TRABALHO TP089

Fazer uma Análise Descritiva e Exploratória de um conjunto de dados no R.

A **análise descritiva** resume e descreve os aspectos importantes de um conjunto de características observadas ou compara tais características entre dois ou mais conjuntos.
A **análise exploratória** de dados é usada para analisar e investigar conjuntos de dados e resumir suas principais características, muitas vezes usando métodos de visualização de dados.

Um pdf deve ser elaborado com as análises, contendo:

- Texto explicativo do que será realizado;
- Código em R, com comandos usados;
- Saída do que os comandos obtiveram;
- Interpretação dos resultados.

Obrigatoriamente, no mínimo, um gráfico deve ser gerado.

Veja um exemplo do que é esperado:
https://acervolima.com/analise-exploratoria-de-dados-no-conjunto-de-dados-iris/
Neste exemplo, o conjunto de dados "Iris" foi analisado no Python.

Cada aluno analisará um conjunto de dados diferentes, obtidos da plataforma Kaggle. Foram escolhidos arbitrariamente 7 conjuntos de dados, sem estudo prévio destes. Todos os conjuntos estão no formato .csv, que podem ser lidos por meio da função read.csv(). É possível que tenham dados faltantes nos conjuntos e também variáveis que não sejam interessantes de se analisar. Alguns conjuntos têm muitas variáveis, portanto pode-se escolher apenas algumas para fazer a análise.

A designação de cada aluno para cada conjunto de dados será feita mediante sorteio realizado em sala de aula no dia 14/12/22.

Enviar o trabalho até **24/01/2023**para o emailmarianakleina11@gmail.com

Conjunto de dados	Aluno(a) sorteado(a)
Emmisionof Air Pollutants	André
Argentina carprices	Maria Vitória
Brazilian League - Players Statistics (2020)	Matheus
COVID-19 Dataset	João
Spotify Tracks Dataset	Ana Carolina
Commodity Prices	Alexia
World PopulationDataset	Luiz Henrique

A seguir, uma breve explicação de cada um dos conjuntos de dados.

Emmision of Air Pollutants

Historical emissions of air pollutants from fuels used by 200+ countries.

About Dataset

Air pollution – the combination of outdoor and indoor particulate matter, and ozone – is a risk factor for many of the leading causes of death including heart disease, stroke, lower respiratory infections, lung cancer, diabetes, and chronic obstructive pulmonary disease (COPD).

Air PollutansList:

- Nitrogen oxide (NOx)
- Sulphurdioxide (SO_2)
- Carbonmonoxide (CO)
- Organiccarbon (OC)
- Non-methane volatile organic compounds (NMVOCs)
- Black carbon (BC)
- Ammonia (NH_3)

Note: Data per tonnes (t) in 2019.

Argentina carprices

About Dataset

This Dataset includes some features of cars. I scraped from a website in Argentina.

Data dictionary

- money: thenumberofprices
- brand: the brand of the car
- model: the model of the car
- year: the year of the car
- color: the color of the car
- fuel_type: the fuel type of the car
- door: the door of thecar
- gear: the gear typeofthecar
- motor: the motor typeofthecar
- body_type: the body typeofthecar
- kilometers: thekilometerofthecar
- currency: thecurrencyofthepriceofcar

Brazilian League - Players Statistics (2020)

Data from players who played in the Brazilian League in the year 2020.

About Dataset

This dataset contains statistics of players who participated in the Brazilian League in the year 2020, they are data from 738 players, in which each of them has 70 distinct characteristics.

COVID-19 Dataset

Number of Confirmed, Death and Recovered cases every day across the globe.

AboutDataset

Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus. Most people infected with COVID-19 virus will experience mild to moderate respiratory illness and recover without requiring special treatment. Older people, and those with underlying medical problems like cardiovascular disease, diabetes, chronic respiratory disease, and cancer are more likely to develop serious illness.

During the entire course of the pandemic, one of the main problems that healthcare providers have faced is the shortage of medical resources and a proper plan to efficiently distribute them. In these tough times, being able to predict what kind of resource an individual might require at the time of being tested positive or even before that will be of immense help to the authorities as they would be able to procure and arrange for the resources necessary to save the life of that patient.

The main goal of this project is to build a machine learning model that, given a Covid-19 patient's current symptom, status, and medical history, will predict whether the patient is in high risk or not.

Content

The dataset was provided by the Mexican government (link). This dataset contains an enormous number of anonymized patient-related information including pre-conditions. The raw dataset consists of 21 unique features and 1,048,576 unique patients. In the Boolean features, 1 means "yes" and 2 means "no". values as 97 and 99 are missing data.

- sex: femaleor male
- age: ofthepatient.
- classification: covid test findings. Values 1-3 mean that the patient was diagnosed with covid in different
  degrees. 4 or higher means that the patient is not a carrier of covid or that the test is inconclusive.
- patient type: hospitalized or not hospitalized.
- pneumonia: whether the patient already have air sacs inflammation or not.
- pregnancy: whether the patient is pregnant or not.
- diabetes: whether the patient has diabetes or not.
- copd: Indicates whether the patient has Chronic obstructive pulmonary disease or not.
- asthma: whether the patient has asthma or not.
- inmsupr: whether the patient is immunosuppressed or not.
- hypertension: whether the patient has hypertension or not.
- cardiovascular: whether the patient has heart or blood vessels related disease.
- renal chronic: whether the patient has chronic renal disease or not.
- other disease: whether the patient has other disease or not.
- obesity: whether the patient is obese or not.
- tobacco: whether the patient is a tobacco user.
- usmr: Indicates whether the patient treated medical units of the first, second or third level.
- medical unit: type of institution of the National Health System that provided the care.
- intubed: whether the patient was connected to the ventilator.
- icu: Indicates whether the patient had been admitted to an Intensive Care Unit.
- death: indicates whether the patient died or recovered.

Spotify Tracks Dataset

A dataset of Spotify songs with different genres and their audio features.

About Dataset

This is a dataset of Spotify tracks over a range of 125 different genres. Each track has some audio features associated with it. The data is in CSV format which is tabular and can be loaded quickly.

Column Description

- track_id: The Spotify ID for the track
- artists: The artists' nameswhoperformedthe track. Ifthereis more thanoneartist, they are separatedbya ;
- album_name: The albumname in whichthe track appears
- track_name: Name of the track
- popularity: The popularity of a track is a value between 0 and 100, with 100 being the most popular. The popularity is calculated by algorithm and is based, in the most part, on the total number of plays the track has had and how recent those plays are. Generally speaking, songs that are being played a lot now will have a higher popularity than songs that were played a lot in the past. Duplicate tracks (e.g. the same track from a single and an album) are rated independently. Artist and album popularity is derived mathematically from track popularity.
- duration_ms: The track length in milliseconds
- explicit: Whether or not the track has explicit lyrics (true = yes it does; false = no it does not OR unknown)
- danceability: Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable
- energy: Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale
- key: The key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on. If no key was detected, the value is -1
- loudness: The overall loudness of a track in decibels (dB)
- mode: Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0
- speechiness: Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks
- acousticness: A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic
- instrumentalness: Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content

- liveness: Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live
- valence: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry)
- tempo: The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration
- time_signature: An estimated time signature. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure). The time signature ranges from 3 to 7 indicating time signatures of 3/4, to 7/4.
- track_genre: The genre in whichthe track belongs

Commodity Prices

Yearly prices of 30 commodities in 1960-2021

AboutDataset

Beverages

- Cocoa ($/kg)
  International Cocoa Organization (ICCO) daily price, an average of the first three positions on the terminal markets of New York and London, nearest three future trading months.
- Coffee ($/kg)
  International Coffee Organization (ICO) indicator price, other mild Arabicas, average New York and Bremen/Hamburg markets, ex-dock
- Tea ($/kg)
  Average three auctions, thearithmetic average of quotations at Kolkata, Colombo, and Mombasa/Nairobi.

Energy

- Crude oil ($/bbl)
  The average spot price of Brent, Dubai, and West Texas Intermediate, equally weighed
- Coal ($/mt)
  Australia, from February 2022, port thermal, f.o.b. Newcastle, 6000 kcal/kg futures price. From 2015 to January 2022, port thermal, f.o.b. Newcastle, 6000 kcal/kg spot price. 2002-2014, thermal GAR, f.o.b. piers, Newcastle/Port Kembla, 6,300 kcal/kg (11,340 btu/lb), less than 0.8%, sulfur 13% ash
- Natural Gas ($/mmbtu)
  US, spot price at Henry Hub, Louisiana

Grains

- Barley ($/kg)
  US. feed, No. 2, spot, 20 days To-Arrive, delivered Minneapolis from May 2012 onwards; during 1980 - 2012 April Canadian, feed, Western No. 1, Winnipeg Commodity Exchange, spot, wholesale farmers' price
- Maize ($/kg)
  US, no. 2, yellow, f.o.b. US Gulf ports

- Rice ($/kg)
  Thailand, 5% broken, white rice (WR), milled, indicative price based on weekly surveys of export transactions, government standard, f.o.b. Bangkok
- Sorghum ($/mt)
  US, no. 2 milo yellow, Texas export bids for grain delivered to export elevators, rail-truck, f.o.b. Gulf ports
- Wheat ($/mt)
  US, no. 2 hard red winter Gulf export price; June 2020 backward, no. 1, hard red winter, ordinary protein, export price delivered at the US Gulf port for prompt or 30 days shipment

Meat

- Beef ($/mt)
  Australia/New Zealand, mixed trimmings 85%, East Coast, 7-45 day deferred delivery, FOB port of entry, beginning January 1995
- Chicken ($/mt)
  US, Urner Barry North East weighted average for broiler/fryer, whole birds, 2.5 to 3.5 pounds, USDA grade "A" from 2013 onwards; 1980-2012, Georgia Dock weighted average, 2.5 to 3 pounds, wholesale
- Lamb ($/mt)
  US, boxed lamb cuts, leg, double, trotter-on, less than truckload (LTL) pricing, from August 2010
- Shrimp ($/kg)
  US, brown, shell-on, headless, in frozen blocks, source Gulf of Mexico, 26 to 30 count per pound, wholesale US beginning 2004

Metals

- Gold ($/kg)
  UK, 99.5% fine, London afternoon fixing, an average of daily rates
- Platinum ($/kg)
  UK, 99.9% refined, London afternoon fixing
- Silver ($/kg)
  UK, 99.9% refined, London afternoon fixing; prior to July 1976 Handy & Harman. Grade prior to 1962 unrefined silver.

Raw Materials

- Cotton ($/troy oz)
  Cotton Outlook "CotlookA index", middling 1-3/32 inch, traded in Far East, C/F beginning 2006
- Rubber ($/troy oz)
  Asia, RSS3 grade, Singapore Commodity Exchange Ltd (SICOM) nearby contract beginning 2004; from 2000 to 2003, Singapore RSS1
- Tobacco ($/troy oz)
  Any origin, unmanufactured, general import, cif, US

Oils

- Coconut Oil ($/kg)
  Philippines/Indonesia, from January 2021, crude, CIF Rotterdam; January 1999 to December 2020, crude, CIF NW Europe
- Groundnut Oil ($/kg)
  Dutch Refined GroundNut Oil A/O, Ex Tank Rotterdam, beginning December 2020; January 1999-November 2020, US crude, FOB South-East
- Palm Oil ($/mt)
  Malaysia, from January 2021, RBD, FOB Malaysia Ports; December 2001 to December 2020, RBD, CIF Rotterdam
- Soybean Oil ($/mt)
  Dutch Soyoil Crude Degummed, EXW Dutch Mills; January 1999 to December 2020, Dutch crude degummed, FOB NW Europe

Other Foods

- Banana ($/mt)
  Central & South America, major brands, US import price, free on truck (f.o.t.) US Gulf ports.
- Sugar ($/mt)
  World, International Sugar Agreement (ISA) daily price, raw, f.o.b. and stowed at greater Caribbean ports
- Orange ($/mt)
  Mediterranean exporters navel, European Union indicative import price, c.i.f. Paris

Woods

- Logs ($/cubic meter)
  Southeast Asia, meranti, Sarawak, sale price charged by importers, Tokyo beginning February 1993
- Sawnwood ($/cubic meter)
  Southeast Asia, dark red seraya/meranti, select and better quality, average 7 to 8 inches; length average 12 to 14 inches; thickness 1 to 2 inch(es); kiln dry, c. & f. UK ports, with 5% agents commission including premium for products of certified sustainable forestbeginningJanuary 2005

World Population Dataset

World population Country by Country

About Dataset

Population rise is currently a major subject of discussion nowadays. Everyday, there is higher birth rate recorded as compare to death rate which is quite alarming for the world. Below is the complete data about the world population , country by country (235 countries). There are some missing values.