**Project Flow**

**Item1: Data Preprocessing**

* Read in data, only consider the tuple (shop\_id, item\_id) that exists in the test file.
* Group data by month, in order to create time series feature later. Year and Month are calculated according to the date\_block\_num

**Item2: Data Visualizing**

* Plot sale data in terms of month, shop and category
* Conclude that month and shop are important features

**Item3: Feature Engineering**

* Consider the sale of next month as a feature of the current month, along with the change in sale
* Calculate average sale in different dimensions (e.g. shop, item, category, month and year)
* Merge all tables with above features together
* Fill in missing value and change data type

**Item4: Model training, Prediction and Ensemble**

* Divide dataset into train\_set, validation\_set and test\_set
* Apply single Linear Regression Model from sklearn
  + - Training model using train & validation set
    - Run prediction on the test set
    - Produce results according to the prediction of one month of all tuples (shop\_id, item\_id)
    - Left merge the result table with the final test set, keeping necessary columns for submission
* Model Ensemble
  + - Use Linear Regression, Random Forest and Decision Tree Models
    - Use the prediction of the first level which contains all three regressors as the input of the second level model which uses only the Linear Regression model.
    - Prediction result is generated by inputting the prediction of the first level to the model we got in the second level
    - Left join to the final test set