Heng Zhou, Hongyang Zheng,

Youyou Xie, Zhengqian Xu

Professor Lisa Singh

Analytics 501 Project Part1

03 October 2018

Data Issue

1. Dataset1
2. **companyName:** The name of company

It is a clean column, because there is no missing value or noise value.

1. **symbol:** Symbol of each company

It is a clean column. Because there is no missing value and noise value, and each value is unique.

1. **sector:** The sector of each company

It is a clean column. Because we generate dataset1 using this sector variable.

1. **avgTotalVolume:** Average volume for each stock

A) There are a few 0: average volume cannot be 0

1. **latestPrice:** Refers to the latest price being the IEX real time price, the 15-minute delayed market price, or the previous close price.

A) There are missing values.

1. **iexAskPrice:** The ask price is what IEX sellers are willing to take for it.
2. **iexAskSize:** The volume of stock expected to be sold by IEX sellers.
3. **iexBidPrice:** The price for IEX sellers finally get.
4. **iexBidSize:** The volume of stock actually sold by IEX sellers.

These columns have 0 but we cannot treat it as noise value for the following reason:

0 is not a noise value because it can be zero when the market percent of some certain. stocks is so small that no one sells or buys this stock on IEX. 0 value should be counted for analyzing the behaviors of stocks among sectors.

A) There are missing values

1. **ytdChange:** Refers to the price change percentage from start of year to previous close.

It is a clean column. Because there is no missing value and no noise value.

1. Dataset2:
2. **Symbol:** The symbol of company.

It is a clean column, because we use the symbol list to generate this dataset. There is no missing or noise value in this column. Keep it for dataset combination in the future

1. **Beta:** Beta is a measurement of market risk or volatility
2. It is impossible to be 0, because only cash has the beta of 0. But in our dataset, beta represents the risk level of company.
3. Beta must less than 100 for stock in the market
4. **MktCap:** Market capitalization
5. There are a few 0. Market capitalization cannot be 0
6. **industry:** Industry
7. There are a few missing values.