First Project - EDA

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

The data

- The datasets can be found in the respective folders in this repository.
- Each folder contains the dataset either as a .csv or .txt file. The description of the column names can be found in the column_names.md files.
- The column names may NOT be 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. (google is your friend ;))

Tasks for you

1. Create a new repo and a new virtual environment.
2. Through EDA/statistical analysis above please come up with AT LEAST 3 insights/recommendations for your stakeholder.

   If you use linear regression in the exploration phase remember that R^2 close to 1 is good.

3. Then, model this dataset with a multivariate linear regression to predict
   a. For ***King County House Sales***: The sale price of houses as accurately as possible. Note, you can take either the perspective of a buyer or a seller.
      i. Split the dataset into a train and a test set. (use the sklearn split method https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html)

    ii. Use Root Mean Squared Error (RMSE) as your metric of success and try to minimize this score on your test data.
b. For *US Bank Wages*: The salary (or log(salary)) as accurately as possible. Note you can take either the perspective of an applicant or company.
    i. Split the dataset into a train and a test set. (use the sklearn split method https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html)
    ii. Use Root Mean Squared Error (RMSE) as your metric of success and try to minimize this score on your test data.

The Deliverables

1. A well **documented Jupyter Notebook** (see **here** for an example) containing the 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 (latest upload: 18.02.2020 12:00). Do not push all the analysis... just the analysis that is relevant!
2. A Python script for training the model, printing out the model statistics and saving the model.
3. 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.
4. A **short Keynote/PowerPoint/Google Slides/Jupyter 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.