Data Types

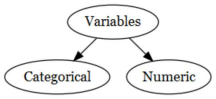
Dr. Mine Dogucu, UCI

Alma Castro, Cypress College

Variables

Variables: Characteristics of the individuals within the population.

Variables can be classified into two groups: categorical or numeric.



Categorical Variables

Categorical (qualitative) variables allows for individuals to be classified based on some attribute or characteristic.

- Variables such as `eye_color` (brown, blue, black, green, other), `own_car` (true, false), `class_year` (freshman, sophomore, junior, senior) are *categorical variables*.

Categorical variables have **levels** (categories). For instance `own_car` has two levels, and `class_year` has four levels.

Numerical Variables

Numeric (quantitative) variables provides numerical measures of individuals.

- Variables such as `n_kids` (number of kids), `height`, and `course_units` are *numerical variables*.

We can do certain analyses using these variables such as finding an average or the maximum or minimum.

Not everything represented by numbers represents a numeric quantity. e.g. student ID, cell phone number, zip code.

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

Categorical variables:

- `chocolate`, `fruity`

Numeric variables:

- `sugarpercent`, `pricepercent`, `winpercent`

Categorical Variables

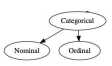
Categorical variables can be further classified as either nominal or ordinal.

- *Nominal* variables have categories that do not follow a natural order. Hence, these can be ordered by alphabetical order or based on highest frequency, or any way that makes sense of the data being presented.

  e.g. `eye_color`, `race`, `gender`

- *Ordinal* variables have categories that have a natural order.

  e.g. `course_grade`, `class_year`, `t-shirt_size`



Numerical Variables

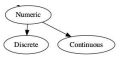
Numerical variables can be further classified as either continuous or discrete.

- *Continuous* variables are not countable and have an infinite number of possible values. Any decimal value make sense. Continuous variables can be measured to a certain precision, e.g. days, hours, minutes, seconds, ...

  e.g. `weight`, `commute_time`

- *Discrete* variables have a finite or countable number of possible values.

  e.g. `n_kids`, `class_units`




```
glimpse(candy_rankings)
```

```
## Rows: 85
## Columns: 13
## $ competitorname   <chr> "100 Grand", "3 Musketeers", "One d
## $ chocolate        <lgl> TRUE, TRUE, FALSE, FALSE, FALSE, TR
## $ fruity           <lgl> FALSE, FALSE, FALSE, FALSE, TRUE, F
## $ caramel          <lgl> TRUE, FALSE, FALSE, FALSE, FALSE, F
## $ peanutyalmondy   <lgl> FALSE, FALSE, FALSE, FALSE, FALSE,
## $ nougat           <lgl> FALSE, TRUE, FALSE, FALSE, FALSE, F
## $ crispedricewafer <lgl> TRUE, FALSE, FALSE, FALSE, FALSE, F
## $ hard             <lgl> FALSE, FALSE, FALSE, FALSE, FALSE,
## $ bar              <lgl> TRUE, TRUE, FALSE, FALSE, FALSE, TR
## $ pluribus         <lgl> FALSE, FALSE, FALSE, FALSE, FALSE,
## $ sugarpercent     <dbl> 0.732, 0.604, 0.011, 0.011, 0.906,
## $ pricepercent     <dbl> 0.860, 0.511, 0.116, 0.511, 0.511,
## $ winpercent       <dbl> 66.97173, 67.60294, 32.26109, 46.11
```



```
glimpse(mariokart)
```

```
## Rows: 143
## Columns: 12
## $ id           <dbl> 150377422259, 260483376854, 320432342985
## $ duration     <int> 3, 7, 3, 3, 1, 3, 1, 1, 3, 7, 1, 1, 1, 1
## $ n_bids       <int> 20, 13, 16, 18, 20, 19, 13, 15, 29, 8, 1
## $ cond         <fct> new, used, new, new, new, new, used, new
## $ start_pr     <dbl> 0.99, 0.99, 0.99, 0.99, 0.01, 0.99, 0.01
## $ ship_pr      <dbl> 4.00, 3.99, 3.50, 0.00, 0.00, 4.00, 0.00
## $ total_pr     <dbl> 51.55, 37.04, 45.50, 44.00, 71.00, 45.00
## $ ship_sp      <fct> standard, firstClass, firstClass, standa
## $ seller_rate <int> 1580, 365, 998, 7, 820, 270144, 7284, 48
## $ stock_photo <fct> yes, yes, no, yes, yes, yes, yes, yes, y
## $ wheels       <int> 1, 1, 1, 1, 2, 0, 0, 2, 1, 1, 2, 2, 2, 2
## $ title        <fct> "~~ Wii MARIO KART &amp; WHEEL ~ NINTEND
```


`character` (`chr`): takes string values (e.g. a person's name, address)
`integer` (`int`): integer (single precision)
`double` (`dbl`): floating decimal (double precision)
`numeric` (`num`): integer or double
`factor` (`fct`): categorical variables with different levels
`logical` (`lgl`): TRUE (1), FALSE (0)

As a data scientist it is **your** job to check the type(s) of data that you are working with. The type of variables you have will determine the type of analysis you will do.

Do **not** assume you will work with clean data frames, with clean names, labels, and types.

Data Wrangling

One of the first steps you'll do as a data scientist, is to determine if you have "clean" data, if you don't then it is a good idea to **tidy** and **transform** it to get the data in a form that is natural to work with.

- **Tidying** data means to put your data in a way that every column is a variable and every row is an observation.
- **Transformation** includes narrowing in on observations of interest, creating new variables that are functions of existing variables (like computing speed from distance and time), changing data types, and cleaning names.

Together, tidying and transforming data is called **data wrangling**.

Activity

```
glimpse(titanic_train)
```

```
## Rows: 891
## Columns: 12
## $ PassengerId <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 1
## $ Survived    <int> 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0
## $ Pclass      <int> 3, 1, 3, 1, 3, 3, 1, 3, 3, 2, 3, 1, 3, 3
## $ Name        <chr> "Braund, Mr. Owen Harris", "Cumings, Mrs
## $ Sex         <chr> "male", "female", "female", "female", "m
## $ Age         <dbl> 22, 38, 26, 35, 35, NA, 54, 2, 27, 14, 4
## $ SibSp       <int> 1, 1, 0, 1, 0, 0, 0, 3, 0, 1, 1, 0, 0, 1
## $ Parch       <int> 0, 0, 0, 0, 0, 0, 0, 1, 2, 0, 1, 0, 0, 5
## $ Ticket      <chr> "A/5 21171", "PC 17599", "STON/O2. 31012
## $ Fare        <dbl> 7.2500, 71.2833, 7.9250, 53.1000, 8.0500
```


Answer the following questions using the `titanic_train` data frame.

1. ※ How many observations are there in this data set?
2. ※ How many variables are there in this data set?
3. ※ Which variables are considered categorical?
4. ※ Which variables are considered numeric?
5. ※ Determine if the type of data for each variable was read correctly by R, if not then specify the most appropriate type of variable that it should be. Use the help feature to find more information about the variables if needed.