[Corporación Favorita](http://www.corporacionfavorita.com/) grocery sales forecasting.

Data can be found at https://www.kaggle.com/c/favorita-grocery-sales-forecasting/data

Due to the large dataset - train.csv has 125 mil records, it is best to perform some data engineering before starting any analysis.

Following are the steps used to reduce memory consumption

* Check the range of values stored in the column
* Check the suitable datatype from the following link <https://docs.scipy.org/doc/numpy-1.13.0/user/basics.types.html>
* Change datatype
* split date col into three columns
  + There are two reasons to do this
    - In pandas any operation on column of type "datetime" is not vectorized. Hence any operations on it will take more time
    - Splitting it into three columns will provide better memory utilization. Eg: in the test dataset date col uses approx. 25 mb while storenbr(uint8) uses approx. 3 mb
* join everything