The population of US cities and cities from fictional sources

The data is called cities.

Description

The data is built to have the count in the number column with the first and last digit separated

The source of this data is < https://github.com/midnightradio/cse140-data-programming and https://simplemaps.com/data/us-cities >

Data format

A data frame with columns:

variable	class	description
country	character	Either US or fiction
city	character	The city within the country
location	character	The region within which the city is located
number	numeric	The population of that city
first	character	The first digit of number
last	character	The last digit of number

The population of US cities

The data is called cities_us.

Description

The data is built to have the count in the number column with the first and last digit separated

The source of this data is < https://simplemaps.com/data/us-cities >

Data format

A data frame with columns:

variable	class	description
city	character	The city within the country
location	character	The region within which the city is located
number	numeric	The population of that city
first	character	The first digit of number
last	character	The last digit of number

The population of cities from fictional sources

The data is called cities_fiction.

Description

The data is built to have the count in the number column with the first and last digit separated

The source of this data is < https://github.com/midnightradio/cse140-data-programming >

Data format

A data frame with columns:

variable	class	description
city	character	The city within the country
location	character	The region within which the city is located
number	numeric	The population of that city
first	character	The first digit of number
last	character	The last digit of number

The count of citizens on waitlists for medical procedures

The data is called waitlist.

Description

The data is built to have the count in the number column with the first and last digit separated

The source of this data is < https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5942457/ >

Data format

A data frame with columns:

variable	class	description
country	character	The source country of the data
type	character	The type of medical procedure
details	character	Further details about the medical procedure
month	character	The month of the year
year	numeric	Year
number	numeric	The number of people on the waitlist
first	character	The first digit of number
last	character	The last digit of number

The count of Finish citizens on waitlists for medical procedures

The data is called waitlist_finland.

Description

The data is built to have the count in the number column with the first and last digit separated

The source of this data is < https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5942457/ >

Data format

A data frame with columns:

variable	class	description
country	character	The source country of the data
type	character	The type of medical procedure
details	character	Further details about the medical procedure
month	character	The month of the year
year	numeric	Year
number	numeric	The number of people on the waitlist
first	character	The first digit of number
last	character	The last digit of number

The count of Spanish citizens on waitlists for medical procedures

The data is called waitlist_spain.

Description

The data is built to have the count in the number column with the first and last digit separated

The source of this data is < https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5942457/ >

Data format

A data frame with columns:

variable	class	description
country	character	The source country of the data
type	character	The type of medical procedure
details	character	Further details about the medical procedure
month	character	The month of the year
year	numeric	Year
number	numeric	The number of people on the waitlist
first	character	The first digit of number
last	character	The last digit of number

The election results for Iran and US presidential elections

The data is called election.

Description

The data is built to have the count in the number column with the first and last digit separated

The source of this data is < https://github.com/midnightradio/cse140-data-programming >

Data format

A data frame with columns:

variable	class	description
country	character	The source country of the data
region	character	The region within which the election votes were tallied
candidate	character	The name of the electoral candidate
number	numeric	The number of votes cast for the candidate
first	character	The first digit of number
last	character	The last digit of number

The election results for the 2009 presidential elections in Iran

The data is called election_iran.

Description

The data is built to have the count in the number column with the first and last digit separated

The source of this data is < https://github.com/midnightradio/cse140-data-programming >

Data format

A data frame with columns:

variable	class	description
region	character	The region within which the election votes were tallied
candidate	character	The name of the electoral candidate
number	numeric	The number of votes cast for the candidate
first	character	The first digit of number
last	character	The last digit of number

The election results for the Obama McCain presidential elections in the US

The data is called election_us.

Description

The data is built to have the count in the number column with the first and last digit separated

The source of this data is < https://github.com/midnightradio/cse140-data-programming >

Data format

A data frame with columns:

variable	class	description
region	character	The region within which the election votes were tallied
candidate	character	The name of the electoral candidate
number	numeric	The number of votes cast for the candidate
first	character	The first digit of number
last	character	The last digit of number

The counts and percentage of first digits for all data objects

The data is called benford.

Description

This data has to counts by first digit for the election, waitlist, and cities data

The source of this data is < https://github.com/midnightradio/cse140-data-programming, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5942457/, and https://simplemaps.com/data/us-cities >

Data format

A data frame with columns:

variable	class	description
data	character	The data object used to calculate digit counts
country	character	The location or group within each data object
first	character	The first digit number
n	integer	The count of numbers that started with that digit
percent	numeric	The percent of the total for each data and country group
benford_percent	numeric	The expected propoprtion under Benford's law

The counts and percentage of last digits for college students asked to pick random numbers

The data is called pick_random.

Description

This data has to counts by last digit for the random guesses

The source of this data is < https://docs.google.com/spreadsheets/d/1TasFdyWr9xN7uWiWw0PkaFDwHYgQiC3y41YKR9C edit#gid=0 and https://www.reddit.com/r/dataisbeautiful/comments/acow6y/asking_over_8500_students_to_pick_a_random_number/ >

Data format

A data frame with columns:

variable	class	description
digit	character	The number of interest between 0-9
n_09	integer	The count of people that picked that digit. Note 10s were changed to 0
percent_09	numeric	The percentage of each digit of the total for the 0-9 digit counts
n_last	integer	The count of the last digit of numbers picked between 0 and 1 million.
percent_last	numeric	The percentage of each digt of the total for the last digit counts.

The counts and percentage of last digits for all data objects

The data is called last_digit.

Description

This data has to counts by last digit for the election, waitlist, and cities data

The source of this data is < https://github.com/midnightradio/cse140-data-programming, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5942457/, and https://simplemaps.com/data/us-cities >

Data format

A data frame with columns:

variable	class	description
data	character	The data object used to calculate digit counts
country	character	The location or group within each data object
last	character	The last digit number
n	integer	The count of numbers that ended with that digit
percent	numeric	The percent of the total for each data and country group
last_percent	numeric	The expected propoprtion under complete randomness

The combined accounting data sets

The data is called accounting.

Description

The data is built to have the count in the number column with the first and last digit separated

The source of this data is < https://github.com/carloscinelli/benford.analysis and https://www.amazon.com/Benfords-Law-Applications-Accounting-Detection/dp/1118152859 >

Data format

A data frame with columns:

variable	class	description
data	character	The data object used to calculate digit counts
number	numeric	The number of votes cast for the candidate

variable	class	description
first	character	The first digit of number
last	character	The last digit of number

The amounts paid to vendors for the 90 days preceding General Motor's 2009 liquidation.

The data is called accounting_gm.

Description

The data is built to have the count in the number column with the first and last digit separated

The source of this data is < https://github.com/carloscinelli/benford.analysis and https://www.amazon.com/Benfords-Law-Applications-Accounting-Detection/dp/1118152859 >

Data format

A data frame with columns:

variable	class	description
number	numeric	The number of votes cast for the candidate
first	character	The first digit of number
last	character	The last digit of number

A dataset containing the card transactions for a government entity - 2010.

The data is called accounting_government.

Description

The data is built to have the count in the number column with the first and last digit separated

The source of this data is < https://github.com/carloscinelli/benford.analysis and https://www.amazon.com/Benfords-Law-Applications-Accounting-Detection/dp/1118152859 >

Data format

A data frame with columns:

variable	class	description
number	numeric	The number of votes cast for the candidate
first	character	The first digit of number
last	character	The last digit of number

Financial Statements numbers of Sino Forest Corporation's 2010 Report.

The data is called accounting_sino.

Description

The data is built to have the count in the number column with the first and last digit separated

The source of this data is < https://github.com/carloscinelli/benford.analysis and https://www.amazon.com/Benfords-Law-Applications-Accounting-Detection/dp/1118152859 >

Data format

A data frame with columns:

variable	class	description
number	numeric	The number of votes cast for the candidate
first	character	The first digit of number
last	character	The last digit of number

A dataset of the 2010's payments data of a division of a West Coast utility company.

The data is called accounting_utility.

Description

The data is built to have the count in the number column with the first and last digit separated

The source of this data is < https://github.com/carloscinelli/benford.analysis and https://www.amazon.com/Benfords-Law-Applications-Accounting-Detection/dp/1118152859 >

Data format

A data frame with columns:

variable	class	description
number	numeric	The number of votes cast for the candidate
first	character	The first digit of number
last	character	The last digit of number

The counts and percentage of first digits for all data objects

The data is called benford_accounting.

Description

This data has to counts by first digit for the accounting data

The source of this data is < https://github.com/carloscinelli/benford.analysis and https://www.amazon.com/Benfords-Law-Applications-Accounting-Detection/dp/1118152859 >

Data format

A data frame with columns:

variable	class	description
data	character	The data object used to calculate digit counts
first	character	The first digit number
n	integer	The count of numbers that started with that digit
percent	numeric	The percent of the total for each data and country group
benford_percent	numeric	The expected propoprtion under Benford's law

The counts and percentage of last digits for all data objects

The data is called last_digit_accounting.

Description

This data has to counts by last digit for the accounting data

The source of this data is < https://github.com/carloscinelli/benford.analysis and https://www.amazon.com/Benfords-Law-Applications-Accounting-Detection/dp/1118152859 >

Data format

A data frame with columns:

variable	class	description
data	character	The data object used to calculate digit counts
last	character	The last digit number
n	integer	The count of numbers that ended with that digit
percent	numeric	The percent of the total for each data and country group
last_percent	numeric	The expected propoprtion under complete randomness

A full dataset of the 2010's payments data of a division of a West Coast utility company.

The data is called utility_data.

Description

This data adds a few more variables beyond accounting_utility

The source of this data is < https://github.com/carloscinelli/benford.analysis and https://www.amazon.com/Benfords-Law-Applications-Accounting-Detection/dp/1118152859 >

Data format

A data frame with columns:

variable	class	description
vendornum	character	Vendor Number
date	Date	Date of the invioce
invnum	character	The invoice number
amount	numeric	The amount on the invoice

A full dataset containing the card transactions for a government entity - 2010.

The data is called government_data.

Description

This data adds a few more variables beyond accounting_government

The source of this data is < https://github.com/carloscinelli/benford.analysis and https://www.amazon.com/Benfords-Law-Applications-Accounting-Detection/dp/1118152859 >

Data format

A data frame with columns:

variable	class	description
cardnum	character	Credit card number used for the purchase
date	Date	The date of the transaction
merchnum	character	The merchant number
merchdescription	character	the merchant name and details
merchstate	character	The state where the merchant is located
merchzip	character	The zipcode of the merchant
transtype	character	The transaction type. A, D, P, Y
amount	numeric	the amount ot the transaction
merch_clean	character	A cleaned merchant name
merch_other200	character	All merchants with less than 200 transactions grouped to other
merch_other100	character	All merchants with less than 100 transactions grouped to other
merch_other50	character	All merchants with less than 50 transactions grouped to other
merch_other10	character	All merchants with less than 10 transactions grouped to other