Package: NHSRdatasets (via r-universe)

October 18, 2024

Type Package

Title NHS and Healthcare-Related Data for Education and Training

Date 2022-11-08

Version 0.3.3

Maintainer Zoë Turner <zoe.turner3@nhs.net>

Description Free United Kingdom National Health Service (NHS) and other healthcare, or population health-related data for education and training purposes. This package contains synthetic data based on real healthcare datasets, or cuts of open-licenced official data. This package exists to support skills development in the NHS-R community: <https://nhsrcommunity.com/>.

License CC0

Language en-GB

Encoding UTF-8

LazyData true

RoxygenNote 7.3.2

Depends R (>= 3.5.0)

BugReports https://github.com/nhs-r-community/NHSRdatasets/issues

Suggests caret, dplyr, e1071, forcats, ggplot2, ggrepel, httr2, knitr, lattice, lme4, lmtest, lubridate, magrittr, MASS, ModelMetrics, rcmdcheck, readr, rmarkdown, rsample, scales, synthpop, tibble, tidyr, varhandle

VignetteBuilder knitr

URL https://github.com/nhs-r-community/NHSRdatasets, https://nhs-r-community.github.io/NHSRdatasets/

Repository https://nhs-r-community.r-universe.dev

RemoteUrl https://github.com/nhs-r-community/NHSRdatasets

RemoteRef main

RemoteSha 158d1db45b04e26b8ed19d2477b741377a6f173d

Contents

`ae_attendances` *NHS England Accident & Emergency Attendances and Admissions*

Description

Reported attendances, 4 hour breaches and admissions for all A&E departments in England for the years 2016/17 through 2018/19 (Apr-Mar). The data has been tidied to be easily usable within the tidyverse of packages.

Usage

```
data(ae_attendances)
```

Format

Tibble with six columns

period The month that this data relates to

org_code The ODS code for this provider

type The department type. either 1, 2 or other

attendances the number of patients who attended this department in this month

breaches the number of patients who breaches the 4 hour target in this month

admissions the number of patients admitted from A&E to the hospital in this month

Details

Data sourced from NHS England Statistical Work Areas which is available under the Open Government Licence v3.0

Source

NHS England Statistical Work Areas

Examples

```
data(ae_attendances)

library(dplyr)
library(ggplot2)
library(scales)

# Create a plot of the performance for England over time
ae_attendances %>%
  group_by(period) %>%
  summarise_at(vars(attendances, breaches), sum) %>%
  mutate(performance = 1 - breaches / attendances) %>%
  ggplot(aes(period, performance)) +
  geom_hline(yintercept = 0.95, linetype = "dashed") +
  geom_line() +
  geom_point() +
  scale_y_continuous(labels = percent) +
  labs(title = "4 Hour performance over time")

# Now produce a plot showing the performance of each trust
ae_attendances %>%
  group_by(org_code) %>%
  # select organisations that have a type 1 department
  filter(any(type == "1")) %>%
  summarise_at(vars(attendances, breaches), sum) %>%
  arrange(desc(attendances)) %>%
  mutate(
    performance = 1 - breaches / attendances,
    overall_performance = 1 - sum(breaches) / sum(attendances),
    rank = rank(-performance, ties.method = "first") / n()
  ) %>%
  ggplot(aes(rank, performance)) +
  geom_vline(xintercept = c(0.25, 0.5, 0.75), linetype = "dotted") +
  geom_hline(yintercept = 0.95, colour = "red") +
  geom_hline(aes(yintercept = overall_performance), linetype = "dotted") +
  geom_point() +
  scale_y_continuous(labels = percent) +
  theme_minimal() +
  theme(
    panel.grid = element_blank(),
    axis.text.x = element_blank()
  ) +
  labs(
    title = "4 Hour performance by trust",
    subtitle = "Apr-16 through Mar-19",
    x = "", y = ""
  )
```

`apha_cpd_survey` *AphA (Association of Professional Healthcare Analysts) CPD Survey Responses*

Description

Full raw data from the AphA CPD Survey

Usage

```
apha_cpd_survey
```

Format

This tidied raw data is available here as a tibble with 38 columns (blank or superfluous columns from the raw data were removed) and 237 rows (1 per respondent ID).

Variables have been named using a "controlled language" approach informed by Emily Riederer's "Column Names as Contracts" https://emilyriederer.netlify.app/post/column-name-contracts/.

***_id** Columns ending in `"_id"` are numeric and represent a unique ID for that response.

***_dttm** Columns ending in `"_dttm"` are in datetime format.

***_cat** Columns ending in `"_cat"` contain categorical data, though in some cases this is mixed with free text responses and may require tidying if you need it to be strictly categorical/factor data.

***_n** Columns ending in `"_n"` are theoretically counts, but in this tibble they may be mixed with non-numeric values and so the columns are in character format.

***_ind** Columns ending in `"_ind"` are theoretically indicator values with 2 main value options (Yes/No). These are in character format, but should be convertible to 1/0 or TRUE/FALSE values, if desired, with minimal wrangling.

***_txt** Columns ending in `"_txt"` contain free text responses and are in character format.

Multi-part questions have column name stubs with sequential letters. For example, `"q20a_"`, `"q20b_"` and so on. For formatting consistency, questions with a single part still have a column name stub with the letter a, for example `"q01a_"`.

Original survey questions (lightly edited) are provided as variable labels using the `{labelled}` package https://larmarange.github.io/labelled/. These labels provide more descriptive context for the "clean" column names. Variable labels can be viewed using `labelled::get_variable_labels(apha_cpd_survey)`.

Survey press release web page: https://www.aphanalysts.org/ltnws/nhs-at-risk-of-losing-a-generation-of-da

Source

https://www.aphanalysts.org/documents/cpd-survey-results-raw-data/

The survey of NHS and other healthcare data analysts was conducted in July 2022. The results data is made available in this package with the permission of AphA.

covid19 *International COVID-19 reported infection and death data*

Description

Reported COVID-19 infections, and deaths, collected and collated by the European Centre for Disease Prevention and Control (ECDC, provided by day and country). Data were collated and published up to 14th December 2020, and have been tidied so they are easily usable within the 'tidyverse' of packages.

Usage

```
data(covid19)
```

Format

Tibble with seven columns

date_reported The date cases were reported

contient A 'factor' for the geographical continent in which the reporting country is located.

countries_and_territories A 'factor' for the country or territory reporting the data.

countries_territory_code A 'factor' for the a three-letter country or territory code.

population_2019 The reported population of the country for 2019, taken from Eurostat for Europe and the World Bank for the rest of the world.

cases The reported number of positive cases.

deaths The reported number of deaths.

Details

Data sourced from European Centre for Disease Prevention and Control which is available under the open licence, compatible with the CC BY 4.0 license, further details available at ECDC.

Source

European Centre for Disease Prevention and Control

Examples

```
data(covid19)

library(dplyr)
library(ggplot2)
library(scales)

# Create a plot of the performance for England over time
covid19 |>
  filter(countries_and_territories ==
```

```
  c("United_Kingdom", "Italy", "France", "Germany", "Spain")) |>
ggplot(aes(
  x = date_reported,
  y = cases,
  col = countries_and_territories
)) +
geom_line() +
scale_color_discrete("Country") +
scale_y_continuous(labels = comma) +
labs(
  y = "Cases",
  x = "Date",
  title = "Covid-19 cases for selected countries",
  alt = "A plot of covid-19 cases in France, Germany, Italy, Spain & the UK"
) +
theme_minimal()
```

`LOS_model` *Hospital Length of Stay (LOS) Data*

Description

Artificially generated hospital data. Fictional patients at 10 fictional hospitals, with LOS, Age and Date status data Data were generate to learn Generalized Linear Models (GLM) concepts, modelling either Death or LOS.

Usage

```
data(LOS_model)
```

Format

Data frame with five columns

ID A fictional patient ID number

Organisation A factor representing one of ten fictional hospital trusts, for example Trust1

Age Age in years of each fictional patient

LOS In-hospital length of stay in days. The difference between admission and discharge date in dates

Death Binary for death status: 0 = survived, 1= died in hospital

Source

Generated by Chris Mainey, Feb-2019

Examples

```
data(LOS_model)

model1 <- glm(Death ~ Age + LOS, data = LOS_model, family = "binomial")
summary(model1)

# Now with an Age, LOS, and Age*LOS interaction.
model2 <- glm(Death ~ Age * LOS, data = LOS_model, family = "binomial")
summary(model2)
```

---

ons_mortality *Deaths registered weekly in England and Wales, provisional*

Description

Provisional counts of the number of deaths registered in England and Wales, by age, sex and region, from week commencing 8th January 2010 to 3rd April 202.

Usage

```
data(ons_mortality)
```

Format

Data frame with five columns

category_1 character, containing the names of the groups for counts, for example "Total deaths", "all ages".

category_2 character, subcategory of names of groups where necessary, for example details of region: "East", details of age bands "15-44".

counts numeric, numbers of deaths in whole numbers and average numbers with decimal points. To retain the integrity of the format this column data is left as character.

date date, format is yyyy-mm-dd; all dates are a Friday.

week_no integer, each week in a year is numbered sequentially.

Details

Source and licence acknowledgement

This data has been made available through Office of National Statistics under the Open Government Licence http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/

Source

Collected by Zoë Turner, Apr-2020 from https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/datasets/weeklyprovisionalfiguresondeathsregisteredinenglandandwale

Examples

```
data(ons_mortality)

library(dplyr)
library(tidyr)

# create a dataset that is "wide" with each date as a column
ons_mortality |>
  select(-week_no) |>
  pivot_wider(
    names_from = date,
    values_from = counts
  )
```

`ons_uk_population_2023`

ONS Mid-2023 Population Estimate for UK

Description

ONS Population Estimates for Mid-year 2023 National and subnational mid-year population estimates for the UK and its constituent countries by administrative area, age and sex (including components of population change, median age and population density).

Usage

```
data(ons_uk_population_2023)
```

Format

Tibble with six columns

sex male or female

Code country/geography code

Name country of the UK

Geography Country

age year of age

count the number of people in this group

Details

ONS Estimates of the population for the UK, England, Wales, Scotland, and Northern Ireland

Source

https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland

Examples

```
data(ons_uk_population_2023)


library(dplyr)
library(tidyr)

# create a dataset that has total population by age groups for England
ons_uk_population_2023 |>
  filter(Name == "ENGLAND") |>
  mutate(age_group = case_when(
    as.numeric(age) <= 17 ~ "0-17",
    as.numeric(age) >= 18 & as.numeric(age) <= 64 ~ "18-64",
    as.numeric(age) >= 65 ~ "65+",
    age == "90+" ~ "65+"
  )) |>
  group_by(age_group) |>
  summarise(count = sum(count))
```

`stranded_data` *Stranded Patient (Patients flagged as having a greater than 7 day LOS) Model*

Description

This model is to be used as a machine learning classification model, for supervised learning. The binary outcome is stranded vs not stranded patients.

Usage

```
data(stranded_data)
```

Format

Tibble with nine columns (1 x outcome and 8 predictors)

stranded.label Outcome variable - whether the patient is stranded or not

age Patient age on admission

care.home.referral Whether than have been referred from a care home

medicallysafe Medically safe for discharge - means the patient is assessed as safe, but has not been discharged yet

hcop Indicates whether they have been triaged from a Health Care for Older People specialty

mental_health_care Flag to indicate whether they need mental health support and care

periods_of_previous_care Count of the number of previous spells of care

admit_date Date they were admitted to hospital

frailty_index An initial index assessment to say if the patient is frail or not. This is needed for alignment of service provision.

Source

Synthetically generated by Gary Hutson, Mar-2021.

Examples

```
library(dplyr)

data(stranded_data)

stranded_data |>
  glimpse()
```

`synthetic_news_data` *Synthetic National Early Warning Scores Data*

Description

Synthetic NEWS data to show as the results of the NHSR_synpop package. These datasets have been synthetically generated by this package to be utilised in the NHSRDatasets package.

Usage

```
data(synthetic_news_data)
```

Format

Tibble with twelve columns

male character string containing gender code

age age of patient

NEWS National Early Warning Score (NEWS)

syst Systolic BP - Systolic BP result

dias Diastolic Blood Pressure - result on NEWS scale

temp Temperature of patient

pulse Pulse of the patient

resp Level of response from the patient

sat SATS(Oxygen Saturation Levels) of the patient

sup Suppressed Oxygen score

alert Level of alertness of patient

died Indicator to monitor patient death

Source

Generated by Dr. Muhammed Faisal and created by Gary Hutson, Mar-2021

Examples

```
library(dplyr)

data("synthetic_news_data")

synthetic_news_data |>
  glimpse()
```

Index