Heng Zhou, Hongyang Zheng,

Youyou Xie, Zhengqian Xu

Professor Lisa Singh

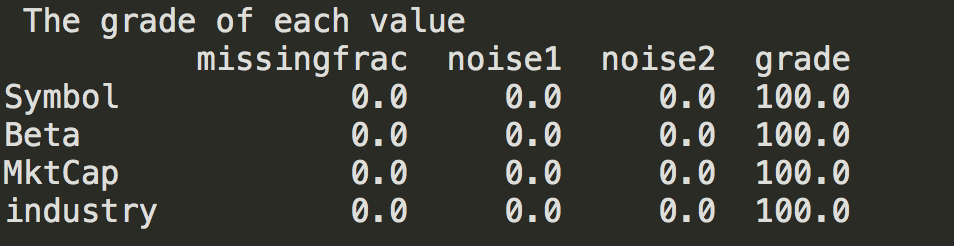
Analytics 501 Project Part1

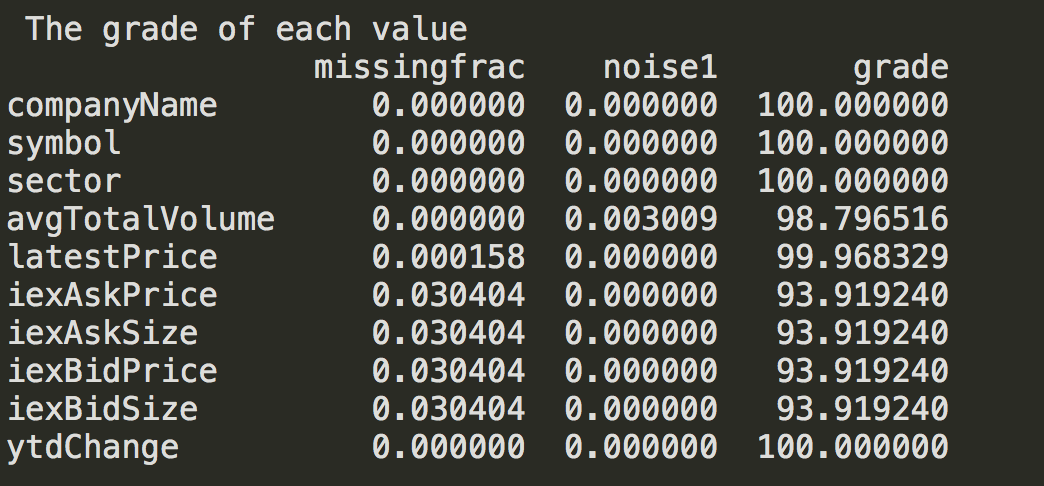
03 October 2018

Explanation of Data Cleaning

Our data quality metric is based on three fractions: the fraction of missing values for each column, the fraction of noise 0 values for each column and the fraction of other noise values for each column. We calculate the fraction by take the ratio of the number of missing values/noise values and the number of total observations in each column. After calculating these fractions, we multiple each with the weight for each part, subtract it from 1, and then multiple 100. The missing value weights 2 and the noise value weights 4, since we think noise value has bigger bad influence on our results. For example, if a column has 0.2% missing value, 0.5% 0 noise value and 0% other noise value, the final grade for this column is 100(1-0.2%\*2-0.5%\*4-0%\*4).

The grade for each column are showed as following tables:





We drop missing values for latest price, iexAskPrice, iexAskSize, iexBidPrice and iexBidSize. We cannot fill it by taking average or prediction of other values, since each observation is independent from another. We cannot fill values by using other website data either, since our target website is IEX. We have four missing values in industry column, so we search on the internet by company name about their industry and then fill by ourselves.

avgTotalVolume has a few 0. We cannot fill it by taking average or prediction of other values, since each observation is independent from another. We cannot use other data source since the time point is different. As a result, we drop them. For Beta, we drop rows which have value 0 since it is impossible. For the rows which have value greater than 100, at first, we want to use the maximum value of beta from that industry. But after comparing the current beta from other website and the average for each noisy company, we found that there was a difference between them. We believe that filling beta with inaccurate values can distort the results, so we drop them. For MktCap, there are some 0. We drop these rows with the same reason for avgTotalVolume. The MktCap looks like a character value, but we make sure it is a float data type. As a result, there is no need to change the data type for it.

After cleaning our dataset based on above description and analysis for all problematic columns, the grade for these columns all increase to 100.

