Analiza i Eksploracja danych

Ostatecznie nie udało mi się znaleźć żadnego dataset'u dobrze współgrającego z tymi od statystykach deficytowych nadwyżkowych zawodów, zamiast tego znalazłem 2 dobrze współgrające dataset'y – ranking szczęścia dla danego kraju i informacje ogólne o kraju Populacja, % sektorów gospodarczych itd. Więc całą analizę przeprowadziłem na nowo. Starą zostawiam jako archiwum.

1. Analiza dataset'ów ranking szczęścia i informacje o krajach

1. Narzędzie do eksploracji danych wgranych do gcs'a

Wydaje mi się że compute engine się wyłącza i nie ma pan jak potem zobaczyć wyników analizy i eksploracji dlate w pkt.3 załączam wygenerowany html jak i plik notebook'owy.

Spróbuje załączyć te 2 pliki również na classroom'ie.

Maszyna n1-standard na europe-north1-a.

https://console.cloud.google.com/compute/instances?project=still-primer-271314&cloudshell=true&supportedpurview=project&instancessize=50

Instrukcja uruchomenia:

1. https://console.cloud.google.com/compute/instances?project=still-primer-271314&cloudshell=true&supportedpurview=project&instancessize=50

2. Cloud shell

3. datalab connect countries-info-exploration

4. Web preview 8081

2. Skrypt do pobrania danych

Pobranie Danych

```
# Builtins
import io

# 3rd party
import google.datalab.storage as storage
from google.datalab import Context
import matplotlib as mpl
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import seaborn as sns
```

```
# Bucket
countries_bucket = storage.Bucket("countries_info")
# Csv's
general_info_csv = countries_bucket.object("country_general_info.csv")
happiness_info_2015_csv = countries_bucket.object("hapiness_by_year/2015.csv")
happiness_info_2016_csv = countries_bucket.object("hapiness_by_year/2016.csv")
happiness_info_2017_csv = countries_bucket.object("hapiness_by_year/2017.csv")
happiness_info_2018_csv = countries_bucket.object("hapiness_by_year/2018.csv")
happiness_info_2019_csv = countries_bucket.object("hapiness_by_year/2019.csv")
```

```
# Dataframes
general = pd.read_csv(io.BytesIO(general_info_csv.read_stream()), delimiter=",", decimal=",")
h_15 = pd.read_csv(io.BytesIO(happiness_info_2015_csv.read_stream()), delimiter=",")
h_16 = pd.read_csv(io.BytesIO(happiness_info_2016_csv.read_stream()), delimiter=",")
h_17 = pd.read_csv(io.BytesIO(happiness_info_2017_csv.read_stream()), delimiter=",")
h_18 = pd.read_csv(io.BytesIO(happiness_info_2018_csv.read_stream()), delimiter=",")
h_19 = pd.read_csv(io.BytesIO(happiness_info_2019_csv.read_stream()), delimiter=",")
dfs = [general, h_15, h_16, h_17, h_18, h_19]
```

Figure 1 Pobranie danych z bucket'a i import potrzebnych bibliotek

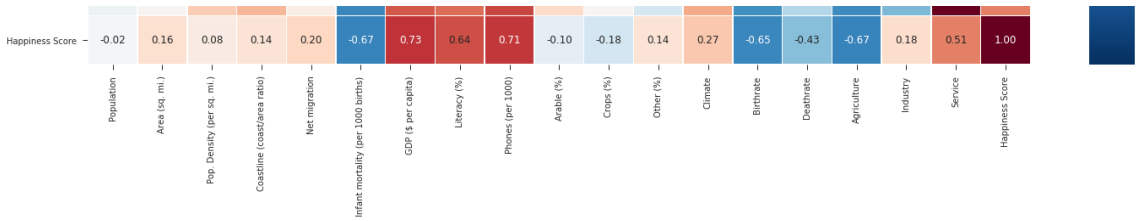
3. Wizualizacja danych


countries_info_exploration.ipynb


countries_info_exploration.html

Pełna analiza + czyszczenie w formie html'a / notebook'a

Diagramy:



Rysunek 1.1 Fragment tablicy korelacji wskazujący gdzie mogą się znajdować ciekawe zależności między Happiness Score'm a pozostałymi wartościami

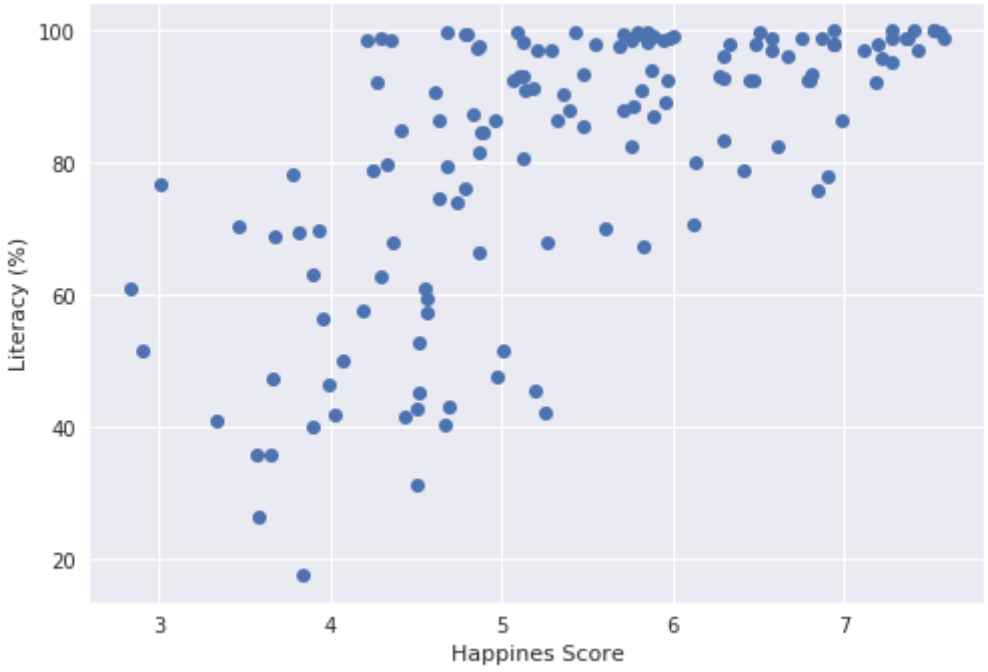
Interpretacja Heatmap'y

Najważniejsza dla tego projektu jest ostatni wiersz pokazujący korelacje Happiness Score / x

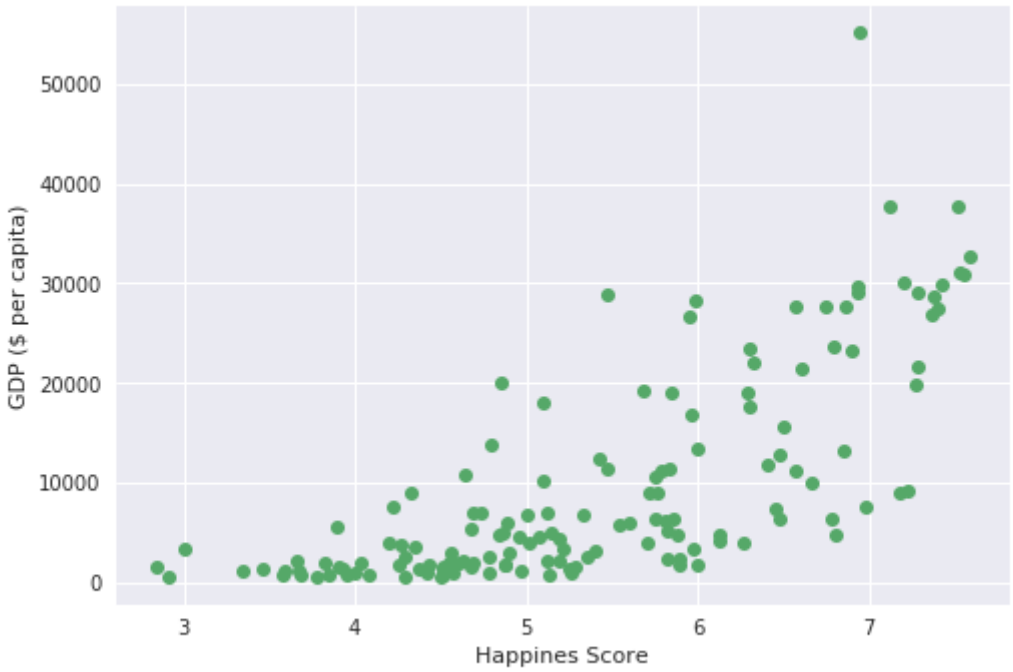
Wysokie dodatnie wartości przy GDP, Literacy i Phones oznaczają "jedna rośnie gdy rośnie druga" tzn. Im wyższe GDP kraju tym szczęśliwsi mieszkańcy

Analogicznie wysokie ujemne wartośći przy Infant Mortality, Deathrate i Agriculture pokazują "im więcej ludzi umiera i im więcej rolników tym ludzie mniej szczęśliwi

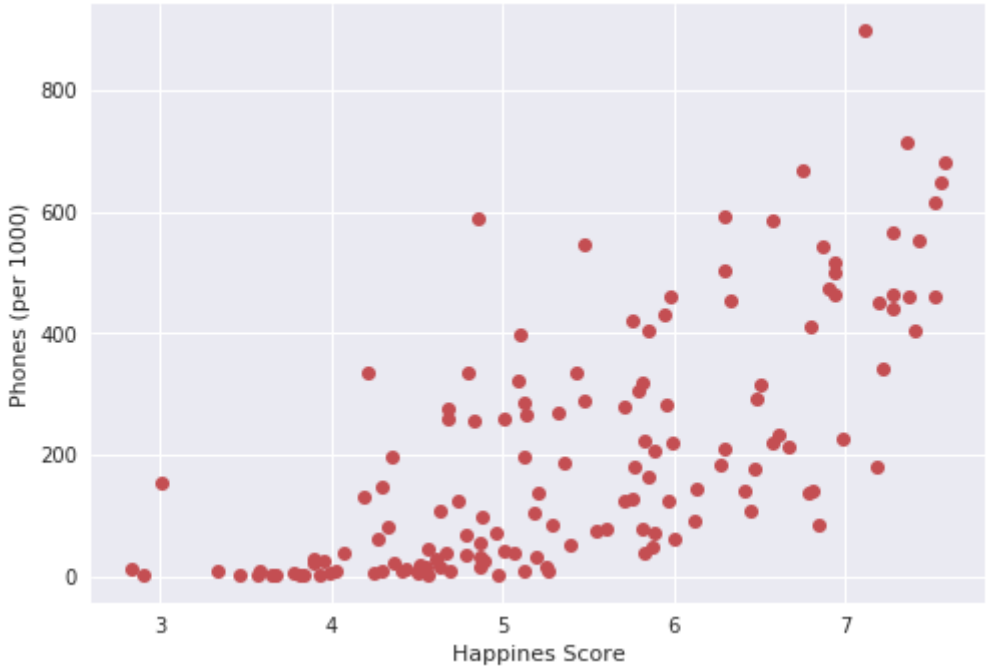
Z tego też można stworzyć następujące wykresy:



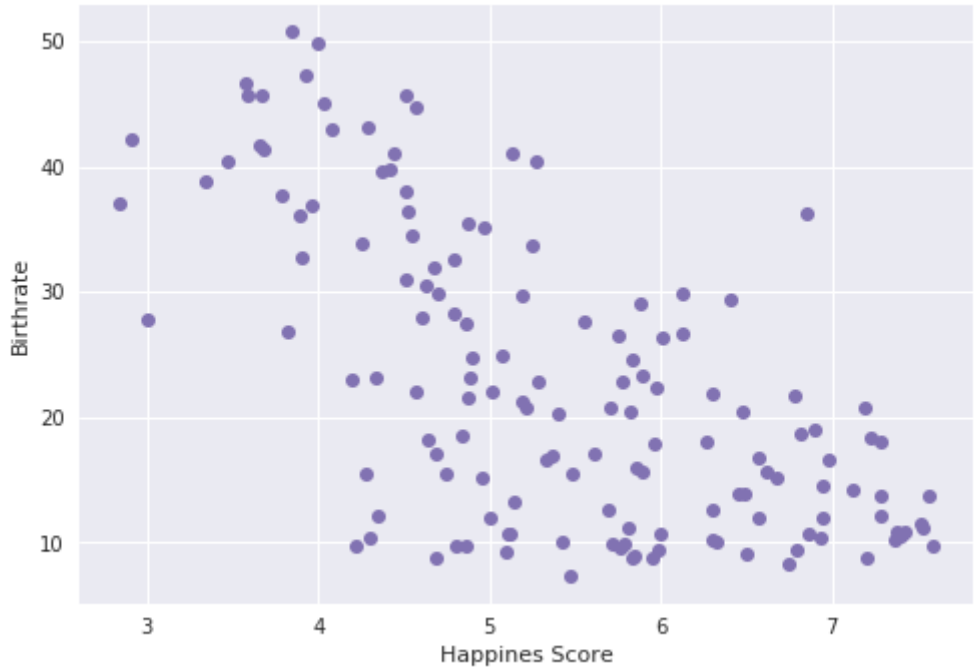
Rysunek 1.2 Wyraźnie widać że analfabetyzm nie sprzyja szczęściu



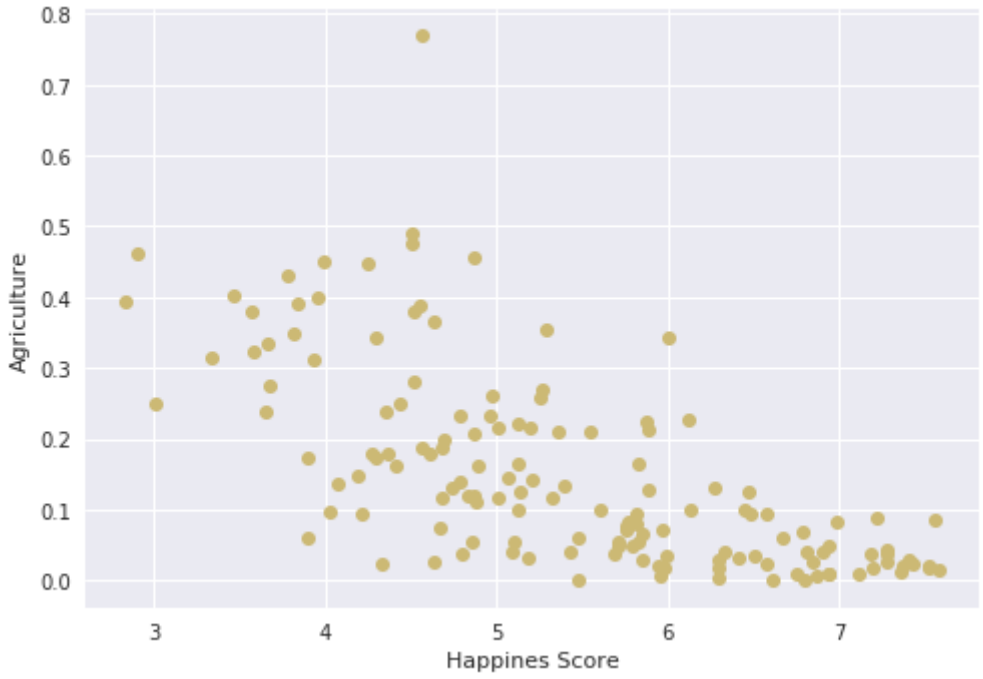
Rysunek 1.3 Wyższy produkt krajowy brutto zazwyczaj oznacza lepiej rozwinięty kraj z lepszą opieką zdrowotną itd. ten wykres nie dziwi



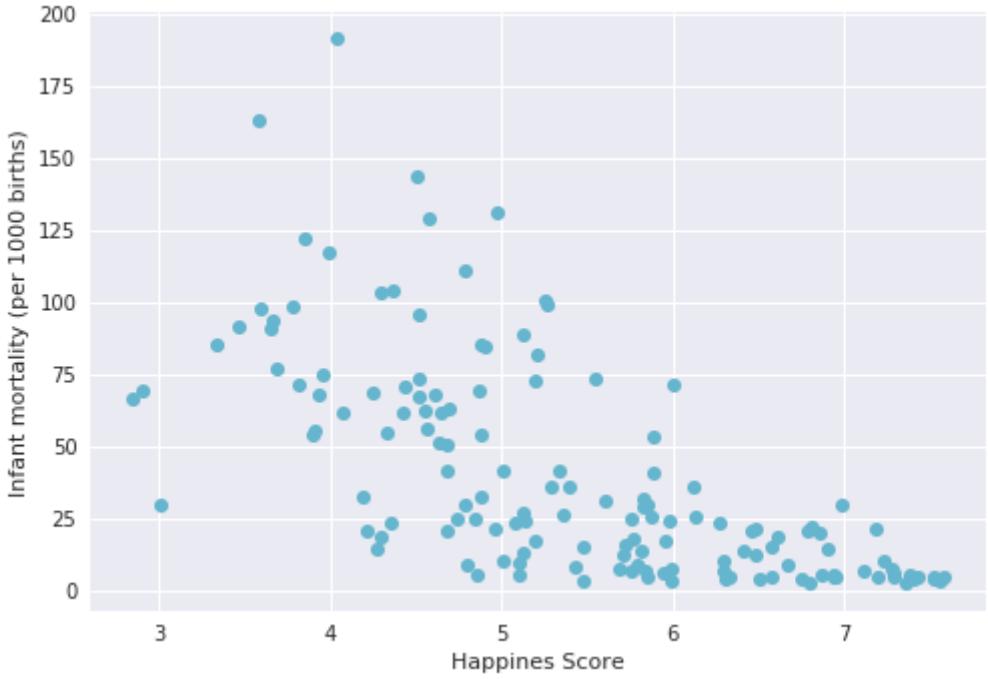
Rysunek 1.4 Nie powiedziałbym żeby posiadanie telefonu czyniło ludzi szczęśliwymi ale wiadomo że również są one czynnikiem wskazującym na wysoki rozwój kraju



Rysunek 1.5 Im więcej narodzin tym mniej szczęśliwy kraj, zdanie na początku kłóci się z intuicją jednak wiadomo że również jest to charakterystyka krajów ubogich nie mających dostępu do środków antykoncepcyjnych



Rysunek 1.6 Im więcej farmerów tym smutniejszy kraj, oczywiście również jest to wyznacznik rozwinięcia kraju i przejścia przez cykl dominacji sektorów rolniczego, przemysłowego oraz usługowego



Rysunek 1.7 Wysoka umieralność noworodków zazwyczaj pokrywająca się z wysoką umieralnością wogóle. Również charakterystyczne na biednych krajów.

4. Wykorzystanie danych

Aplikacja ma zbierać dane o szczęśliwości kraju do przygotowania raportu na rok 2020 oraz wyświetlać historyczne wyniki tych badań. Zbieranie i pokazywanie tych danych i korelacji nimy jest więc przeznaczeniem aplikacji.

2. Stara analiza dataset'u o zawodach deficytowych

3. Stara Konfiguracja Datalab'a

Maszyna n1-standard na europe-north1-a.

https://console.cloud.google.com/compute/instancesDetail/zones/europe-north1-a/instances/deficit-jobs-exploration?project=still-primer-271314

Instrukcja uruchomienia środowiska:

1. https://console.cloud.google.com/compute/instancesDetail/zones/europe-north1-a/instances/deficit-jobs-exploration?project=still-primer-271314

2. Cloud shell

3. datalab connect --zone europe-north1-a --port 8081 deficit-jobs-exploration

4. Web preview 8081

Raport graficzny znaduje się w notebook'u explore-deficit-and-surplus-jobs.ipynb

Dane zostały oczyszczone

explore-deficit-and-surplus-jobs.html

explore-deficit-and-surplus-jobs.ipynb