To download the data that we'll be working with in this video, click on the hyperlink given in the text above this video.

Don't use Internet Explorer.

Chrome, Safari, or Firefox should all work fine.

After you click on the hyperlink, it will take you to a page that looks like this.

Go ahead and copy all the text on this page by first selecting all of it and then hitting Control C on a PC or Command C on a Mac.

Then go to a simple text editor, like Notepad on a PC or Text Edit on a Mac, and paste what you just copied into the text editor with Control V on a PC or Command V on a Mac.

Then go ahead and save this file as the name movielens.txt, for text.

Save this somewhere that you can easily navigate to in R. Now, let's switch to R and load our data.

First, in your R console, navigate to the directory where you just saved that file.

And click OK.

Now, to load our data, we'll be using a slightly different command this time.

Our data is not a CSV file.

It's a text file, where the entries are separated by a vertical bar.

So we'll call it data set movies, and then we'll use the read.table function, where the first argument is the name of our data set in quotes.

The second argument is header=FALSE.

This is because our data doesn't have a header or a variable name row.

And then the next argument is sep="|" , which can be found above the Enter key on your keyboard.

We need one more argument, which is quote="(backslash)" ".

Close the parentheses, and hit Enter.

That last argument just make sure that our text was read in properly.

Let's take a look at the structure of our data using the str function.

We have 1,682 observations of 24 different variables.

Since our variables didn't have names, header equaled false, R just labeled them with V1, V2, V3, et cetera.

But from the Movie Lens documentation, we know what these variables are.

So we'll go ahead and add in the column names ourselves.

To do this, start by typing colnames, for column names, and then in parentheses, the name of our data set, movies, and then equals, and we'll use the c function, where we're going to list all of the variable names, each of them in double quotes and separated by commas.

So first, we have "ID", the ID of the movie, then "Title", "ReleaseDate", "VideoReleaseDate", "IMDB", "Unknown"-- this is the unknown genre-- and then our 18 other genres-- "Action", "Adventure", "Animation", "Children's, "Comedy", "Crime", "Documentary", "Drama", "Fantasy", "FilmNoir", "Horror", "Musical", "Mystery", "Romance", "SciFi", "Thriller", "War", and "Western".

Go ahead and close the parentheses, and hit Enter.

Let's see what our data looks like now using the str function again.

We can see that we have the same number of observations and the same number of variables, but each of them now has the name that we just gave.

We won't be using the ID, release date, video release data, or IMDB variables.

So let's go ahead and remove them.

To do this, we type the name of our data set-- movies$-- the name of the variable we want to remove, and then just say =NULL, in capital letters.

This would just remove the variable from our data set.

Let's repeat this with release date, video release date, and IMDB.

And there are a few duplicate entries in our data set, so we'll go ahead and remove them with the unique function.

So just type the name of our data set, movies = unique(movies).

Let's take a look at our data one more time.

Now, we have 1,664 observations, a few less than before, and 20 variables-- the title of the movie, the unknown genre label, and then the 18 other genre labels.

In this video, we've seen one example of how to prepare data taken from the internet to work with it in R.

In the next video, we'll use this data set to cluster our movies using hierarchical clustering.