# Project with Python

1. I choose “Filipino Family Income and Expenditure” data (link: <https://www.kaggle.com/grosvenpaul/family-income-and-expenditure>). This data set have 60 columns and 41544 rows. Problem that I need to solve is create best model in predicting household income.

* This is supervised regression problem.
* I will use root mean square error (RMSE) for performance measure in linear regression.

1. Created Jupyter notebook “view of data set” for data visualization and basic overview.

* In “view of data set” you can find graphs, correlations between attributes and comments.
* Transformation that could help improve model:
  1. count sum of things which in data set start with “Number of ...” (for example: “Number of Television”, “Number of CD/VCD/DVD” and so on.);
  2. sum all expenses;
  3. sum just other expenses, excluding food expenses;
  4. do not use specific food expenses, use only “Total Food Expenditure”.