Forecasting Patient Enrolment for Clinical Trials

3rd Team Project Sprint Review
May 4th, 2020

01

Overview

Statistics about the individual attributes (missing values,...)

02

Trials Data

First Analysis of the clinical trials data

03

Hospital Data

Handling new hospital sources like wikidata and Google Places

04

Country Data

Country Data merged within the MongoDB

05

Next Steps

Plans on how to proceed in this project

Clinical Trials Overview

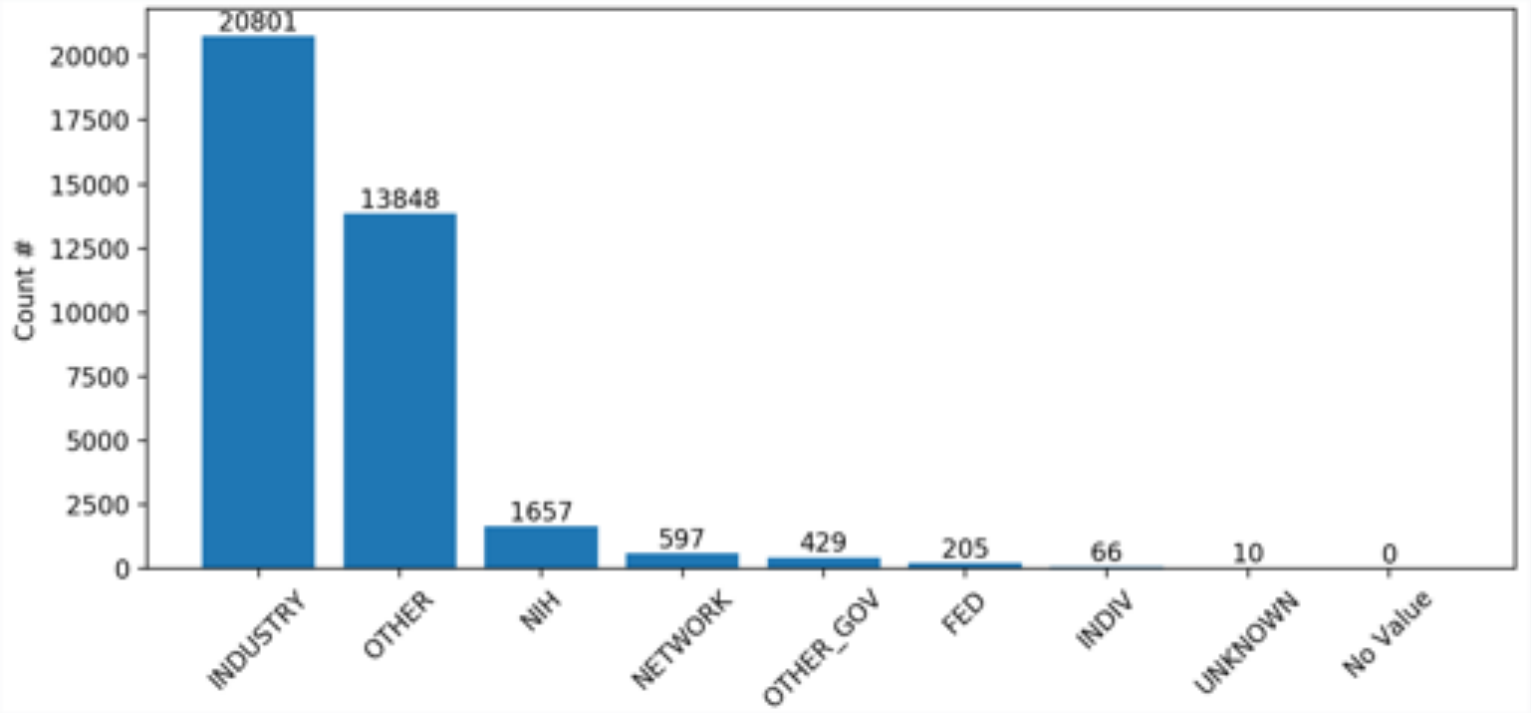
Overview

Attribute	Missing (#)	Missing (%)	Data Structure	Data Type	Distinct Values	Examples
NCTId	0	0	single value	free text	37604	NCT01658995, NCT01436500, NCT01048242, NCT01986933, NCT01
OrgFullName	0	0	single value	categorical (huge diversity)	4536	Vicus Therapeutics, Phosphate Therapeutics, University of Colo
OrgClass	0	0	single value	categorical	8	UNKNOWN, FED, NIH, INDIV, NETWORK, OTHER_GOV, INDUSTRY
BriefTitle	0	0	single value	free text	37313	'A Trial Evaluating the Efficac...', 'Determination of Safe and Effe
BriefSummary	0	0	single value	free text	0	
StudyType	0	0	single value	categorical	1	Interventional
OverallStatus	0	0	single value	categorical	2	Terminated, Completed
Phase	0	0	list	categorical	6	Phase 3, Not Applicable, Phase 4, Phase 2, Phase 1, Early Phase 1
StartDate	0	0	single value	categorical	360	May 2004, September 2001, June 2005, January 2016, October 20
StatusVerifiedDate	0	0	list	categorical	402	November 29, 2017, May 2004, June 2005, January 2016, October
CompletionDate	0	0	single value	categorical	316	May 2004, September 2001, June 2005, January 2016, October 20
Condition	0	0	list	categorical (huge diversity)	12562	Intraoperative Ureter Injury, Stage II Mycosis Fungoides and Sez
LeadSponsorName	0	0	list	categorical (huge diversity)	5444	Phosphate Therapeutics, University of Colorado, Boulder, JANSS
LeadSponsorClass	0	0	list	categorical	8	UNKNOWN, FED, NIH, INDIV, NETWORK, OTHER_GOV, INDUSTRY
EnrollmentCount	0	0	single value	integer	2005	2, 3, 4, 5, 6, 7, 8, 9, 10, 11
StdAge	0	0	list	categorical	3	Child, Older Adult, Adult
InterventionType	0	0	list	categorical	11	Biological, Diagnostic Test, Radiation, Behavioral, Drug, Combin
InterventionName	0	0	list	free text	35194	Standard-of-Care plus Dexmedetomidine, Adrenocorticotropin,
EligibilityCriteria	5	0,01	list	---	0	
Gender	28	0,07	single value	categorical	4	Female, Male, None, Female
HealthyVolunteers	173	0,46	single value	categorical	3	Accepts Healthy Volunteers, None, None
DesignPrimaryPurpose	420	1,12	single value	categorical	11	Device Feasibility, Treatment, Diagnostic, Supportive Care, Non
OfficialTitle	685	1,82	single value	free text	36601	'A Multicenter, Double-blind, R...', 'HIGH INTENSITY, BRIEF DURA

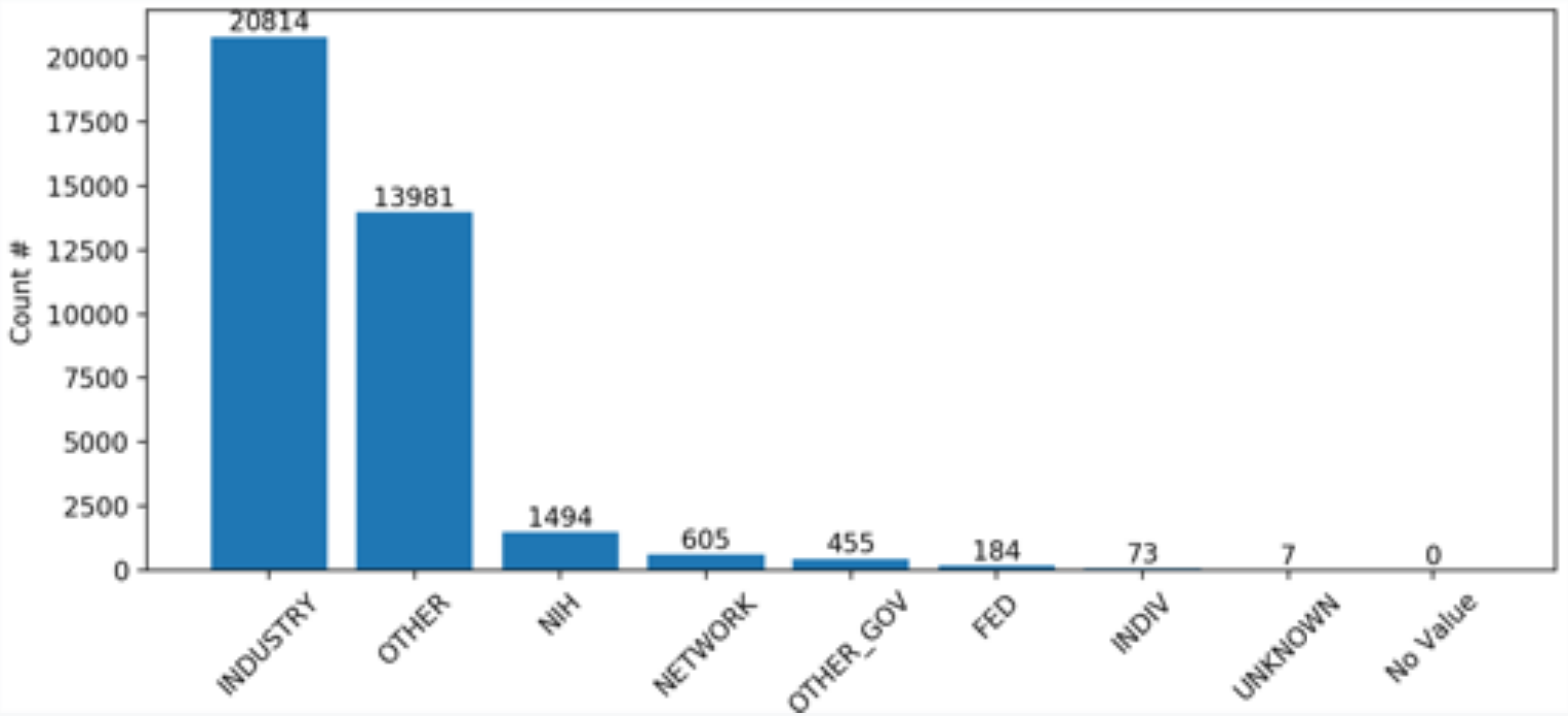
"Missingness" of critical Info

Attribute	Missing (#)	Missing (%)	Data Structure	Data Type	Distinct Values	Examples
OrgFullName	0	0	single value	categorical (huge diversity)	4536	Vicus Therapeutics, Phosphate Therapeutics, University o
OrgClass	0	0	single value	categorical	8	UNKNOWN, FED, NIH, INDIV, NETWORK, OTHER_GOV, IND
StartDate	0	0	single value	categorical	360	May 2004, September 2001, June 2005, January 2016, Octo
CompletionDate	0	0	single value	categorical	316	May 2004, September 2001, June 2005, January 2016, Octo
LeadSponsorClass	0	0	list	categorical	8	UNKNOWN, FED, NIH, INDIV, NETWORK, OTHER_GOV, IND
EnrollmentCount	0	0	single value	integer	2005	2, 3, 4, 5, 6, 7, 8, 9, 10, 11
LocationCity	3406	9,06	list	categorical (huge diversity)	25989	Lugansk, Oakwood, Kagawa, Longjumeau, Goudi, Sumperl
LocationCountry	3406	9,06	list	categorical	166	Portugal, Former Serbia and Montenegro, Jordan, Hungar
LocationFacility	6542	17,39	list	free text	160953	Virginia Commonwealth University, School of Medicine, 2
CollaboratorClass	25917	68,9	list	categorical	9	UNKNOWN, FED, NIH, INDIV, NETWORK, OTHER_GOV, AM

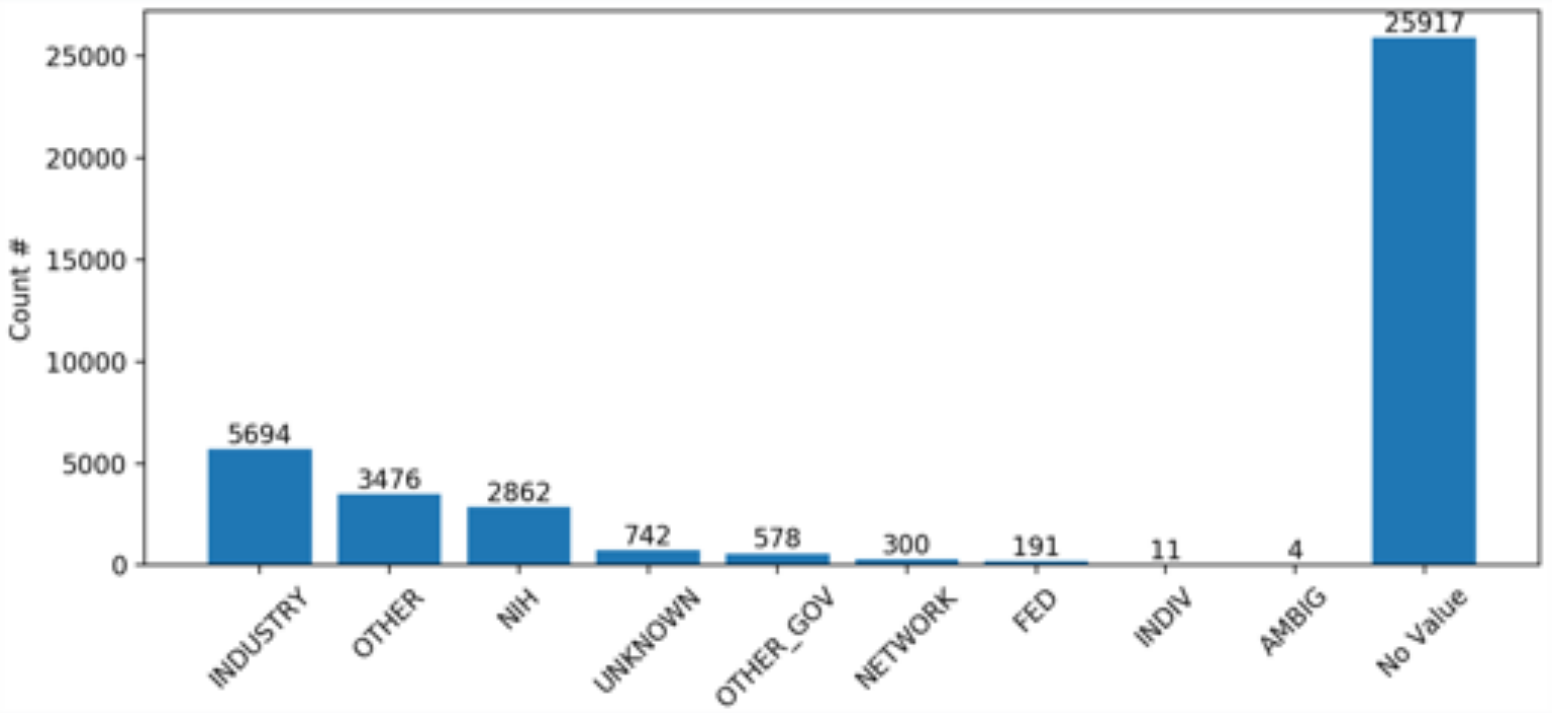
Distribution of OrgClass



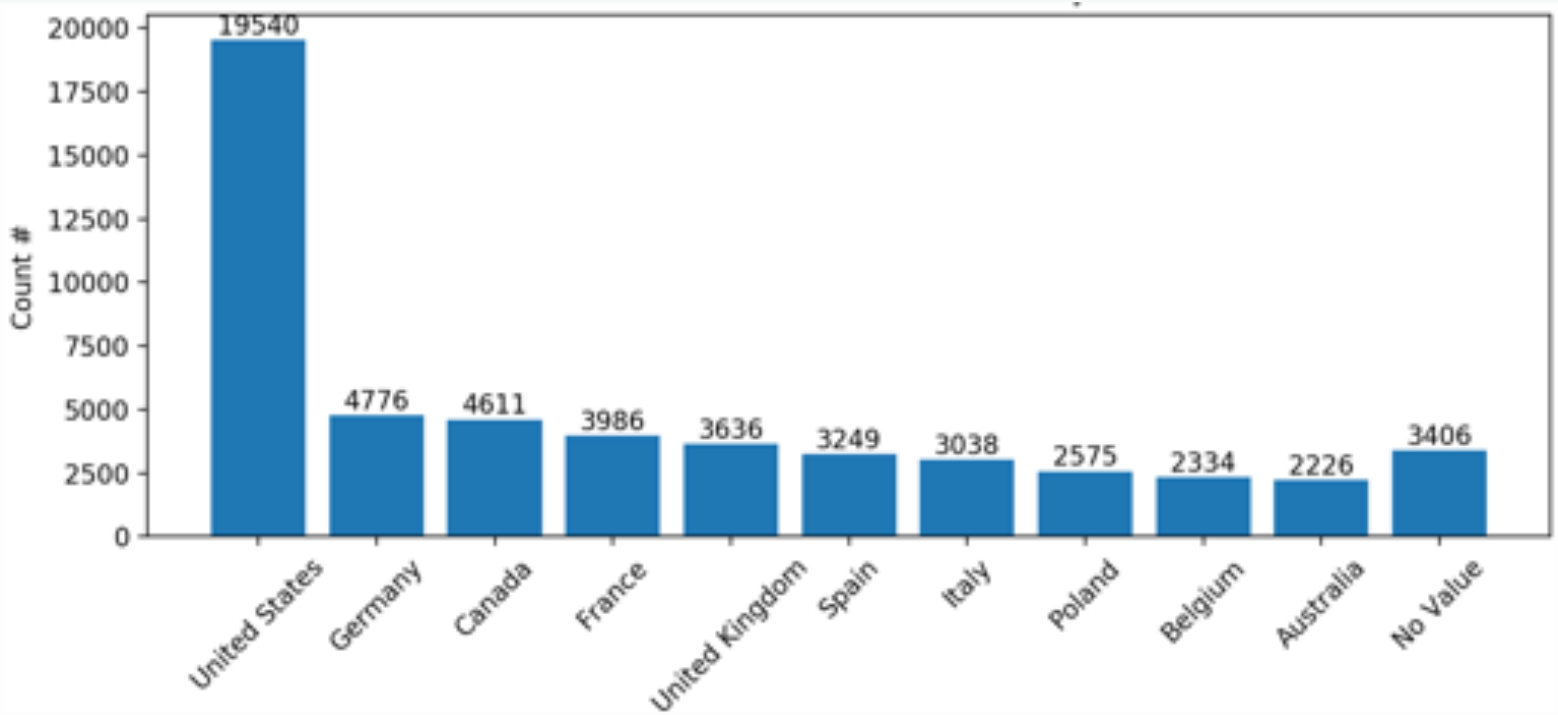
Distribution of LeadSponsorClass



Distribution of CollaboratorClass



Distribution of LocationCountry (Top 10)



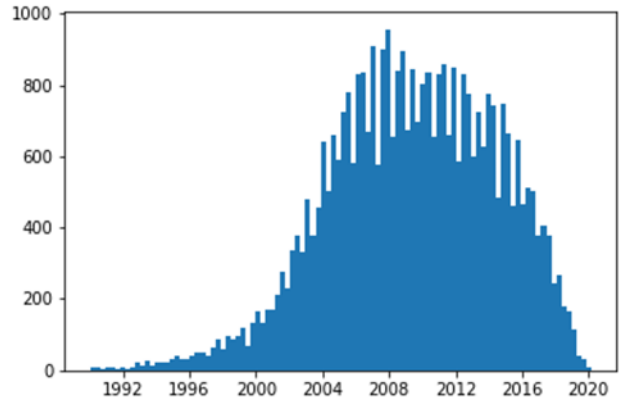
Clinical Trials Analysis

- Duration
- Average Enrollment
- Enrollment per month per country

Clinical Trials Data - Duration

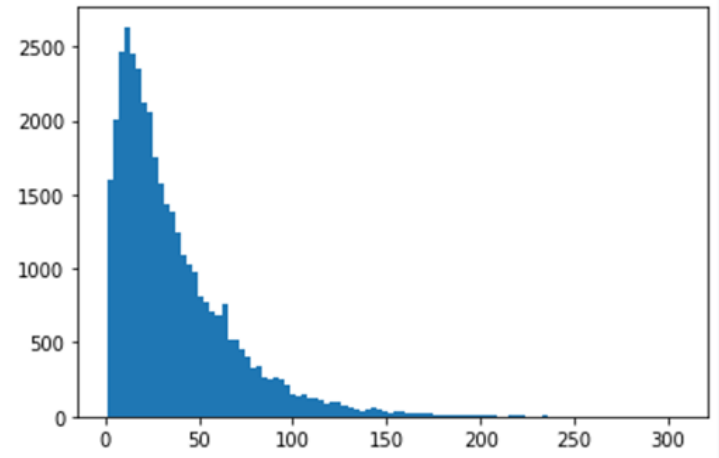
- Saved attributes “StartDate” and “CompletionDate” in standardized datetime Format “Month, Year”
- Dropped Studies with StartDate before 1990 as recent studies are more representative

```
Earliest date: 1990-01-01 00:00:00
Latest date: 2020-03-01 00:00:00
```



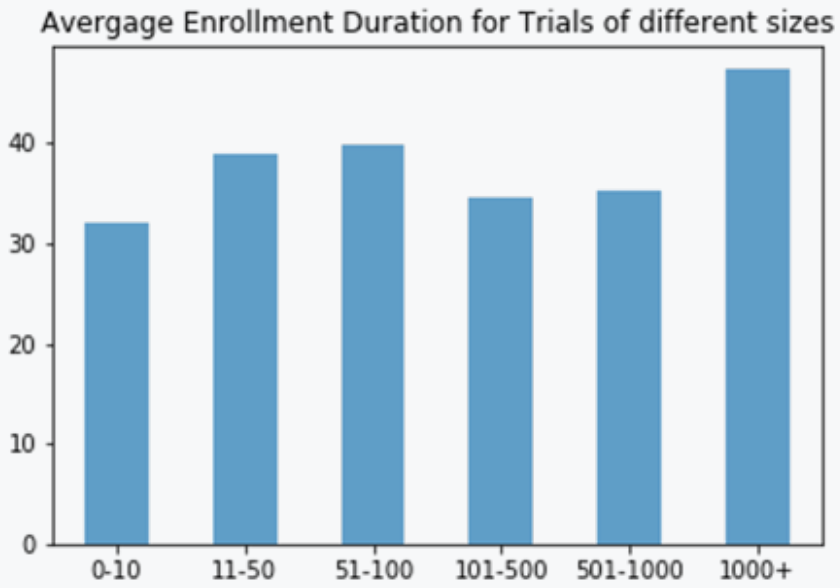
Clinical Trials Data - Duration

- Created new attribute “EnrollmentDuration” from StartDate and CompletionDate in number of months

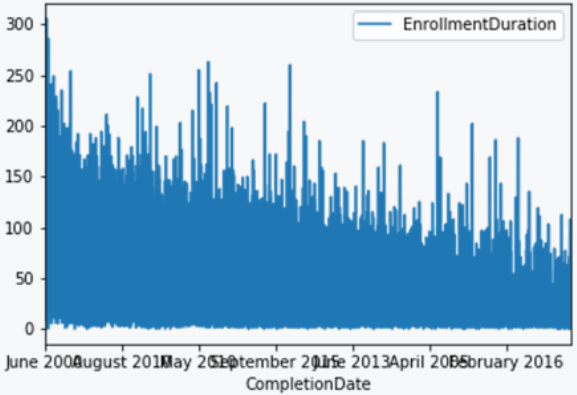


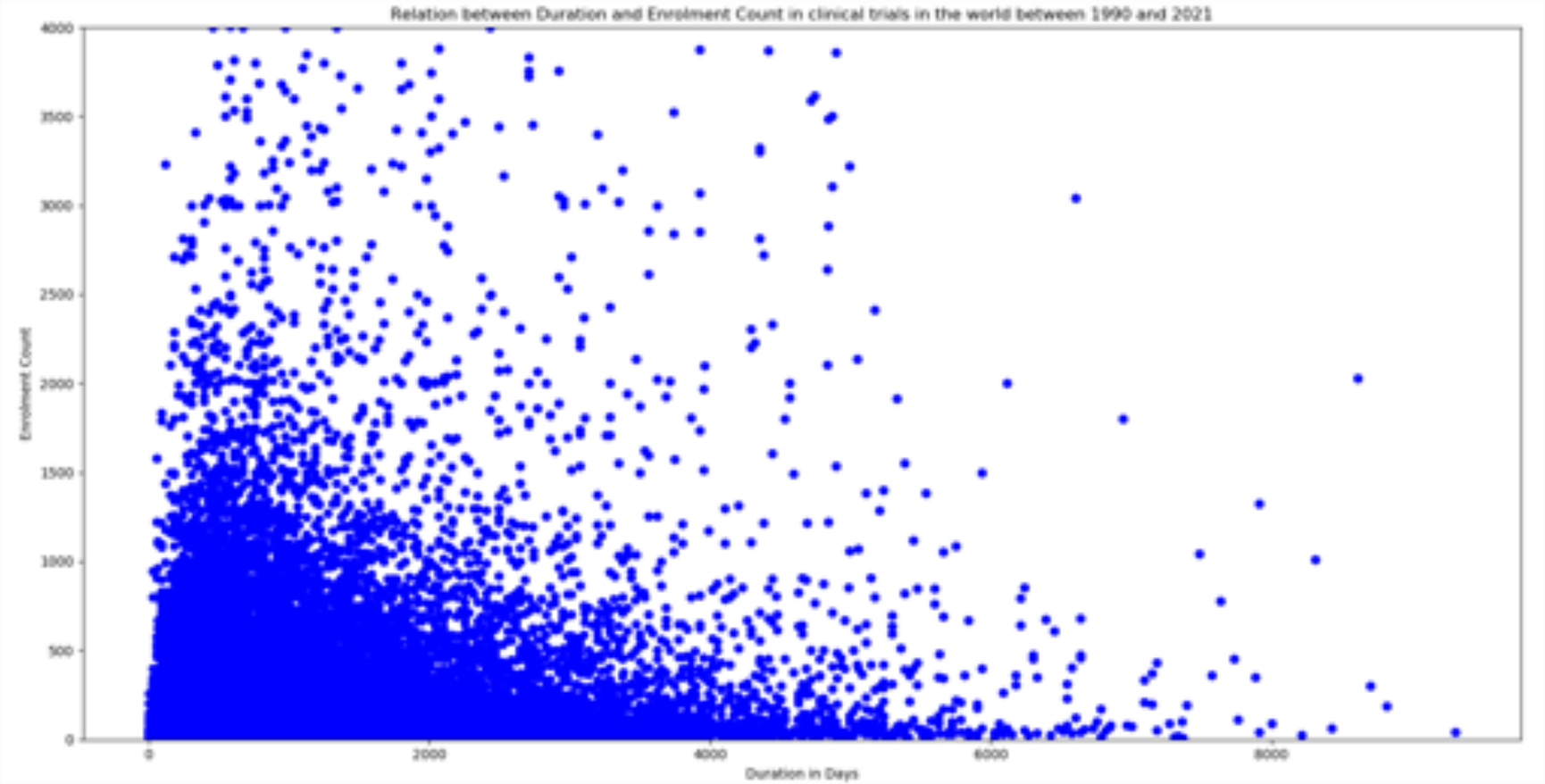

```
Quickest Duration: 1
Longest Duration: 306
Mean Duration: 36.649176050901154 Months
```

Average Duration for different Trial sizes

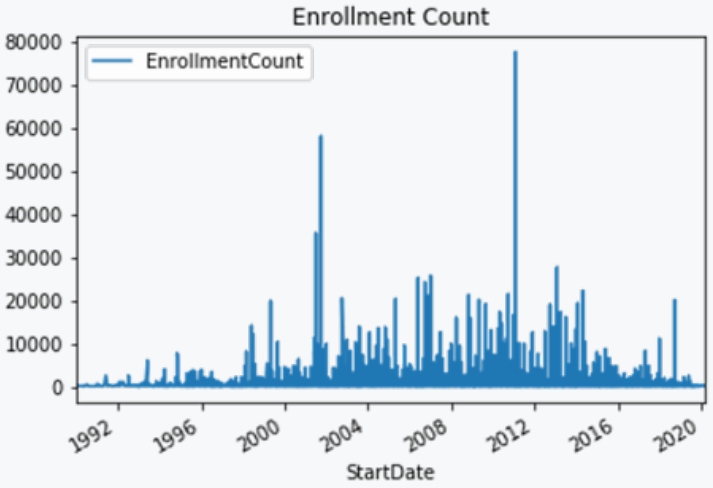


- only a slight tendency towards longer enrollment times for bigger studies




Relation between Duration and Enrolment Count in clinical trials in the world between 1990 and 2021
Enrolment Count
Duration in Days
0
500
1000
1500
2000
2500
3000
3500
4000
0
2000
4000
6000
8000

Analysing the Patient Enrollment on a time axis



April 19	Denmark	100
	Egypt	161
	Iran, Islamic Republic of	40
	Japan	254
	New Zealand	24
	Syrian Arab Republic	66
	United States	2567
Mai 19	France	94
	Mexico	160
	Portugal	66
	United States	2073

Juni 19	China	50
	Egypt	121
	Japan	81
	Mexico	13
	Netherlands	103
	Taiwan	472
	United Kingdom	37
	United States	3253
Juli 19	Pakistan	92
	Tanzania	399
	United States	1222

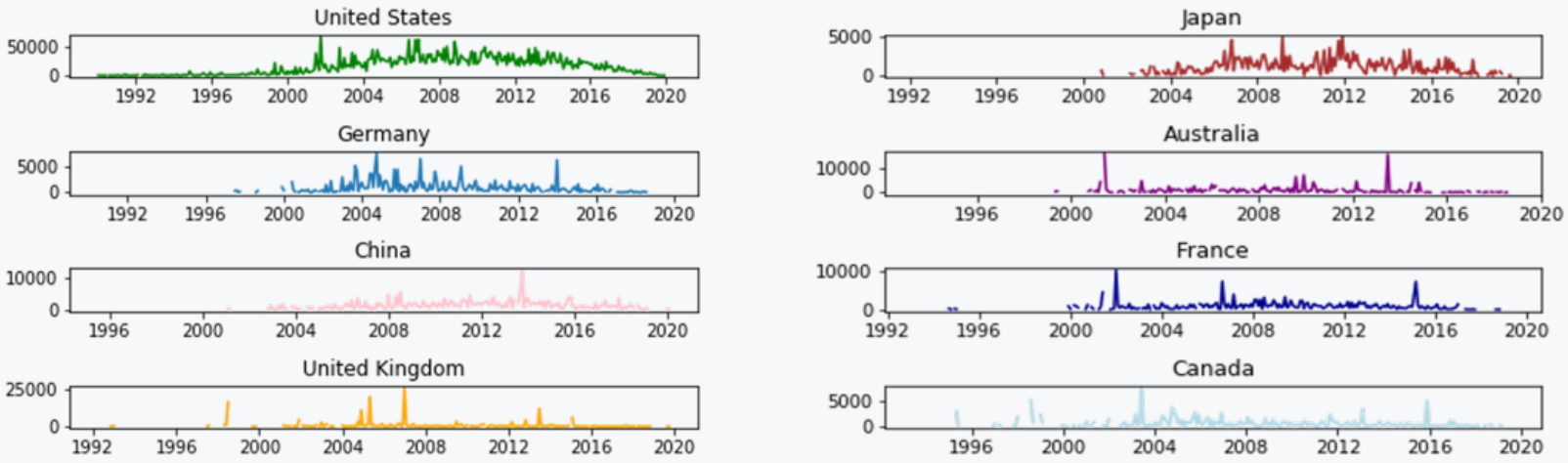
Analysing the Patient Enrollment on a time axis

Take the countries with the most enrolled patients for visualization

LocationCountry	EnrollmentCount
United States	5574727
China	272808
Japan	247079
Germany	214296
Australia	202589
France	201350
United Kingdom	200332
Austria	163626
Argentina	152104
Canada	145292

Analysing the Patient Enrollment on a time axis

Choose countries of interest



Problems/ideas to improve:

1) how to handle studies with several countries
2) how to show duration of enrollment

Hospital Data

- The Google Places API
- Wikidata

Hospital Data - Wikidata

- Enhanced hospital database with additional attributes from wikidata by comparing Name and Link of hospitals
- Inserted additional hospitals from wikidata to enlarge hospital dataset

Hospital Data: Duplication check - trials facility

Remove digits, and abnormal characters: '\(.*\)|\(|\)|\$|#|\d+|\'|\/|ID|"|%|@|!|\^|:|>|<|、 |-|_|\•|*|\.\.+|;|\+|^,|^\.|^A\s|^B\s|^C\s|^D\s|^E\s|^F\s|^G\s|^H\s|^I\s|^J\s|^K\s|^L\s|^M\s|^\s|^\s*|\s+$

Lower cases, remove all spaces and . while comparing, restore the whole name afterwards

SPSK nr 1
SPSK Nr 1 im. Prof. Tadeusza Sokolowskiego
SPSK Nr 1 im. Prof. Tadeusza Sokolowskiego Pomorskiej Akademii Medycznej Klinika Gastroenterologii i Chorob Wewnetrznych
SPSK nr 2 Pomorskiej Akademii Medycznej w Szczecinie
SPSK nr 4, Klinika Ortopedii, Traumatologii i Rehabilitacji AM
SPSK NR 4; Reumatologii i Ukladowych Chorob Tkanki
SPSK Nr 7 Klaskiego Uniwersytetu Medycznego w Katowicach, Gornoslaskie Centrum Medyczne im. Prof. Leszka Gieca, III Oddzial Kardiologii, Zkla
SPSK Nr 7 Slaskiego Uniwersytetu Medycznego w Katowicach, Gornoslaskie Centrum Medyczne im. Prof. Leszaka Gieca, I Oddzial Kardiologii, ul. Z
SPSK nr.1 im. Prof. T. Sokołowskiego PAM Szczecin
SPSK, Klinika Hematologii Akademii Medycznej

Hospital Data: Duplication check with trials facility

Spelling errors?!
Abbreviations problems too much. I could just found the most likely appear ones.
e.g. Ctr.|Ctr|ctr to center, dept.|Dept.|dept|Dept to department

```
'C.I.C. Maurice Inc.',
'C.I.C. Mauricie Inc.',
```

IASO General Hospital of Athens
IASO General Hospital of Athens Athens, Greece
IASO General Hospital of Athens, st Dep of Medical Oncology
IASO General Hospital of Athens, st department of Medical Oncology
IASO General Hospital of Athnes, Dep of Medical Oncology

Reduced from 177k to 107k

Still, 107956 has left

Hospital Data: Google API script improvement

Example Search: Columbia University

Google retrieves:
Columbia Central University, 8am - 17 pm, 68163
Columbia University, 8am - 17 pm, 68168
Columbia University, 8am - 17 pm, 68163
Columbia University, 8am - 17 pm, 68163

Case:
If more than one retrieval for "Columbia University":
** Priorities to retrieve one result only: (Nested look)
1) Compare zip codes → if there are more than one same zip code or no zip code → Compare names
2) Choose from above all with the higher user_ratings_totals_total

Hospital Data: Check matches of hospitals and facilities

Method:
Some data look like: Shiley Eye Center, 0946, University of California
→ if it matches anyone in hospital collection, we say it's match and it works well
(still running to get the results)

```
for i in range(len(unique_facilities["0"])):
    facilities_list = unique_facilities["0"][i].split(',')
    #Returns collection where the field contains the element __ or __
    #e.g. if the name is: Shiley Eye Center Center or 0946 or University of California
    found_match = (hospitalsCollection.find( { 'Name': { '$in': facilities_list} } ))
```

Next step:
Since we already have retrieved Google API in hospital collection, the more API efficient way is to deduct the collection that google API already found and retrieve further for the Trial collection

Country Data

- Merging two sources (Wikidata and World Bank API)
- Prepare for additional attributes

Country Data

	Name String	Code String	Population String	PopulationDate String	LifeExpectancy Mixed	CapitalCity Mixed	Continent Mixed
1	"Afghanistan"	"AFG"	"34940835"	"2018-01-01T00:00:00Z"	null	null	null
2	"Algeria"	"DZA"	"41318142"	"2017-01-01T00:00:00Z"	null	null	null
3	"American Samoa"	""	"55689"	"2018-01-01T00:00:00Z"	null	null	null
4	"Angola"	"AGO"	"29784193"	"2017-01-01T00:00:00Z"	null	null	null
5	"Antigua and Barbuda"	"ATG"	"103812"	"2017-01-01T00:00:00Z"	null	null	null
6	"Argentina"	"ARG"	"44938712"	"2019-01-01T00:00:00Z"	null	null	null
7	"Australia"	"AUS"	"24511800"	"2017-03-01T00:00:00Z"	"82.5"	"Canberra"	"Oceania"
8	"Austria"	"AUT"	"8840212"	"2017-01-01T00:00:00Z"	"80.69024"	"Vienna"	"Europe"
9	"The Bahamas"	""	"395361"	"2017-01-01T00:00:00Z"	null	null	null
10	"Bahrain"	"BHR"	"1492184"	"2017-01-01T00:00:00Z"	"76.9"	"Manama"	"Asia"
11	"Bangladesh"	"BGD"	"164689752"	"2017-01-01T00:00:00Z"	null	null	null
12	"Barbados"	"BRB"	"285719"	"2017-01-01T00:00:00Z"	null	null	null
13	"Benin"	"BEN"	"11175692"	"2017-01-01T00:00:00Z"	null	null	null
14	"Bhutan"	"BTN"	"807610"	"2017-01-01T00:00:00Z"	null	null	null
15	"Bolivia"	""	"11051600"	"2017-01-01T00:00:00Z"	null	null	null
16	"Botswana"	"BWA"	"2291661"	"2017-01-01T00:00:00Z"	null	null	null
17	"Brazil"	"BRA"	"208494900"	"2018-01-01T00:00:00Z"	null	null	null
18	"Brunei"	""	"428697"	"2017-01-01T00:00:00Z"	null	null	null
19	"Bulgaria"	"BGR"	"6951482"	"2019-12-31T00:00:00Z"	null	null	null
20	"Burkina Faso"	"BFA"	"19193382"	"2017-01-01T00:00:00Z"	null	null	null

Wikidata

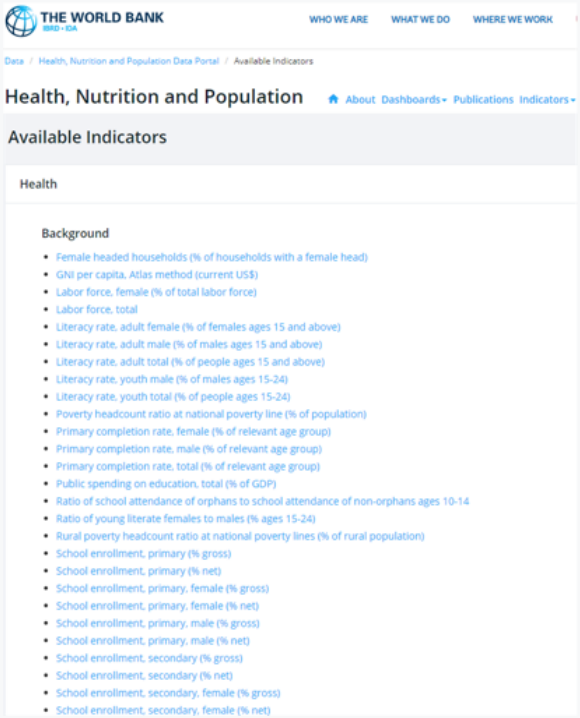
Props:

- Have basic countries' attributes
- Some attributes have up-to-date data

Cons:

- A lot of missing data

Country Data



Providing 16000 time-series indicators =>1000 health-related

=> Consider the wikidata set as the foundation

=> Use World Bank API to fill missing data (Only take latest values)

All features come along with their year

```
MongoDB Enterprise ClinicalTrials-shard-0:PRIMARY> db.country.find({'countryCode': 'DEU'}).pretty()
{
        "_id" : ObjectId("5ea1f0f4804f487dacf28ea9"),
        "countryCode" : "DEU",
        "countryName" : "Germany",
        "capitalCity" : "Berlin",
        "population" : 83149300,
        "populationYear" : 2019,
        "lifeExpectancy" : 80,
        "lifeExpectancyYear" : 2018,
        "GDP" : NumberLong("3947620162502"),
        "GDPYear" : 2018,
        "unemploymentRate" : 3,
        "unemploymentRateYear" : 2019,
        "hospitalBed" : 8,
        "hospitalBedYear" : 2013,
        "healthExpenditure" : 11,
        "healthExpenditureYear" : 2017
}
```

https://datatopics.worldbank.org/health/available-indicators

Country Data

Created a script for getting any new country attributes from the list of World Bank indicators

```python

try:
    # Request population data
    resp = requests.get(url=URL, params=payload)
    if resp.status_code == 200:
        data = resp.json()
        for i in range(len(data[1])):

            if data[1][i]['value'] not in (None, ''):
                # Check with country list
                checkCountry = country.find({'countryCode': data[1][i]['countryiso3code']})
                # If this country exists in country database
                if checkCountry.count() > 0:

                    # Check with current list
                    for item in result:
                        if data[1][i]['countryiso3code'] == item['countryCode']:
                            if int(data[1][i]['date']) > item['year']:
                                item['year'] = int(data[1][i]['date'])
                                item['value'] = int(data[1][i]['value'])
                            break
                    else:
                        print('Inserting new country: {}'.format(data[1][i]['countryiso3code']))
                        schema = {
                            'countryCode': data[1][i]['countryiso3code'],
                            'year': int(data[1][i]['date']),
                            'value': int(data[1][i]['value'])
                        }
                        result.append(schema)
```

Next Steps

Next Steps

- Get the facility type of a trial
- Retrieve information about companies having clinical trials
- Add Chinese studies