EDA-Project

This is your first project during your bootcamp. You'll be working with the King County House Sales dataset. Here, the focus is on EDA though you are required to demonstrate an entire Data Science Lifecycle.

The data

- The dataset can be found in the file "kc_house_data.csv", in this folder.
- The description of the column names can be found in the column_names.ipynb file in this repository.
- The column names are NOT clear at times.

  In the real world we will run into similar challenges. We would then go ask our business stakeholders for more information. In this case, let us assume our business stakeholder who would give us information, left the company. Meaning we would have to identify and look up what each column names might actually mean.

Tasks for you

1. Through statistical analysis/EDA, above please come up with AT LEAST 3 (you can definitely get bonus points for more than 3) recommendations for home sellers and/or buyers in King County (business questions).
2. Then model this dataset with a multivariate linear regression to predict the sale price of houses as accurately as possible.
3. Use the Mean Absolute Percentage Error (MAPE) to evaluate your linear regression model.
4. **Note for modeling:**
    a. Split the dataset into a train and a test set (hint: scikitlearn).
    b. Use Mean Absolute Percentage Error (MAPE) as your metric of success and try to minimize this score on your test data.

The Process

Each day you start with a short Stand-Up at 09:00 am with your teammates where you answer the following questions:

1. What have you done yesterday
2. What are you going to do today?
3. Do you have any blockers?

On Wednesday each of you will have a short stakeholder review meeting (5 minutes showing what you have done and what are your business questions) and 5 minutes stakeholder feedback. Use your time wisely!

The Deliverables

1. A well documented Jupyter Notebook containing any code you've written for this project and comments explaining it. This work will need to be pushed to your GitHub repository in order to submit your project.
2. An organized README.md file in the GitHub repository that describes the contents of the repository. This file should be the source of information for navigating through the repository.
3. A short Keynote/PowerPoint/Google Slides presentation giving a high-level overview of your methodology and recommendations for non-technical stakeholders. The duration of the presentation should be 10 minutes, then the discussion will continue for 5 minutes. Also put your slides (delivered as a PDF export) on Github to get a well-rounded project.