Problem Set 4

Ana Gallart

February 19, 2023

1 Data Interested in Scraping

I would be interested in scraping data from the Wall Street Journal, possibly about the bank market, car dealerships, or something else (I'm pretty open). I also thought credit card transaction data could be interesting, but maybe it would be too complicated– with a quick google search I found Yodlee.com which has transaction data. Will have to look further into that. I think looking at standardized test scores could be cool too, not sure if those are available. I feel like I could be interested in a lot, so narrowing it down may come down to the quality and availability of data if that's fair to say.

2 Questions from 6

2.1 Verify that the two dataframe are different types: type class(df1) and class(df). What is the class of each?

df is class ”tbl_df”,”tbl”,”$data.frame$” - a dataframe table

df1 is class ”tbl_spark”,”$tabl_sql$”,”tbl_lazy”,”tbl” - a spark table

2.2 Are the column names any different across the two objects? If so, why might that be?

The columns of df1 have periods seperating the words ex : ”$Sepal.Length$” , while the columns of df have underscores ex : ”$Sepal_Length$. This may be because of the class types. I know in class we mentioned how in some (most) languages you can't use periods in the names of dataframes or variables, so that's probably why the tbl_spark data frame (df) does not have periods the way the R dataframe does (df1).