Problem Set 4

adam spohn

February 2024

1 Data Frame Questions

what type of object is "mydf(dollar sign)date? Character.

list the first n rows (i chose 5) of mydf. date description lang category1 category2 granularity ¡chr¿ ¡chr¿ ¡chr¿ ¡chr¿ ¡chr¿ ¡chr¿ 1 1 Tiberius, under order of Augustus... en By place Roman Em... year 2 1 Gaius Caesar and Lucius Aemilius ... en By place Roman Em... year 3 1 Gaius Caesar marries Livilla, dau... en By place Roman Em... year 4 1 Quirinius becomes a chief advisor... en By place Roman Em... year 5 1 Areius Paianeius becomes Archon o... en By place Roman Em... year

What are the class of df and df1? df is a data frame, df1 is a spark data frame.

Are the column names any different across the two objects? If so, why might that be? Yes, df1 uses . in between words, df replaces this with an undescore so that spark can use it to be compatible with spark sql because it cannot use dots.

List the first 6 rows of the SepalLength and Species columns of df.

5.1 setosa 4.9 setosa 4.7 setosa 4.6 setosa 5 setosa 5.4 setosa

List the first 6 rows of all columns of df where SepalLength is larger than 5.5.

SepalLength SepalWidth PetalLength PetalWidth Species ¡dbl¿ ¡dbl¿ ¡dbl¿ ¡dbl¿ ¡chr¿ 1 5.8 4 1.2 0.2 setosa 2 5.7 4.4 1.5 0.4 setosa 3 5.7 3.8 1.7 0.3 setosa 4 7 3.2 4.7 1.4 versicolor 5 6.4 3.2 4.5 1.5 versicolor 6 6.9 3.1 4.9 1.5 versicolor

2 Data Sources

Some data sources I might be interested in scraping from might include baseball stats sites like baseball savant or baseball reference. Baseball savant has a lot of really interesting in depth stats like diffent pitchers spin rates, that many other sites does not have. I also might be intersted in more economy based data sites like FRED.