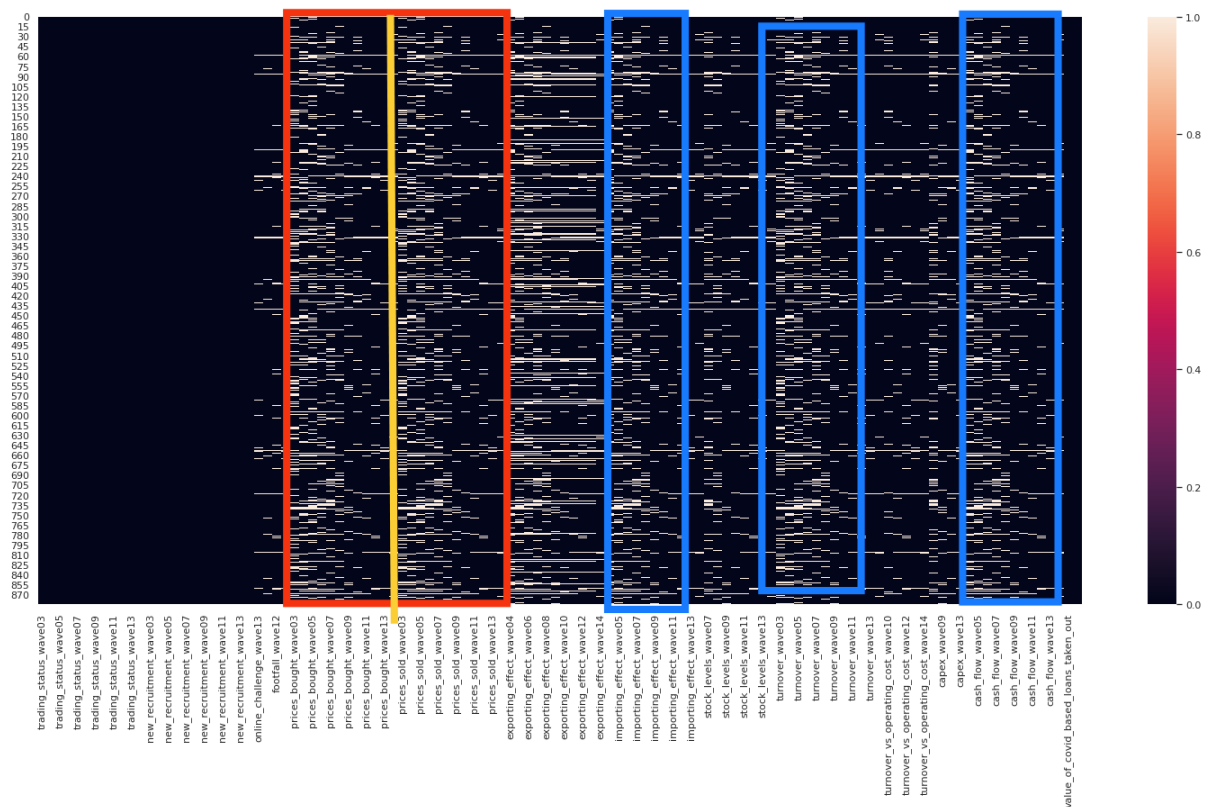COVID-19 data report

1. The dataset has 884 rows by 117 columns

2. The data has over 85 categorical columns

3. The index column should be dropped as further data preprocessing

4. 90 columns have missing values

5. There is a regular pattern to the way the missing values are generated; this can easily be exploited
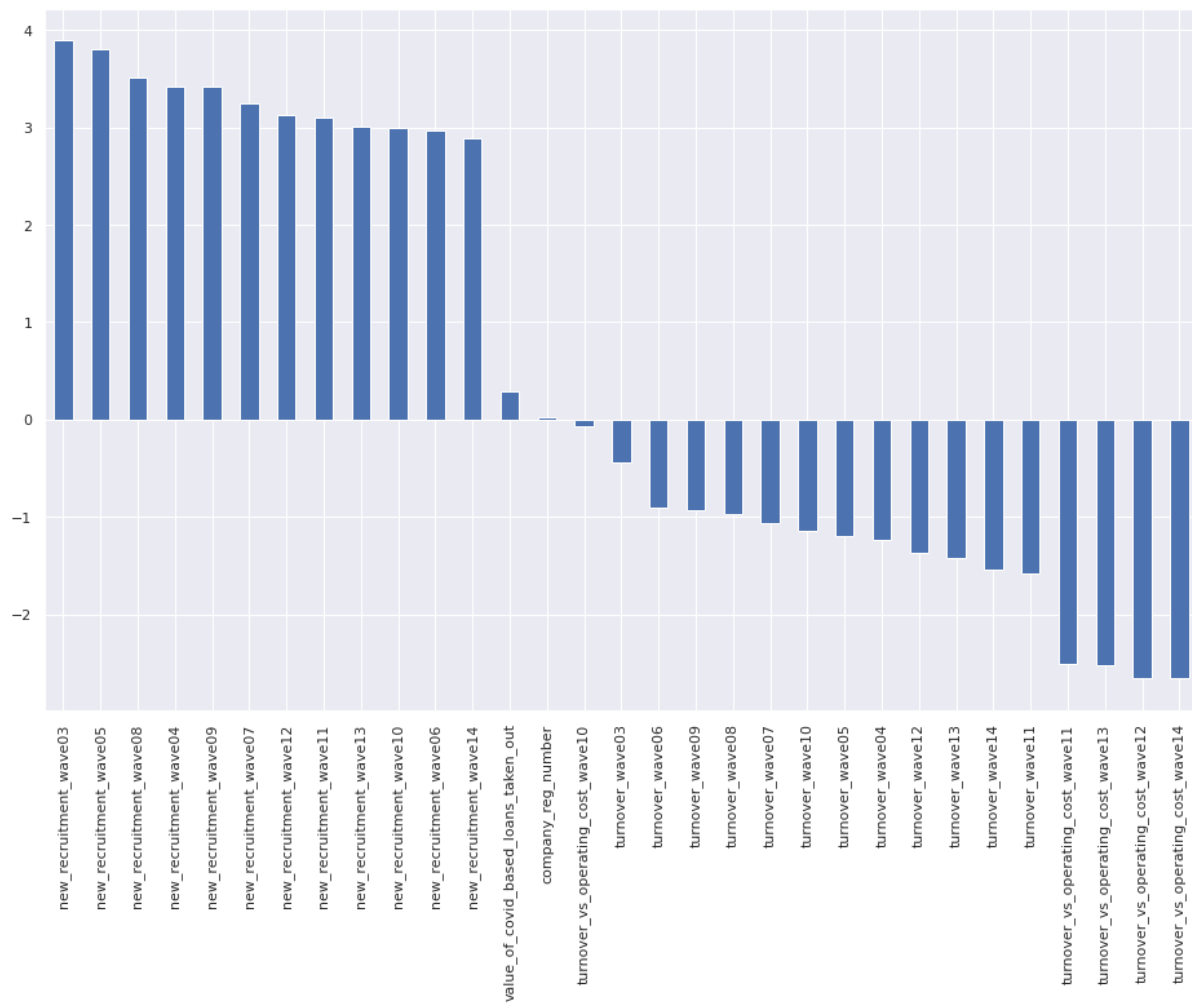
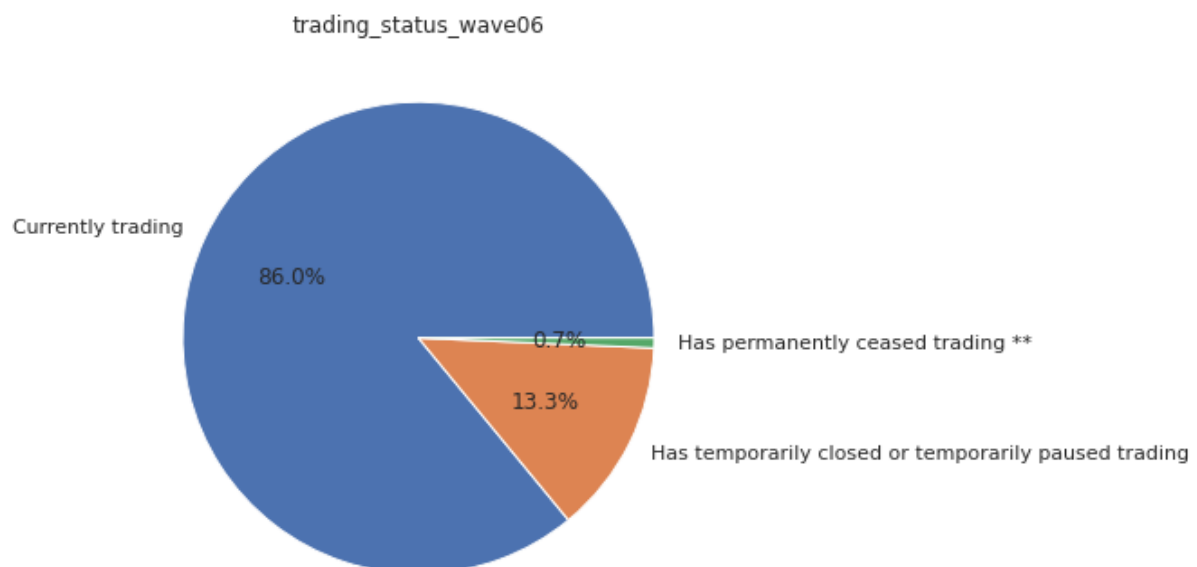6. There is no much multicollinearity between the columns except a few
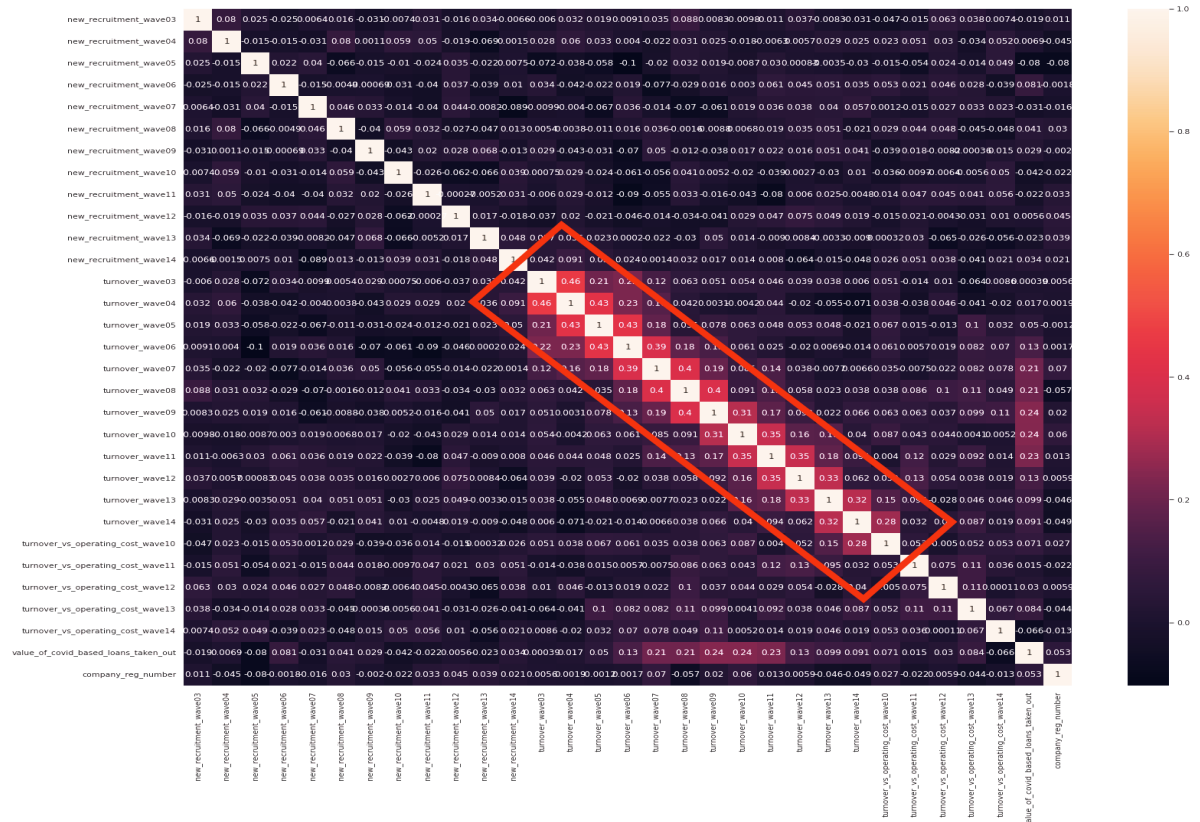
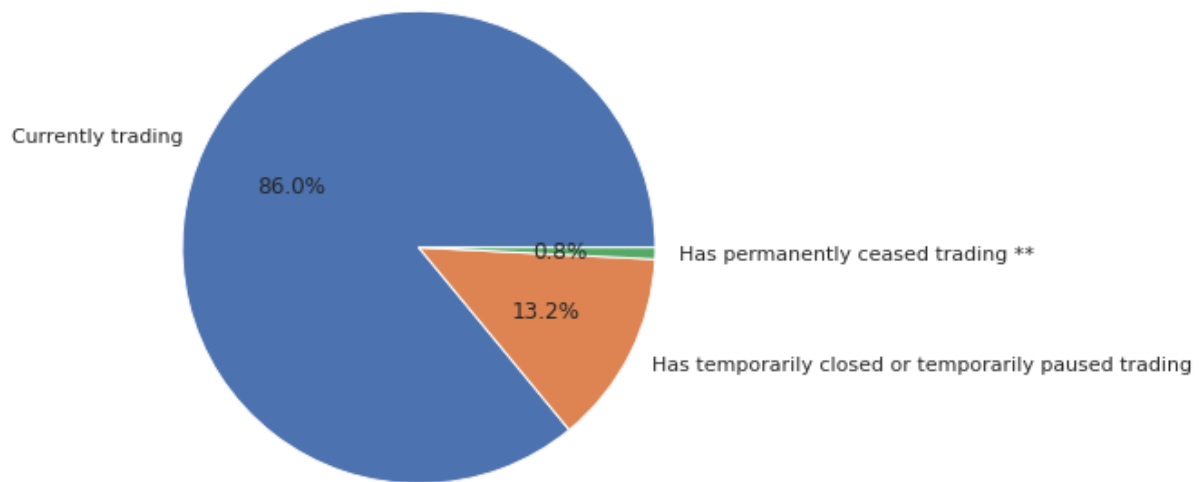7. The distribution of some numeric columns shows tight similarity
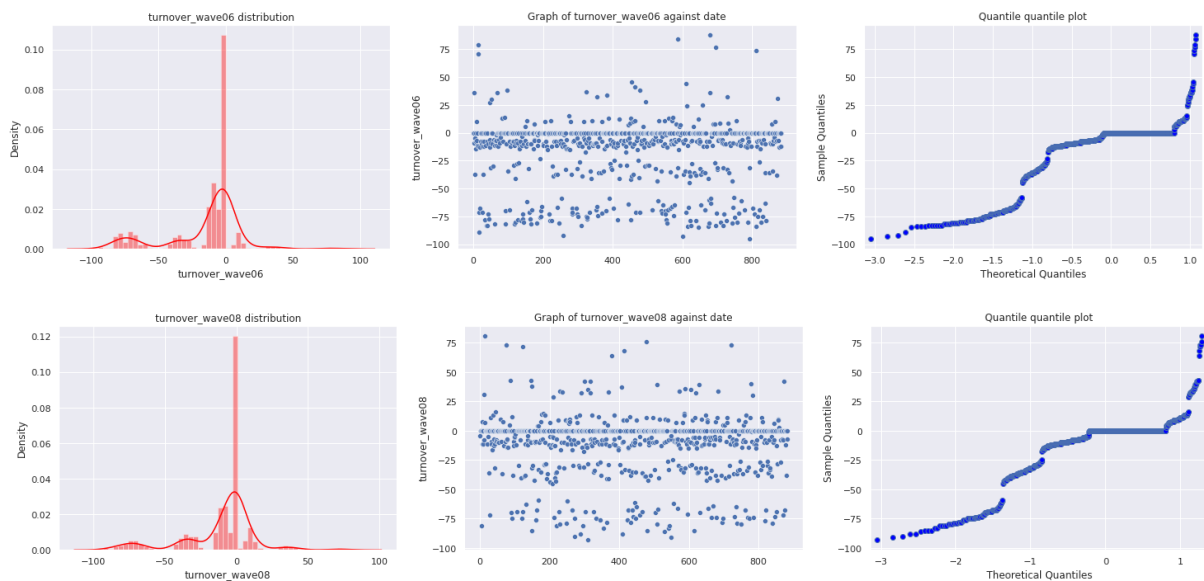
8. Most columns are skewed

9. Most categorical column show semblance in the ratio of the categorical values
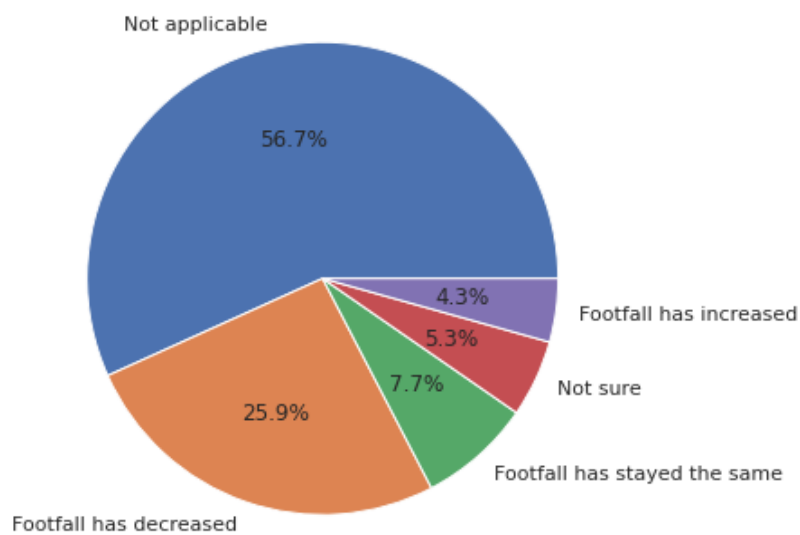
trading_status_wave07

Currently trading — 86.0%

Has permanently ceased trading ** — 0.8%

Has temporarily closed or temporarily paused trading — 13.2%

Another categorical column

footfall_wave11

Not applicable — 56.7%

Footfall has decreased — 25.9%

Footfall has stayed the same — 7.7%

Not sure — 5.3%

Footfall has increased — 4.3%

footfall_wave12

Not applicable — 57.6%

Footfall has decreased — 22.1%

Footfall has stayed the same — 9.5%

Not sure — 5.7%

Footfall has increased — 5.1%
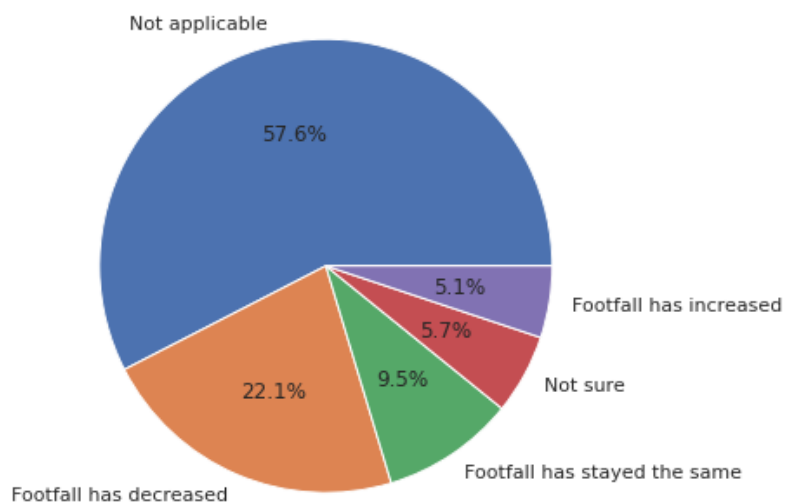
Property	Score
Accuracy	Adept
Complete	Moderately
Timely	It would be great if the meta data addressed this i.e add date data was collected
Auditable	No private/personal information leakage
Metadata	None ; column names not explained
Plausibility	The data does not defy logic

FINAL NOTE

The depth of the data generated is great but some duplication in some of the columns makes the data unnecessarily bloated