Introduction to Data

Dr. Mine Dogucu, UCI

Alma Castro, Cypress College

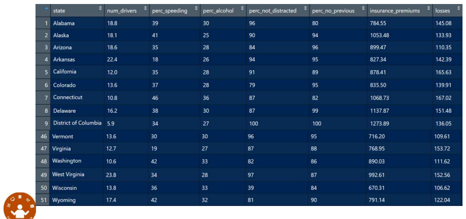
Terminology

- **Population**: The entire group to be studied.
- **Sample**: Subset of the population that is being studied.
- **Individual** or **case**: Person or object that is a member of the population being studied.
- **Parameter**: A numerical summary of a population.
- **Statistic**: A numerical summary of a sample.
- **Variable**: The characteristic of the individual.
- **Descriptive statistics**: Describing the data we have at hand using numerical summaries and graphs.
- **Inferential statistics**: Using results from sample data to make conclusions about the population and reporting the reliability of the result.

Data Frames

Dear Mona, Which State Has the Worst Drivers?

Data Frame `bad_driver`



	state	num_drivers	perc_speeding	perc_alcohol	perc_not_distracted	perc_no_previous	insurance_premiums	losses
1	Alabama	18.8	39	30	96	80	784.55	145.08
2	Alaska	18.1	41	25	90	94	1053.48	133.93
3	Arizona	18.6	35	28	84	96	899.47	110.35
4	Arkansas	22.4	18	26	94	95	827.34	142.39
5	California	12.0	35	28	91	89	878.41	165.63
6	Colorado	13.6	37	28	79	95	835.50	139.91
7	Connecticut	10.8	46	36	87	82	1068.73	167.02
8	Delaware	16.2	38	30	87	99	1137.87	151.48
9	District of Columbia	5.9	34	27	100	100	1273.89	136.05
46	Vermont	13.6	30	30	96	95	716.20	109.61
47	Virginia	12.7	19	27	87	88	768.95	153.72
48	Washington	10.6	42	33	82	86	890.03	111.62
49	West Virginia	23.8	34	28	97	87	992.61	152.56
50	Wisconsin	13.8	36	33	39	84	670.31	106.62
51	Wyoming	17.4	42	32	81	90	791.14	122.04

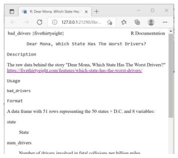
Data Frame `bad_driver`

- The data frame has 8 **variables** (`state`, `num_drivers`, `perc_speeding`, `perc_not_distracted`, `perc_no_previous`, `insurance_premiums`, `losses`).
- The data frame has 51 **cases**. Each case represents a US state (or District of Columbia).

Description of dataset and its variables

1. As an html

```
?bad_drivers
```



bad_drivers {fivethirtyeight} R Documentation

Dear Mona, Which State Has The Worst Drivers?

Description

The raw data behind the story "Dear Mona, Which State Has The Worst Drivers?" https://fivethirtyeight.com/features/which-state-has-the-worst-drivers/

Usage

`bad_drivers`

Format

A data frame with 51 rows representing the 50 states + D.C. and 8 variables:

state

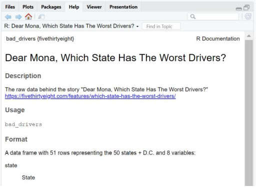
State

num_drivers

Number of drivers involved in fatal collisions per billion miles

Description of dataset and its variables

2. In the "Help" tab



R: Dear Mona, Which State Has The Worst Drivers?

bad_drivers {fivethirtyeight} R Documentation

Dear Mona, Which State Has The Worst Drivers?

Description

The raw data behind the story "Dear Mona, Which State Has The Worst Drivers?" https://fivethirtyeight.com/features/which-state-has-the-worst-drivers/

Usage

`bad_drivers`

Format

A data frame with 51 rows representing the 50 states + D.C. and 8 variables:

state

State

Functions for data frames

Getting to know the dataset

`head()`: Shows the first few rows of the data frame.

```
head(bad_drivers)
```

```
## # A tibble: 6 × 8
##   state      num_drivers perc_speeding perc_alcohol perc_not
##   <chr>            <dbl>         <int>        <int>
## 1 Alabama           18.8            39           30
## 2 Alaska            18.1            41           25
## 3 Arizona           18.6            35           28
## 4 Arkansas          22.4            18           26
## 5 California        12              35           28
## 6 Colorado          13.6            37           28
## # ℹ 3 more variables: perc_no_previous <int>, insurance_prem
## #   losses <dbl>
```

`tail()`: Shows the last few rows of the data frame.

```
tail(bad_drivers)
```

```
## # A tibble: 6 × 8
##   state         num_drivers perc_speeding perc_alcohol perc_
##   <chr>               <dbl>         <int>        <int>
## 1 Vermont              13.6            30           30
## 2 Virginia             12.7            19           27
## 3 Washington           10.6            42           33
## 4 West Virginia        23.8            34           28
## 5 Wisconsin            13.8            36           33
## 6 Wyoming              17.4            42           32
## # ℹ 3 more variables: perc_no_previous <int>, insurance_prem
## #   losses <dbl>
```

`glimpse()`: Displays the number of rows (observations or cases) and columns (variables) along with the list of variables and the first few data values of those variables.

```
glimpse(bad_drivers)
```

```
## Rows: 51
## Columns: 8
## $ state              <chr> "Alabama", "Alaska", "Arizona",
## $ num_drivers        <dbl> 18.8, 18.1, 18.6, 22.4, 12.0, 13
## $ perc_speeding      <int> 39, 41, 35, 18, 35, 37, 46, 38,
## $ perc_alcohol       <int> 30, 25, 28, 26, 28, 28, 36, 30,
## $ perc_not_distracted <int> 96, 90, 84, 94, 91, 79, 87, 87,
## $ perc_no_previous   <int> 80, 94, 96, 95, 89, 95, 82, 99,
## $ insurance_premiums  <dbl> 784.55, 1053.48, 899.47, 827.34,
## $ losses             <dbl> 145.08, 133.93, 110.35, 142.39,
```

`ncol()`: Shows the number of columns of the data frame.

```
ncol(bad_drivers)
```

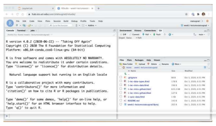
```
## [1] 8
```

`nrow()`: Shows the number of rows of the data frame.

```
nrow(bad_drivers)
```

```
## [1] 51
```

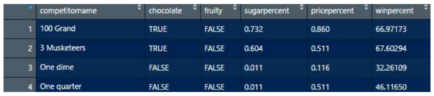
Getting to Know the Data Frame in Action



```
R version 4.0.2 (2020-06-22) -- "Taking Off Again"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-conda_cos6-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

Activity

Data Frame for You to Try Out `candy_rankings`

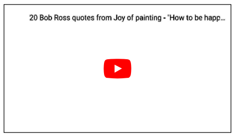


	competitorname	chocolate	fruity	sugarpercent	pricepercent	winpercent
1	100 Grand	TRUE	FALSE	0.732	0.860	66.97173
2	3 Musketeers	TRUE	FALSE	0.604	0.511	67.60294
3	One dime	FALSE	FALSE	0.011	0.116	32.26109
4	One quarter	FALSE	FALSE	0.011	0.511	46.11650

Answer the following questions about the `candy_rankings` data frame using functions for data frames when appropriate.

1. ⁎ Use the help feature to find more information about the variables in the data set. And display the first few rows of the data set.
2. ⁎ What does the variable "pluribus" describe?
3. ⁎ How many observations are there in this data set?
4. ⁎ How many variables are there in this data set? Name 3 of them.
5. ⁎ How many rows and columns does the data set have?

Bob Ross



20 Bob Ross quotes from Joy of painting - "How to be happ...

```
glimpse(bob_ross)
```

```
## Rows: 403
## Columns: 71
## $ episode            <chr> "S01E01", "S01E02", "S01E03", "S0
## $ season             <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
## $ episode_num        <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11
## $ title              <chr> "A WALK IN THE WOODS", "MT. MCKIN
## $ apple_frame        <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ aurora_borealis    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ barn               <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ beach              <int> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,
## $ boat               <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ bridge             <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ building           <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ bushes             <int> 1, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0,
## $ cabin              <int> 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0,
```

`candy_rankings` **VS** `bob_ross`

False - 0
True - 1