the one with the maximum retweets.

```
GET /trump_tweets_analysis/_search
{
  "size": 0,
  "query": {
    "bool": {
      "must": [
        {
          "bool": {
            "should": [
              { "match": { "text": "covid" } }},
              { "match": { "text": "virus" } }},
              { "match": { "text": "vaccine" } }
            ]
          }
        }
      ],
      "term": { "is_retweet": "FALSE" }},
      "range": { "datetime": { "gte": "2020-01-01" }}}
```



```

200 - OK 178 ms
{
  "key_as_string": "2020-10",
  "key": 1601510400000,
  "doc_count": 24,
  "top_retweets": {
    "doc_count_error_upper_bound": -1,
    "sum_other_doc_count": 23,
    "buckets": [
      {
        "key": 1311892190680014800,
        "doc_count": 1,
        "tweet_details": {
          "hits": {
            "total": {
              "value": 1,
              "relation": "eq"
            },
            "max_score": 6.8468285,
            "hits": [
              {
                "_index": "trump_tweets_analysis",
                "_id": "Y-LlkowBRpt3Hc6VgL3f",
                "_score": 6.8468285,
                "_source": {
                  "datetime": "2020-10-02T04:54:06Z",
                  "text": "Tonight, @FLOTUS and I tested positive for COVID-19. We will begin our quarantine and recovery process immediately. We will get through this TOGETHER!"
                }
              }
            ]
          }
        },
        "max_retweets": {
          "value": 408866
        }
      }
    ]
  }
}

```

Figure 2.9: Query 7 partial outcome

In Figure 2.10 it is shown the count of tweets per month that were about covid pandemic. We can see that there was a peak in March 2020 when a lot of states had to go in lockdown, there was a decrease during summer and again a new increase during autumn when there was a second wave of Covid cases during which former president Trump got infected too.

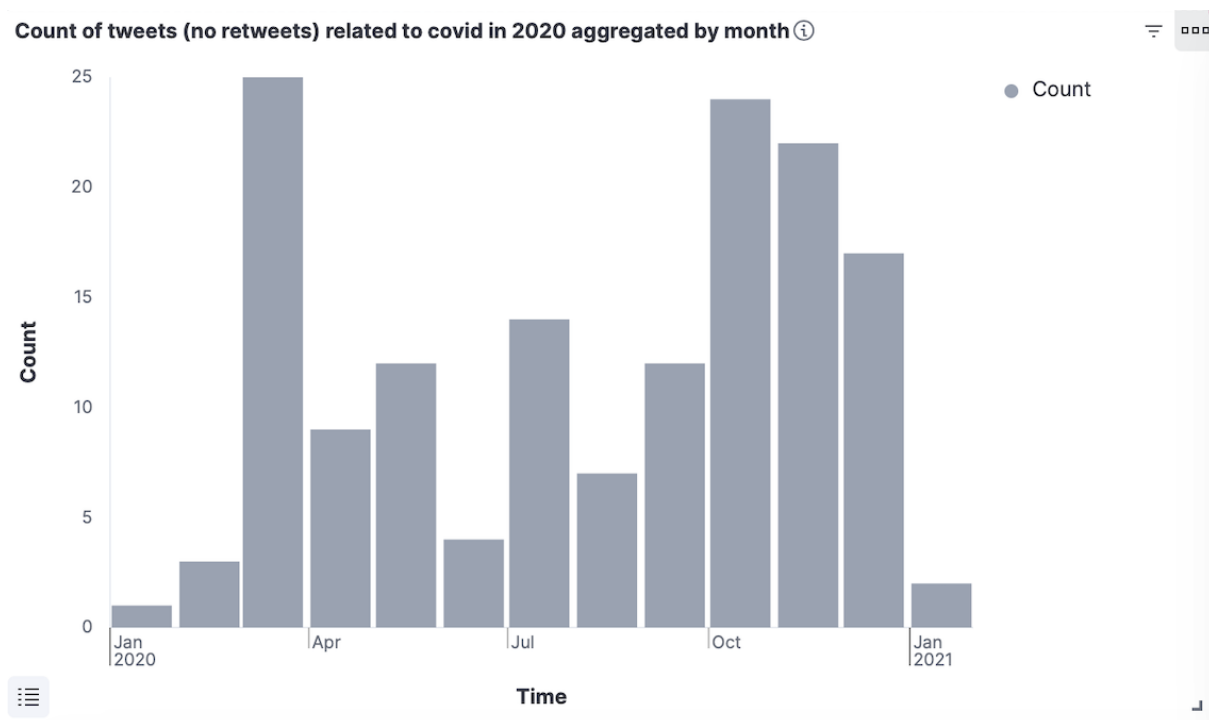


Figure 2.10: Dashboard query 7

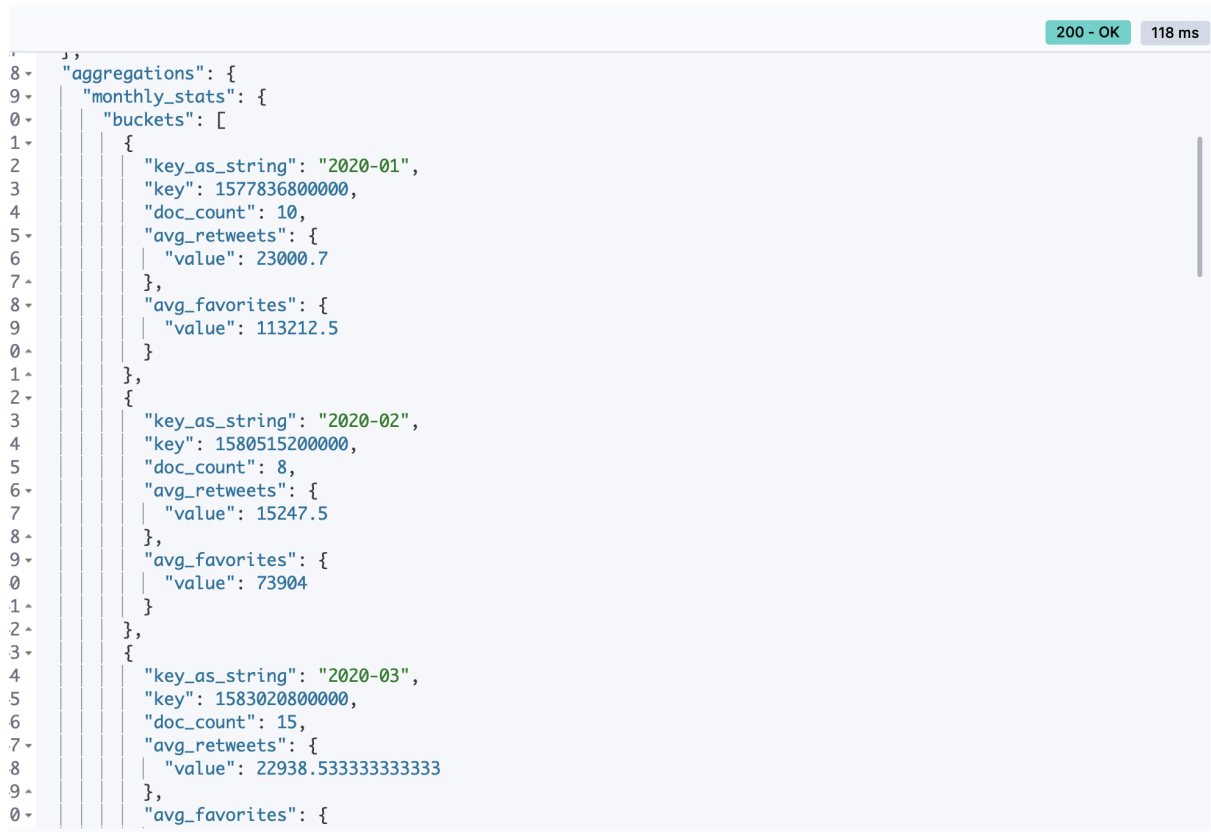
### 2.1.8. Query Eight

The eight query goal is to get the average retweet and favorite per tweet per month related to China and virus and vaccine. That is not a retweet and that was written since 2020.

```
GET trump_tweets_analysis/_search
{
  "size": 0,
  "query": {
    "bool": {
      "must": [
        {
          "bool": {
            "should": [
              { "match": { "text": "covid" } }},
              { "match": { "text": "china" } }},
              { "match": { "text": "vaccine" } }
            ]
          }
        }
      ],
      "term": { "is_retweet": "FALSE" }},
      "range": { "datetime": { "gte": "2020-01-01" }}}
  }
}
```

```
    }  
  },  
  "aggs": {  
    "monthly_stats": {  
      "date_histogram": {  
        "field": "datetime",  
        "calendar_interval": "month",  
        "format": "yyyy-MM",  
        "order": { "_key": "asc" }  
      },  
      "aggs": {  
        "avg_favorites": {  
          "avg": {  
            "field": "favorites"  
          }  
        },  
        "avg_retweets": {  
          "avg": {  
            "field": "retweets"  
          }  
        }  
      }  
    }  
  }  
}
```

In Figure 2.11 it is shown the number of documents retrieved for each month with the corresponding average retweets and favorites.



```
8  "aggregations": {
9    "monthly_stats": {
10     "buckets": [
11       {
12         "key_as_string": "2020-01",
13         "key": 1577836800000,
14         "doc_count": 10,
15         "avg_retweets": {
16           "value": 23000.7
17         },
18         "avg_favorites": {
19           "value": 113212.5
20         }
21       },
22       {
23         "key_as_string": "2020-02",
24         "key": 1580515200000,
25         "doc_count": 8,
26         "avg_retweets": {
27           "value": 15247.5
28         },
29         "avg_favorites": {
30           "value": 73904
31         }
32       },
33       {
34         "key_as_string": "2020-03",
35         "key": 1583020800000,
36         "doc_count": 15,
37         "avg_retweets": {
38           "value": 22938.533333333333
39         },
40         "avg_favorites": {
```

Figure 2.11: Query 8 partial outcome

In Figure 2.12 it is shown the visualization contained in the Kibana dashboard for this query. It is possible to see how the values changed over time.



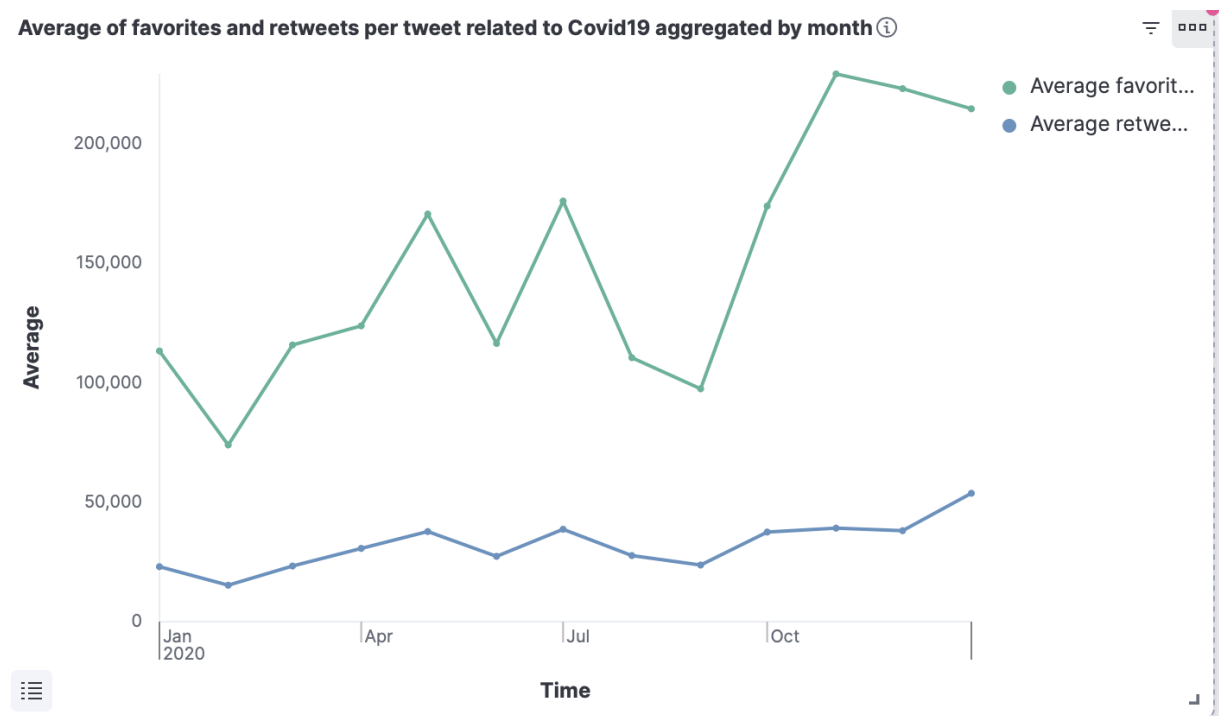


Figure 2.12: Dashboard query 8

### 2.1.9. Query Nine

The ninth query goal is to get the number of tweets per year that were deleted. And for each year, get the text and datetime of the one with the most favorites.

```
GET /trump_tweets_analysis/_search
{
  "size": 0,
  "query": {
    "bool": {
      "must": [
        { "term": { "is_deleted": "TRUE" } }
      ]
    }
  },
  "aggs": {
    "deleted_tweets_by_year": {
      "date_histogram": {
        "field": "datetime",
        "calendar_interval": "year",
        "format": "yyyy",
        "order": { "_key": "asc" }
      },
    }
  }
}
```

```

"aggs": {
  "top_favorites": {
    "terms": {
      "field": "id",
      "order": { "max_favorites.value": "desc" },
      "size": 1
    },
    "aggs": {
      "max_favorites": {
        "max": {
          "field": "favorites"
        }
      },
      "tweet_details": {
        "top_hits": {
          "size": 1,
          "_source": {
            "includes": ["text", "datetime", "favorites"]
          }
        }
      }
    }
  }
}

```

In figure 2.13 it is shown that in 2017 the most liked tweet that was then eliminated is a tweet in which there was a clear error because it contains non-sense word like "covfefe".

```

200 - OK 116 ms
{
  "key_as_string": "2017",
  "key": 1483228800000,
  "doc_count": 103,
  "top_favorites": {
    "doc_count_error_upper_bound": -1,
    "sum_other_doc_count": 102,
    "buckets": [
      {
        "key": 869766994899468300,
        "doc_count": 1,
        "tweet_details": {
          "hits": {
            "total": {
              "value": 1,
              "relation": "eq"
            },
            "max_score": 3.9469748,
            "hits": [
              {
                "_index": "trump_tweets_analysis",
                "_id": "huPlkowBRpt3Hc6VjT_C",
                "_score": 3.9469748,
                "_source": {
                  "favorites": 162788,
                  "datetime": "2017-05-31T04:06:25Z",
                  "text": "Despite the constant negative press covfefe"
                }
              }
            ]
          }
        }
      }
    ]
  },
  "max_favorites": {
    "value": 162788
  }
}

```

Figure 2.13: Query 9 partial outcome

### 2.1.10. Query Ten

The tenth query goal is to get the tweet count per month of tweets containing 'Hillary' or 'Clinton' or 'democrats' made in 2016 and get the tweet count per month of tweets containing 'Biden' or 'Sleepy Joe' or 'democrats' in 2020.

```

GET /trump_tweets_analysis/_search
{
  "size": 0,
  "query": {
    "bool": {
      "should": [
        {
          "bool": {
            "must": [
              { "match": { "text": "Hillary" } },
              { "range": { "datetime": { "gte": "2016-01-01", "lt": "2017-01-01" } } }
            ]
          }
        }
      ]
    }
  }
}

```

```

},
{
  "bool": {
    "must": [
      { "match": { "text": "Clinton" }},
      { "range": { "datetime": { "gte": "2016-01-01", "lt": "2017-01-01" }}}
    ]
  }
},
{
  "bool": {
    "must": [
      { "match": { "text": "democrats" }},
      { "range": { "datetime": { "gte": "2016-01-01", "lt": "2017-01-01" }}}
    ]
  }
},
{
  "bool": {
    "must": [
      { "match": { "text": "Biden" }},
      { "range": { "datetime": { "gte": "2020-01-01", "lt": "2021-01-01" }}}
    ]
  }
},
{
  "bool": {
    "must": [
      { "match": { "text": "Sleepy Joe" }},
      { "range": { "datetime": { "gte": "2020-01-01", "lt": "2021-01-01" }}}
    ]
  }
},
{
  "bool": {
    "must": [
      { "match": { "text": "democrats" }},
      { "range": { "datetime": { "gte": "2020-01-01", "lt": "2021-01-01" }}}
    ]
  }
}

```

```

        }
      }
    ]
  }
},
"aggs": {
  "monthly_tweets_2016": {
    "filter": {
      "range": {
        "datetime": {
          "gte": "2016-01-01",
          "lt": "2017-01-01"
        }
      }
    },
    "aggs": {
      "monthly_tweets": {
        "date_histogram": {
          "field": "datetime",
          "calendar_interval": "month",
          "format": "yyyy-MM",
          "order": { "_key": "asc" }
        }
      }
    }
  },
  "monthly_tweets_2020": {
    "filter": {
      "range": {
        "datetime": {
          "gte": "2020-01-01",
          "lt": "2021-01-01"
        }
      }
    },
    "aggs": {
      "monthly_tweets": {
        "date_histogram": {
          "field": "datetime",
          "calendar_interval": "month",
          "format": "yyyy-MM",
          "order": { "_key": "asc" }
        }
      }
    }
  }
}

```

```

    }
  }
}
}

```

In Figure 2.14 it is shown the number of tweets related to Trump's opponent in the elections grouped by month. In that case the election were the ones that happened in 2020, there is in the image also the indication of the number of tweets happened in 2016 for those elections which is 562.



```

{
  "hits": [],
  "aggregations": {
    "monthly_tweets_2016": {
      "doc_count": 562,
      "monthly_tweets": {
        "buckets": []
      }
    },
    "monthly_tweets_2020": {
      "doc_count": 1510,
      "monthly_tweets": {
        "buckets": [
          {
            "key_as_string": "2020-01",
            "key": 1577836800000,
            "doc_count": 141
          },
          {
            "key_as_string": "2020-02",
            "key": 1580515200000,
            "doc_count": 88
          },
          {
            "key_as_string": "2020-03",
            "key": 1583020800000,
            "doc_count": 74
          },
          {
            "key_as_string": "2020-04",
            "key": 1585699200000,
            "doc_count": 72
          }
        ]
      }
    }
  }
}

```

Figure 2.14: Query 10 partial outcome

In Figure 2.15 we can see the tweet count about opponents grouped by month. We can see across 2020 Trump published way more tweets than what he did in 2016. We can see that in both case he reached a peak in the tweets count during October so in the middle of the electoral campaign.

This visualization is available in the Kibana dashboard with also the other visualizations presented before.

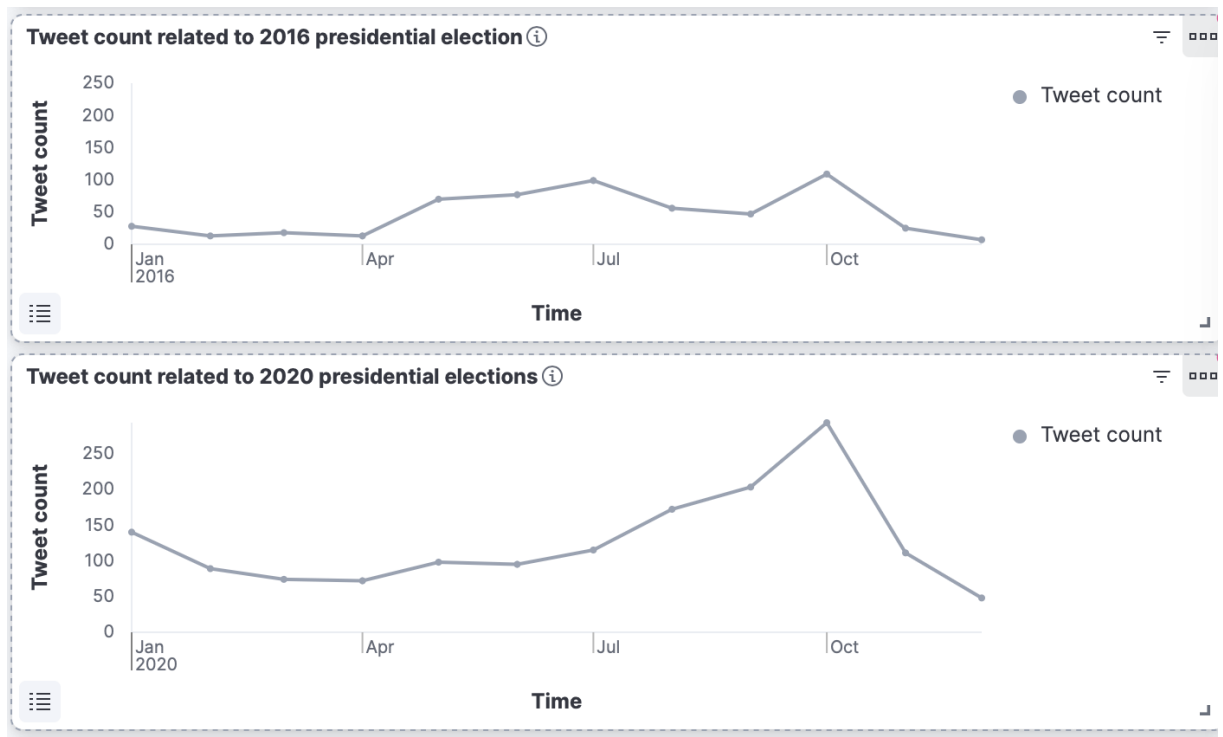


Figure 2.15: Dashboard query 10 divided into two parts





# 3 | Chapter 3

## 3.1. Kibana

### 3.1.1. Kibana introduction

Kibana is an open-source data visualization and exploration tool used for log and time-series analytics, application monitoring, and operational intelligence use cases.

In Kibana, dashboards are user interfaces that can contain various visualizations and panels. They provide an at-a-glance view of data and insights derived from that data which is real-time data from Elasticsearch indices.

Users can create different types of visualizations such as charts, maps, tables, and more. These visualizations help in understanding complex queries by representing them in graphical formats.

Dashboards are interactive. Users can quickly change the data view by applying filters, changing time ranges, or modifying queries. This allows for deep, real-time analysis. Moreover users can customize dashboards according to their needs. They can arrange visualizations, resize panels, and add text annotations for better understanding.

### 3.1.2. Trump's tweets analysis dashboard

In Figure 3.1 it is shown the way the dashboard is presented. It contains some filters related to the time in the top right. It also contains some other filters that if applied affect all the visualizations present in the dashboard.

There are filters about the device used to send the tweet, a filter to get only the deleted tweets, a filter to get only the retweets, a filter to decide the range of favorites or retweets number of the tweet to fetch and a filter to get only the flagged tweets.

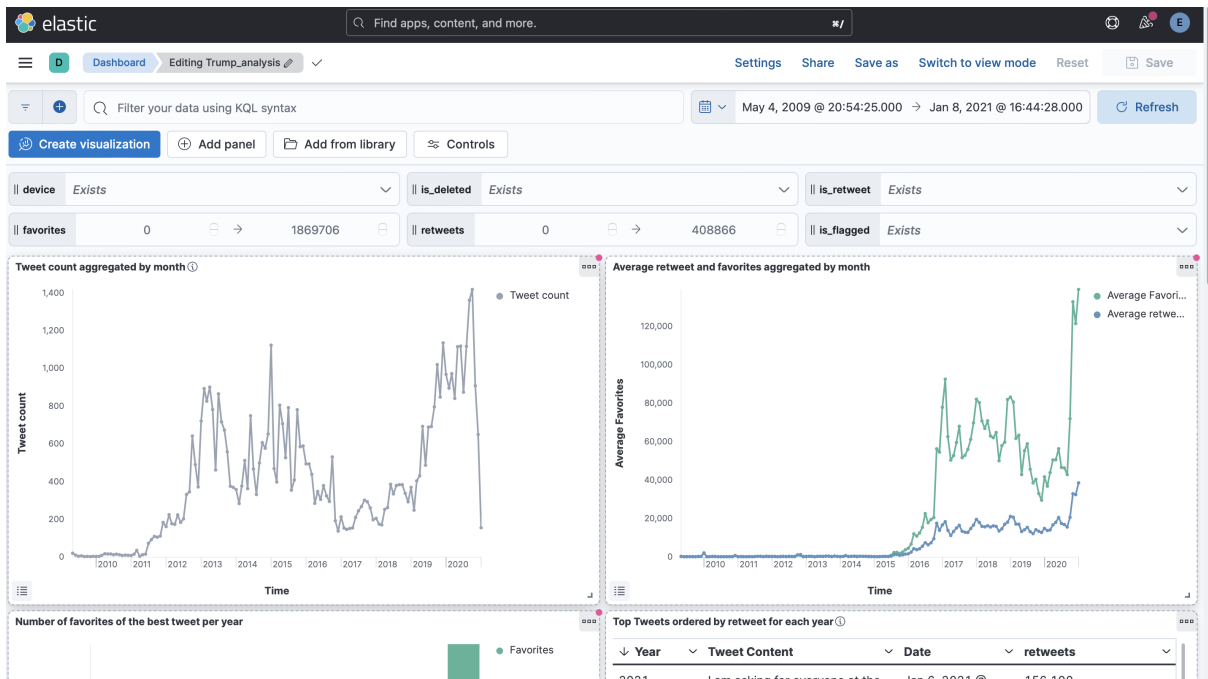


Figure 3.1: Dashboard Overview

In Figure 3.2 it is shown a chart showing the tweet count grouped by month.



Figure 3.2: Dashboard tweet count aggregated by month

In Figure 3.3 it is shown the trend in the average likes and retweets of Trump tweets aggregated by month across all his "Twitter Experience". We can see that he started to get more and more attention starting from 2016. He had then a peak during the last months of 2020 in which he was particularly under the spotlight due to everything that happened during the elections and after them.

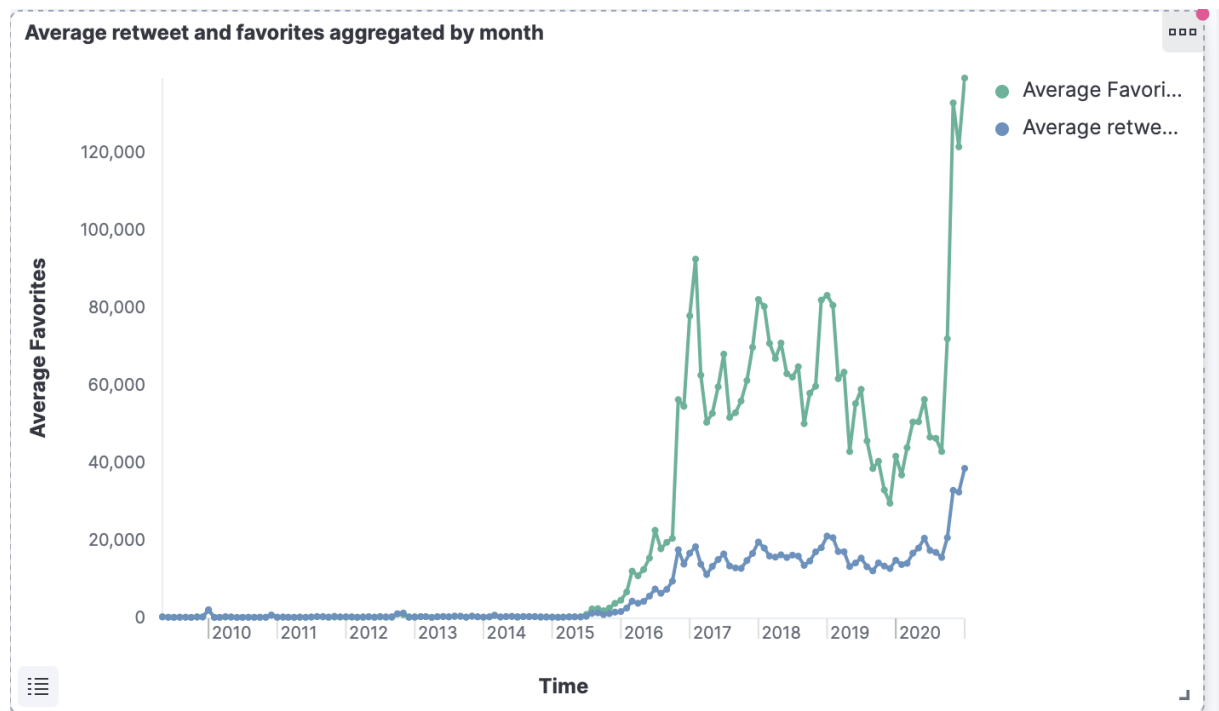


Figure 3.3: Dashboard avg of retweet and favorites aggregated by month

In Figure 3.4 it is shown the number of favorites of his 'best' tweet for every year.

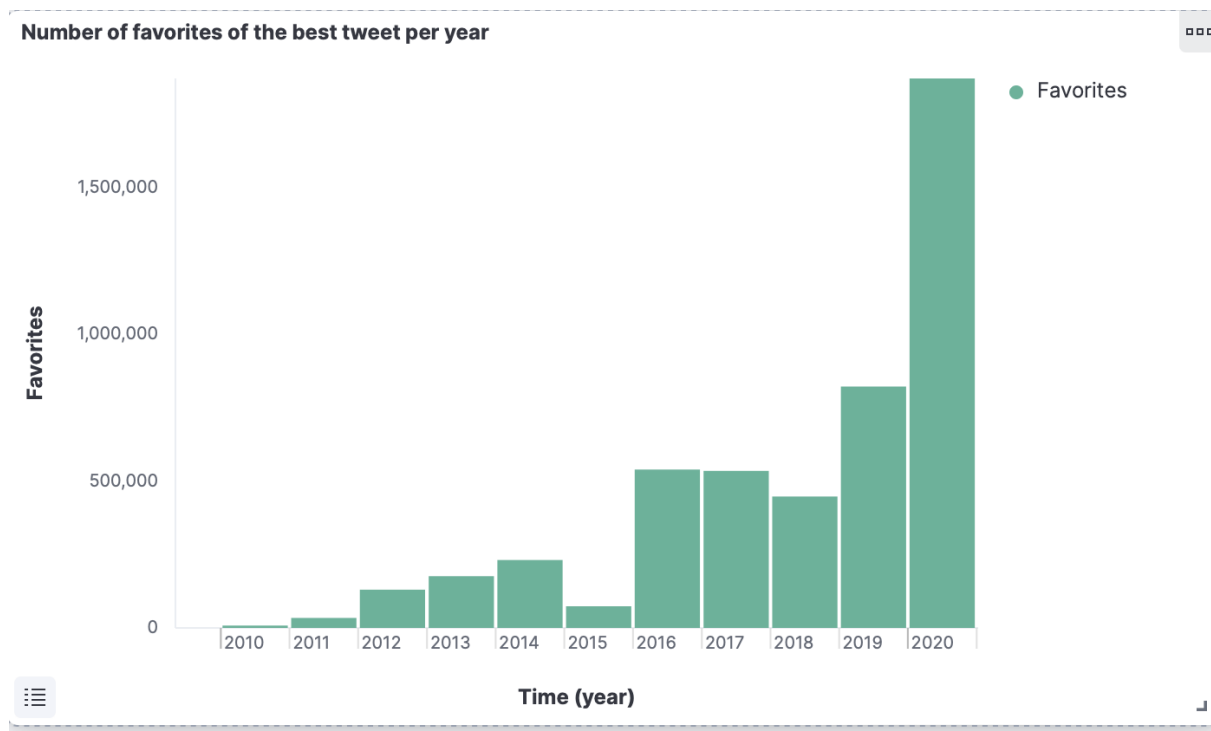


Figure 3.4: Dashboard favorites of best tweet

## List of Figures

2.1	Query 1 partial outcome . . . . .	6
2.2	Query 2 partial outcome . . . . .	8
2.3	Query 2 dashboard visualization . . . . .	8
2.4	Query 3 outcome . . . . .	10
2.5	Query 4 partial outcome . . . . .	11
2.6	Query 5 partial outcome . . . . .	12
2.7	Query 6 partial outcome . . . . .	14
2.8	Dashboard query 6 . . . . .	15
2.9	Query 7 partial outcome . . . . .	17
2.10	Dashboard query 7 . . . . .	18
2.11	Query 8 partial outcome . . . . .	20
2.12	Dashboard query 8 . . . . .	21
2.13	Query 9 partial outcome . . . . .	23
2.14	Query 10 partial outcome . . . . .	26
2.15	Dashboard query 10 divided into two parts . . . . .	27
3.1	Dashboard Overview . . . . .	30
3.2	Dashboard tweet count aggregated by month . . . . .	30
3.3	Dashboard avg of retweet and favorites aggregated by month . . . . .	31
3.4	Dashboard favorites of best tweet . . . . .	32



## List of Tables

