# From the Expert – Week 2 Assignment

By Zach Adair

MSDS 662 – Exploratory Data Analysis

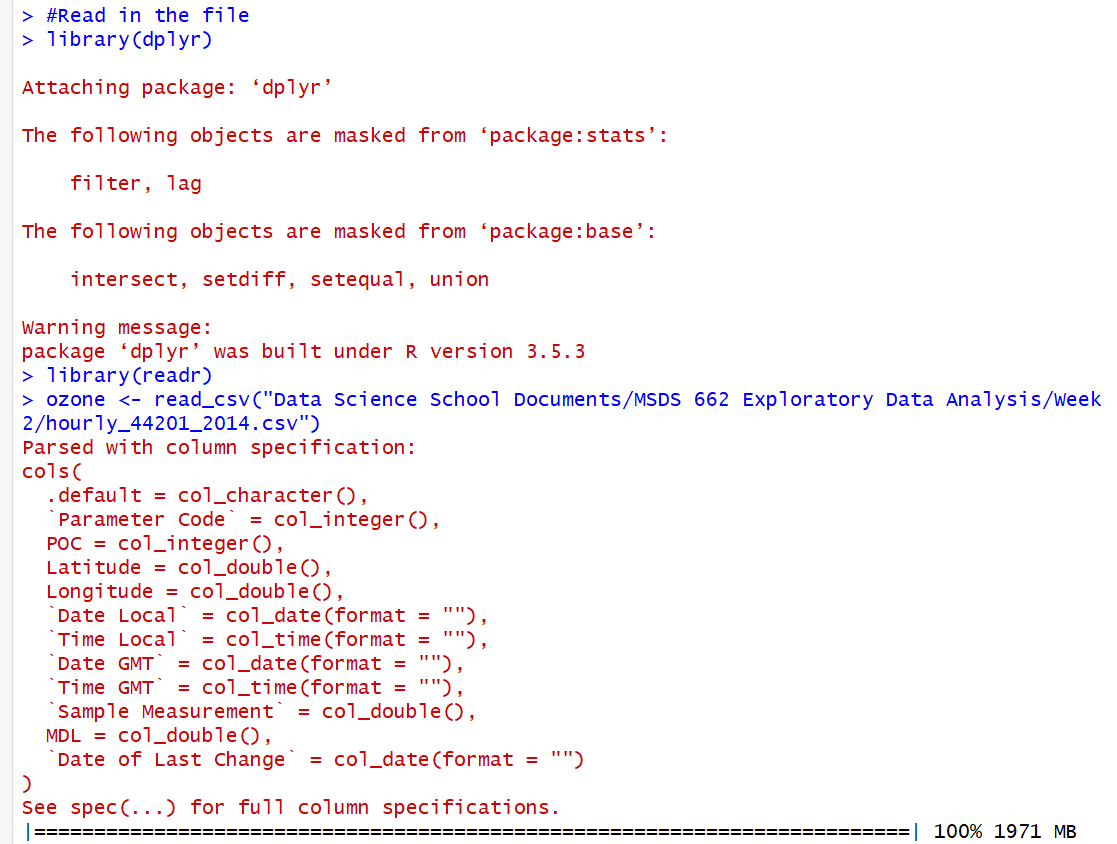
Regis University

9/8/2019

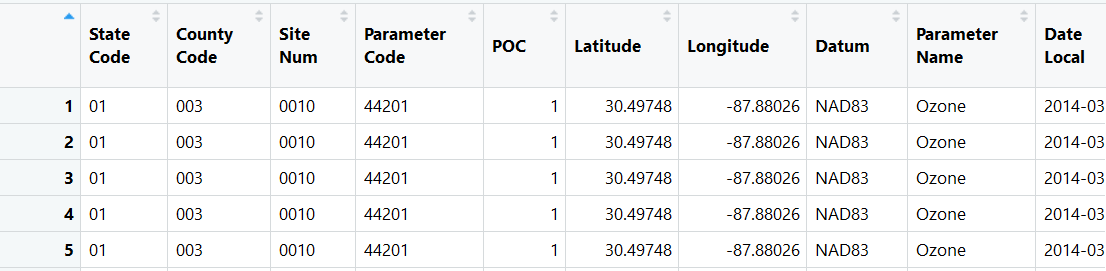
## Introduction

The data is referring to the ozone and the From The Expert walk-thru. It will explore the observations, their location and their measurements and we will learn a little bit more about this data that way as well as more about the dplyr library in RStudio.

## Load the Data and Dplyr Library

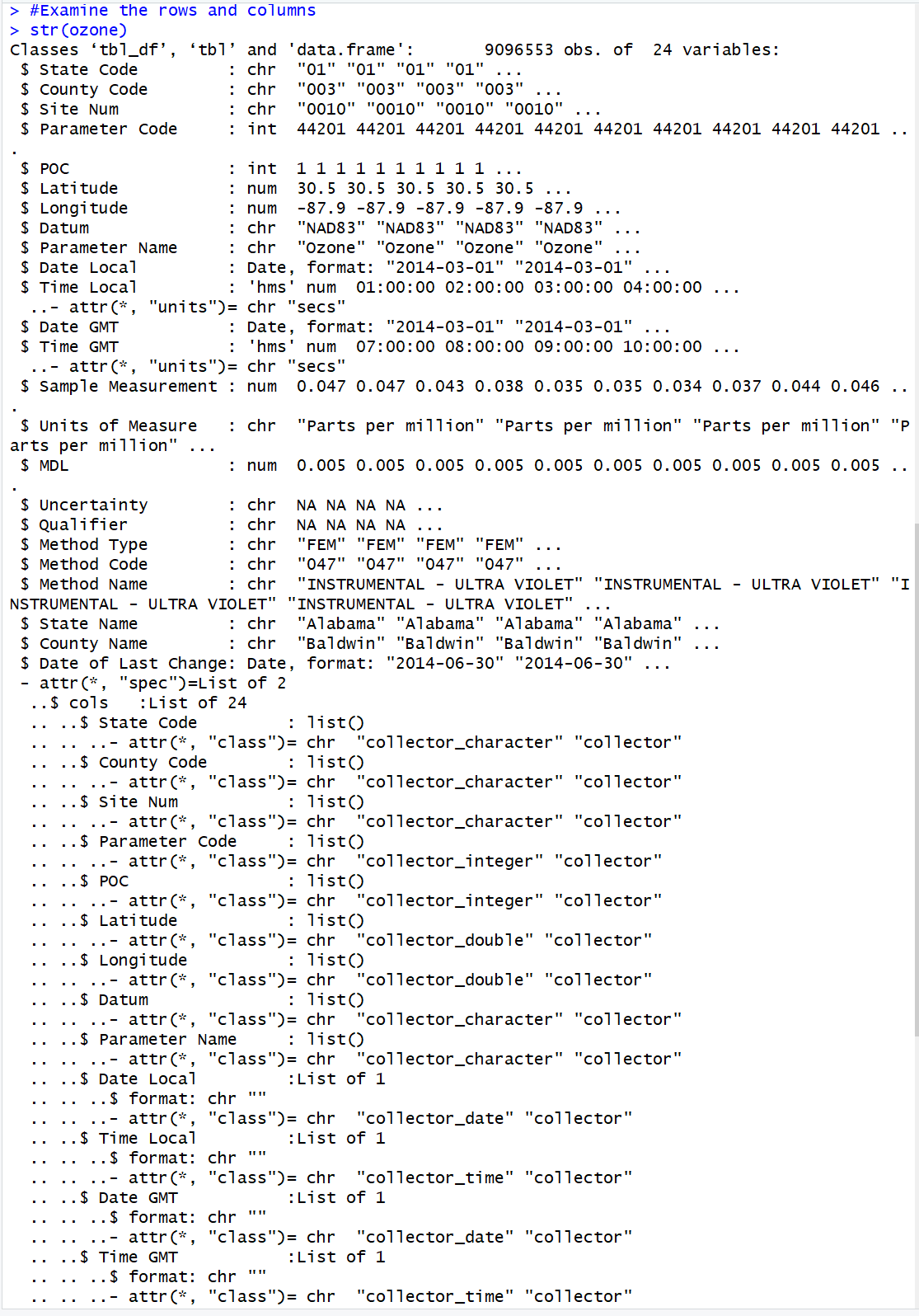


Now that the data and dplyr package are in, lets view the ozone data.



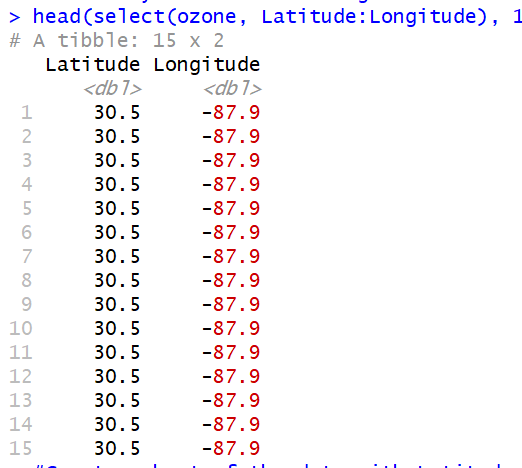
## Examine the Data

Let’s take a look at the data and see what information we can find from this dataset.



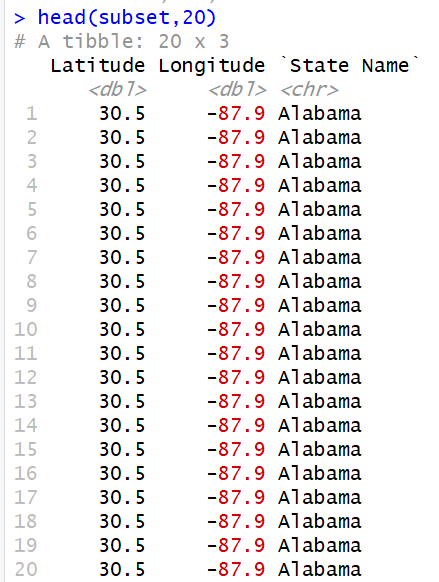
This gives us a good look at the data attributes and some of the characteristics about them like the type as well as the number of rows and variables in the entire dataset which are 9096553 and 24 total variables.

In particular let’s look at the Latitude and Longitude variables, lets use the head function and look at the first 15 variables.



All these records are from the same latitude and longitude meaning they are all from the same location, let’s do the same thing but create a subset with state name as well and add an extra 5 rows to see what our results are now. Start with making a subset of the data with those three variables.





With adding the state name we know that the data points in question are all from the state of Alabama. Next what we will do the From the Expert commands from the text, followed by going through the Exploratory Data Analysis Checklist from the textbook.

## The Exploratory Data Analysis Checklist

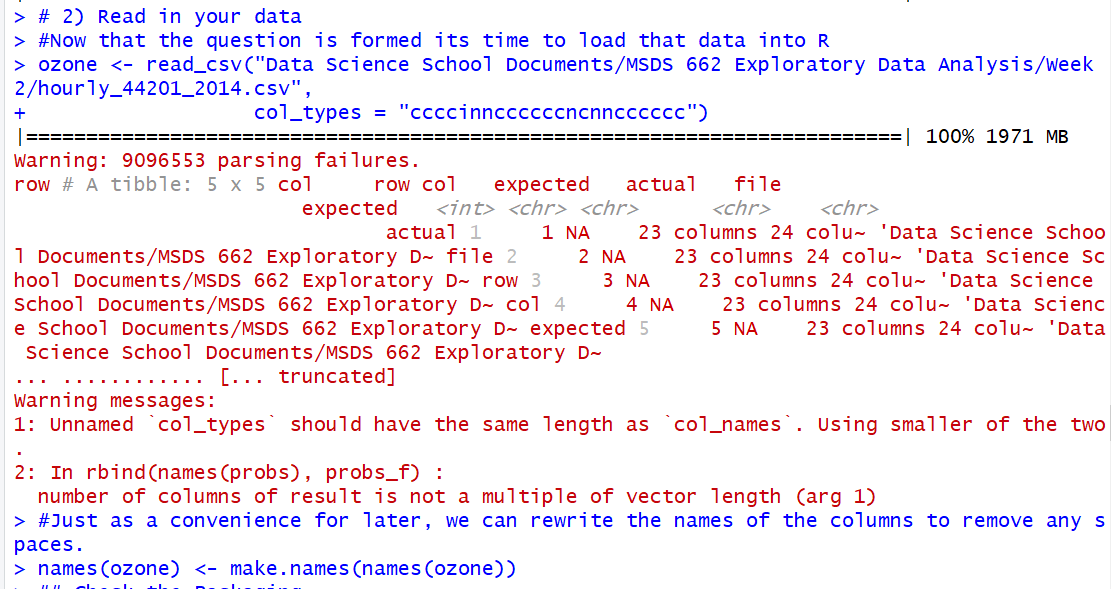
The Exploratory Data Analysis checklist is an informal list to help analysis go through the EDA process. The steps to the process are: 1) Formulate your question, 2) Read in your data, 3) Check the packaging, 4) Run str(), 5) Look at the top and the bottom of your data, 6) Check your n’s, 7) Validate with at least one external data source, 8) Try the easy solution first, 9) Challenge your solution, 10) Follow up.

1. Formulate your question

The question of interest we are trying to solve with this data is Which counties in the USA have the highest level of ambient ozone pollution?

1. Read in your Data

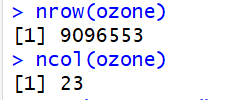
Now we will run the data into RStudio. At the same time, we will load our necessary libraries.



Now, we have our ozone data loaded, as well as setting the proper column types.

1. Check the Packaging

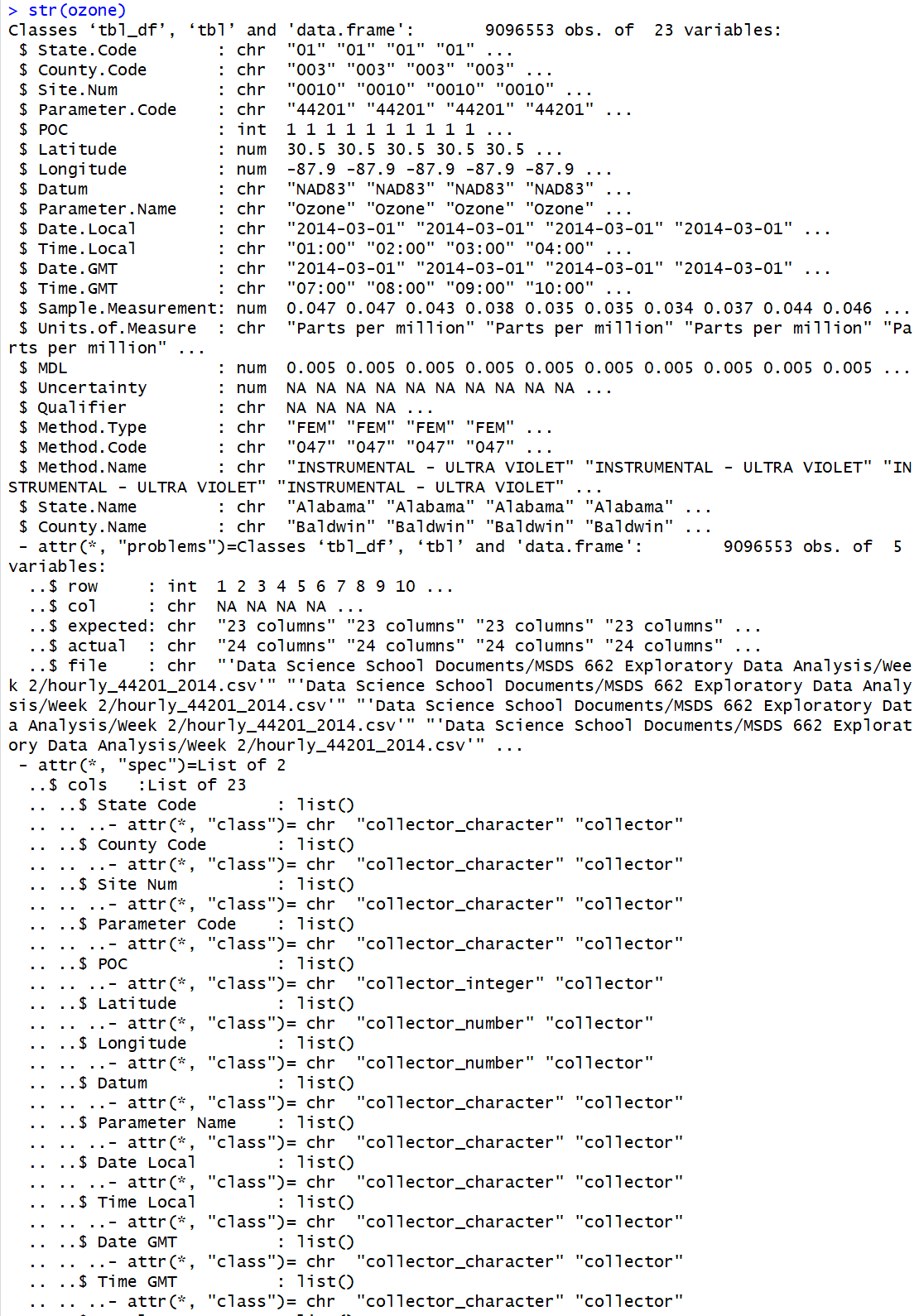
Checking the packaging is when the analyst will look at the number of columns and rows in the dataset before learning more about it.

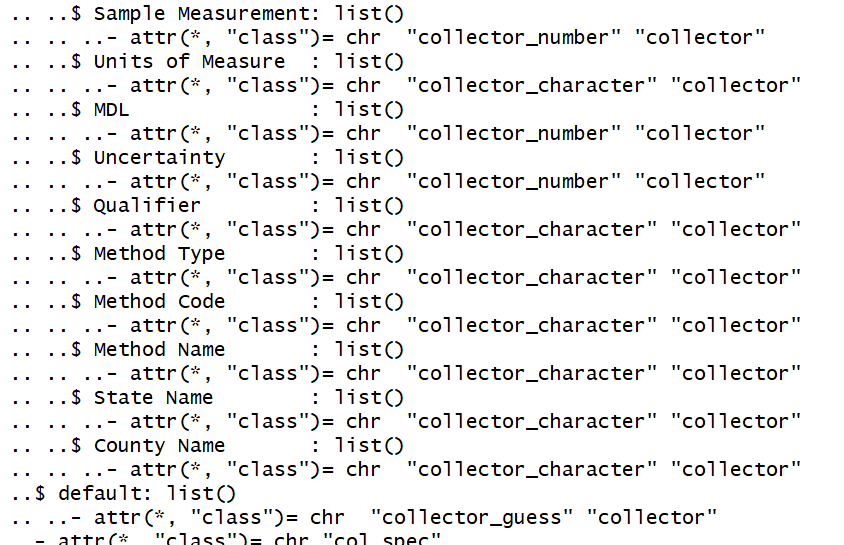


The number of records in the ozone data is 9,096,553 and the number of attributes are 23.

1. Run str()

The str() is one of the best commands in R for learning about all of the characteristics that a dataset entails, so let’s do it to the ozone data now.

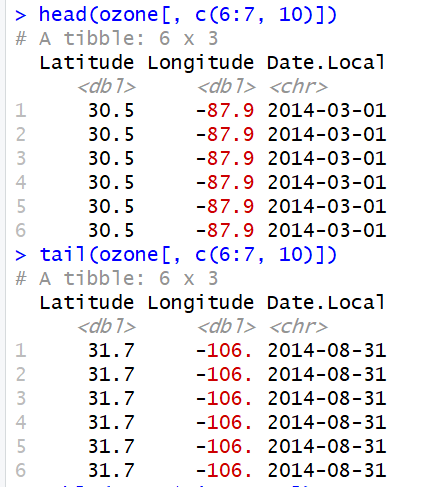




The str() command allows us to again learn about the number of rows and columns as well as the type of columns the dataset has. In this situation, we set the column types, so this is a good way to make sure they were set correctly.

1. Look at the top and bottom of your data

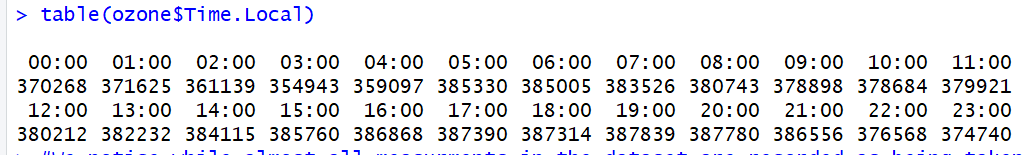
Looking at the top and bottom of your dataset, you can make sure the dataset was loaded correctly into R. Let’s do that now with the ozone dataset using the head and tail commands.



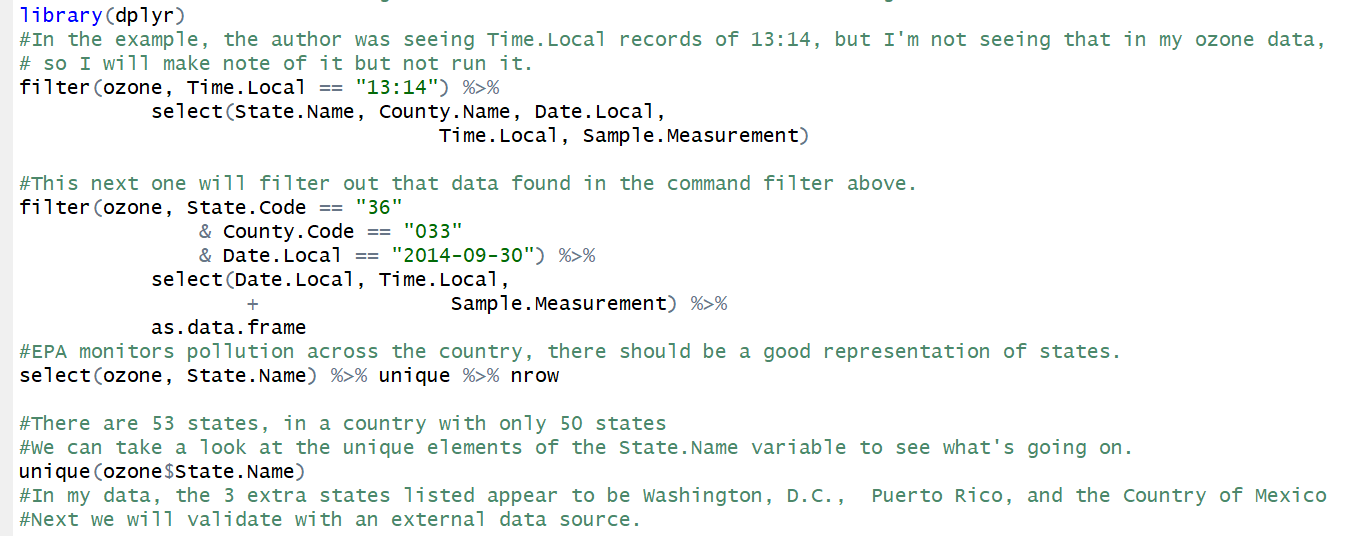
From what we are seeing, it appears our data is loaded correctly with no errors, when you have a lot of variables like we do with the ozone dataset it is smart to just take a subset of columns and that can show us what we are looking for. In this scenario, we are seeing that the data was loaded correctly and we should good moving forward with our EDA checklist.

1. Check your “n”s

Checking your “n”s is good for also figuring out if anything is wrong with the dataset. We will start by doing a table of the Time Local column and see if there are any issues with it. There should an aggregate number the hours in which the measure occurred.



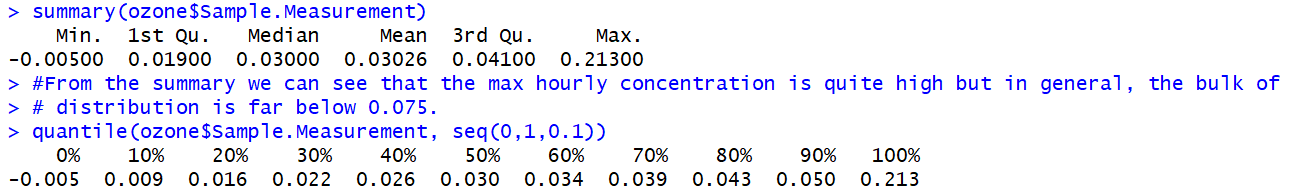
In the example from the text, there was some non-hourly aggregate data, so they filter out those rows, we don’t need to do that with this dataset but I will screen shot those commands. To do these commands an analyst will use the dplyr library.



Now we will validate this ozone dataset with an external data source.

1. Validate with at least one external data source

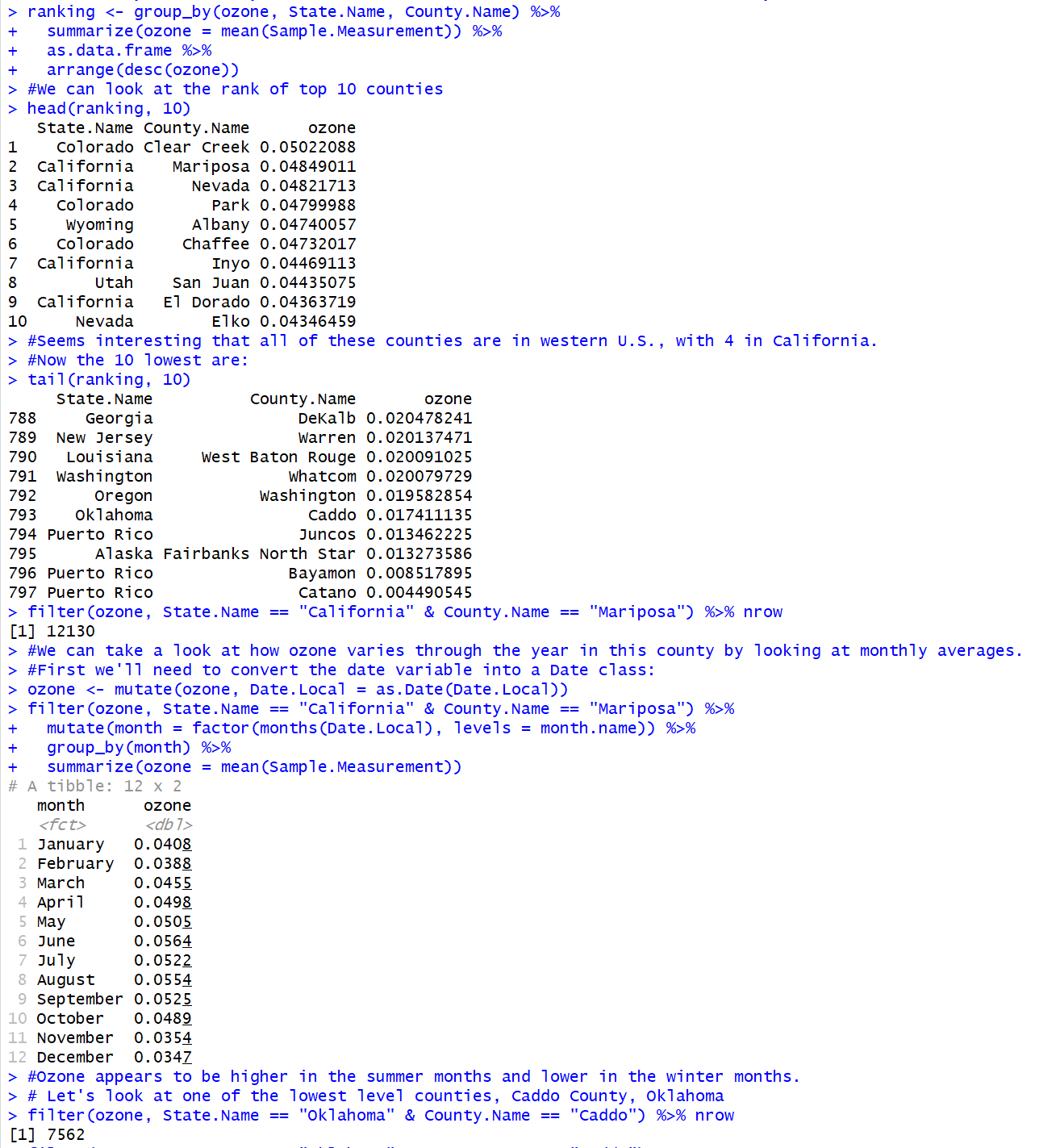
The validity of a data source are very important when analyzing data. So what we need to do next is compare that to standard set in 2008 that “annual fourth-highest daily maximum 8 hour concentration, averaged over 3 years” should not exceed 0.075 parts per million (ppm).

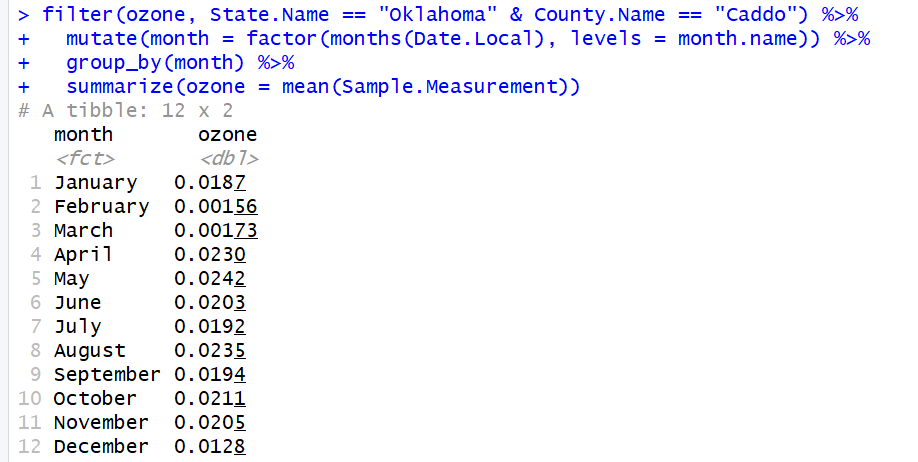


Knowing that the national standard for ozone is something like 0.075, we can see from the data that: the data are at least of the right order of magnitude (the units are correct), the range of the distribution is roughly what we’d expect, given the regulation around ambient pollution levels, some hourly levels (less than 10%) are above 0.075 but this may be reasonable given the wording of the standard and the averaging involved.

1. Try the easy solution first

The simplest answer we can do provided the question is to find the counties that have the highest levels of pollution, we will need a list of counties that are ordered from highest to lowest by levels of ozone.

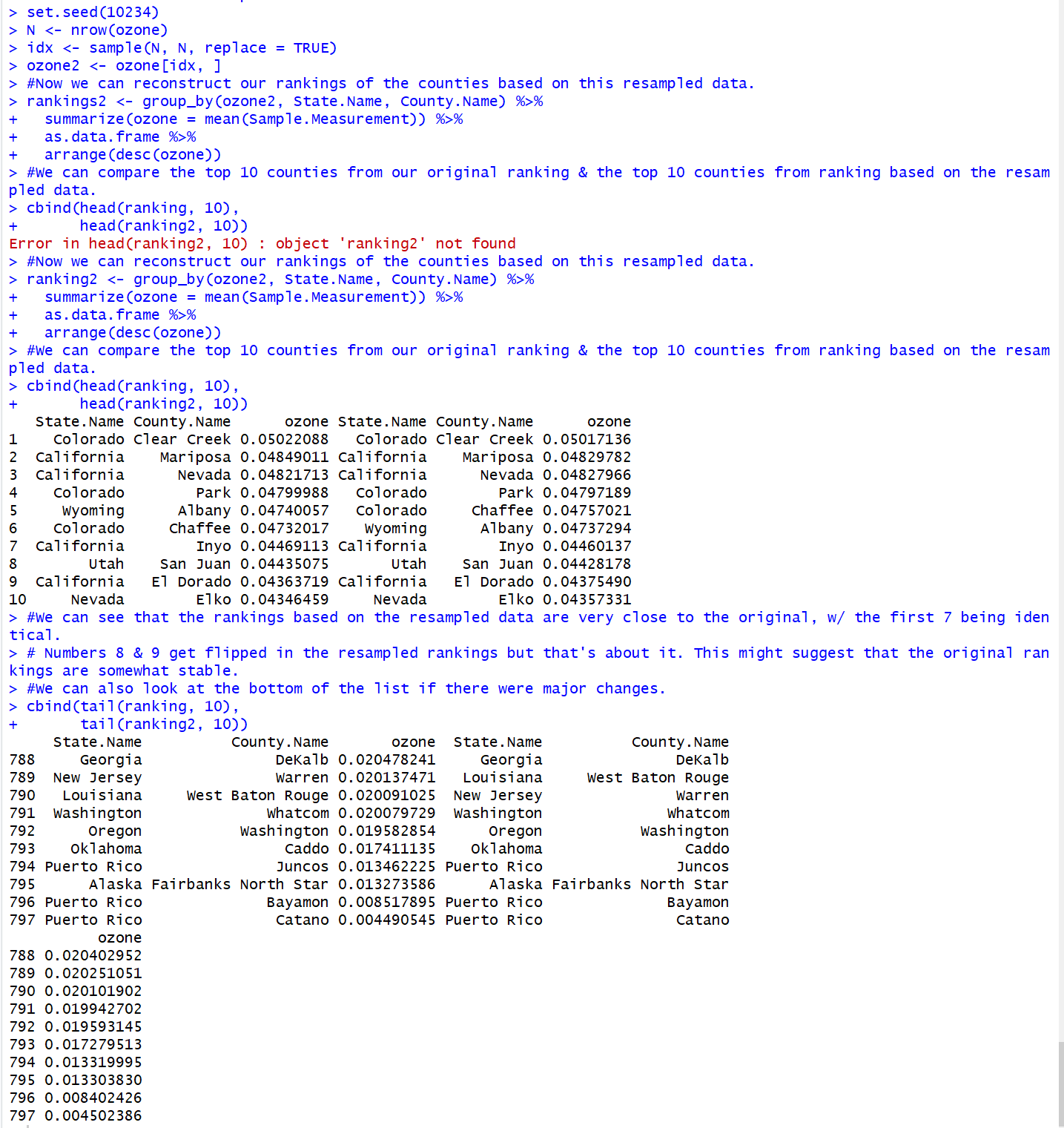




What we did in this step is we created a top 10 and bottom 10 list of counties and ozone levels, than we grabbed one county from each list and tested their monthly output of ozone levels. Doing this test what areas are outputting the highest a lowest ozone levels as well as seeing what monthly output it takes to be in each of those lists.

1. Challenge your solution

Challenging your solution is important in EDA, it adds validity to the data and helps with answering questions that much further. With the ozone dataset, we will set a random number and rank the counties again and then compare that from our answer above.



There’s a majority of counties are similar between the two rankings, the non-random and random, which is reassuring for our dataset and outcomes from before.

1. Follow up questions

After going through the entire process we can now circle back to question one here and start asking more questions about the data. And with the strength of the knowledge we gained we can form better questions which can in-turn lead to more meaningful insights overall. Some questions we can ask to dive deeper into the knowledge of this dataset would be: do we have the right data? Do we need other data sources? And do we have the right question(s)? Depending on the dataset and what your are trying to answer you could have more questions to ask but starting with this as a base can go a long way in analyzing data and pulling out a lot of meaningful information.