Projekt Statystyka

January 20, 2020

1 Analiza zależności między emisją CO2 i średnią roczną temperaturą

Dane zostały pobrane ze strony keggle.com CO2 and GHG emission data Climate Change: Earth Surface Temperature Data

Następnie zostały przerzucone z formatu .csv do bazy SQLite. Wstępne oczyszczanie danych nie jest w tym zbiorze potrzebne.

W celu analizy danych używamy środowiska `jupyter` oraz bibliotek: `pandas`, `seaborn`, `sqlite3`, `matplotlib`

1.1 Wczytywanie danych

[1]:
```
import pandas as pd
import numpy as np
pd.plotting.register_matplotlib_converters()
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
import sqlite3
import warnings
warnings.filterwarnings('ignore')
print("Setup Complete")
```

```
Setup Complete
```

[2]:
```
!ls
```

```
Untitled.ipynb
Untitled.pdf
climate-change-earth-surface-temperature-data
emission data.csv
emission_to_sql.py
podział_pracy.pdf
project_data.db
tempretures_to_sql.py
```

[3]:
```
emission_data = None
global_land_temperatures_by_city = None
global_land_temperatures_by_country = None
global_land_temperatures_by_major_city = None
global_land_temperatures_by_state = None
global_temperatures = None

with sqlite3.connect('project_data.db') as conn:
    emission_data = pd.read_sql('select * from emissions', conn)
    global_temperatures = pd.read_sql('select * from GlobalTemperatures', conn)
```

Zmienna `emission_data` zawiera `pandas.DataFrame` zawierająca dane o emisji CO2 przez każdy kraj w latach `1751 - 2017`

[4]:
```
emission_data.head()
```

[4]:

	Country	1751	1752	1753	1754	1755	1756	1757	1758	1759	\
0	Afghanistan	0	0	0	0	0	0	0	0	0	
1	Africa	0	0	0	0	0	0	0	0	0	
2	Albania	0	0	0	0	0	0	0	0	0	
3	Algeria	0	0	0	0	0	0	0	0	0	
4	Americas (other)	0	0	0	0	0	0	0	0	0	

	…	2008	2009	2010	2011	2012	\
0	…	8.515264e+07	9.191295e+07	1.003652e+08	1.125912e+08	1.233332e+08	
1	…	3.183077e+10	3.301904e+10	3.421283e+10	3.541120e+10	3.664504e+10	
2	…	2.287948e+08	2.331696e+08	2.377643e+08	2.430001e+08	2.479062e+08	
3	…	2.894820e+09	3.015005e+09	3.132819e+09	3.252626e+09	3.380736e+09	
4	…	7.746025e+10	7.961787e+10	8.187178e+10	8.416656e+10	8.654197e+10	

	2013	2014	2015	2016	2017
0	1.333337e+08	1.431228e+08	1.532303e+08	1.654882e+08	1.785029e+08
1	3.789569e+10	3.918617e+10	4.047518e+10	4.178583e+10	4.311757e+10
2	2.529662e+08	2.586784e+08	2.646261e+08	2.708990e+08	2.772782e+08
3	3.513171e+09	3.656348e+09	3.806940e+09	3.957319e+09	4.107870e+09
4	8.894874e+10	9.139192e+10	9.382747e+10	9.624253e+10	9.864116e+10

[5 rows x 268 columns]

Zmienna `global_temperatures` zawiera średnie miesięczne wartości pomiarów temperatury w latach `1750 - 2015` wraz z niepewnościami.

[5]:
```
global_temperatures.head()
```

[5]:

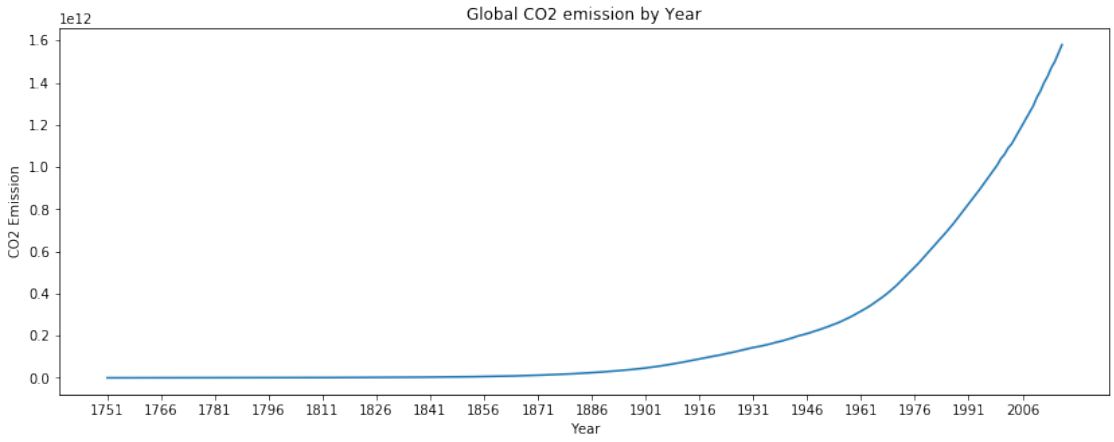
	dt	LandAverageTemperature	LandAverageTemperatureUncertainty	\
0	1750-01-01	3.034	3.574	
1	1750-02-01	3.083	3.702	
2	1750-03-01	5.626	3.076	

```
3  1750-04-01                   8.490                                2.451
4  1750-05-01                  11.573                                2.072

   LandMaxTemperature  LandMaxTemperatureUncertainty  LandMinTemperature  \
0                 NaN                            NaN                 NaN
1                 NaN                            NaN                 NaN
2                 NaN                            NaN                 NaN
3                 NaN                            NaN                 NaN
4                 NaN                            NaN                 NaN

   LandMinTemperatureUncertainty  LandAndOceanAverageTemperature  \
0                            NaN                             NaN
1                            NaN                             NaN
2                            NaN                             NaN
3                            NaN                             NaN
4                            NaN                             NaN

   LandAndOceanAverageTemperatureUncertainty
0                                        NaN
1                                        NaN
2                                        NaN
3                                        NaN
4                                        NaN
```

1.2 Emisja CO2

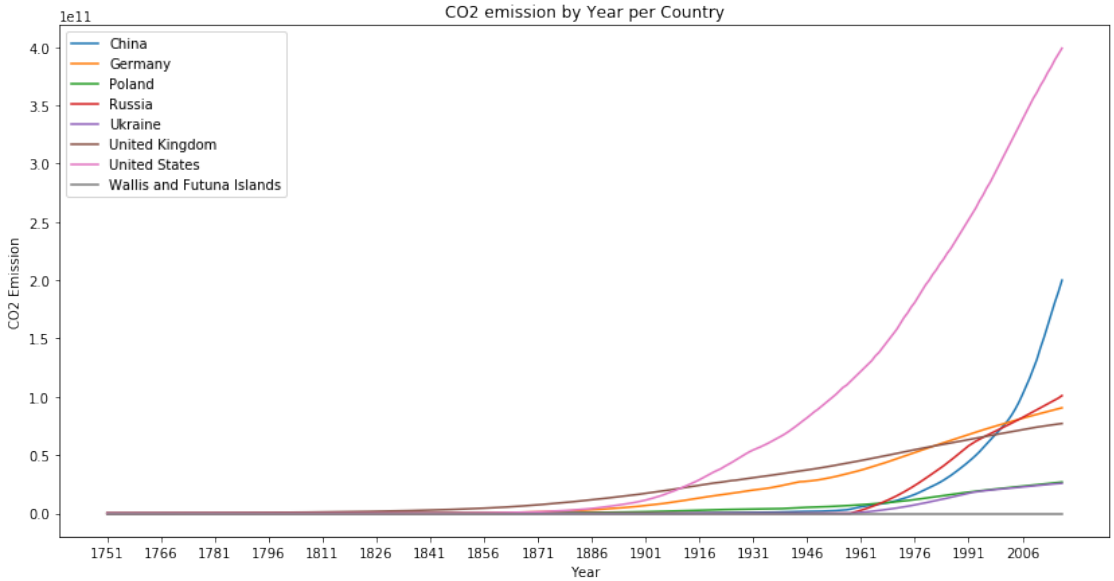
Poniższy wykres przedstawia emisję CO2 na świecie. Możemy zauważyć wyraźny wzrost tego wskaźnika z upływem czasu

[6]:
```
years = emission_data.columns[1:]

plt.figure(figsize=(14, 5))
plt.title('Global CO2 emission by Year')
sns.lineplot(x=years, y=np.squeeze(emission_data[emission_data.Country.
 ↪isin(["World"])].values)[1:].astype(np.int64))
plt.xlabel('Year')
plt.ylabel('CO2 Emission')
t = plt.xticks(years[::15])
```



Krótko możemy przeanalizować największe czynniki powodujące taki drastyczny wzrost.

[7]:
```
plt.figure(figsize=(14, 7))
plt.title('CO2 emission by Year per Country')
country_list = ["United States", "China", "Russia", "Germany", "United␣
 ↪Kingdom", "Poland", "Ukraine", "Wallis and Futuna Islands"]
for country_data in emission_data[emission_data.Country.isin(country_list)].
 ↪values:
    sns.lineplot(x=years, y=country_data[1:].astype(np.int64), legend='brief',␣
 ↪label=country_data[0])

plt.xlabel('Year')
plt.ylabel('CO2 Emission')
t = plt.xticks(years[::15])
```



Jak możemy zauważyć na powyższym wykresie głównymi emiterami **CO2** są państwa wysokorozwinięte, może świadczy o zależności między emisją **CO2** a rozwojem państwa?

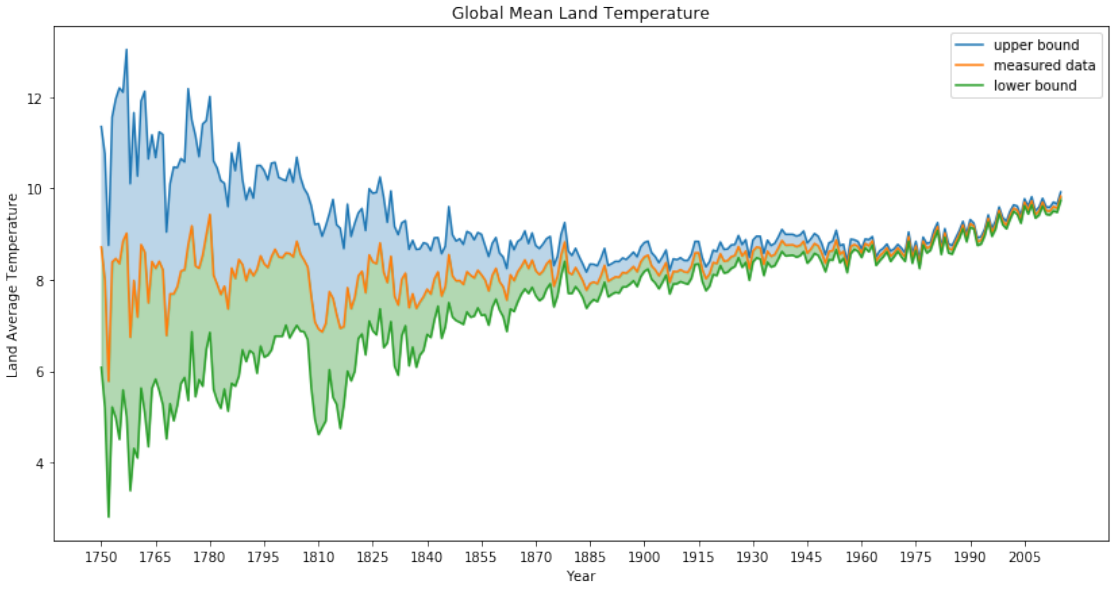
1.3 Globalne temperatury

[8]:
```
temp = global_temperatures['LandAverageTemperature']
temp_upper = global_temperatures['LandAverageTemperature'] +␣
↪global_temperatures['LandAverageTemperatureUncertainty']
temp_lower = global_temperatures['LandAverageTemperature'] -␣
↪global_temperatures['LandAverageTemperatureUncertainty']
```

[9]:
```
temps = global_temperatures[['dt', 'LandAverageTemperature',␣
↪'LandAverageTemperatureUncertainty']]
temps['dt']=pd.to_datetime(temps.dt).dt.strftime('%d/%m/%Y')
temps['dt'] = temps.dt.apply(lambda row: row[6:])
mean_temp_by_year = temps.groupby('dt')[['LandAverageTemperature',␣
↪'LandAverageTemperatureUncertainty']].mean().reset_index()
mean_temp_by_year_upper = mean_temp_by_year['LandAverageTemperature'] +␣
↪mean_temp_by_year['LandAverageTemperatureUncertainty']
mean_temp_by_year_lower = mean_temp_by_year['LandAverageTemperature'] -␣
↪mean_temp_by_year['LandAverageTemperatureUncertainty']
```

Na poniższym wykresie przedstawione są średnie wartości pomiarów rocznej temperatury wraz z niepewnościami pomiarowymi, wyliczonymi wyżej.

[10]:
```
plt.figure(figsize=(14, 7))
plt.title('Global Mean Land Temperature')
sns.lineplot(x=mean_temp_by_year['dt'], y=mean_temp_by_year_upper,␣
↪legend='brief', label='upper bound')
sns.lineplot(x=mean_temp_by_year['dt'],␣
↪y=mean_temp_by_year['LandAverageTemperature'], legend='brief',␣
↪label='measured data')
ax = sns.lineplot(x=mean_temp_by_year['dt'], y=mean_temp_by_year_lower,␣
↪legend='brief', label='lower bound')
ax.fill_between(mean_temp_by_year['dt'], mean_temp_by_year_upper,␣
↪mean_temp_by_year['LandAverageTemperature'], alpha=0.3)
ax.fill_between(mean_temp_by_year['dt'], mean_temp_by_year_lower,␣
↪mean_temp_by_year['LandAverageTemperature'], color='green', alpha=0.3)
plt.xlabel('Year')
plt.ylabel('Land Average Temperature')
t = plt.xticks(mean_temp_by_year['dt'][::15])
```



Jak możemy zauważyć dane historyczne są obarczone sporym błędem pomiarowym z tego powodu w dalszej analizie nie uwzględniamy danych do roku 1850

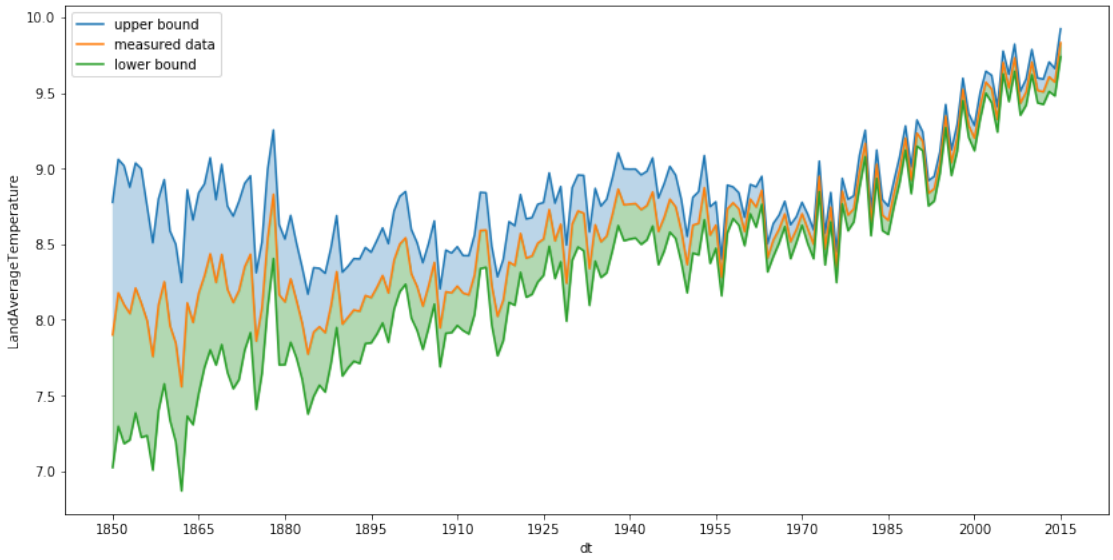
[11]:
```
mean_temp_by_year_shorten = mean_temp_by_year.iloc[100:]
mean_temp_by_year_upper_shorten = mean_temp_by_year_upper[100:]
mean_temp_by_year_lower_shorten = mean_temp_by_year_lower[100:]
mean_temp_by_year_shorten.head()
```

[11]:

	dt	LandAverageTemperature	LandAverageTemperatureUncertainty
100	1850	7.900667	0.876417
101	1851	8.178583	0.881917
102	1852	8.100167	0.918250
103	1853	8.041833	0.835000
104	1854	8.210500	0.825667

[12]:
```
plt.figure(figsize=(14, 7))
sns.lineplot(x=mean_temp_by_year_shorten['dt'],␣
 ↪y=mean_temp_by_year_upper_shorten, legend='brief', label='upper bound')
sns.lineplot(x=mean_temp_by_year_shorten['dt'],␣
 ↪y=mean_temp_by_year_shorten['LandAverageTemperature'], legend='brief',␣
 ↪label='measured data')
ax = sns.lineplot(x=mean_temp_by_year_shorten['dt'],␣
 ↪y=mean_temp_by_year_lower_shorten, legend='brief', label='lower bound')
ax.fill_between(mean_temp_by_year_shorten['dt'],␣
 ↪mean_temp_by_year_upper_shorten,␣
 ↪mean_temp_by_year_shorten['LandAverageTemperature'], alpha=0.3)
```

```
ax.fill_between(mean_temp_by_year_shorten['dt'],␣
 ↪mean_temp_by_year_lower_shorten,␣
 ↪mean_temp_by_year_shorten['LandAverageTemperature'], color='green', alpha=0.
 ↪3)
t = plt.xticks(mean_temp_by_year_shorten['dt'][::15])
```



Powyższy wykres pokazuje wyraźny wzrost temperatury, który w ostatnim półwieczu znacząco przyspieszył.

Możliwą przyczyną takiego zjawiska może być znaczący wzrost emisji **CO2** na świecie. Z tego powodu w dalszej części projektu sprawdzamy czy istnieje **silna** korelacja między tymi danymi.

1.4 Analiza zależności

Aby sprawdzić czy zachodzi liniowa zależność danych, użyjemy regresji liniowej. Dla zobrazowania tego użyjemy `seaborn.regplot` i funkcję `pandas.corr` dla zobrazowania krzywej dobranej regresją liniową i wyznaczenia współczynika korelacji odpowiednio.

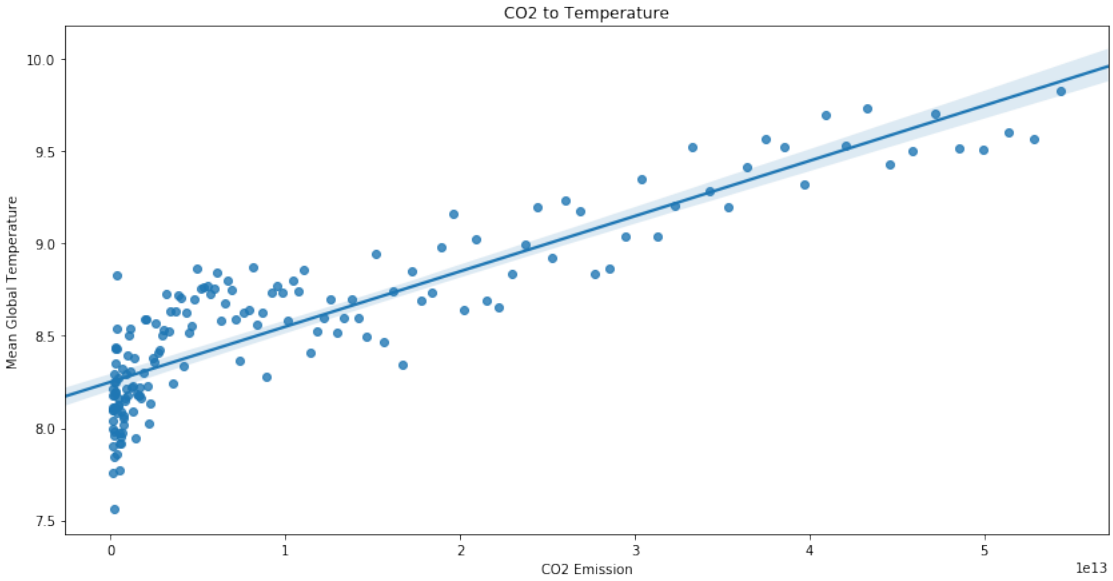
Uwzględniając to, że **CO2** wyemitowane w poprzednich latach także wpływa na stan obecny atmosfery (CO2 dość długo trzyma się w atmosferze), dane emisyjne kumulujemy:

[13]:
```
emission_by_world = np.squeeze(emission_data[emission_data.Country.
 ↪isin(["World"])].values)[1:-2].astype(np.int64)
emission_by_world = np.cumsum(emission_by_world)
```

[14]:
```
emission_by_world = emission_by_world[(-mean_temp_by_year_shorten.shape[0]):]
```

[15]:
```
plt.figure(figsize=(14, 7))
plt.title('CO2 to Temperature')
```

```
p = sns.regplot(x=emission_by_world,␣
 ↪y=mean_temp_by_year_shorten['LandAverageTemperature'])
plt.xlabel('CO2 Emission')
t = plt.ylabel('Mean Global Temperature')
```



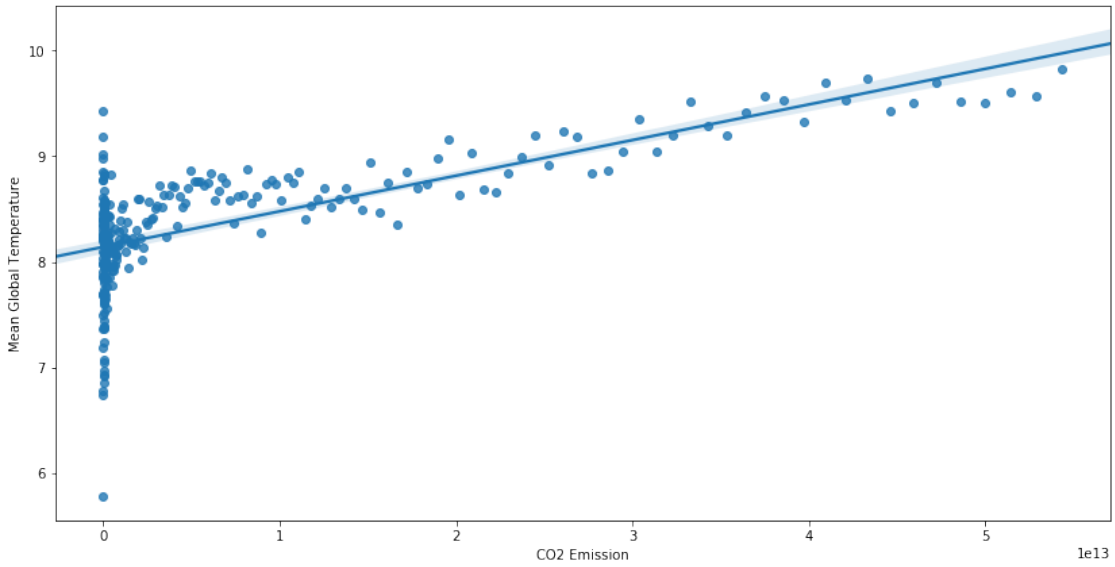
[16]:
```
mean_temp_by_year_shorten['LandAverageTemperature'].corr(pd.
 ↪Series(emission_by_world))
```

[16]:
```
0.32651878123923495
```

Powyższa analiza pokazała że zachodzi jedynie słaba zależność pomiędzy danymi. Co również mogło być spowodowan wcześniejszą decyzją o zmniejszeniu ilości danych. Poniżej sprawdzimy zależność uwzględniając również wcześniej usunięte dane temperaturowe.

[17]:
```
emission_by_world = np.squeeze(emission_data[emission_data.Country.
 ↪isin(["World"])].values)[1:-2].astype(np.int64)
emission_by_world = np.cumsum(emission_by_world)
#emission_by_world
```

[18]:
```
plt.figure(figsize=(14, 7))
p = sns.regplot(x=emission_by_world,␣
 ↪y=mean_temp_by_year['LandAverageTemperature'][1:])
plt.xlabel('CO2 Emission')
t = plt.ylabel('Mean Global Temperature')
```



[19]:
```
mean_temp_by_year['LandAverageTemperature'].corr(pd.Series(emission_by_world))
```

[19]: 0.6926937358750206

Powyższy test pokazuje silną zależność danych, nie mniej jednak głobalne ocieplenie jest złożonym procesem, który nie może zależeć tylko od jednego parametru.