Dataset suggestions for Python Tsunami Datathon

1. birthwt.csv

This dataset contains information on risk factors associated with low infant birth weight (189 rows and 10 columns). The data were collected at Baystate Medical Center, Springfield, Mass during 1986.

Source: This is one of the built-in datasets in R.

** Note: We use this for some exercises in the HeaDS course *From Excel to R*.

Columns:

1. `low`: indicator of birth weight less than 2.5 kg.
2. `age`: mother's age in years.
3. `lwt`: mother's weight in pounds at last menstrual period.
4. `race`: mother's race (1 = white, 2 = black, 3 = other).
5. `smoke`: smoking status during pregnancy.
6. `ptl`: number of previous premature labours.
7. `ht`: history of hypertension.
8. `ui`: presence of uterine irritability.
9. `ftv`: number of physician visits during the first trimester.
10. `bwt`: birth weight in grams.

2. Correct_Dataset.csv

This dataset contains variables on patients with heart diseases.

Source: https://www.kaggle.com/ronitf/heart-disease-uci/discussion/173003

** Note: The dataset originally uploaded to Kaggle apparently contained mistakes and someone therefore provided a corrected dataset in the discussion section. This is not ideal and potentially confusing to figure out, but I thought I'd leave this in just so that there is some variations in topics.

Columns:

1. `age`: age
2. `sex`: sex
3. `cp`: chest pain type (4 values)
4. `trestbps`: resting blood pressure
5. `chol`: serum cholestoral in mg/dl
6. `fbs`: fasting blood sugar > 120 mg/dl
7. `restecg`: resting electrocardiographic results (values 0,1,2)
8. `thalach`: maximum heart rate achieved
9. `exang`: exercise induced angina
10. `oldpeak`: ST depression induced by exercise relative to rest
11. `slope`: the slope of the peak exercise ST segment
12. `ca`: number of major vessels (0-3) colored by flourosopy
13. `thal`: 3 = normal; 6 = fixed defect; 7 = reversable defect
14. `target`: presence of heart disease in patient, can take values from 0 (no presence) to 4

3. Covid-19 related datasets:

3.1. country_vaccinations.csv

This dataset details the vaccination progress per country worldwide.

Source: https://www.kaggle.com/gpreda/covid-world-vaccination-progress

Columns

1. `country`: country information;
2. `iso_code`: ISO code for the country;
3. `date`: date for the data entry
4. `total_vaccinations`: total vaccinations per date and country
5. `people_vaccinated`: number of people
6. `people_fully_vaccinated`: total number of people fully vaccinated
7. `daily_vaccinations_raw`: raw data, the number of vaccination for that date/country
8. `daily_vaccinations`: daily vaccinations - for a certain data entry, the number of vaccination for that date/country
9. `total_vaccinations_per_hundred`: ratio (in percent) between vaccination number and total population up to the date in the country
10. `people_vaccinated_per_hundred`: total number of people vaccinated per hundred - ratio (in percent) between population immunized and total population up to the date in the country
11. `people_fully_vaccinated_per_hundred` - ratio (in percent) between population fully immunized and total population up to the date in the country
12. `daily_vaccinations_per_million`: daily vaccinations per million - ratio (in ppm) between vaccination number and total population for the current date in the country
13. `vaccines`: total number of vaccines used in the country (up to date)
14. `source_name`: source of the information (national authority, international organization, local organization etc.)
15. `source_website`: website of the source of information

3.2. worldometer_coronavirus_daily_data.csv

This file contains data on the the daily development of corona cases and deaths per country worldwide.

Source: https://www.kaggle.com/josephassaker/covid19-global-dataset

** Note: This might be interesting to combine with the vaccination data.

Columns:

1. `date`: Date of observation of the row's data in YYYY-MM-DD format
2. `country`: Country in which the the row's data was observed
3. `cumulative_total_cases`: Cumulative number of confirmed cases as of the row's date, for the row's country
4. `daily_new_cases`: Daily new number of confirmed cases on the row's date, for the row's country
5. `active_cases`: Number of active cases (i.e., confirmed cases that still didn't recover nor die) on the row's date, for the row's country
6. `cumulative_total_deaths`: Cumulative number of confirmed deaths as of the row's date, for the row's country

7. `daily_new_deaths_deaths`: Daily new number of confirmed deaths on the row's date, for the row's country

3.3. covid_impact_on_airport_traffic.csv

This dataset shows traffic to and from different airports worldwide as the percentage of traffic volume during the baseline period (1 Feb – 15 March 2020).

Source: https://www.kaggle.com/terenceshin/covid19s-impact-on-airport-traffic

** Note: This is obviously not at all health data science related, but I thought it might still be interesting since academia is such an international field of work and most people either go abroad a lot for conferences etc. or to visit their family outside of Denmark.

Columns:

1. `AggregationMethod`: Aggregation period used to compute this metric
2. `Date`: Date in format YYYY-MM-DD
3. `Version`: Version # of this dataset
4. `AirportName`: Name of airport
5. `PercentofBaseline`: Proportion of trips on this date as compared to Avg number of trips on the same day of week in baseline period i.e 1st
6. `Centroid`: Geography representing centroid of the Airport polygon
7. `City`: City within which the Airport is located
8. `State`: State within which the Airport is located
9. `ISO_3166_2`: ISO-3166-2 code representing Country and Subdivision
10. `Country`: Country within which the Airport is located
11. `Geography`: Polygon of the Airport that is used to compute this metric