=====================================================================

**LAPORAN MILESTONE 2:**- Eksplorasi dengan Materi Statistik yang telah diajarkan pada minggu terkait  
- Research Question  
Submit pada kolom tugas di sidebar kanan

=====================================================================

**Kelas BIG DATA A, Kelompok 3 :**

1. Bagus Kharisma Nusantara
2. Bakhtiar Amaludin
3. Darmawan Nur Kusuma
4. Eman Sukmana
5. Febrina Puspa S

**Github Repository:** [**https://github.com/ProjectA3BD/milestone**](https://github.com/ProjectA3BD/milestone)

**JUDUL PROJECT**

**Pengaruh Citra Pemberitaan Media Internasional terhadap Indeks Ekonomi Negara ASEAN**

**RINGKASAN PROJECT**

Terdapat dua macam data dalam project ini, yaitu data pemberitaan media online yang diperoleh dari GDELT ([www.gdeltproject.org](http://www.gdeltproject.org)) serta data indeks ekonomi yang diperoleh dari dataset World Development Indikator (WDI) yang disediakan oleh World Bank ([www.worldbank.org](http://www.worldbank.org)). Data yang diperoleh dari situs GDELT membutuhkan beberapa perlakuan (Data Cleansing Process) sebelum siap diolah secara statistik bersama dengan data indeks ekonomi beberapa Negara ASEAN. Periode penelitian selama 15 tahun, yakni sejak 2003 hingga 2017.

**EKSPLORASI STATISTIK**

1. **Data Preparation and Preprocessing**
2. Gambaran Umum Data

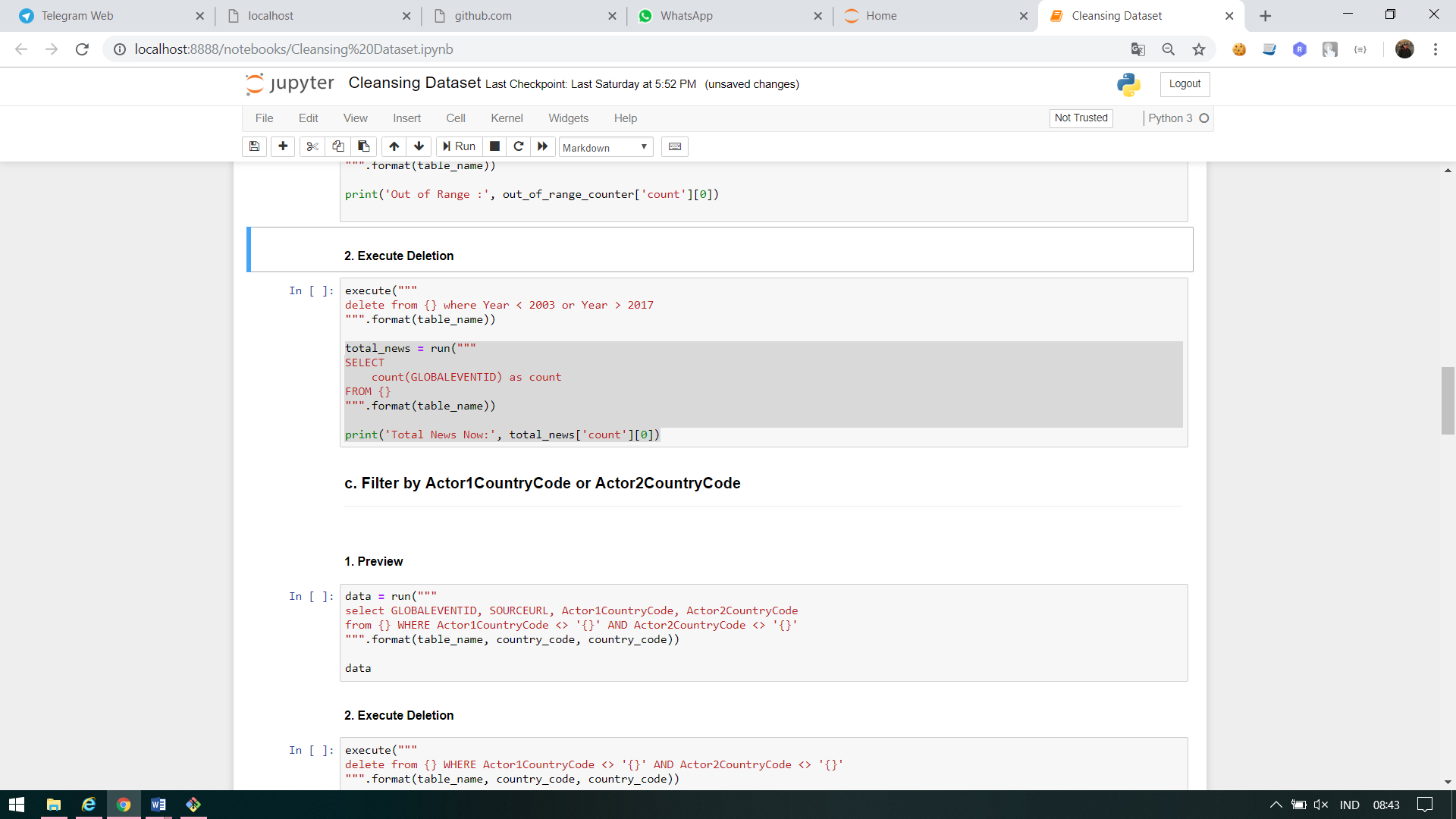
Dalam rangka melakukan ekslorasi statistik, maka persiapan dan pemrosesan awal data dilakukan sebagai tindak lanjut hasil eksplorasi pada milestone 1. Terdapat beberapa hal penting **khususnya pada data GDELT** yang kami konfirmasi juga dengan beberapa artikel yang memuat permasalahan yang sama yakni sebagai berikut:

|  |  |  |
| --- | --- | --- |
| No | Uraian | Sumber Pendukung |
| 1 | Tidak terdapat SOURCEURL untuk data berita pada tahun 2013 ke bawah. *(missing value)* | GDELT Event Database Data Format Codebook |
| 2 | Terdapat banyak data berita dengan SOURCEURL yang redundan. Sementara untuk tahun 2013 ke bawah berita yang redundan tidak dapa diketahui karena kolom SOURCEURL tidak tersedia. | Wang, Wei (2017) |
| 3 | Terdapat SOURCEURL dengan yang menunjukkan alamat berita yang bersumber bukan dari URL (http/https). | Wang, Wei (2017) |
| 4 | Berita dengan sentimen (AvgTone) negatif hanya ada di tahun 2013 ke atas, sehingga hampir seluruh sentimen berita pada tahun 2013 ke bawah memiliki nilai AvgTone positif. | Haren, Juul Van (2017) |
| 5 | Terdapat perbedaan range penghitungan AvgTone untuk berita tahun 2013 ke bawah dan tahun 2013 ke atas. Dalam hal ini nilai AvgTone rata-rata untuk tahun 2013 ke atas dihitung lebih rendah dibandingkan dengan tahun 2013 ke bawah. | Lammers, Loes (2017)  Haren, Juul Van (2017) |

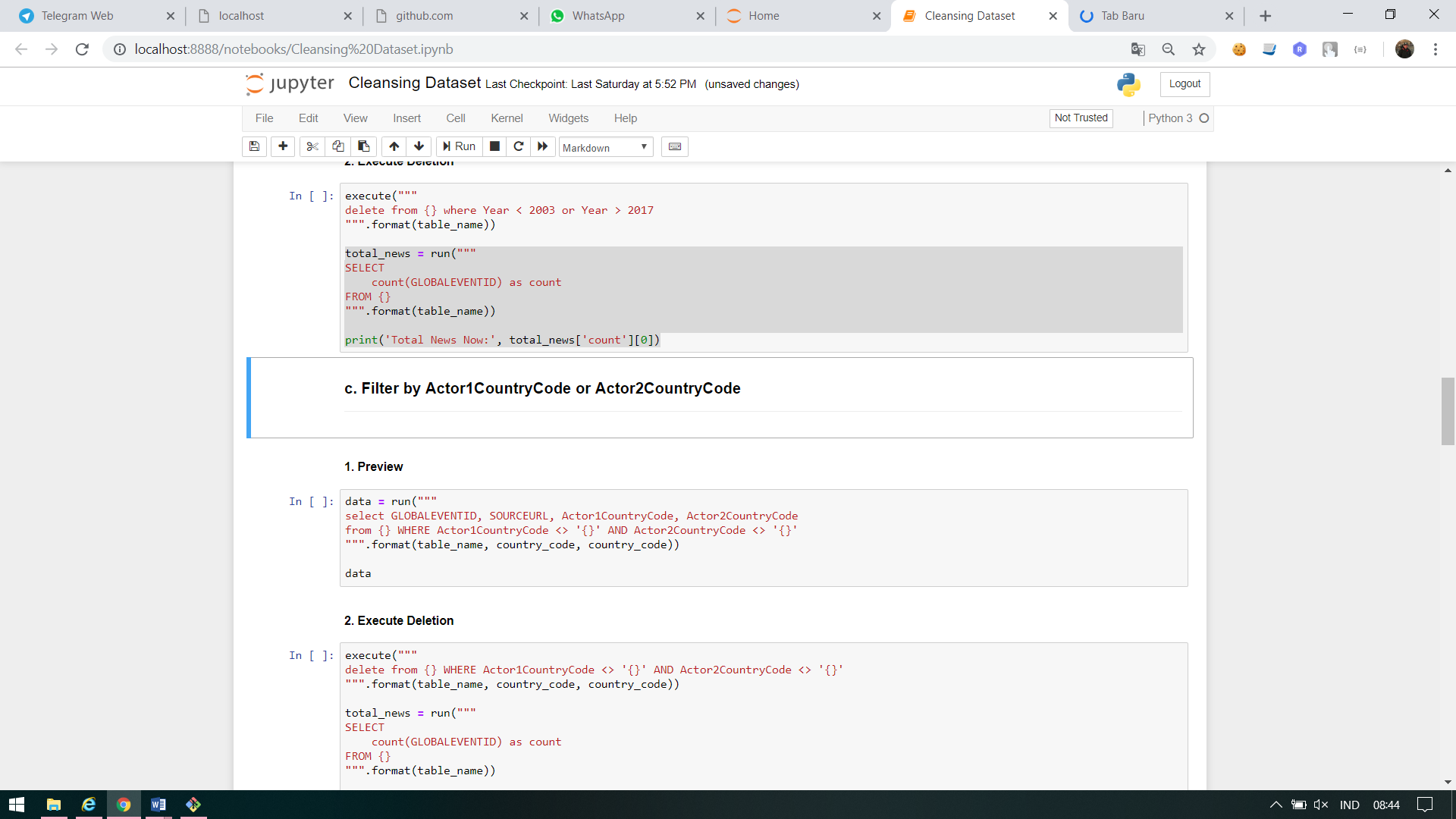
1. Data Filtering

Pada milestone 1, pemerolehan data dilakukan dengan mengambil seluruh data dari website GDELT berupa data mendah *(raw data)* dengan kriteria filter semua berita yang melibatkan negara sample dari tahun 2003 – 2018. Untuk menyesuaikan dengan scope ini perlu dilaukan penyaringan data lebih lanjut yakni dengan hal berikut:

1. memastikan range data berapa dalam kurun waktu 2002 – 2017 (Filter 1)



1. memastikan Actor1 atau Actor2 dari berita yang ada terkait langsung dengan negara sample (Filter 2).



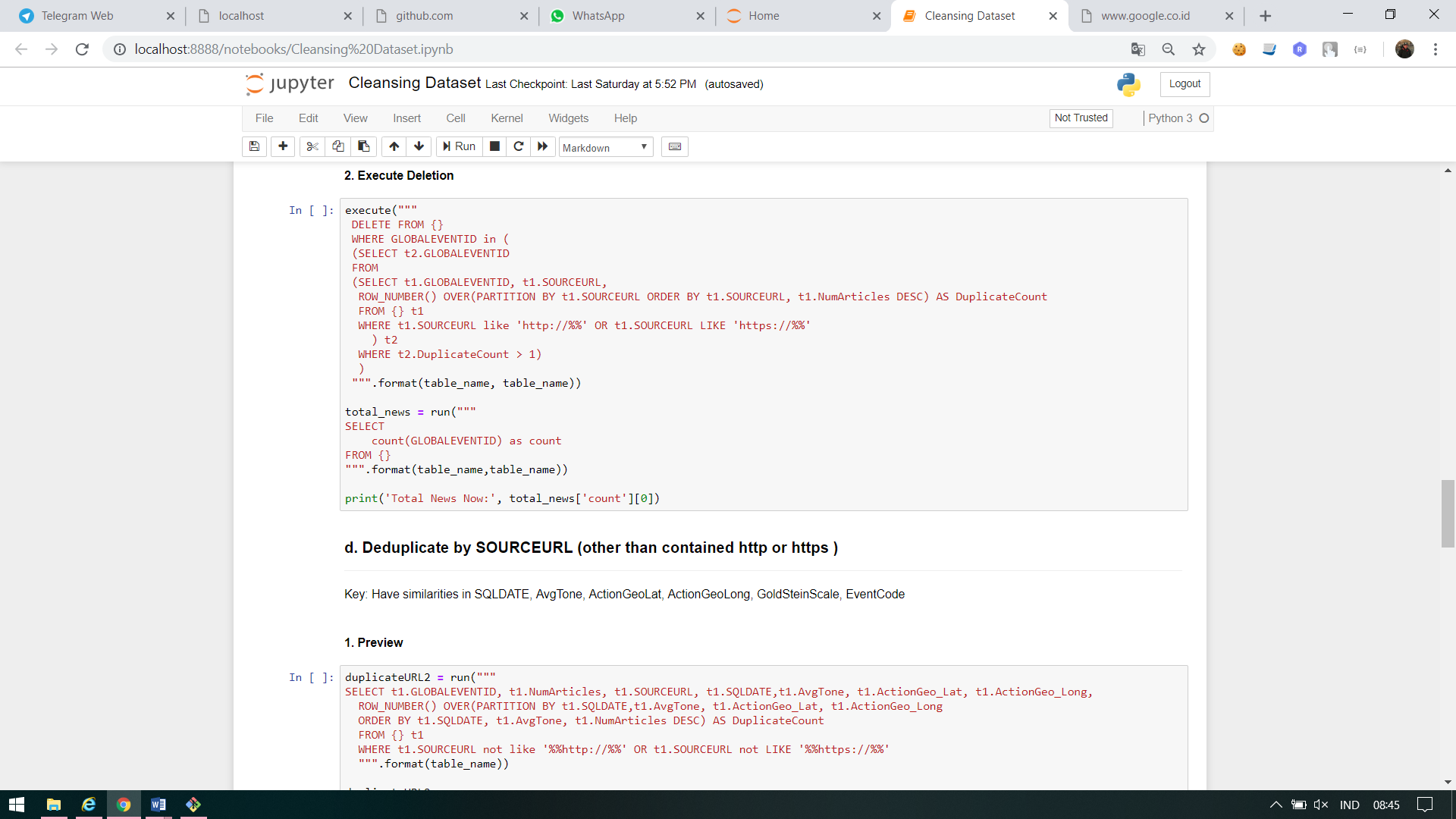
Berikut ini adalah rekap hasil filtering data dimaksud:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Jumlah Berita GDELT | Negara | | | | |
| Indonesia | Malaysia | Thailand | Singapura | Filipina |
| Jumlah Berita Awal. | 4.336.756 | 4.184.703 | 3.019.178 | 1.913.966 | 6.674.871 |
| Berita Setelah Filter 1 | 3.426.161 | 3.802.732 | 2.807.155 | 1.628.075 | 6.031.074 |
| Berita Setelah Filter 2 | 2.123.411 | 2.249.577 | 1.655.186 | 872.095 | 3.784.568 |

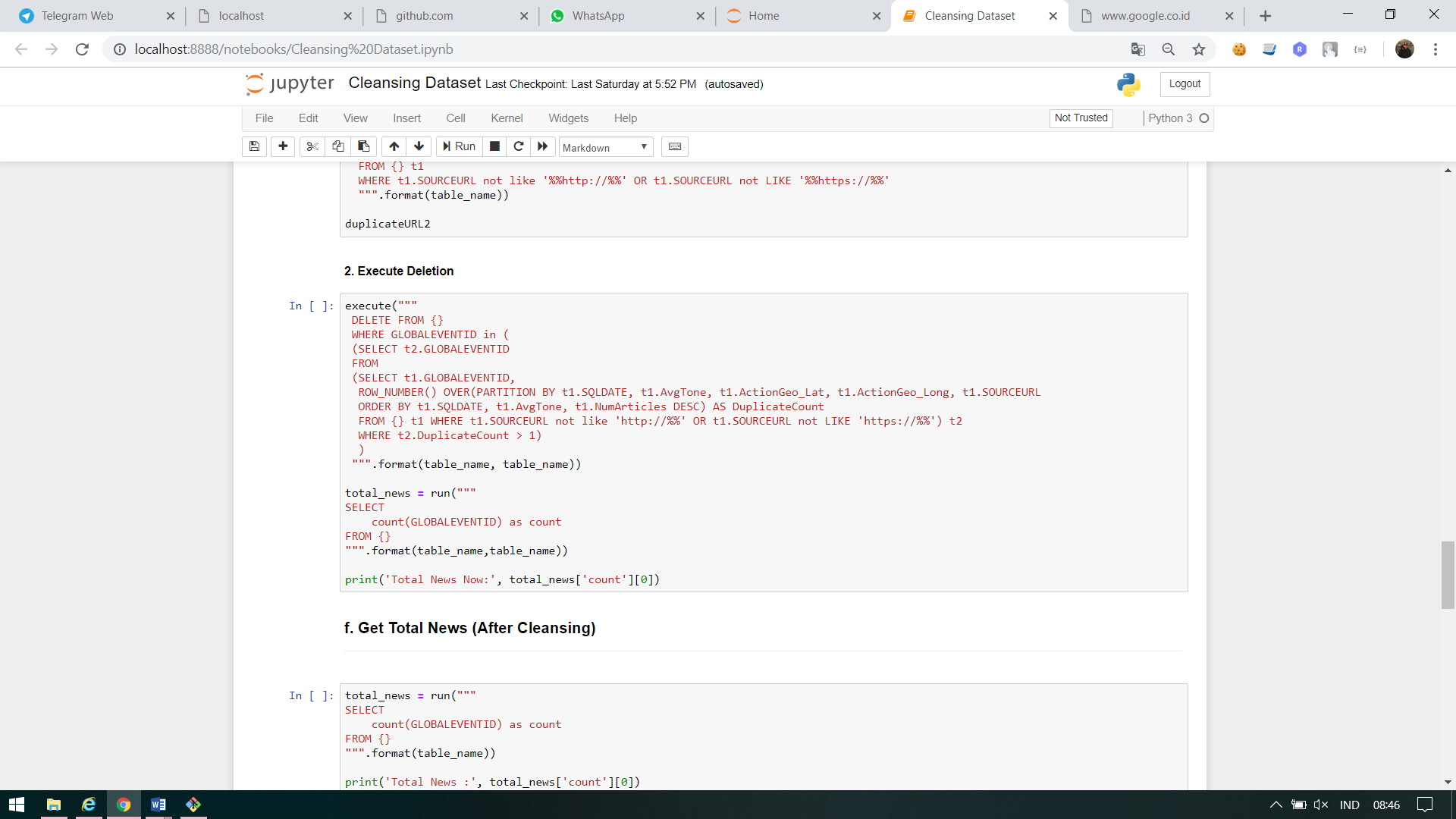
1. Deduplikasi Data

Sesuai dengan permasalahan pada data mentah GDELT yang telah diurakan di bagian sebelumnya yakni terkait dengan adanya duplikasi data, dilakukan beberapa langkah penanganan yakni:

1. untuk data berita tahun 2013 ke atas, deduplikasi dilakuan dengan melihat nilai kolom SOURCEURL yang sama. (Deduplikasi 1)



1. untuk data berita tahun 2013 ke bawah, deduplikasi dilakuan dengan melihat kesamaan pada beberapa field yakni kesamaan tanggal, kesamaan lokasi, kesamaan nilai pada SQLDATE, AvgTone, ActionGeoLat, ActionGeoLong, GoldSteinScale, EventCode. (Deduplikasi 2)



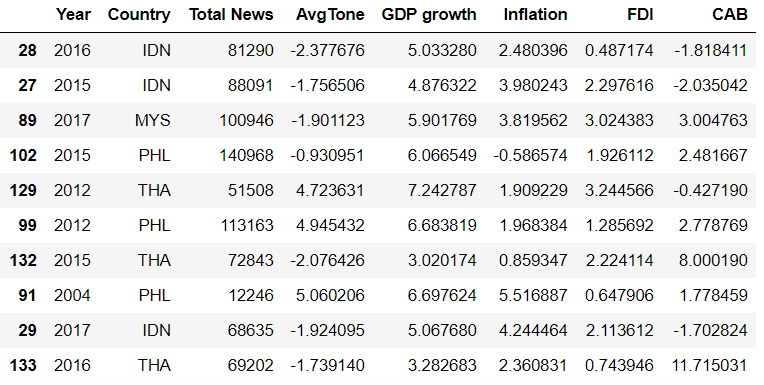
Berikut ini adalah rekap hasil deduplikasi data dimaksud:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Jumlah Berita GDELT | Negara | | | | |
| Indonesia | Malaysia | Thailand | Singapura | Filipina |
| Berita Setelah Filtering. | 2.123.411 | 2.249.577 | 1.655.186 | 872.095 | 3.784.568 |
| Berita Setelah Deduplikasi 1 | 1.489.281 | 1.244.157 | 1.039.945 | 516.507 | 1.934.451 |
| Berita Setelah Deduplikasi 2 | 1.249.137 | 866.895 | 628.365 | 370.718 | 1.247.234 |

1. Penyusunan dataset

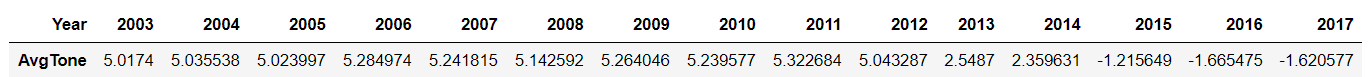
Setelah semua data mentah GDELT yang disimpan dalam MySQL dilakukan *cleansing*  dan filtering, selanjutnya berdasarkan data tersebut kami menyusun rekapitulasi data per tahun dan kami simpan dalam sebuah dataset yang digunakan untuk analisis lanjutan. Dataset tersebut lalu digabungkan dengan data indikator pada WDI dan disimpan dalam file csv agar mempermudah analisis datanya (*script* python tersedia di repositor github proyek ini).

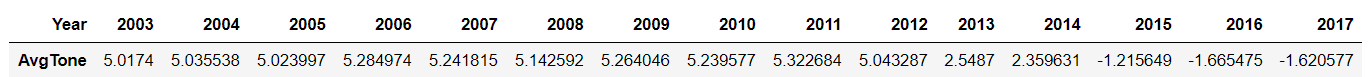
Preview dari dataset tersebut adalah:



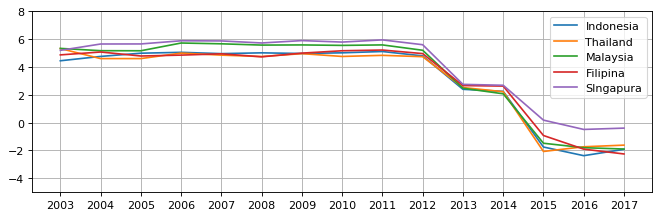
1. Normalisasi Data

Sebagaimana diuraikan sebelumnya, terdapat isu mengenai perbedaan range penghitungan AvgTone pada data GDELT untuk berita tahun 2013 ke bawah dan tahun 2013 ke atas dimana nilai AvgTone rata-rata untuk tahun 2013 ke atas dihitung lebih rendah dibandingkan dengan tahun 2013 ke bawah. Hal tersebut dapat dilihat dari tabel dan grafik dibawah ini:

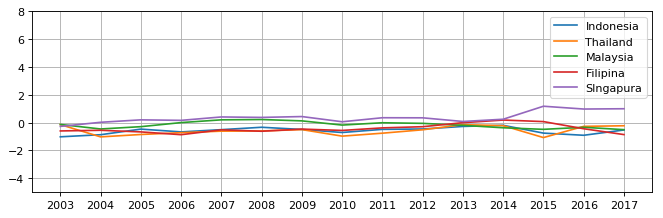




Untuk mengatasi hal ini, kami melakuan transformasi data dengan cara melakukan pengurangan mean *(mean substraction)* data AvgTone setiap tahunnya. Pertama, kami menghitung *grand mean* AvgTone per tahun untuk seluruh negara lalu hasilnya digunakan untuk mengurangi nilai AvgTone per tahun masing-masing negara (Mt). Adapun hasil dari transformasi data tersebut tergambar dalam grafik di bawah ini:



*AvgTone* per negara tahun 2013-2017



*AvgTone* setelah transformasi

1. **Statistik Deskriptif**
2. Olah Data Deskriptif

Data yang telah melalui proses Data Preparation diolah lebih lanjut untuk melihat statistik deskriptifnya. Pada notebook, digunakan library scope.describe untuk menunjukkan jumlah n, nilai mean, standar deviasi, mean, min, maks, Q1, Q2, dan Q3.

1. Visualisasi

Visualisasi yang akan ditampilkan berupa histogram dan plot yang diperoleh dengan menggunakan library .

1. Outliers

Saat visualisasi data, akan terlihat ada data yang terpencil atau berjarak sangat jauh dari data lain. Ada beberapa pilihan perlakuan yang akan diterapkan pada data ini. Data ini nantinya akan dianalisis lebih lanjut

1. **Statistik Inferensial**
2. Analisis Korelasi (Pearson)

Tujuan dilakukan analisis ini adalah untuk melihat kebersamaan antara variabel sentimen pemberitaan dan variabel indikator ekonomi. Pada negara yang tidak ditemukan kebersamaan (atau nilai korelasi mendekati dan sama dengan 0) tidak akan dianalisis lebih lanjut. Range nilai pada analisis ini adalah -1 hingga 1. Jika diperoleh nilai 0, maka disimpulkan bahwa tidak ada hubungan atau korelasi antar variabel yang diuji. Uji yang digunakan adalah Korelasi Pearson. Library Notebook Jupiter yang digunakan pada uji ini adalah scope.corr, dengan output terlihat seperti berikut.

1. Analisis Regresi

Jika terdapat kebersamaan antara variabel citra pemberitaan online internasional dengan indeks ekonomi Negara ASEAN, maka akan dicari bagaimana bentuk kebersamaan tersebut. Atau dapat dibentuk sebuah model persamaan.

**RESEARCH QUESTION**

1. Bagaimana perlakuan yang dilakukan pada data GDELT yaitu variabel citra pemberitaan (variabel x) sebelum dilakukan uji statistik?
2. Statistik Deskriptif apa yang akan ditampilkan sebagai analisis statistik awal?
3. Apakah ada data yang hilang? Bagaimana perlakuannya?
4. Apakah data terdistribusi normal, jika tidak bagaimana cara menormalkannya?
5. Adakah korelasi antara citra pemberitaan internasional di GDELT dengan indeks ekonomi di tiap Negara ASEAN? Bagaimana bentuknya?
6. Bagaimana model regresi citra pemberitaan dan indeks ekonominya?