1.5 - Downloading data from World Bank

The first data set I tried to download was:

Country: Croatia and Slovenia
Series: 006.Export Value per Exporter: Mean, 021.Export Value per Surviving Entrant: Mean, and 031.Growth of Incumbents: Mean
Years: 2012, 2013, 2014

This downloaded without a problem, but only had data for two fields (HRV 2012 006 and 031). From there I first tried a different set of yeas with the same countries and series:

Years: Past 10 years

This was better since the years 2007 to 2011 were almost complete. However the years before and after were blank.
Since I decided that this is the data I am interested in, I decided to look that those 4 years and try to add more countries in the region that may have data for that same period:

Country: Croatia, Slovenia, Albania, Bulgaria, Kosovo, Macedonia, and Romania

This worked slightly better, but some of these countries seemed to not report any data at all. Bulgaria and Kosovo did not report any data in this 4 year period. In the case of Kosovo this is my mistake since it was not a formal country before 2008 and probably did not declare any information on it's own before 2013 Brussels Agreement. Therefore I went back to my original two countries and tried to find some more in the more Central European region.That turned out to be rather difficult since those countries did not report numbers.
After this challenge I decide to pick 10 European countries that were on the list regardless of geographical location. This data was more complete, although some countries like Sweden had no data and others like Germany were missing years.
This exercise has whos me that is it time consuming to find a good sample to use for data analysis even from a reputable source.

1.4 - Scraping data about used cars (Ford Focus)

I used an webscaping tool called webscraper.io. to scrape information about Ford Kugas from cars.com It was very easy to use. I needed to build a sitemap which the program would use to scrape the information I needed. I added the root, the product link, plus all in information I wanted from that product. It was simple to do and worked well. I ran into a problem with the pages, but the tool showed me how to work this into the sitemap. I somehow only ended up with 94 records, which almost all the cars on the first page. The records seem complete and good for further analysis.

2.2 - Clean car data

I loaded and cleaned my data in R. Please look at the script for more details on the process. I ended up with a very small data set, but it is clean and tidy and easy to work with. The most controversial decision is the dropping of mileage which would have been great to keep, but the data was so corrupt that it would have been impossible to analyse.