Collecting and analyzing dataset connected to

**brain strokes**

**1.dataset**

**Link**:

<https://data.world/chhs/06ed38d3-b047-4ae2-aa00-2e43b5491d6e?fbclid=IwAR0pOS8Tn2z7ZTzttxhOQbQyv9LSzkAJrVSZDPw_W7EK8lke6lgB1fAAit4>

**Published by:**

**California Health and Human Services (**[**https://data.world/chhs**](https://data.world/chhs)**)**

**Description:**

This dataset contains risk-adjusted 30-day mortality and 30-day readmission rates, quality ratings, and number of deaths / readmissions and **cases for ischemic stroke** treated in **California hospitals** from the years 2011-12 to 2014-15.

**Data analysis:**

**There are 2188 instances with 10 features.**

**Year**

(from what year is the feature). This feature doesn’t have null values.

**County**

(where is the hospital located). Includes null values when the feature is for the whole California or has the value **“AAAA”**

**Hospital**

(hospitals name). This feature doesn’t have null values. When the instances is for the whole California the feature has value **“Statewide”.**

**OSHPDID**

(hospital ID). When the instance is about whole California the value of the feature is **“.”** Or null.(4 null values in total)

**Measure**

(30-day Mortality or 30-day Readmission) This feature doesn’t have null values.

**Risk Adjusted Rate**

This feature has 10 null values.

**Number of Deaths/Readmissions**

This feature has 10 null values.

**Number of Cases**

This feature has 10 null values.

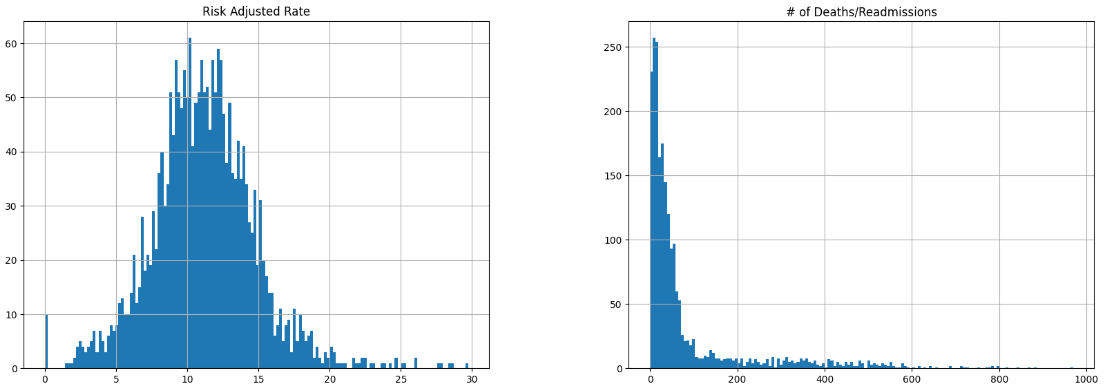
**Hospital Ratings**

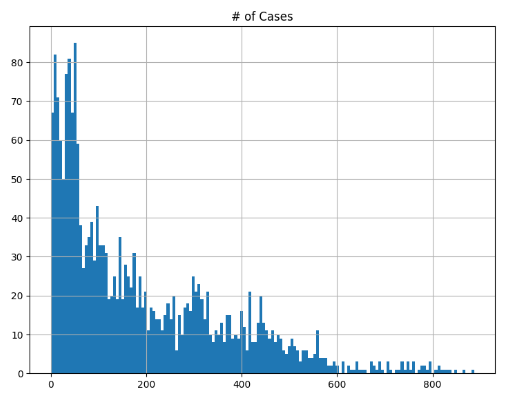
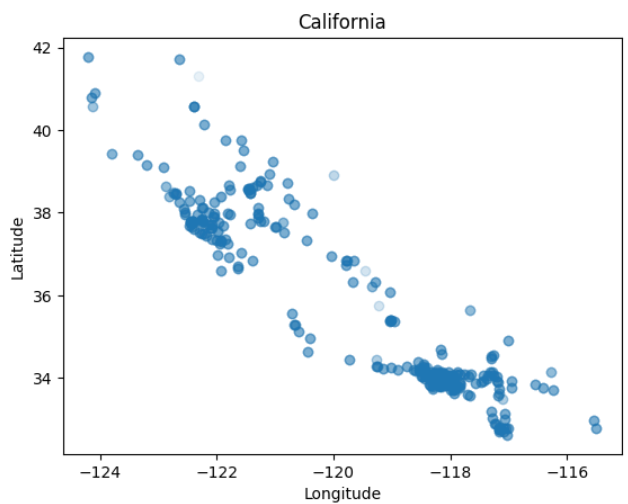
Hospital Rating don’t have the 10 above mentioned instances and the 8 where the instances are about the whole California. (18 null instances in total)

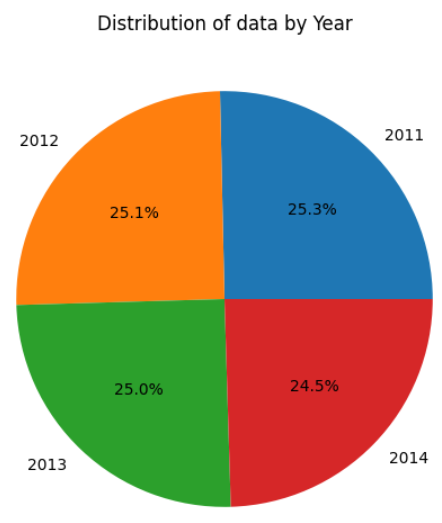
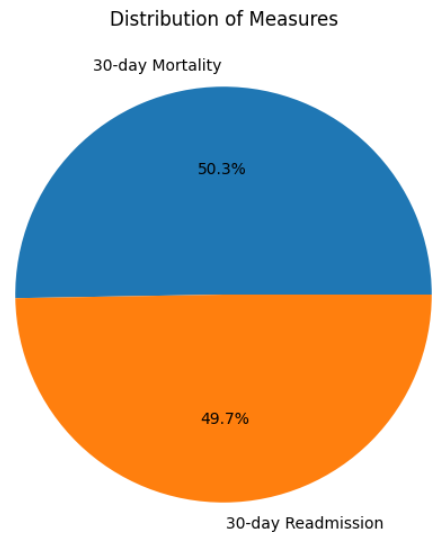
**Location**

Longitude and latitude for the hospital. (8 null instances when the instance is for the whole California)

By removing 18 instances we have clean dataset.



**2.Dataset**

**Link:**

<https://data.world/datagov-uk/0cb6045e-f44f-4dcb-814b-b97840cc80c3?fbclid=IwAR3MEd33szJsu-Sv3aDVuByvmwaBBQhQQw4WYgQG1swlApnYxYKUJBYD7ck>

**Published by:**

<https://ckan.publishing.service.gov.uk/dataset/stockport-local-health-characteristics>

**Description:**

This dataset contains information on the prevalence of a variety of health conditions amongst Stockport residents, aggregated by LSOA. The count of individuals affected by each condition is provided, along with the GP registered population for each LSOA. The data represents a **snapshot taken in June 2016**. Conditions covered include: Hypertension, Anxiety, Depression, Asthma, Obesity, Diabetes, Coronary Heart Disease, Falls, Cancer, Chronic Kidney Disease, Chronic Obstructive Pulmonary Disease**, Stroke/Trans-Ischaemic Attack** and Atrial Fibrillation

**Data analysis:**

There are 190 instances with 18 features. There are no null values in any of the datasets.

**lsoa11nm** and **lsoa11nmw** have the same values with each other for every instances.

**ogc\_fid, lsoa11cd, lsoa11nm, lsoa11nmw** have different value for every instances( 190 different values) so it doesn’t give us any more info than a basic indexer.

The 18 features are the following:

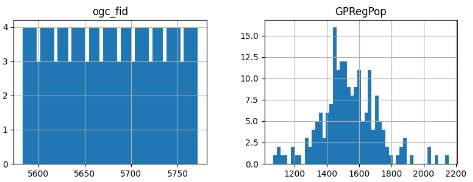
**ogc\_fid lsoa11cd lsoa11nm lsoa11nmw GPRegPop** (GP registered population)

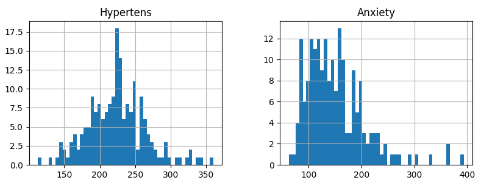
**Hypertens Anxiety Depression Asthma Obesity Diabetes**

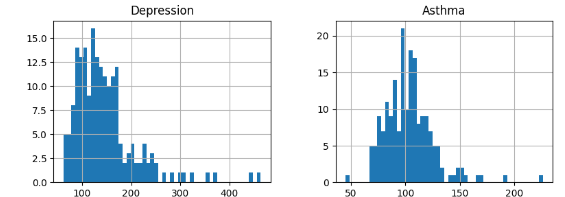
**CHD** (Coronary Heart Disease) **Fall Cancer CKD** (Chronic Kidney Disease)

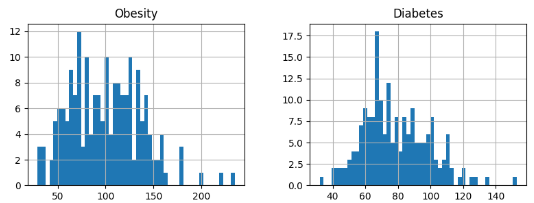
**COPD** (Chronic Obstructive Pulmonary Disease)  **Stroke\_TIA** (Stroke/Trans-Ischaemic Attack)

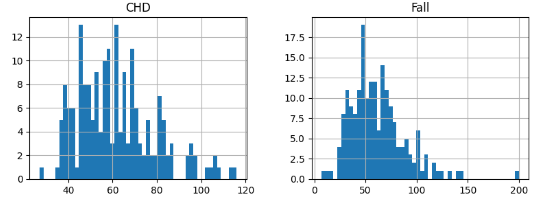
**AF** (Atrial Fibrillation)

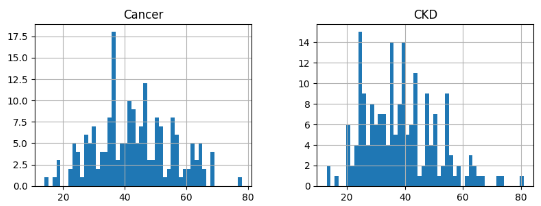


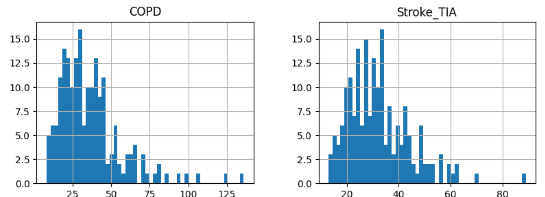


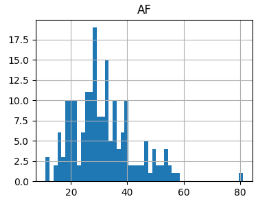












|  |  |
| --- | --- |
| Stroke\_TIA | 1.00000 |
| CHD | 0.711947 |
| Hypertens | 0.657975 |
| AF | 0.618815 |
| CKD | 0.599411 |
| Diabetes | 0.493134 |
| COPD | 0.451992 |
| Cancer | 0.445393 |
| Fall | 0.394422 |
| Depression | 0.332552 |
| Asthma | 0.323972 |
| GPRegPop | 0.328847 |
| Anxiety | 0.388459 |
| Obesity | 0.248199 |
| ogc\_fid | 0.002752 |

The table represents the correlation between Stroke\_TIA and all the other features.

**3.Dataset**

**Title:**

All Payer In-Hospital/30-Day Acute Stroke Mortality Rates by Hospital (SPARCS):

**Link:**

<https://data.world/healthdatany/r29i-yr49?fbclid=IwAR03liwBhR_XWfdnkj3tWBKjdHDljDTiY9YiSDSsTXdgwVR7OOxfBQuPNa0>

**Published by:**

<https://data.world/healthdatany>

**Description:**

The dataset contains **hospital stroke designation and Coverdell registry participation status, acute stroke discharges counts** (numerators, denominators), observed, expected and risk-adjusted **acute stroke in-hospital/30-day post admission mortality rates with corresponding 95% confidence intervals**. Mortality rates risk adjustment was based on the methodology developed by the New York State Department of Health. The **purpose of this data set is reporting of hospital-specific risk adjusted acute stroke mortality rates** (RAMR) to inform hospitals, to aid initiatives to improve hospital quality performance and measurement, and to identify performance outliers for public reporting.

The data is from the year 2013.

**Data analysis:**

The dataset includes 137 instances each having 14 features. The instances have the following features:

|  |  |
| --- | --- |
|  |  |

**Year Facility Id Hospital Name Hospital County Stroke Designated Center**

**Coverdell Hospital Stroke Cases Died Observed Rate Expected Rate Risk Adjusted Rate**

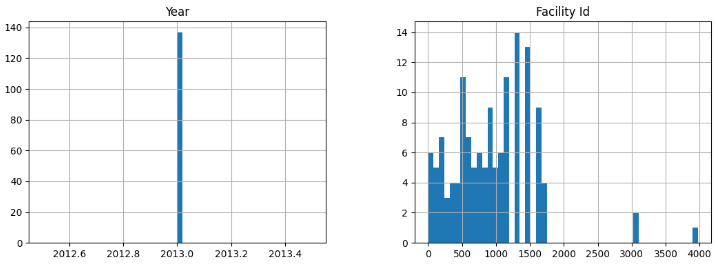
**Lower 95CI RAR Upper 95CI RAR Compare to State**

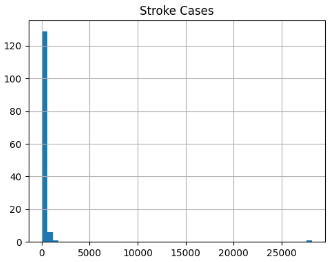
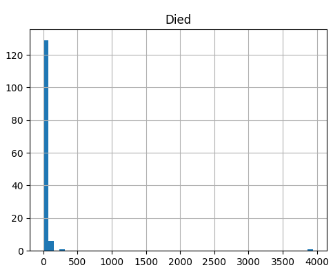
There is only one instance where the instance if **referring for the whole New York** with null values for the Stroke Designated Center, Coverdell Hospital, Expected Rate, Lower 95CI RAR, Upper 95CI RAR, Compare to State. The values for the Hospital Name and Hospital County are “**Statewide”** and the Facility Id value is **“0”**.

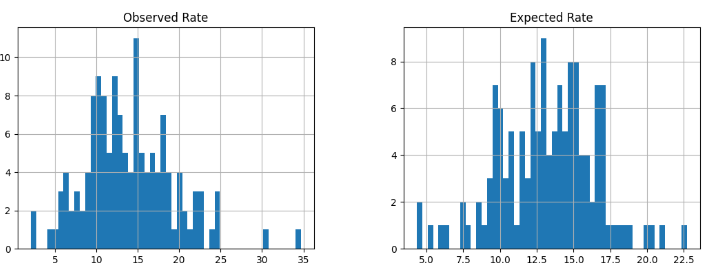
**All the features in all the other instances have non-null values.**

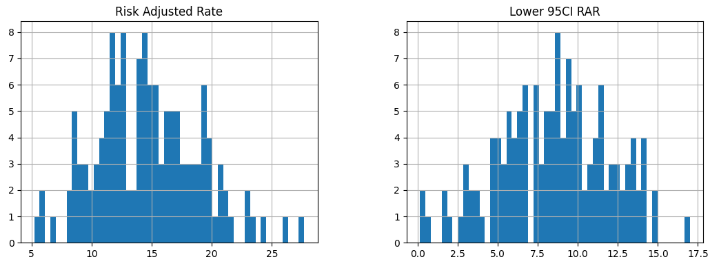
In every instance the value for the feature Year is 2013.

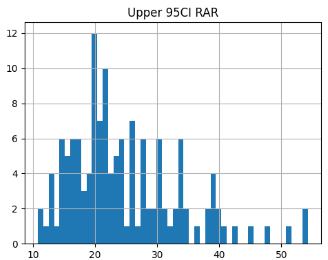
(Disclaimer: I think that this dataset is not very useful in our research but I included it just in case)



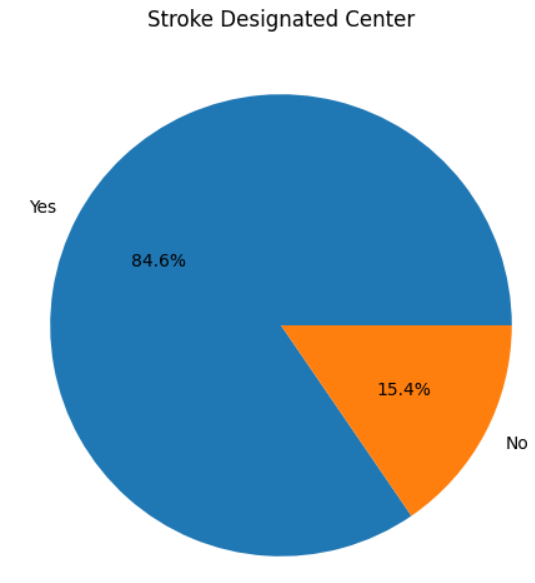
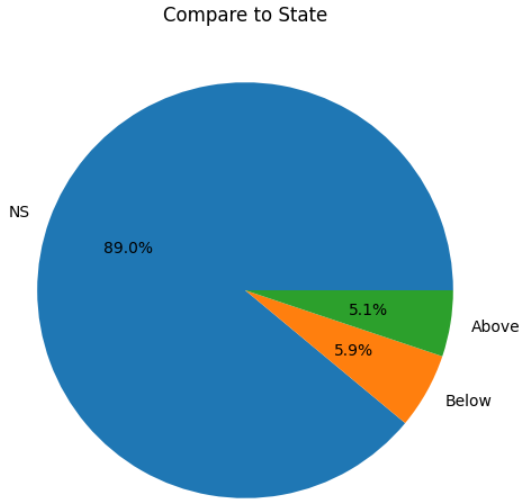


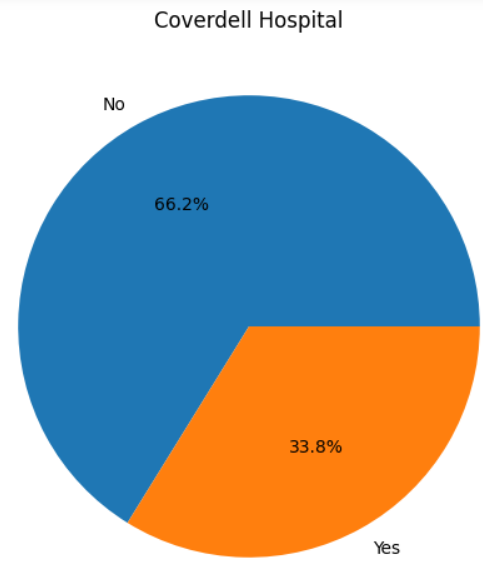
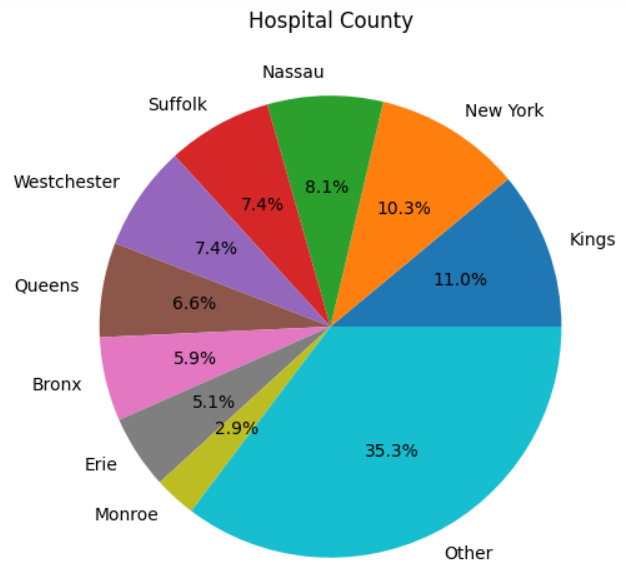




|  |  |
| --- | --- |
| Stroke Cases | 1 |
| Died | 0.950054 |
| Lower 95CI RAR | 0.315589 |
| Expected Rate | 0.250939 |
| Observed Rate | -0.00618 |
| Facility Id | -0.03457 |
| Risk Adjusted Rate | -0.18081 |
| Upper 95CI RAR | -0.54039 |

The table represents the correlation between **Stroke Cases** and all the other numeric features.



**4.Dataset**

**Title:**

NCHS - Potentially Excess Deaths from the Five Leading Causes of Death

**Link:**

<https://data.world/us-hhs-gov/112731d8-6a9c-4835-a4cd-598f21f13a39?fbclid=IwAR3u0z3XT-PptT1TZBcgEQKL8uinfPU7jkeV2A1g566U18Sa7OyWGVOc1YU>

**Published by:**

U.S. Department of Health & Human Services

**Description:**

MMWR Surveillance Summary 66 (No. SS-1):1-8 found that **nonmetropolitan areas have significant numbers of potentially excess deaths from the five leading causes of death**. These figures accompany this report by presenting information on potentially excess deaths in nonmetropolitan and metropolitan areas at the state level. They also add **additional years of data and options for selecting different age ranges and benchmarks**. Potentially excess deaths are defined in MMWR Surveillance Summary 66(No. SS-1):1-8 as deaths that exceed the numbers that would be expected if the death rates of states with the lowest rates (benchmarks) occurred across all states. They are calculated by subtracting expected deaths for specific benchmarks from observed deaths. **Not all potentially excess deaths can be prevented;** some areas might have characteristics that predispose them to higher rates of death. However, **many potentially excess deaths might represent deaths that could be prevented through improved public health programs that support healthier behaviors and neighborhoods or better access to health care services**. Mortality data for U.S. residents come from the National Vital Statistics System. Estimates based on fewer than 10 observed deaths are not shown and shaded yellow on the map. Underlying cause of death is based on the International Classification of Diseases, 10th Revision (ICD-10) Heart disease (I00-I09, I11, I13, and I20–I51) Cancer (C00–C97) Unintentional injury (V01–X59 and Y85–Y86) Chronic lower respiratory disease (J40–J47) Stroke (I60–I69) **Locality (nonmetropolitan vs. metropolitan) is based on the Office of Management and Budget’s 2013 county-based classification scheme.** Benchmarks are based on the three states with the lowest age and cause-specific mortality rates. Potentially excess deaths for each state are calculated by subtracting deaths at the benchmark rates (expected deaths) from observed deaths. Users can explore three benchmarks: “**2010 Fixed**” is a fixed benchmark **based on the best performing States in 2010**. “**2005 Fixed**” is a fixed benchmark **based on the best performing States in 2005**. “**Floating” is based on the best performing States in each year so change from year to year.** SOURCES CDC/NCHS, National Vital Statistics System, mortality data (see <http://www.cdc.gov/nchs/deaths.htm>); and CDC WONDER (see [http://wonder.cdc.gov](http://wonder.cdc.gov/)). REFERENCES Moy E, Garcia MC, Bastian B, Rossen LM, Ingram DD, Faul M, Massetti GM, Thomas CC, Hong Y, Yoon PW, Iademarco MF. Leading Causes of Death in Nonmetropolitan and Metropolitan Areas – United States, 1999-2014. MMWR Surveillance Summary 2017; 66(No. SS-1):1-8. Garcia MC, Faul M, Massetti G, Thomas CC, Hong Y, Bauer UE, Iademarco MF. Reducing Potentially Excess Deaths from the Five Leading Causes of Death in the Rural United States. MMWR Surveillance Summary 2017; 66(No. SS-2):1–7.

**Data analysis:**

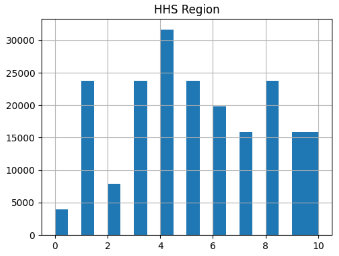
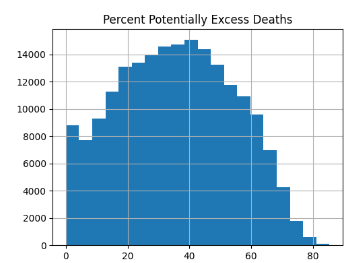
The dataset has 205920 instances with 13 features each. Each instance has the following features:

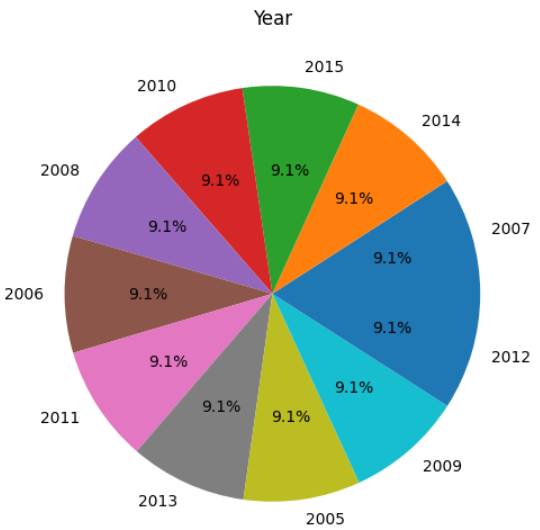
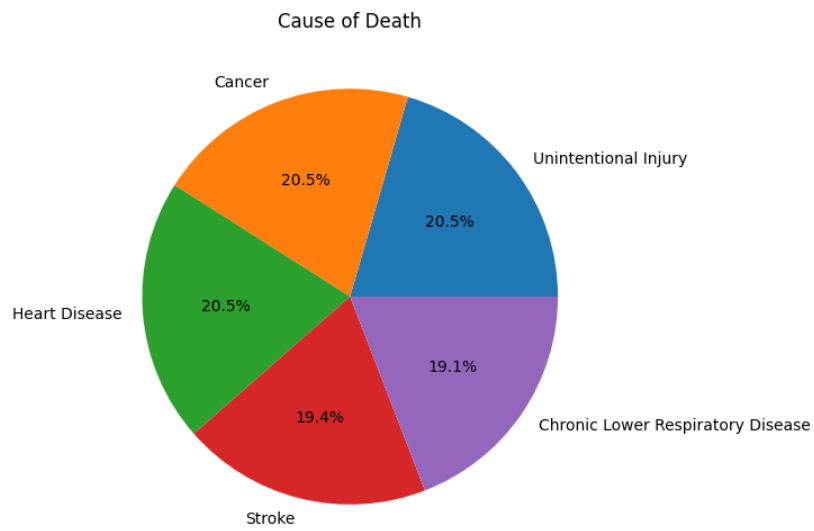
**Year, Cause of Death, State, State FIPS Code, HHS Region, Age Range, Benchmark, Locality, Observed Deaths, Population, Expected Deaths, Potentially Excess Deaths, Percent Potentially Excess Deaths.**

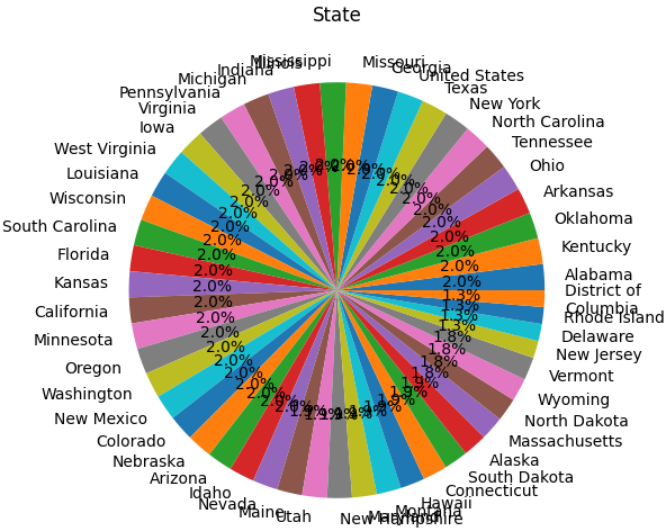
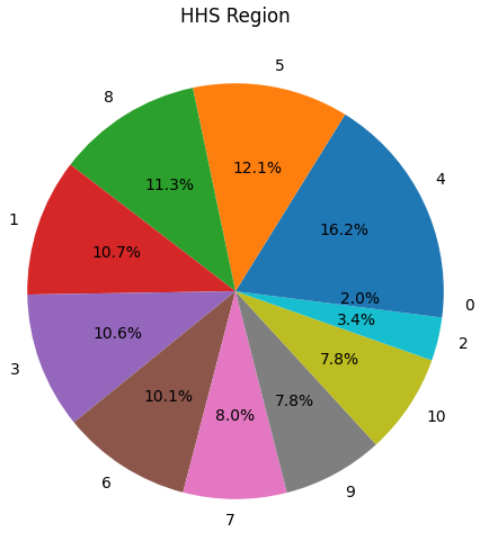
10212 instances have null values for the features: Observed Deaths, Expected Deaths, Potentially Excess Deaths, Percent Potentially Excess Deaths, and the rest (195708 instances) don’t have any features with null values. So in conclusion 10212 instances don’t have any significant data apart from the obvious default data (place, year, age group and so on).

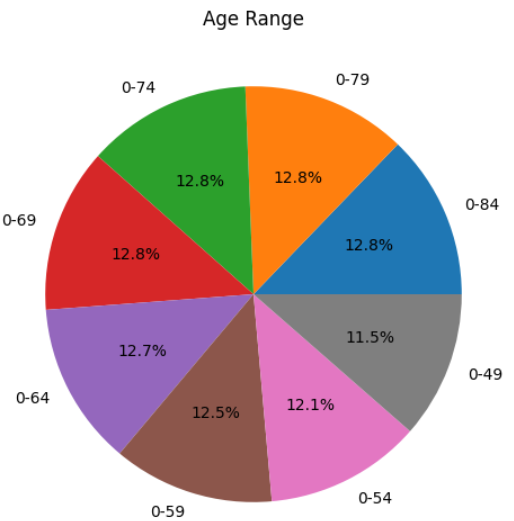
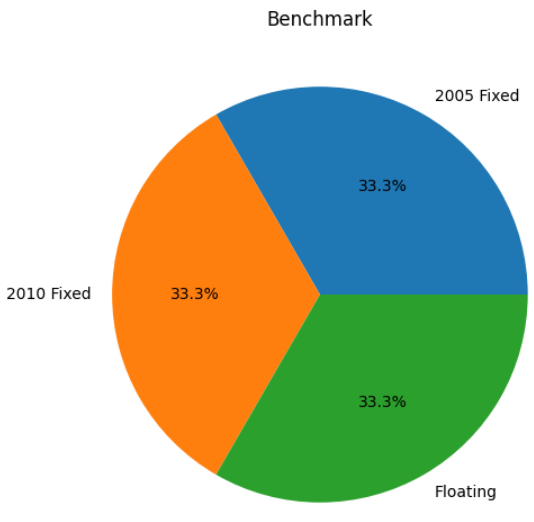
After removing the instances with null values the data ends up fairly balanced having almost the same number or instances for every different value for all of the features as you can see from the pie charts bellow.

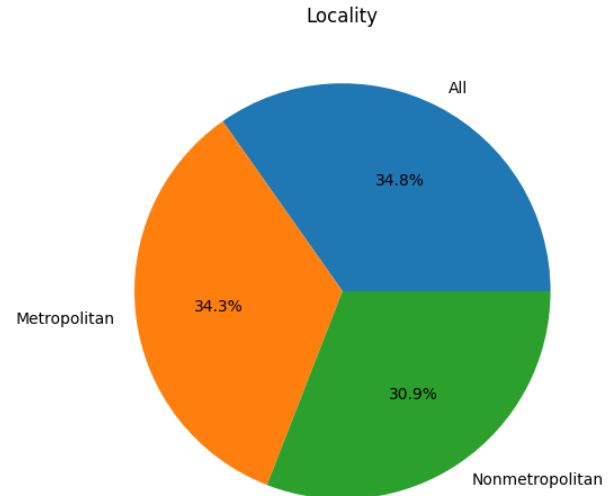
(Tip: If needed for the age range we can subtract the smaller category and we will have deaths for age range like 64-69 instead of the 0-69.)

** **

** **

** **

** **

****

**5.Dataset**

**Title:**

Brain Stroke Dataset

**Link:**

<https://data.world/researchersj/brain-stroke-dataset?fbclid=IwAR3Y-rrWMYck5OoP15HJVJiihvZVvzVyUj8B7cijBO-Q3XbmQX8fMAd-n0o>

**Published by:**

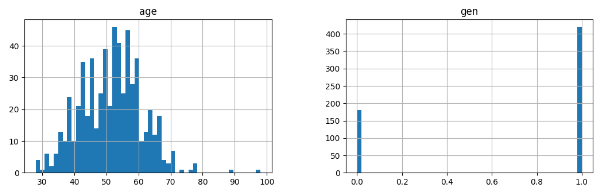
<https://data.world/researchersj>

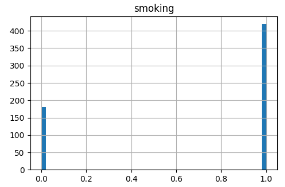
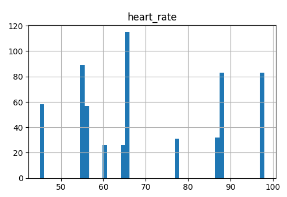
**Data analysis:**

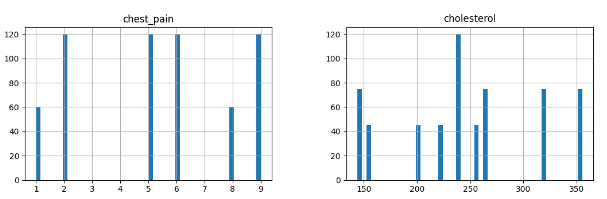
This dataset has 600 instances each with 9 features. None of the features have null values for any instance. Each instance contains the following features: **age, gen, smoking, heart\_rate, chest\_pain, cholesterol, bloodpressure, bloodsugar, stroke.**

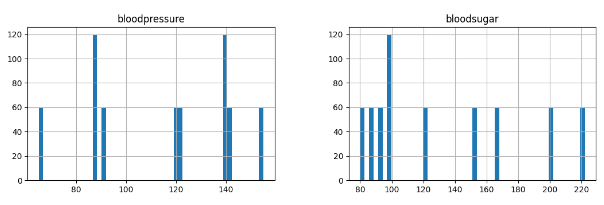
The dataset contains instances with 1,2,3 strokes without having any instance with 0 strokes(regular people).

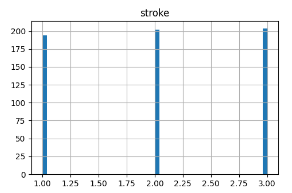
I couldn’t find a description for the features. Almost all of the features are self-describing but the feature **“gen”** is feature with 2 values 1 and 0 (true, false) which means is type of gene in the patients but without any description on what gene it is.



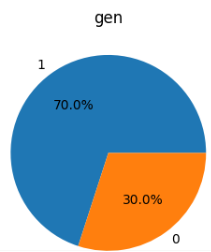
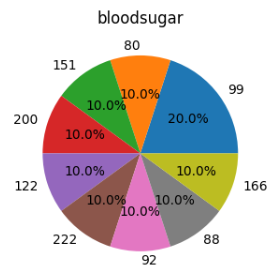
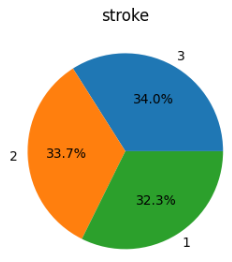
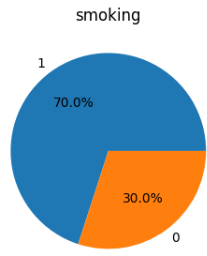
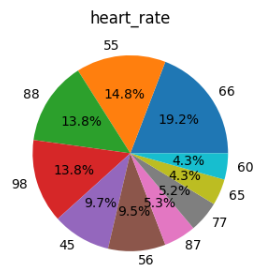
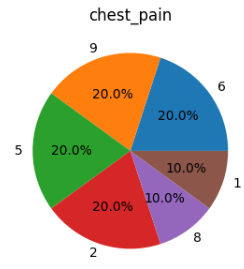
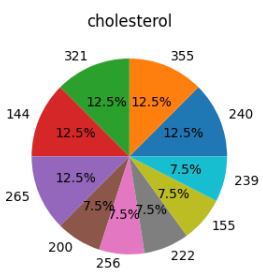
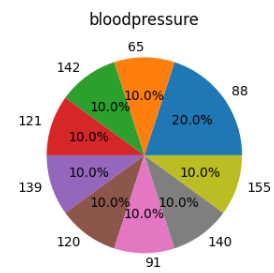
 







(Warning: The dataset is from a person without any description, and the values for the features are very clustered, for example there are 9 different values for the bloodpresure witch is a bit odd given there are exact same number of people having bloodpresure 65,91,120,121,139,140,142,155 and nobody having bloodpresure the numbers in between. Also for the mesured bloodpresures there are the same number of instances for each value exept one. Which is a little suspecious and we need to take the dataset with a grain of salt)

|  |  |
| --- | --- |
| stroke | 1.00E+00 |
| cholesterol | 3.92E-02 |
| heart\_rate | 9.89E-03 |
| bloodpressure | 4.16E-03 |
| smoking | -4.54E-16 |
| chest\_pain | -2.97E-03 |
| bloodsugar | -3.65E-03 |
| gen | -4.47E-03 |
| age\_group | -2.86E-01 |
| age | -2.88E-01 |

From the correlation matrix we can see that none of the features are in direct linear corelation with the number of strokes

**6.Dataset**

**Title:**

Brain Stroke Dataset

**Link:**

<https://www.kaggle.com/datasets/jillanisofttech/brain-stroke-dataset>

**Published by:**

Open knowledge foundation (<https://okfn.org/>)

**Description:**

The purpose of the dataset is predicting first strokes of patient based on few simple features. The data is oversampled with 248 true (brain stroke) instances and 4733 false (not brain stroke) instances.

**Data analysis:**

The dataset contains 11 features. All the features apart from smoking don’t have null values.

**gender**: "Male", "Female" or "Other";

**age** of the patient;

**hypertension**: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension;

**heart** **disease**: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease;

**Ever-married**: "No" or "Yes";

**work** **type**: "children", "Govtjov", "Never worked", "Private" or "Self-employed";

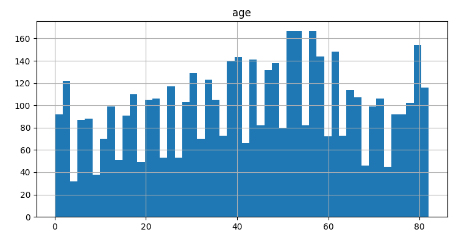
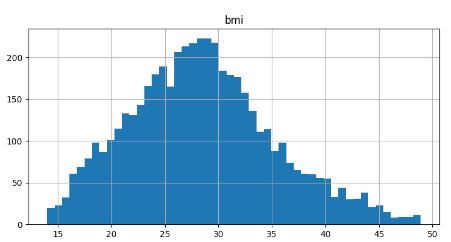
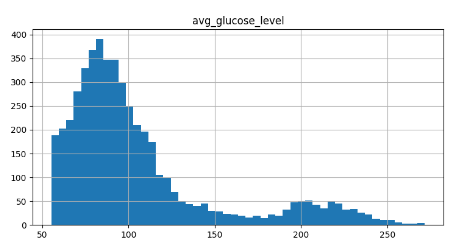
**Residencetype**: "Rural" or "Urban";

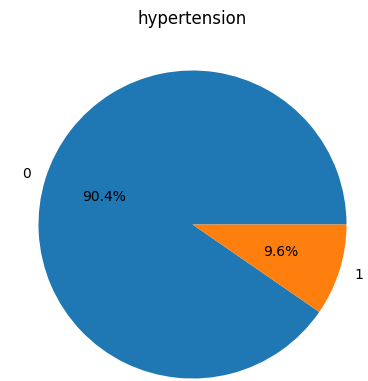
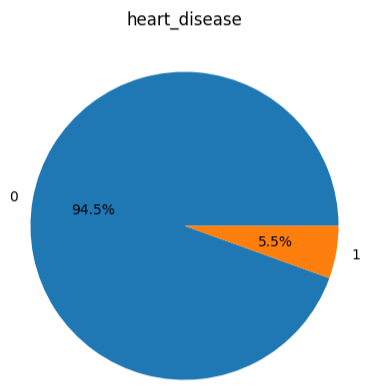
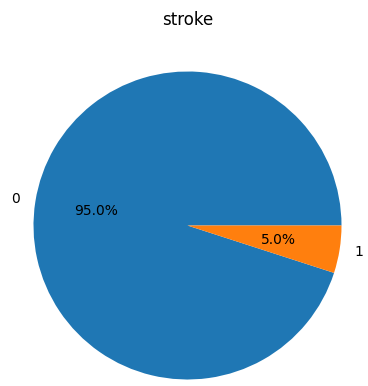
**average glucose level in blood;**

**BMI** (body mass index);

**smoking\_status**: "formerly smoked", "never smoked", "smokes" or "Unknown";

**stroke**: 1 if the patient had a stroke or 0 if not.

By mapping **gender** from **'Male'** and **'Female'** to 1 and 0 respectivly, **smoking\_status** from **'never** **smoked'**, **'Unknown'**, **'formerly** **smoked'** and **'smokes'** to -1,0,1,2, **ever\_married** from **'No'** and **'Yes'** to 0,1 and **Residence\_type** from ‘**Urban’** and ‘**Rural’** to 0,1 we got linear **correlation** **matrix** with the following values.

|  |  |
| --- | --- |
| stroke | 1 |
| age | 0.246478 |
| heart\_disease | 0.13461 |
| avg\_glucose\_level | 0.133227 |
| hypertension | 0.131965 |
| ever\_married | 0.108398 |
| bmi | 0.056926 |
| smoking\_status | 0.031013 |
| gender | 0.00887 |
| Residence\_type | -0.01649 |

**7.Dataset**

**Title:**

Cerebral Stroke Prediction-Imbalanced Dataset

**Link:**

<https://www.kaggle.com/datasets/shashwatwork/cerebral-stroke-predictionimbalaced-dataset>

**Published by:**

Creative commons (<https://creativecommons.org/>)

**Description:**

The Electronic Health Record (EHR) controlled by McKinsey & Company was used as the dataset in our research which was a part of their healthcare hackathon. The dataset is easily accessible as a free dataset repository. The gathered data contained information of 29, 072 patients having 12 common attributes. Out of the 12 attributes, 11 of them are input features including age, gender, marital status, patient identifier, work type, residence type (urban/rural), binary attribute heart disease condition, body mass index, smoking status of patient, glucose level and binary attribute hypertension indicating a patient is suffering from hypertension or not. The 12th attribute is the binary output attribute indicating a patient is suffered stroke or not.

Data for A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical-datasets

**Data analysis:**

The data contains 43400 instances. From the 43400 there are 29072 with non-null values.

Almost all 12 features have all non-null values except **bmi** (with 1462 instances with null values) and **smoking\_status** (with 13292 instances with null values).

The dataset has the following 12 features:

**id, gender, age, hypertension, heart\_disease, ever\_married, work\_type, Residence\_type, avg\_glucose\_level, bmi, smoking\_status, stroke.**

**gender** has values “**Male**”(25665 instances), “**Female**”(17724 instances) and “Other”(11 intances)

**hypertension** has values “**0**” (false, 39339 instances) and “**1**” (true, 4061 instances).

**heart\_disease** has values “**0**” (false, 41338 instances) and “**1**” (true, 2062 instances).

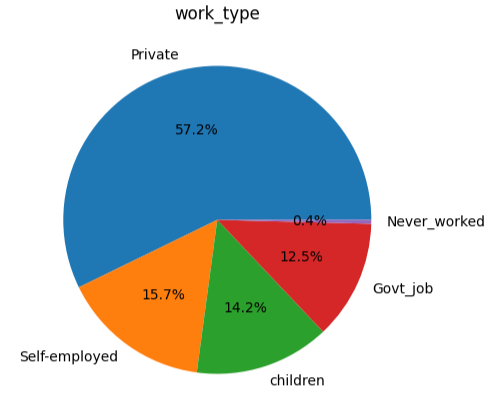
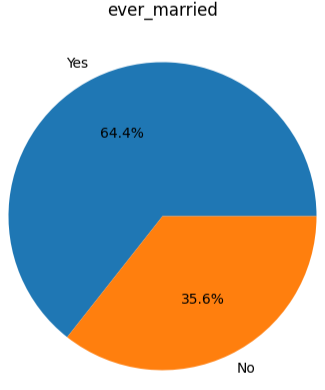
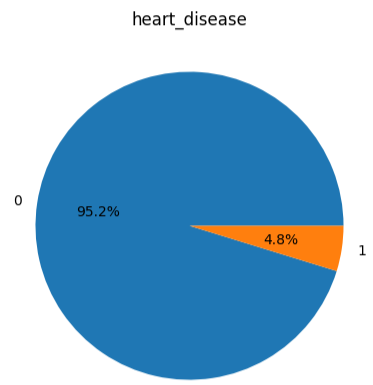
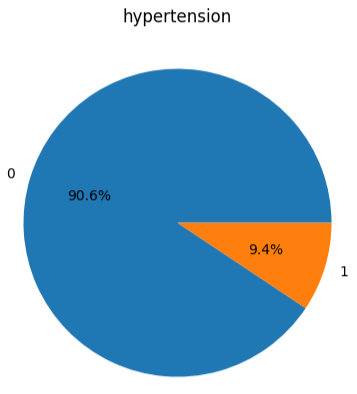
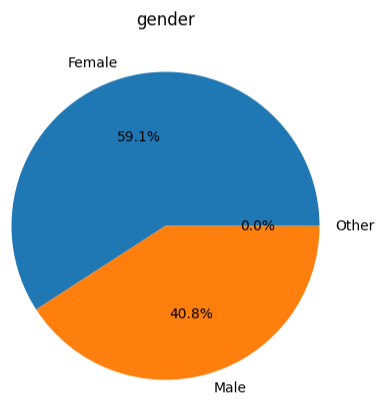
**ever\_married** has values “**Yes**” (27938 instances) and “**No**” (15462 instances).

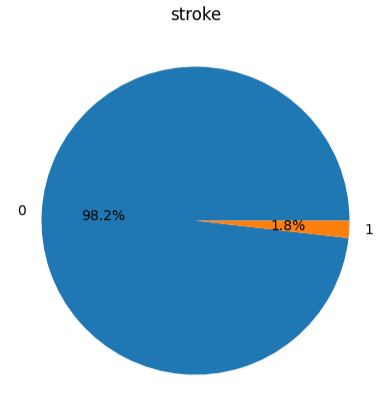
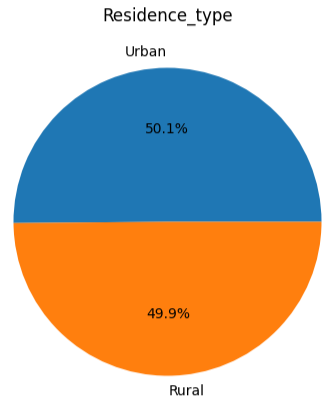
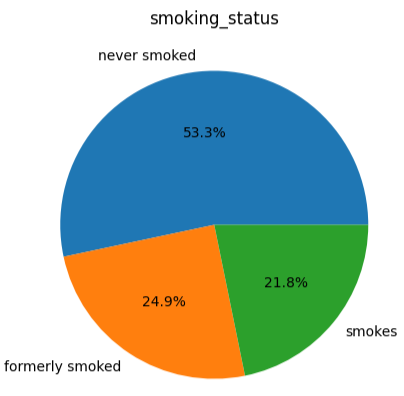
**work\_type** has values "**Private**" (24834 instances), "**Self-employed**" (6793 instances), "**children**" (6156 instances), "**Govt\_job**" (5440 instances), "**Never\_worked**" (177 instances).

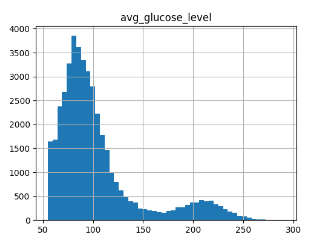
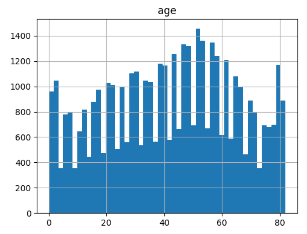
**Residence\_type** has values “**Urban**” (21756 instances) and “**Rural**” (21644 instances).

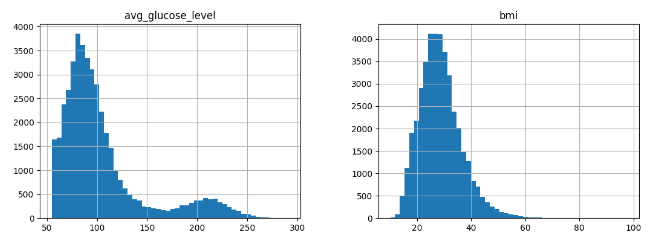
**Stroke** has values “**0**” (false, 42617 instances) and “**1**” (true, 783 instances).

**smoking\_status** has values "**never smoked**" (16051 instances), "**formerly smoked**" (7487 instances), "**smokes**" (6561 instances).







|  |  |
| --- | --- |
| stroke | 1 |
| age | 0.15605 |
| heart\_disease | 0.113756 |
| avg\_glucose\_level | 0.078908 |
| hypertension | 0.075322 |
| ever\_married | 0.071917 |
| bmi | 0.020284 |
| smoking\_status | 0.014329 |
| gender | 0.011324 |
| id | 0.002975 |
| Residence\_type | -0.00224 |

The correlation matrix was generated in a simmular way as the 6.Dataset only here there is no value “Unknown” in the feature **smoking\_status** and I have removed the 11 instances with value Other for gender.

**8.Dataset**

**Title:**

Brain Stroke CT Image Dataset

**Link:**

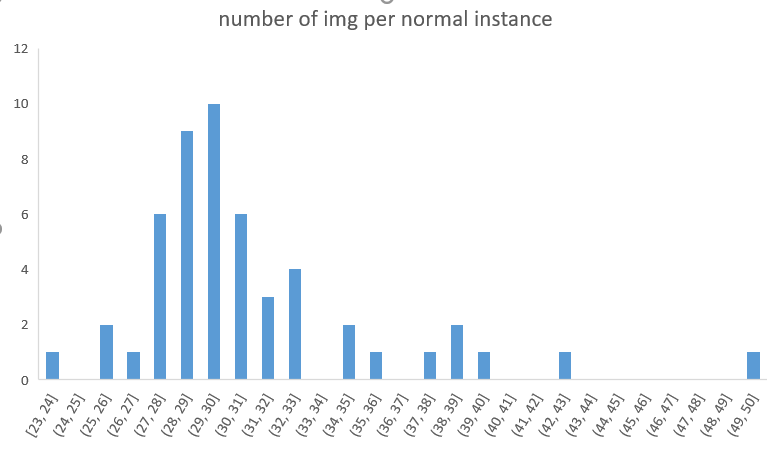
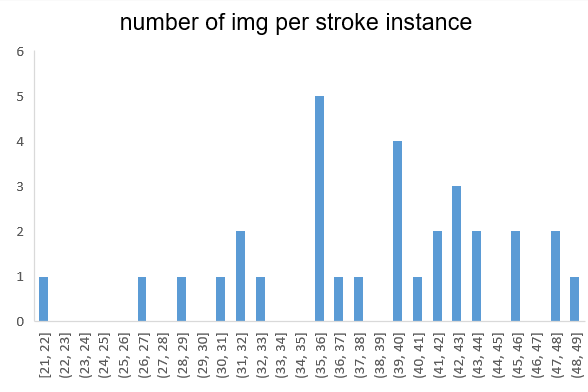
<https://www.kaggle.com/datasets/afridirahman/brain-stroke-ct-image-dataset>

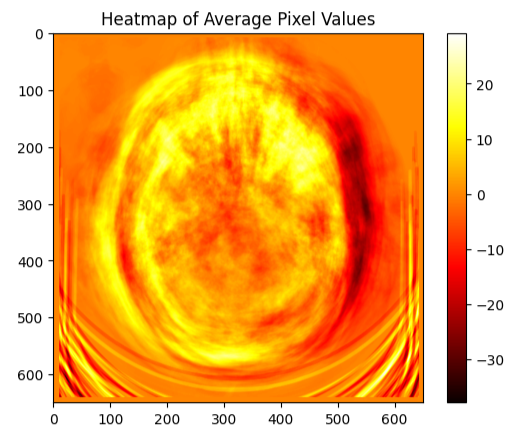
**Published by:**

Unknown

**Data analysis:**

There are **2501 images of CT**[[1]](#footnote-1) in this dataset from which **1551 are from normal** (without stroke) people and **950 are from people who had stroke**. There are **51 different normal instances** each having **on average 31 pictures** of the brain and **31 instances witth stroke having on average 39 pictures**. Each picture has 422500 features (pixels) with **dimensions 650x650 pixels** and is in the **JPG format**. This means that on avera for each (head/person/instance) we have 14 365 000 features/pixels.





**9.Dataset**

**Title:**

Acute Ischemic Stroke MRI

**Link:**

<https://www.kaggle.com/datasets/buraktaci/mri-stroke>

**Published by:**

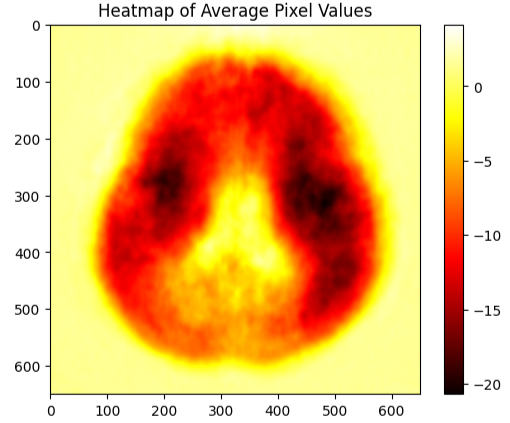
PDRNet, Biomedical Signal Processing and Control (<https://www.pdr.net/> )

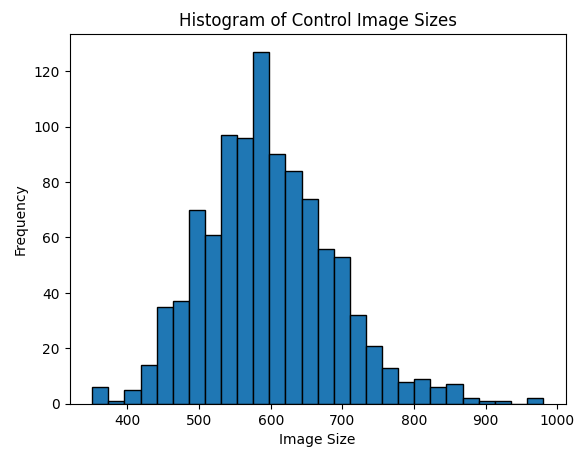
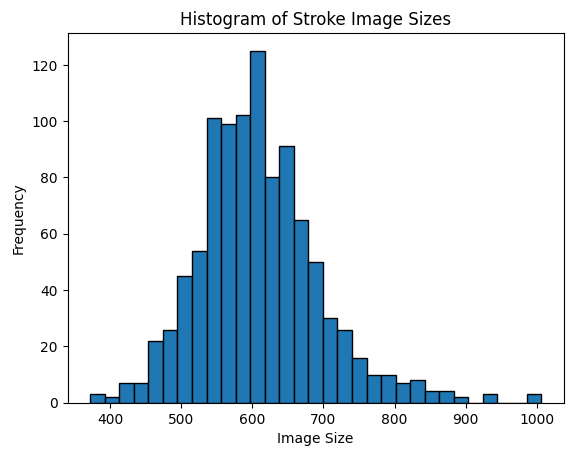
**Description:**

In this research, three **brain magnetic resonance image datasets were used** to test the proposed model. A deep feature engineering model has been proposed to deploy the raw MRI and four preprocessing algorithms: GradCAM, histogram-matching, canny edge detection, and **Locally Interpretable Model-Agnostic Explanations(LIME**). The deep features have been extracted using Resnet101 and DenseNet201 pre-trained convolutional neural networks (CNN). Thus, this model is titled preprocessing based **DenseNet and ResNet (PDRNet).**

**Data analysis:**

There are **2009 images of MRI**[[2]](#footnote-2) in this dataset from which **1008 are control** (from people without stroke) and **1002 are from people who had Acute Ischemic Stroke**. From all control images 203 are in the JPG format and the rest are PNG format with various dimensions ranging from 348 to 980 pixels. 130 images from those with stroke are JPG and the rest PNG format. The dimensions of the images also vary from 372 to maximum of 1006 pixels.





**10.Dataset**

**Title:**

Mortality from stroke

**Link:**

<https://digital.nhs.uk/data-and-information/publications/statistical/compendium-mortality/current/mortality-from-stroke/mortality-from-stroke-crude-death-rate-by-age-group-3-year-average-mfp>

**Published by:**

https://digital.nhs.uk/

**Description:**

The purpose of the study is to reduce deaths from stroke. The study aims to measure the crude death rate for different age groups using a 3-year average, **specifically employing the MFP (Mean of Future Projections**) method. The geographic coverage of the study includes **England and Wales**. The geographical granularity of the data is presented at both the country and regional levels.

**Data analysis:**

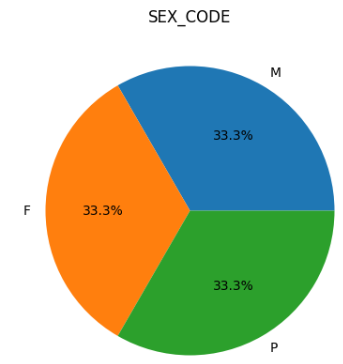
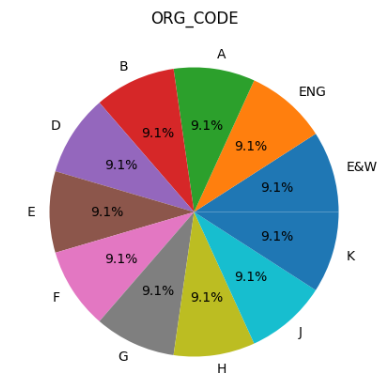
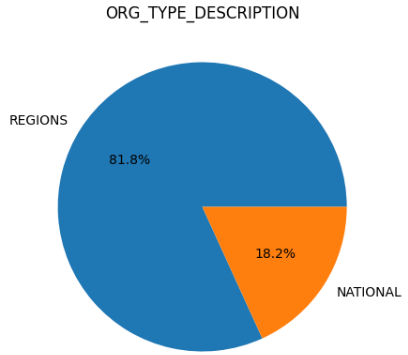
The dataset contains 231 instances each containing 9 features witch don’t have any null values.

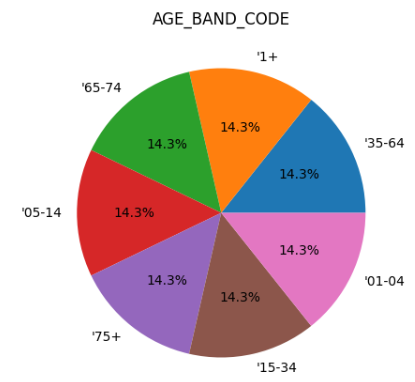
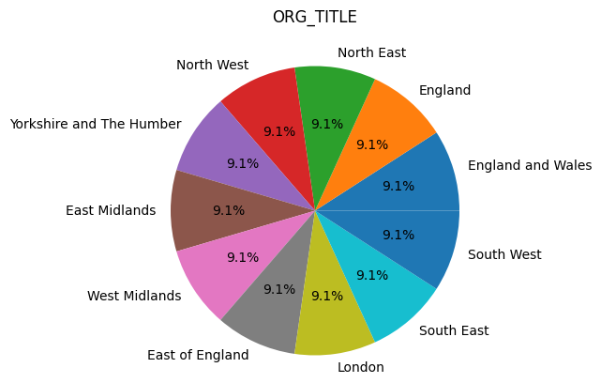
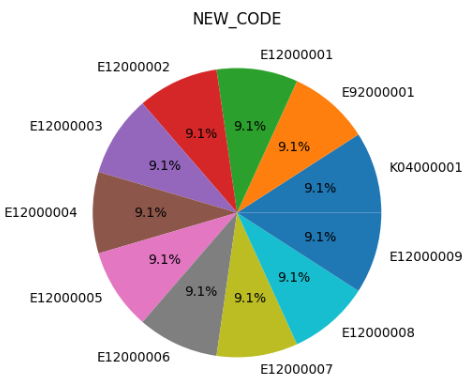
The 9 features are: **YEAR, Filename, ORG\_TYPE\_DESCRIPTION, ORG\_CODE, NEW\_CODE, ORG\_TITLE, SEX\_CODE, AGE\_BAND\_CODE, Rate**.

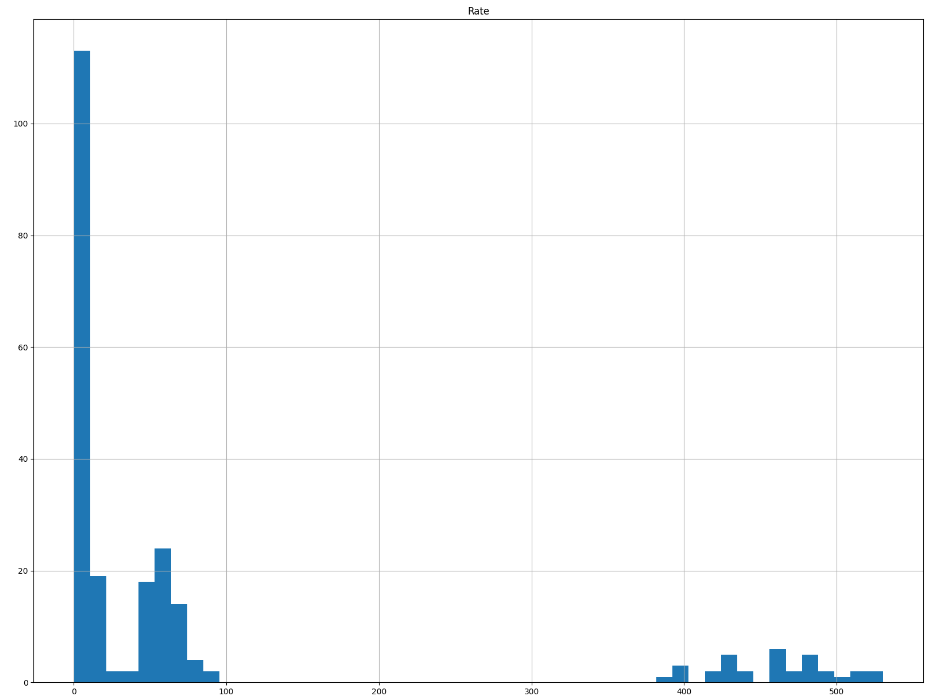
"Crude death rate"(Rate feature) refers to the number of deaths from stroke in a population per unit of population.

For all the instances features Year and Filename have one constant value.

(Disclaimer: nowhere in the study does it say if the stroke rate is brain or heart stroke)







**11.Dataset**

**Title:**

Lesion-Symptom Mapping in Brain Tumor and Stroke Patients

**Link:**

<https://data.mendeley.com/datasets/k2847vw9gg/1>

**Published by:**

<https://plu.mx/plum/a?mendeley_data_id=k2847vw9gg&theme=plum-bigben-theme>

(Contributors: Eva van Grinsven ,Anouk R. Smits)

**Description:**

Data accompanying the paper: The impact of etiology in lesion-symptom mapping – A **direct comparison between tumor and stroke** Authors: E.E. van Grinsven, A.R. Smits, E. van Kessel, M.A.H. Raemaekers, E.H.F. de Haan, I.M.C. Huenges Wajer, V.J. Ruijters, M.E.P. Philippens, J.J.C. Verhoeff, N.F. Ramsey, P.A.J.T. Robe, T.J. Snijders and M.J.E. van Zandvoort Background: The behavioral consequences of lesions from different etiologies may vary because of how they affect brain tissue and how they are distributed. Therefore, **the main objective of the present study was to directly compare lesion-symptom maps for memory and language functions from two populations, a tumor versus a stroke population.** Methods: Data from **two different studies were combined**. Both the **brain tumor (N = 196)** and **stroke (N = 147)** patient populations underwent **neuropsychological testing and an MRI,** **pre-operatively for the tumor** population and **within three months after stroke**. For this study, we selected two internationally widely used standardized cognitive tasks, the **Rey Auditory Verbal Learning Test** and the **Verbal Fluency Test**. We used a state-of-the-art machine learning-based, multivariate voxel-wise approach to produce lesion-symptom maps for these cognitive tasks for both populations separately and combined. To substantiate the results from the multivariate lesion-symptom mapping, additional univariate lesion-symptom mapping was performed for each cognitive task for the tumor and stroke data separately Results: Our **lesion-symptom mapping results for the separate patient populations largely followed the expected neuroanatomical pattern based on previous literature**. Substantial differences in lesion distribution hindered direct comparison. Still, in brain areas with adequate coverage in both groups, considerable **LSM** differences between the two populations were present for both memory and fluency tasks. Conclusion: The differences in the lesion-symptom maps between the stroke and tumor population could partly be explained by differences in lesion volume and topography. Despite these methodological limitations, **our results confirmed that etiology matters when investigating the cognitive consequences of lesions with lesion-symptom mapping.** Therefore, caution is advised with generalizing lesion-symptom results across etiologies.

Folders:

(1) Lesion overlap maps: contains the etiology-specific lesion overlap maps.

(2) Multivariate lesion-symptom maps: contains the thresholded p-value maps for the multivariate analysis for each cognitive task and for each etiology, separately.

(3) Power maps for Univariate analyses: contains all power maps for each univariate lesion-symptom mapping analysis.

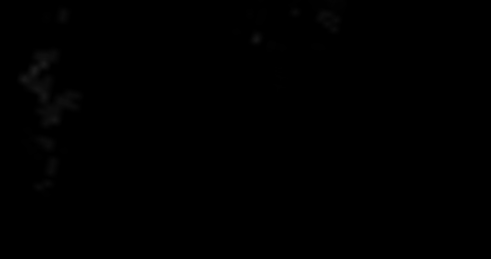
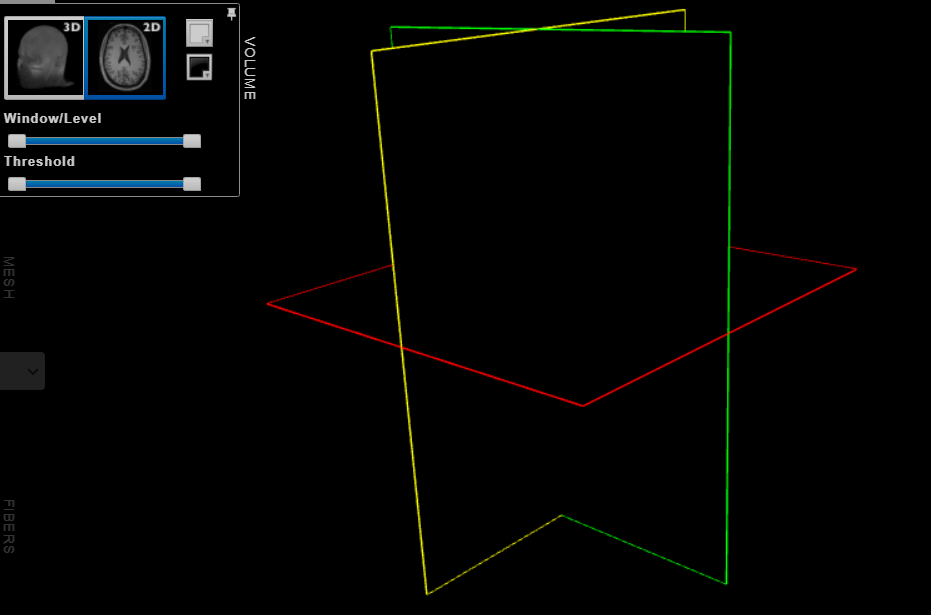
(4) Univariate lesion-symptom maps: contains the thresholded Z-score maps for the univariate analysis for each cognitive task and for each etiology, seperately.

All data is registered to the 2mm MNI standard brain. LVC = lesion volume correction.

lesion refers to an area of damaged tissue or abnormality in a specific part of the body.

**Data analysis:**

The data is in NII format. I downloaded a NII viewer and tried online NII viewer but for almous all the data I got nothing that is visible.



**12 .Dataset**

**Title:**

Prognostication of Recovery from Acute Stroke (PRAS Dataset)

**Link:**

<https://data.mendeley.com/datasets/y86srgks26/1>

**Published by:**

<https://plu.mx/plum/a?mendeley_data_id=y86srgks26&theme=plum-bigben-theme>

Contributors: Yauhen Statsenko, Fatmah Al Zahmi, Miklos Szolics, Jamal Al Koteesh.

**Description:**

The file titled "Stroke\_ICH\_Data" contains a table which is labeled the PRAS dataset after the project title “Prognostication of Recovery from Acute Stroke” (6,7). The table holds records for 2016-2019 years from the stroke registry of Al Ain Hospital which serves as a tertiary level of care clinic. The dataset consists of de-identified patients data and weather parameters. We retrieved information on the following clinicodemographic risk factors of hemorrhagic stroke from medical histories: **age, sex, body mass index, smoking status, history of cardiovascular diseases, and ethnicity.** From the website National Oceanic and Atmospheric Administration for Al Ain city we requested **the weather parameters for seven days before the stroke onset.**

**Data analysis:**

This dataset has 161 features with 110 features each witch are listed and explained down bellow.

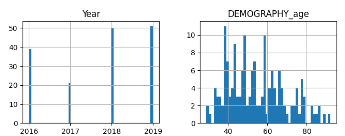
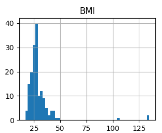
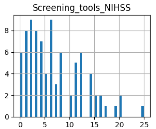
1. Year: The year when the data was recorded or collected.
2. DEMOGRAPHY\_age: Age of the patient, capturing demographic information.
3. DEMOGRAPHY\_sex: Sex of the patient (male or female), also a demographic attribute.
4. DEMOGRAPHY\_nationality: Nationality of the patient.
5. History\_OldStroke: Indicates if the patient has a history of old strokes.
6. History\_DM: History of Diabetes Mellitus (DM).
7. History\_HyperTension: History of hypertension (high blood pressure).
8. History\_IschemicHeartDisease: History of Ischemic Heart Disease (coronary artery disease).
9. History\_ArterFibrillation: History of Atrial Fibrillation (a type of irregular heart rhythm).
10. History\_HyperLypidAemia: History of hyperlipidemia (high cholesterol or lipids).
11. History\_Smoking: Indicates if the patient has a history of smoking.
12. BMI: Body Mass Index, a measure of body fat based on height and weight.
13. ONSET\_LKW\_time: Time of onset of symptoms, possibly related to Last Known Well (LKW) time.
14. ONSET\_Date: Date of onset of symptoms.
15. Screening\_tools\_NIHSS: National Institutes of Health Stroke Scale (NIHSS), a tool for assessing stroke severity.
16. Lab\_Investigation\_Trop I: Lab investigation result for Troponin I, a protein indicating heart muscle damage.
17. Lab\_Investigation\_international\_norm\_ratio: Lab investigation result for International Normalized Ratio (INR), used to monitor blood clotting.
18. Lab\_Investigation\_C-reactive protein: Lab investigation result for C-reactive protein, indicating inflammation.
19. Lab\_Investigation\_TotalCholeserol: Lab investigation result for total cholesterol level.
20. Lab\_Investigation\_low-density\_lipoprotein: Lab investigation result for low-density lipoprotein (LDL) cholesterol.
21. Lab\_Investigation\_POC\_Random blood sugar: Lab investigation result for random blood sugar, indicating glucose levels.
22. Lab\_Investigation\_Creatinine: Lab investigation result for creatinine, a marker of kidney function.
23. Discharge\_Plan\_Modified\_Rankin\_Score: Modified Rankin Scale score at discharge, used to assess the level of disability after a stroke.
24. DEMOGRAPHY\_agerange: Age range of the patient, another demographic attribute.
25. Clinical\_Diagnosis: Clinical diagnosis of the patient.
26. MIMICS: Not specified in the given list, but it might refer to medical imaging data or diagnostic tests.
27. ICH: Intracranial hemorrhage, a type of stroke caused by bleeding within the brain.
28. IS: Ischemic Stroke, a type of stroke caused by a blocked blood vessel in the brain.
29. IS\_verified: Verification status of Ischemic Stroke.
30. TIA\_verified: Verification status of Transient Ischemic Attack (TIA), a temporary stroke-like episode.
31. IS - outOfWindow, IS - rtpA, IS - withinWindow: Not specified in the given list, but they might be related to specific categories or treatments for Ischemic Stroke.
32. Day\_Time: Time of day when data was recorded.
33. TEMP, STP, WDSP, RH, HUMIDEX: Meteorological parameters related to temperature, atmospheric pressure, wind speed, relative humidity, and the combination of temperature and humidity.
34. TEMP1, TEMP2... TEMP7: Temperature readings at different day in a 7day period.
35. STP1, STP2... STP7: Atmospheric pressure readings at different day in a 7day period.
36. WDSP1, WDSP2... WDSP7: Wind speed readings at different day in a 7day period.
37. RH1, RH2... RH7: Relative humidity readings at different day in a 7day period.
38. HUMIDEX1, HUMIDEX2... HUMIDEX7: Humidex values at different day in a 7day period.
39. TDIF1, TDIF2... TDIF7: Temperature difference values at different day in a 7day period.
40. PDIF1, PDIF2... PDIF7: Pressure difference values at different day in a 7day period.
41. WDIF1, WDIF2... WDIF7: Wind difference values at different day in a 7day period.
42. RHDIF1, RHDIF2... RHDIF7: Relative humidity difference values at different day in a 7day period.
43. HDIF1, HDIF2... HDIF7: Humidex difference values at different day in a 7day period.
44. NIHSS\_group: Group classification based on NIH Stroke Scale scores, used for assessing stroke severity.

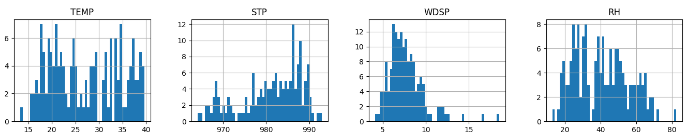
From all the instances 3 of them have null values for the following features: **TEMP**, **STP, WDSP, RH, HUMIDEX TEMP1-TEMP7, STP1-STP7, WDSP1-WDSP7, RH1-RH7,HUMIDEX1-HUMIDEX7, TDIF1-TDIF7, PDIF1-PDIF7, WDIF1-WDIF7, RHDIF1-RHDIF7, HDIF1-HDIF7.**

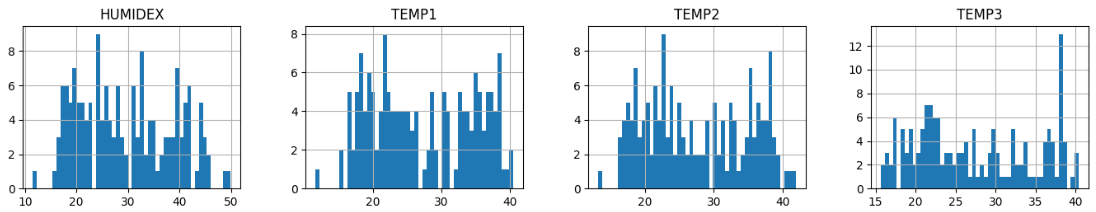
The features: **Lab\_Investigation\_Trop I, Lab\_Investigation\_international\_norm\_ratio, Lab\_Investigation\_C-reactive protein, Lab\_Investigation\_TotalCholeserol, Lab\_Investigation\_low-density\_lipoprotein, Lab\_Investigation\_POC\_Random blood sugar and Lab\_Investigation\_Creatinine** have NaN values for all 161 instances. Maybe I read them incorrectly with pandas and exel. But making error on two places I think is highly unlikely.

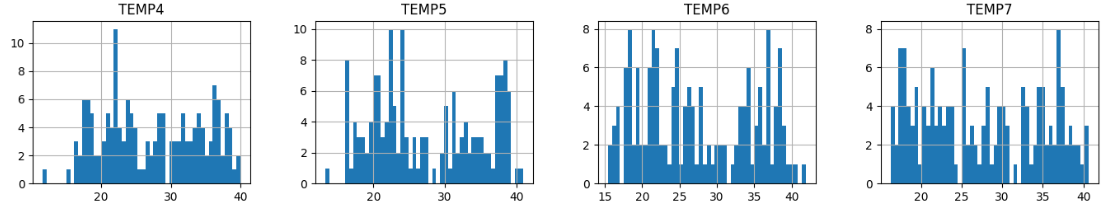
Also null values have the features: **ONSET\_LKW\_time, Screening\_tools\_NIHSS, Discharge\_Plan\_Modified\_Rankin\_Score, Clinical\_Diagnosis, Day\_Time, NIHSS\_group,**

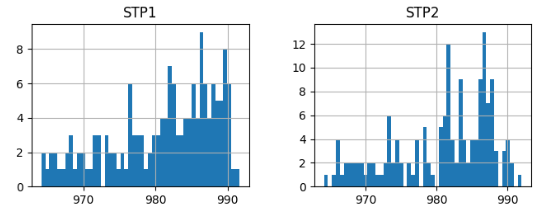
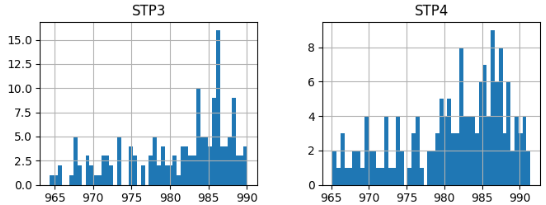
In total staggering 88 features have at least one null value for an instance.

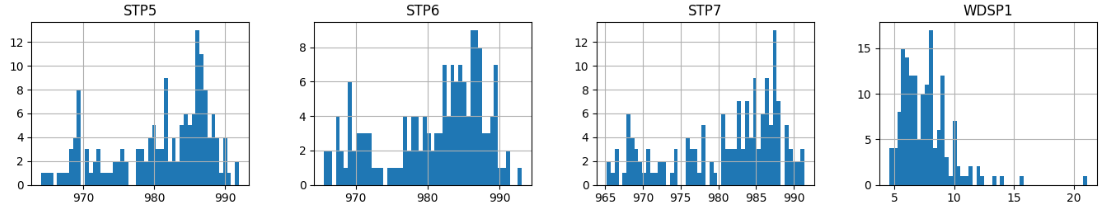
  

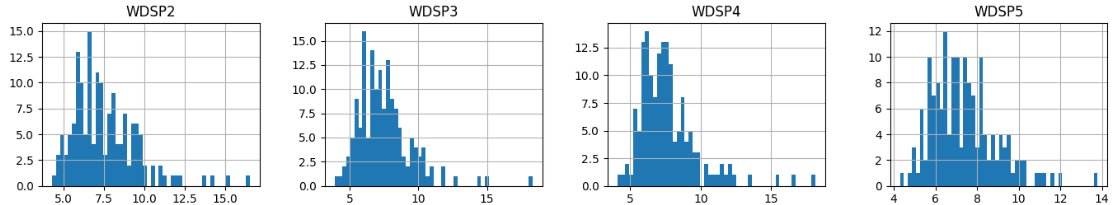


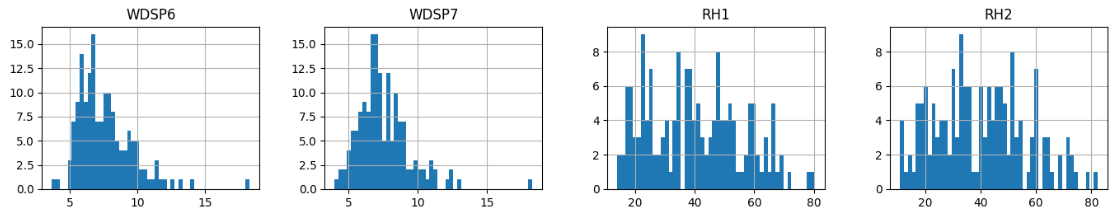


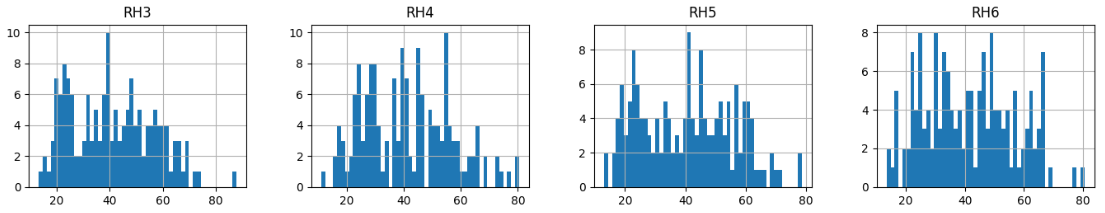


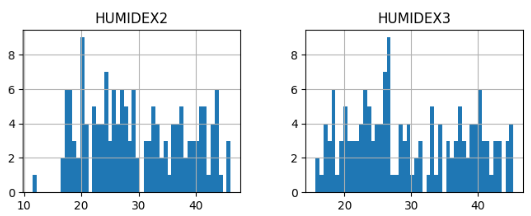
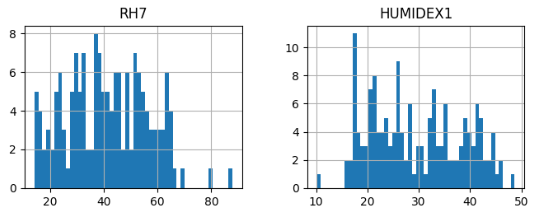
 

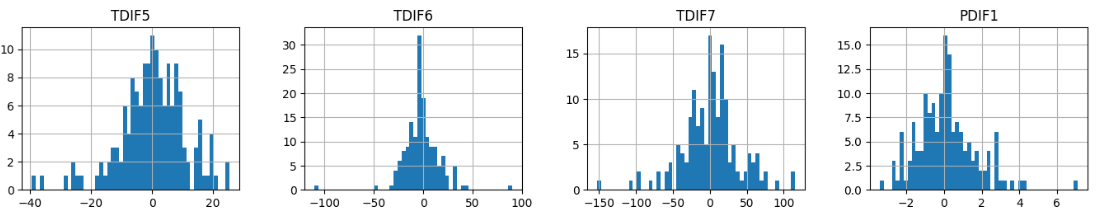
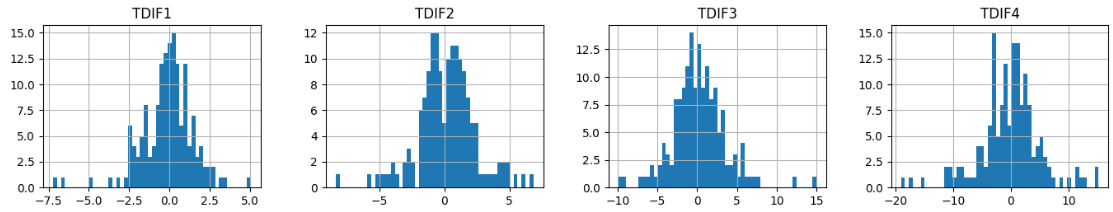
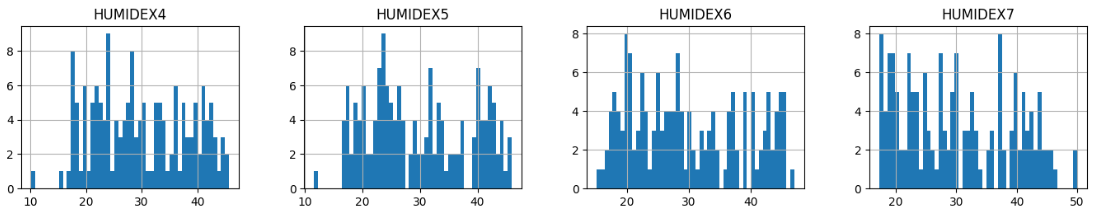


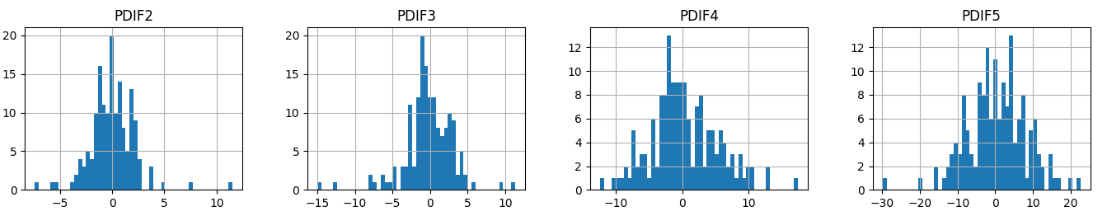


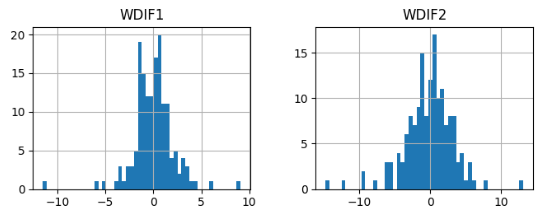
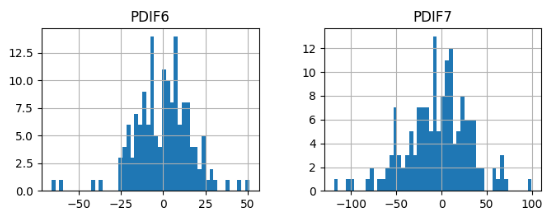


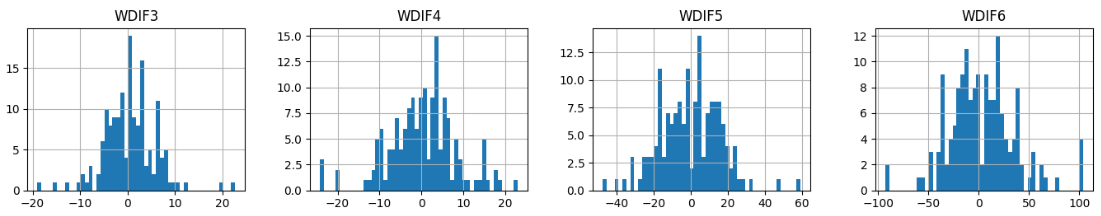


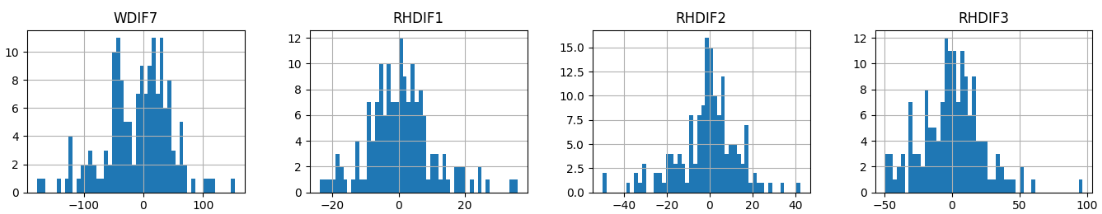


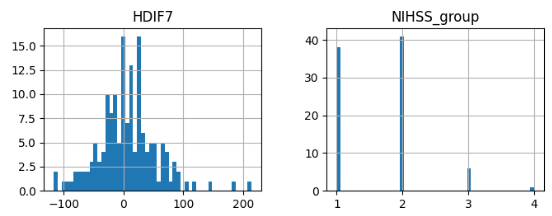
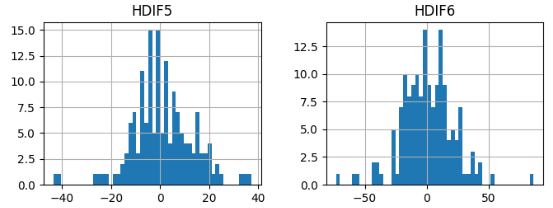
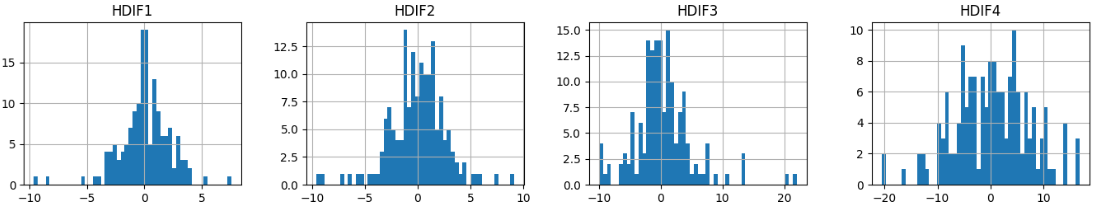
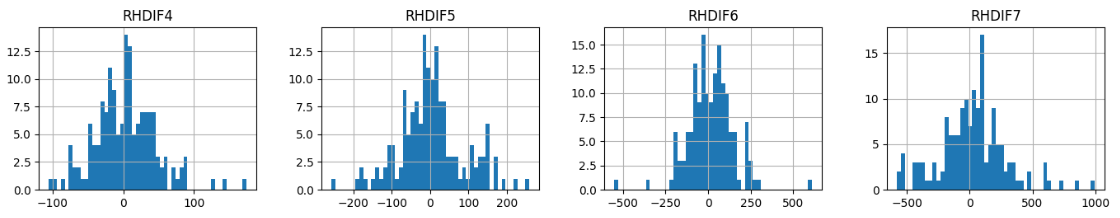












**12 .Dataset**

**Title:**

Data for: Prognostic Model of In-hospital Ischemic Stroke Mortality Based on an Electronic Health Record Cohort in Indonesia

**Link:**

<https://data.mendeley.com/datasets/rvhbhyht2s/1>

**Published by:**

<https://plu.mx/plum/a?mendeley_data_id=y86srgks26&theme=plum-bigben-theme>

Contributors: Nizar Yamanie, Yuli Felistia, Nugroho Harry Susanto, Aly Lamuri, Muhammad Miftahussurur, Anwar Santoso.

**Description:**

Background: Stroke patients rarely have satisfactory survival, which worsens further if comorbidities develop in such patients. Limited data availability from South-east Asia countries, especially Indonesia, has impeded the disentanglement of post-stroke mortality determinants. This study aimed to investigate predictors of in-hospital mortality in patients with ischemic stroke (IS).

Methods: This retrospective observational study used IS medical records from the National Brain Centre Hospital, Jakarta, Indonesia. A **theoretically driven logistic regression model was established by controlling for age and sex to calculate the odds ratio of each plausible risk factor for predicting post-stroke mortality**.

Findings: This study included **3,479 patients** with IS, 999 (28.72%) of whom had cardiovascular disease, 421 (12.1%) had renal disease, and 511 (14.69%) were verbally incoherent. Bivariate exploratory analysis revealed **lower blood levels of triglycerides**, **low density lipoprotein**, and **total cholesterol in patients with post-stroke mortality**. The average age of patients with **post-stroke mortality was 64 ± 12 years**, with a mean body mass index **(BMI) of 24 ± 3.5** kg/m2 and a median Glasgow Coma Scale (GCS) score of 12 ± 5. **Cardiovascular disease was more prevalent than renal disease** (28.72% vs. 12.1%), and **both contributed to a 4.5-times increase in the mortality risk**. Comorbidities, such as cardiovascular disease (odds ratio [OR]=2.66, 95% confidence interval [CI]: 1.82–3.91) and renal disease (OR=2.63, 95% CI: 1.77–3.89), caused higher odds of post-stroke mortality. However, **the factors contributing to lower odds of mortality were BMI** (OR=0.94, 95% CI: 0.89–0.99) and **GCS** (OR=0.67, 95% CI: 0.67–0.72).

Conclusion: After controlling for age and sex, our study reported that cardiovascular diseases, renal disease, BMI, and GCS on admission were strong predictors of in-hospital mortality in patients with IS.

**Data analysis:**

**The dataset contains 3561 instances with total of 81 features. The features are the following:**

1. sex\_ps: The sex or gender of the stroke patient ("laki-laki" translates to "male" and "perempuan" translates to "female.").
2. umur\_ps: The age of the stroke patient (in years).
3. tgl\_admisi: The date of admission for the stroke patient.
4. jam\_admisi: The time of admission for the stroke patient,(in hours I guess).
5. st\_nikah[[3]](#footnote-3): Marital status of the patient (**menikah**: This value indicates that the patient is currently married. **belum menikah**: This value indicates that the patient is not married, i.e., they are single**. duda/janda**: This value indicates that the patient is a widow (janda) or widower (duda), meaning they were previously married, but their spouse passed away.).
6. etnis[[4]](#footnote-4): The ethnic background of the patient.
7. pekerjaan: The occupation of the patient.

(pekerjaan has the following values:

IRT: Housewife

Pekerja swasta: Private sector employee

Pensiunan: Pensioner (Retiree)

Wiraswasta: Self-employed (Entrepreneur)

Tidak bekerja: Unemployed

ASN/PNS/POLRI: Civil servant / Government employee / Member of the Indonesian National Police

Lainnya: Other (Unspecified or unclassified category)

Mahasiswa/Pelajar: Student / School-going individual

9: Undefined or unspecified category (possibly indicating missing or unknown information)

60: Undefined or unspecified category with the value '60'

58: Undefined or unspecified category with the value '58')

1. pendidikan: The educational level of the patient.

Tidak sekolah: No Formal Education, Akademi: Academy, SD: Elementary School , SMP: Junior High School , SMA: High School, D3: Diploma Degree, S1: Bachelor's Degree , S2: Master's Degree , S3: Doctorate Degree

1. alamat: The address of the patient.
2. kelurahan: The neighborhood or local area where the patient resides.
3. kecamatan: The district or sub-district where the patient resides.
4. kota: The city or municipality where the patient resides.
5. diagnosa\_sek: Secondary diagnosis of the patient.
6. DIAGN0: Primary diagnosis of the patient.
7. onset: The time from the onset of symptoms to admission.
8. tindakan: Medical interventions or procedures performed for the patient.
9. dtn: Duration of the patient's illness
10. riw\_stroke\_tia: Stroke or transient ischemic attack (TIA) history of the patient.
11. thn\_riw\_stroke: The year of the patient's stroke history
12. jenis\_riw\_stroke: Type of stroke experienced by the patient.
13. riw\_ht: Hypertension (high blood pressure) history of the patient.
14. riw\_dm: Diabetes mellitus (diabetes) history of the patient.
15. obt\_rutin: Routine medications or treatments given to the patient.
16. riw\_jantung: Heart-related medical history of the patient.
17. riw\_ginjal: Kidney-related medical history of the patient.
18. merokok: Smoking status of the patient. (has values **tidak** **merokok**: Non-smoker. **Current** **smokers**: Current smokers, **Pernah** **merokok**: Former smokers)
19. alkohol: Alcohol consumption status of the patient.
20. stroke\_klg: Stroke classification or severity.
21. E, M, V: Specific features represented as integers or objects.
22. sistol: Systolic blood pressure of the patient
23. diastol: Diastolic blood pressure of the patient
24. GDS: Glasgow Coma Scale score of the patient
25. komplikasi\_rawat: Complications that occurred during treatment.
26. d\_dimer: D-dimer level in the patient's blood
27. trigliserida: Triglyceride level in the patient's blood
28. hdl: High-density lipoprotein (HDL) cholesterol level in the patient's blood
29. ldl: Low-density lipoprotein (LDL) cholesterol level in the patient's blood
30. kol\_total: Total cholesterol level in the patient's blood
31. as\_urat: Uric acid level in the patient's blood
32. GDP: Glucose level in the patient's blood
33. G2PP: Another feature related to glucose
34. HBA1C: Hemoglobin A1c level in the patient's blood
35. Hb: Hemoglobin level in the patient's blood
36. Ht: Hematocrit level in the patient's blood
37. Leukosit: White blood cell count in the patient's blood
38. Trombosit: Platelet count in the patient's blood
39. nihss\_msk: National Institutes of Health Stroke Scale (NIHSS) score on admission.
40. mrs\_keluar: Modified Rankin Scale (mRS) score at discharge.
41. imt: Body Mass Index (BMI) of the patient
42. ekg: Electrocardiogram (ECG) results.
43. lama\_rawat: Length of hospital stay for the patient, represented as an integer.
44. outcome: Outcome of the patient's treatment.
45. ct\_scan: CT scan results.
46. CT\_SC0: Another feature related to CT scan.
47. foto\_thorax: Chest X-ray results.
48. FOTO\_0: Another feature related to chest X-ray.
49. mri\_brain: MRI brain scan results.
50. MRI\_B0: Another feature related to MRI brain scan.
51. transformasi: A transformation feature (nature not specified).
52. stroke\_in\_evolution: Indicates whether the stroke is in evolution or not.
53. kelas\_rawat: Class of treatment.
54. pembayaran: Payment method for the treatment.
55. kelas\_bpjs: Class of treatment covered by BPJS (social security agency in Indonesia).
56. covid: Indicates whether the patient had COVID-19.
57. riw\_sakit\_lainnya: Other medical history of the patient.
58. RIW\_S0: Another feature related to other medical history.
59. keterangan: Additional information or comments.
60. death: Boolean value indicating whether the patient died during treatment.
61. DM: Diabetes mellitus status.
62. DM.uncontrolled: Indicates whether diabetes mellitus is uncontrolled.
63. heart.disease: Boolean value indicating whether the patient has heart disease.
64. HT: Hypertension status.
65. HT.uncontrolled: Indicates whether hypertension is uncontrolled.
66. renal.disease: Boolean value indicating whether the patient has renal (kidney) disease.
67. V.coherent: Boolean value related to a coherent feature.
68. V.num: An integer value related to the V feature.
69. GCS: Glasgow Coma Scale score (an alternative representation).
70. GCS.cat: Categorized Glasgow Coma Scale score.
71. GCS.cat2: Another categorization of Glasgow Coma Scale score.

The following features have null values:

sistol, diastol, transformasi, stroke\_in\_evolution have 1 null values;

riw\_stroke\_tia, riw\_dm, DM, DM.uncontrolled have 2 null values;

komplikasi\_rawat have 4 null values;

kelas\_rawat, pembayaran, kelas\_bpjs have 5 null values;

ekg has 6 null values;

onset has 14 null values;

imt has 82 null values;

Hb, Trombosit have 98 null values;

Leukosit has 101 null values;

Ht has 104 null values;

ldl has 131 null values;

trigliserida has 132 null values;

kol\_total has 135 null values;

hdl has 137 null values;

GDS has 140 null values;

GDP has 167 null values;

as\_urat has 219 null values;

G2PP has 257 null values;

pekerjaan has 289 null values;

pendidikan has 941 null values;

HBA1C has 1130 null values;

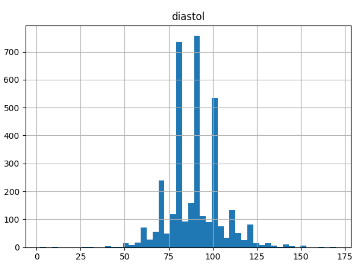
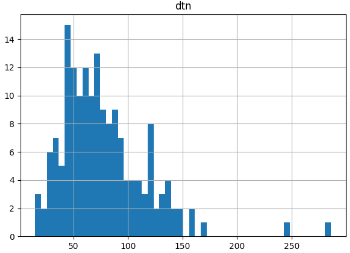
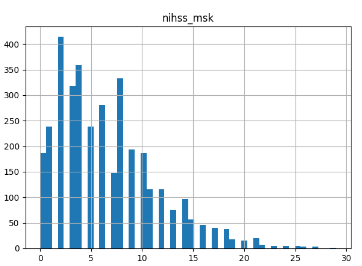
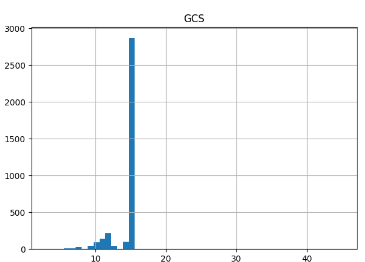
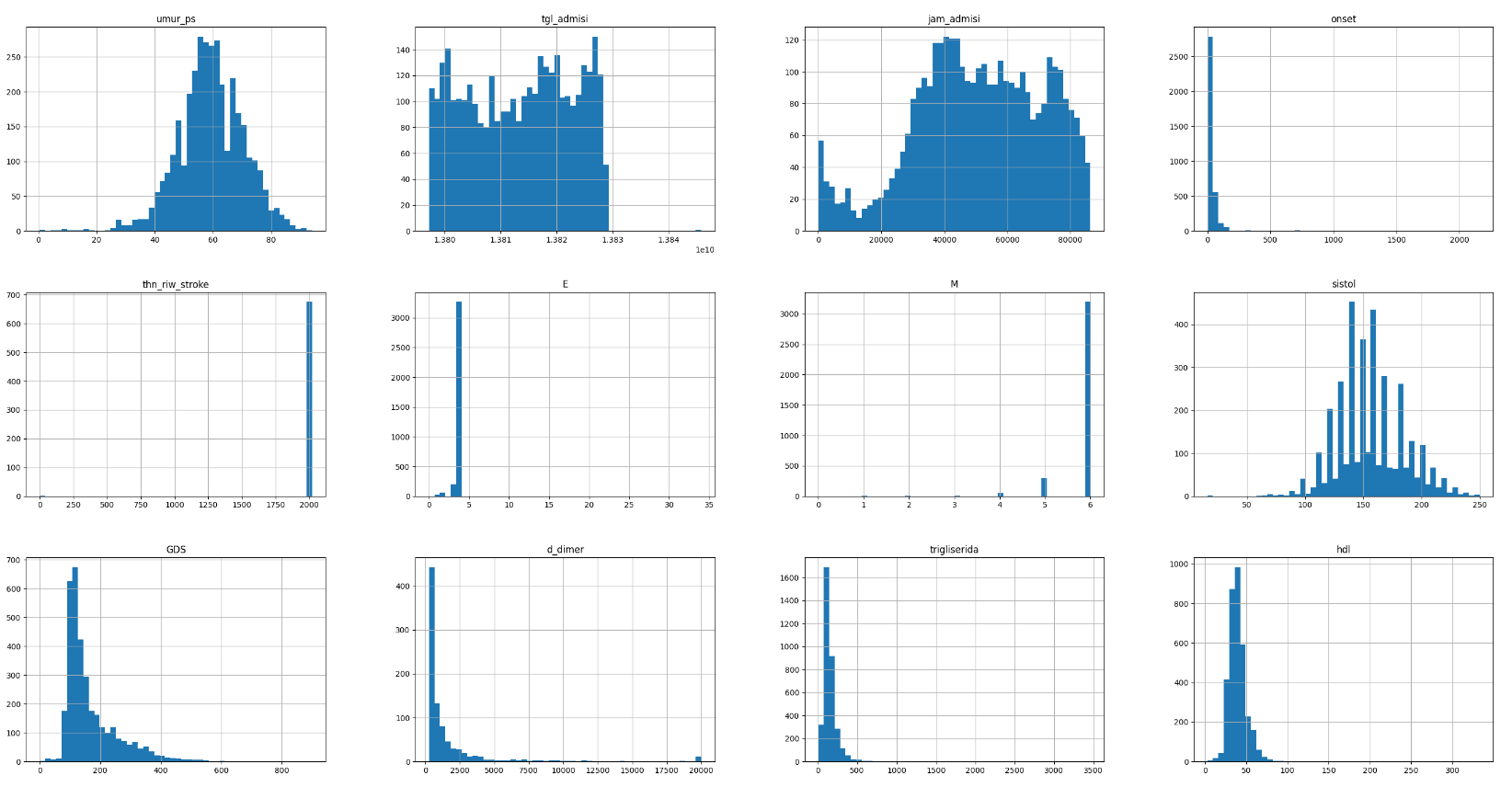
etnis has 2304 null values;

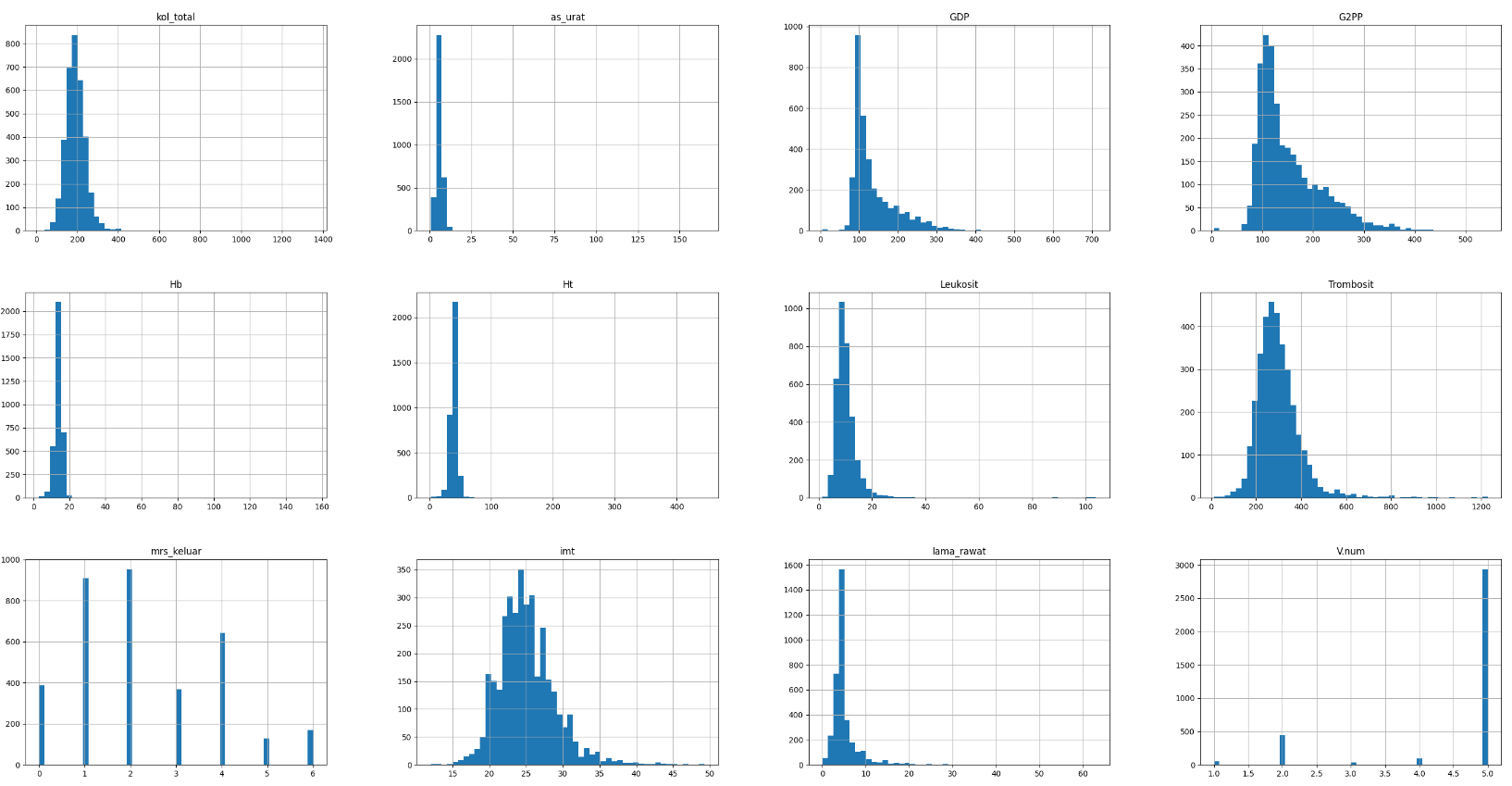
d\_dimer has 2688 null values;

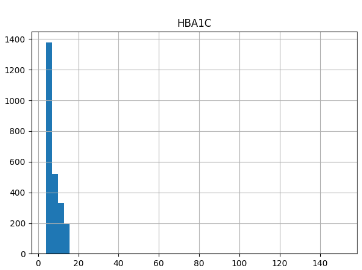
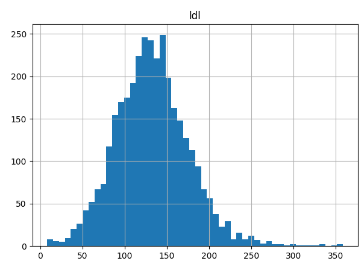
jenis\_riw\_stroke has 2697 null values;

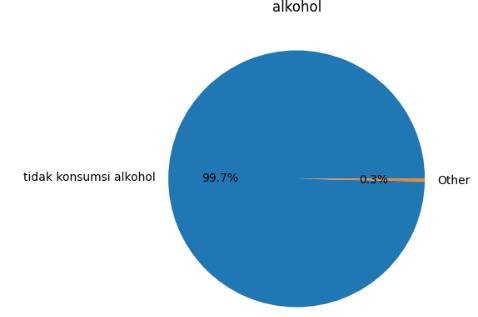
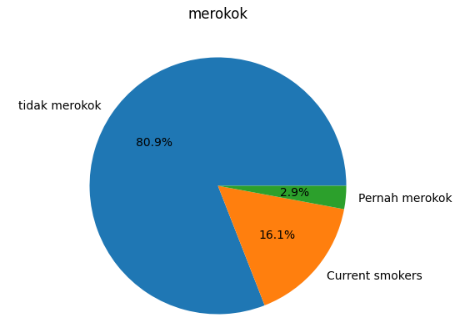
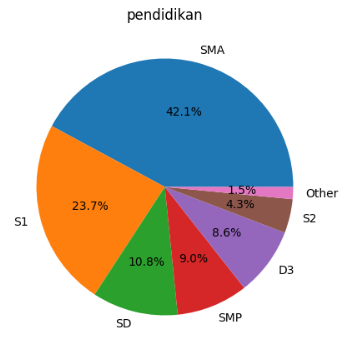
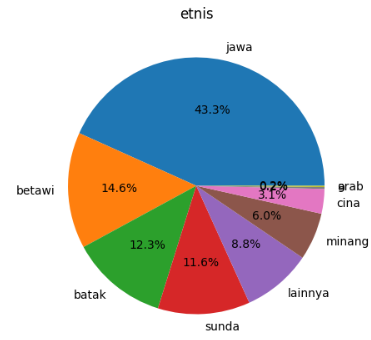
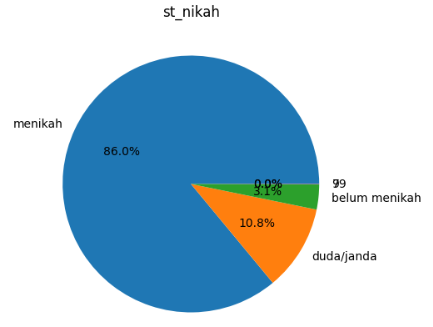
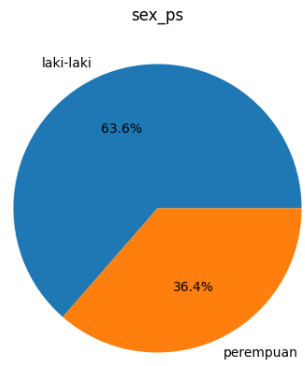
thn\_riw\_stroke has 2885 null values;

dtn has 3392 null values.









1. Computed Tomography. CT scans use a combination of X-rays and computer technology to create detailed cross-sectional images of the brain. [↑](#footnote-ref-1)
2. MRI (Magnetic Resonance Imaging) is a non-invasive medical imaging technique that uses strong magnetic fields and radio waves to generate detailed, high-resolution images of the internal structures of the body. [↑](#footnote-ref-2)
3. There are also values 7 and 99 which probebly mean unknown. [↑](#footnote-ref-3)
4. This feature has value **lainnya** which means "others" in Indonesian. This category might include patients from various smaller ethnic groups or those whose specific ethnicity is not listed separately. There is also value **9** which indicates missing or unknown information. [↑](#footnote-ref-4)