UK Companies House is a public dataset which contains profile information for public & private companies in the UK. Attached is a sample of 1000 companies from that dataset.

For your interview please be prepared to discuss and share your thoughts on these following points:

Determine the viability to aggregate geographically and by sic\_code.

Explain why you believe it will or will not work.

List concerns or issues about this dataset.

If you make any attempts to cleanup this dataset, explain why and then how you did it.

What columns seemed more useful than others?

1. Dataset Information

1000 data points and 8 variables

1. Geography

Missing values:

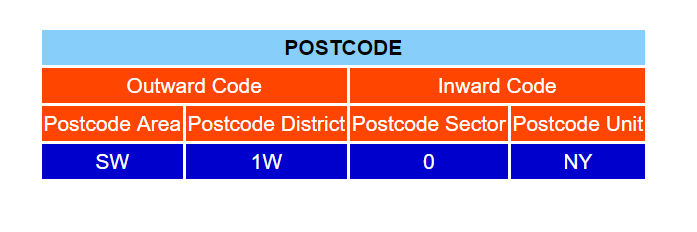
Reg\_address\_post\_town: 985/1000 from London

Reg\_address\_county: 937

Reg\_address\_country: 535

Reg\_address\_post\_code: 50

* 1. Post\_code



Split by blank

Outward code

Inward code

1. Sic-code

Current Standard Industrial Classification (SIC) used in classifying business establishments and other statistical units by the type of economic activity in which they are engaged.

Merge with Sic-code table: 50 NAs

Most from Section T (238 out of 950):

Activities of households as employers; undifferentiated goods- and services-producing activities of households for own use

1. Company status

Active Active - Proposal to Strike off

916 35

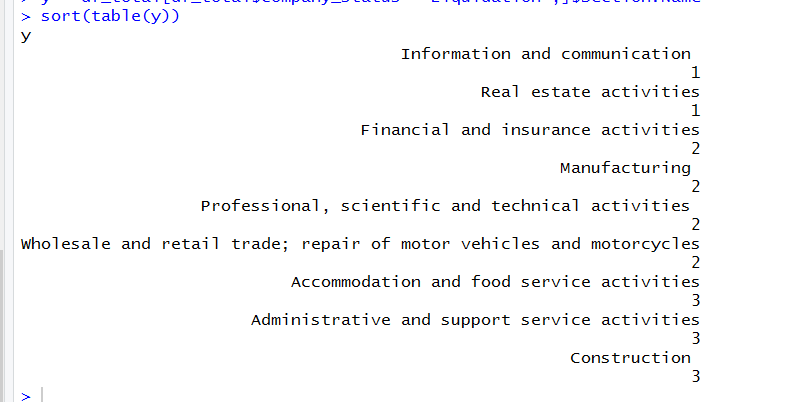
In Administration/Administrative Receiver Liquidation

1 47

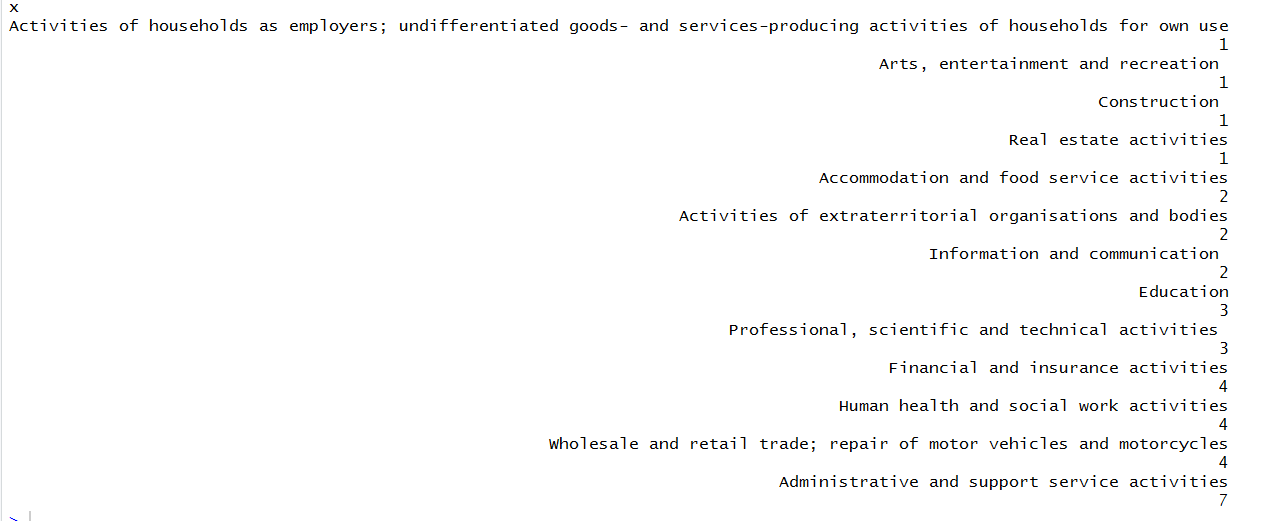
Voluntary Arrangement

1

Company status with liquidation: 28 invalid sic-code (NAs)



Company status with active – proposal to strike off: (35)



Active

