CSE 519 -- Data Science (Fall 2017)
Prof. Steven Skiena
Homework 2: Exploratory Data Analysis in IPython
Due: Monday, September 25, 2017

This homework will investigate doing exploratory data analysis in IPython. The goal is to get you fluent in working with the standard tools and techniques of exploratory data analysis, by working with a data set where you have some basic sense of familiarity.

This homework is based Zillow Prize Challenge on Kaggle, revolving around predicting the price that a particular real estate property (usually a home) will sell for. More than just data exploration, you must also join the challenge and submit your model before the deadline, to get a score feedbacked from Kaggle. You are to explore the data and uncover interesting observations about the Zillow data. You will need to return all your results in a single, well-documented IPython notebook documenting your methods and the exact sequence of operations you needed to produce the resulting tables and figures.

Data downloading

First of all, you need to join the challenge and download the data here. The description of the data can also be found at this page.

Python Installation

Instead of installing python and other tools manually, we suggest to install Anaconda, which is a Python distribution with package and environment manager. It simplifies a lot of common problems when installing tools for data science. More introduction can be found at here. Installing instruction can be found here. A useful instruction about Anaconda in Youtube can be found here.

Note that if your shell is not bash, you need to add an environment variable by yourself. For example, if you use zsh, you need to open file [~/.zshrc] and add this line into the end: [export PATH="path_to_your_home/anaconda/bin:$PATH"]

If you are an expert of Python and data science, what you need to do is install some packages relevant to data science. Some packages I believe you will definitely use for this homework are as following:

- python
- panda
- scikit-learn
- numpy
- seaborn

Installation on Windows may be a little harder, so you may consider installing Ubuntu on a virtual machine on your Windows machine in case you do not have it installed on your laptop. You may work together with your buddies to figure out the installation process, but you must have this done within a week of receiving this assignment.

Tasks

1. Do a pairwise Pearson correlation analysis on all interesting pairs of variables. Show the result with heat map and find out most positive and negative correlations. You can use the seaborn library to plot the heatmap, with instructions found here. (10%)
2. Produce five other informative plots revealing aspects of this data. For each plot, write a paragraph in your notebook describing what interesting properties your visualization reveals. These must include:
    - at least one line chart
    - at least one scatter plot or data map
    - at least one histogram or bar chart (20%)
3. Set up a simple linear regression model on one or more variables to predict the logerror as a function of other variables. How well/badly does it work? Which variable is the most important one? (20%)
4. Then try to build a better prediction model that works somewhat harder to solve the task. Perhaps it will preprocess features better (e.g. normalize or scale the input vector, convert non-numerical value into float, or do a special treatment of missing values). Perhaps it will use a different machine learning approach (e.g. nearest neighbors, random forests, etc). Which of your models minimizes the squared error? (20%)
5. Predict all the logerror for instances at file “sample_submission.csv”. Write the result into a csv file and submit it to the website. Report the score you get. Your submission file should look like: (10%)

ParcelId	201610	201611	201612	201710	201711	201712
10754147	0.1659	0.1659	0.1659	0.1659	0.1659	0.1659
10759547	-0.0236	-0.0236	-0.0236	-0.0236	-0.0236	-0.0236
10843547	0.1243	0.1243	0.1243	0.1243	0.1243	0.1243
10859147	0.3111	0.3111	0.3111	0.3111	0.3111	0.3111
10879947	0.0745	0.0745	0.0745	0.0745	0.0745	0.0745
10898347	-0.0757	-0.0757	-0.0757	-0.0757	-0.0757	-0.0757
10933547	-0.0075	-0.0075	-0.0075	-0.0075	-0.0075	-0.0075
10940747	-0.4351	-0.4351	-0.4351	-0.4351	-0.4351	-0.4351
10954547	-0.1973	-0.1973	-0.1973	-0.1973	-0.1973	-0.1973
10976347	0.0708	0.0708	0.0708	0.0708	0.0708	0.0708
11073947	0.1371	0.1371	0.1371	0.1371	0.1371	0.1371
11114347	0.1192	0.1192	0.1192	0.1192	0.1192	0.1192
11116947	-0.043	-0.043	-0.043	-0.043	-0.043	-0.043

6. Write a 2-3 page report about your favorite model. The report should include:
    - A description of how it works.
    - An evaluation of how well it works.
    - Any interesting experiences or surprises you had over the course of these experiments.

    Be honest. This is your first modelling experience, and I am hoping to see you learned something rather than that you are now ranked on the leaderboard. (20%)

Rules of the Game

This assignment must be done **individually by each student**. It is not a group activity.

1. If you do not have much experience with Python and the associated tools, this homework will be a substantial amount of work. Get started on it as early as possible!
2. Your written analysis should be embedded in the notebook next to the figures and output.
3. We will discuss topics like linear regression in detail only after the HW is due. Muddle along for now, and we will understand the issues better when we discuss them in the course.
4. There are some public discussions and demos relevant to this problem on Kaggle. It is okay for students to read these discussions, but they must write the code and analyze the data by themselves.
5. Note that HW3 will also be based on this Kaggle challenge, so please treat this homework seriously.
6. Our class Piazza account is an excellent place to discuss the assignment. Check it out at piazza.com/stonybrook/fall2017/cse519. You can also send email to the TA at cse519@cs.stonybrook.edu if you have questions about IPython, github, etc.

Submission

Actually there is no "submission" for this homework. You will need to use GitHub to manage your homework and all we need to know will be recorded there. So watch this video if you are a beginner of Git or GitHub.

To submit your homework successfully, you need:

1. Sign up a GitHub account and install git.
2. Apply for student package using your stony brook email or your cs department email. Click the "Request a discount" button on the top right of the page and fill the forms. This process may need 5 days (or ten minutes). So do it ASAP! It will enable you to have unlimited private repositories.
3. Create a new private repository named "CSE519-2017-IDNumber". Here the IDNumber is you student ID number. For example, I need to create a repo named "CSE519-2017-111753600".
4. Invite me as your collaborator. Click the "Settings"->"Collaborators", search my email "cse519@cs.stonybrook.edu" and click "Add collaborators".
5. Clone this repo to your computer.
6. Create a folder named "HW2". Coding, debugging and saving your IPython notebook (named "HW2.ipynb") in folder "HW2".
7. Add, commit and push the code before deadline. Note that:

    a. You may need to add an ssh key to push to your repo if you get an error saying that it is unable to access your repo. You can learn SSH key and how to add it here.
    b. Add a ".gitignore" file to ignore big data files like *.csv files. More information about gitignore can be found here.
8. Only the commits before the deadline will be considered to be graded. Thus make sure you have successfully pushed you files to the server (use git command "pull" to test).
9. Note that GitHub can show the IPython notebook directly, which means the content showing when you open HW2/HW2.ipynb should the same with what in your computer (browser). So we will grade directly based on the content of this file on you repo.
10. Remember to submit your predicting results to Kaggle and save the screenshot of your ranking and score into HW2 directory.
11. Please search Google for a solution your problem first or consult a fellow student before asking the TA. It will help you a lot.

Git Tutorial & references

Some useful links includes:

- GitHub help
- A ten minutes read about GitHub
- Codecademy's git tutorial found

If you're interested in learning more information about git or expanding your knowledge, refer to these references:

- git-book - Chapter 2 is a MUST read chapter, checkout git aliases!
- An interactive tutorial on git branching - Learn Git Branching
- git cheat sheet