Data Analysis: A Practical Introduction for Absolute Beginners

Module 3, Lab 1: Data Structures

Learning Objectives

- Explore the anatomy of a data set.
- Identify which variables in a provided data set are nominal, ordinal, interval, and ratio.

Data Set

Mod3Lab1.csv

What You'll Need

To complete the lab, you will need the online version of Microsoft Excel.

Overview

In this lab, we're going to the movies. Using data on a number of different films, we'll get familiar with the basic structure and anatomy of a data set. We'll also explore and identify four basic types of data measurement: nominal, ordinal, interval, and ratio.

Exercise 1: Reading Data

For starters, we'll get our feet wet and look at the basic structure of a data set.

1. Open the data set in Excel. You should see info on 104 different movies. Here's a snapshot of what our tidy data set looks like, with a few different variables:

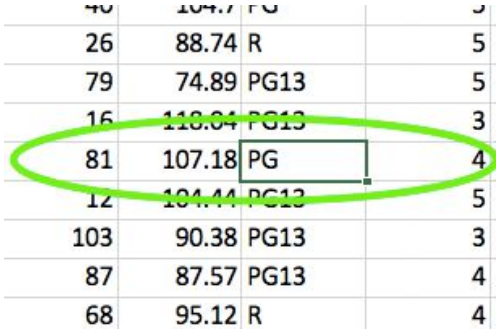
	A	B	C	D
1	movieid	runtime	rating	liking
2	6	112.74	PG13	3
3	76	96.68	PG	3
4	39	81.14	PG13	5
5	89	104.07	PG	4
6	93	101.38	G	4
7	78	102.75	R	3
8	31	92.05	G	3
9	47	114.65	R	3
10	41	98.86	R	4
11	104	85.41	PG	5
12	75	94.98	PG	3
13	60	97.85	PG13	3
14	32	109.98	PG13	3
15	77	110.27	G	5
16	71	104.81	PG13	4

2. Identify the variables. In this case, there are four different vertical **columns**: movie ID, runtime (in minutes), rating, and liking (which gives the audience's average ranking of the movie on a scale of 1–5, with 1 being the lowest). Those are the variables because those values vary from movie to movie.

3. Identify the observations. Since each horizontal **row** corresponds to a single entry from each of the variable columns, each row must represent an individual movie. For example, Movie 6 at the top has a runtime of 112.74 minutes, is rated PG-13, and has an average likability of 3 out of 5 from the audience.

	A	B	C	D
1	movieid	runtime	rating	liking
2	6	112.74	PG13	3
3	76	96.68	PG	3
4	39	81.14	PG13	5
5	89	104.07	PG	4
6	93	101.38	G	4
7	78	102.75	R	3
8	31	92.05	G	3
9	47	114.65	R	3
10	41	98.86	R	4
11	104	85.41	PG	5
12	75	94.98	PG	3
13	60	97.85	PG13	3
14	32	109.98	PG13	3
15	77	110.27	G	5
16	71	104.81	PG13	4

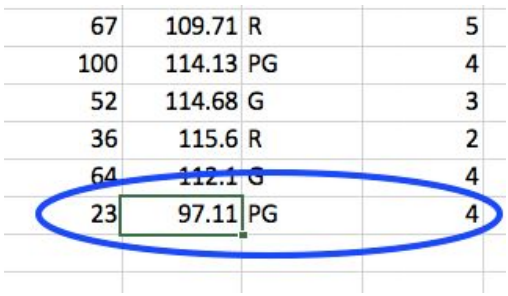
4. Now you can answer simple questions about the data set. For example, what's the rating of Movie 81? To figure it out, hunt down 81 in the Movie ID column, then move along that movie's row to the rating variable.



26	88.74	R	5
79	74.89	PG13	5
16	118.04	PG13	3
81	107.18	PG	4
12	104.44	PG13	5
103	90.38	PG13	3
87	87.57	PG13	4
68	95.12	R	4

Movie 81 is rated PG.

5. How long is Movie 23? Head to the row for Movie 23 (it's at the very bottom of the list), and check the runtime variable.



67	109.71	R	5
100	114.13	PG	4
52	114.68	G	3
36	115.6	R	2
64	112.1	G	4
23	97.11	PG	4

Boom. Movie 23 is 97.11 minutes long.

Exercise 2: Variable Properties

Now we'll look at the different types of measurement used in each of the four variables from the previous exercise. Here are the four basic levels of measurement:

Categorical data types:

Nominal data are grouped into distinct categories. There's no order or ranking to nominal data, and the only information is the category itself. Examples include colors, car models, etc.

Ordinal data are also grouped into distinct categories, but the categories *do* have an order/ranking. For example, if T-shirt sizes include Small, Medium, Large, and Extra Large, then those sizes are ordinal categories because they are ordered from Small to Extra Large.

Numeric data types:

Interval data have numeric values at regular intervals, but the number 0 (zero) is arbitrary and doesn't actually mean "none." For example, temperature in degrees Fahrenheit is a type of interval measurement because 0 degrees does not mean "no heat" or "no temperature at all." That zero is

arbitrary — 0 degrees is just one point along the scale, not a lack of temperature. Another feature of interval data is that you can't make useful comparisons between the numbers/values. For instance, a temperature of 40 degrees Fahrenheit is *not* twice as hot as 20 degrees Fahrenheit — if you converted them both to Celsius, you'd have about -16.7 degrees Celsius and 4.44 degrees Celsius, and 4.44 is clearly not double -16.7.

Ratio data also have numeric values, but the number 0 (zero) is absolute and *does* mean "none of this thing." Examples include age, weight, and height — age 0 means you haven't been born yet; a weight of 0 means there's nothing there; and a height of 0 means no height at all. With ratio data, you *can* make comparisons between values. For instance, a 200-pound person does weigh twice as much as a 100-pound person.

Using these four types, let's take another look at our movie data.

1. Which variable from the movie data set is nominal?

	A	B	C	D
1	movieid	runtime	rating	liking
2	6	112.74	PG13	3
3	76	96.68	PG	3
4	39	81.14	PG13	5
5	89	104.07	PG	4
6	93	101.38	G	4
7	78	102.75	R	3
8	31	92.05	G	3
9	47	114.65	R	3
10	41	98.86	R	4
11	104	85.41	PG	5
12	75	94.98	PG	3
13	60	97.85	PG13	3
14	32	109.98	PG13	3
15	77	110.27	G	5
16	71	104.81	PG13	4

Hmm. At first glance, it might seem like the movie ratings are a good match for nominal data because that's the only variable that doesn't involve numbers. But the ratings system, which runs from G to PG to PG-13 to R, actually *does* have an order to it. A movie rated PG, for example, is definitely in between G and PG-13 in terms of the recommended age groups.

Nominal data have distinct categories and no ranking. You can definitely eliminate the runtime variable because it involves a numerical measurement, not categories. You can also eliminate the likability score because it's a ranking system. But what about the movie ID?

Yep, the **movie ID** is the nominal variable here. It does have numbers, but those numbers don't actually refer to any numeric values. For example, Movie 8 isn't somehow double the value of Movie 4. The numbers are really just identifiers — they don't have any actual value here. We could just as easily identify them by letters, symbols, or titles.

2. Which variable from the movie data is ordinal?

   Ordinal data have distinct categories, and the categories have an order/ranking.

   Like we mentioned above, the **movie ratings** have distinct categories (G, PG, PG-13, R) and are classified according to rank: G is clearly the lowest (general audiences), followed by PG (parental guidance suggested), PG-13 (parents strongly cautioned), and R (restricted). So that's your ordinal variable.

3. Which variable from the movie data uses an interval scale?

   Interval data have numeric, measurable values and an arbitrary definition of zero. We have runtime and likability left as possible variables, so it has to be one of those. But they both involve numbers, so which one is interval data?

   Well, it can't be runtime, because think about what zero means in terms of a movie's length: A length of zero minutes means there's no movie, right? Zero is absolute.

   The "**liking**" variable is measured on an interval scale. There's a definite order to the numbers — 5 is a higher score than 1 — but zero is arbitrary. A likability score of 0 wouldn't mean that some elusive, measurable factor called "likability" is completely empty. It's an arbitrary scale; you could just as easily rank each movie from 10 to 20 instead, with 10 as the lowest score.

4. Which variable from the movie data uses a ratio scale?

   By process of elimination, it's obviously the **runtime**. But why?

   Ratio data have numeric values and an absolute zero. Like we mentioned above, if a movie is 0 minutes long, then there's no movie. Zero really does mean zero here, and a 2-hour movie really is twice as long as a 1-hour movie, so the runtime uses a ratio scale.