Lord of the Rings example

I will give you a concrete example of some untidy data I created from this data from the Lord of the Rings Trilogy.

The Fellowship Of The Ring

Race	Female	Male
Elf	1229	971
Hobbit	14	3644
Man	0	1995

The Two Towers

Race	Female	Male
Elf	331	513
Hobbit	0	2463
Man	401	3589

The Return Of The King

Race	Female	Male
Elf	183	510
Hobbit	2	2673
Man	268	2459

We have one table per movie. In each table, we have the total number of words spoken, by characters of different races and genders.

You could imagine finding these three tables as separate worksheets in an Excel workbook. Or hanging out in some cells on the side of a worksheet that contains the underlying data raw data. Or as tables on a webpage or in a Word document.

This data has been formatted for consumption by *human eyeballs* (paraphrasing Murrell; see Resources). The format makes it easy for a *human* to look up the number of words spoken by female elves in The Two Towers. But this format actually makes it pretty hard for a *computer* to pull out such counts and, more importantly, to compute on them or graph them.

Exercises

Look at the tables above and answer these questions:

What's the total number of words spoken by male hobbits?

Does a certain Race dominate a movie? Does the dominant Race differ across the movies?

How well does your approach scale if there were many more movies or if I provided you with updated data that includes all the Races (e.g. dwarves, orcs, etc.)?

Now download this file as a worksheet, open it in Excel, LibreOffice or… and make this set Tidy.

- Each variable forms a column and contains values
- Each observation forms a row
- Each type of observational unit forms a table

https://drive.google.com/file/d/104IySyZicqnuBOY_5pNCJ1eHgHCGabh2/view?usp=sharing

Tidy Lord of the Rings data

Here's how the same data looks in tidy form:

Notice that tidy data is generally taller and narrower. It doesn't fit nicely on the page. Certain elements get repeated a lot, e.g. `Hobbit`. For these reasons, we often

instinctively resist tidy data as inefficient or ugly. But, unless and until you're making the final product for a textual presentation of data, ignore your yearning to see the data in a compact form.