Problem Set 4

Collin DeVore

February 2023

Question 6

Part 7

df1 is a tibble dataframe and df is a lazy dataframe or a spark dataframe.

Part 8

The names do, in fact, change. Specifically, the underscore symbols in df become periods in df1. This makes sense because the names may be delimited by a different symbol, or the underscores may now have a different specific meaning so they have to be changed.

Question 7

I would be interested in scraping from Project Gutenberg to try performing analytics on the complete Sherlock Holmes stories. I would be curious what would happen if this is compared to Agatha Christie's Hercule Poirot stories. Another set of data I would be interested in scraping is the COVID-19 vaccine data. I will be helping to do this for Yue, who needs this information for his research. The website I would be scraping from is from democratandchronicle.com. Other than this, I am not sure exactly what I would like to try scraping. Sports information and stock information sounds fun to scrape, but I do not yet have an idea for a project utilizing them.