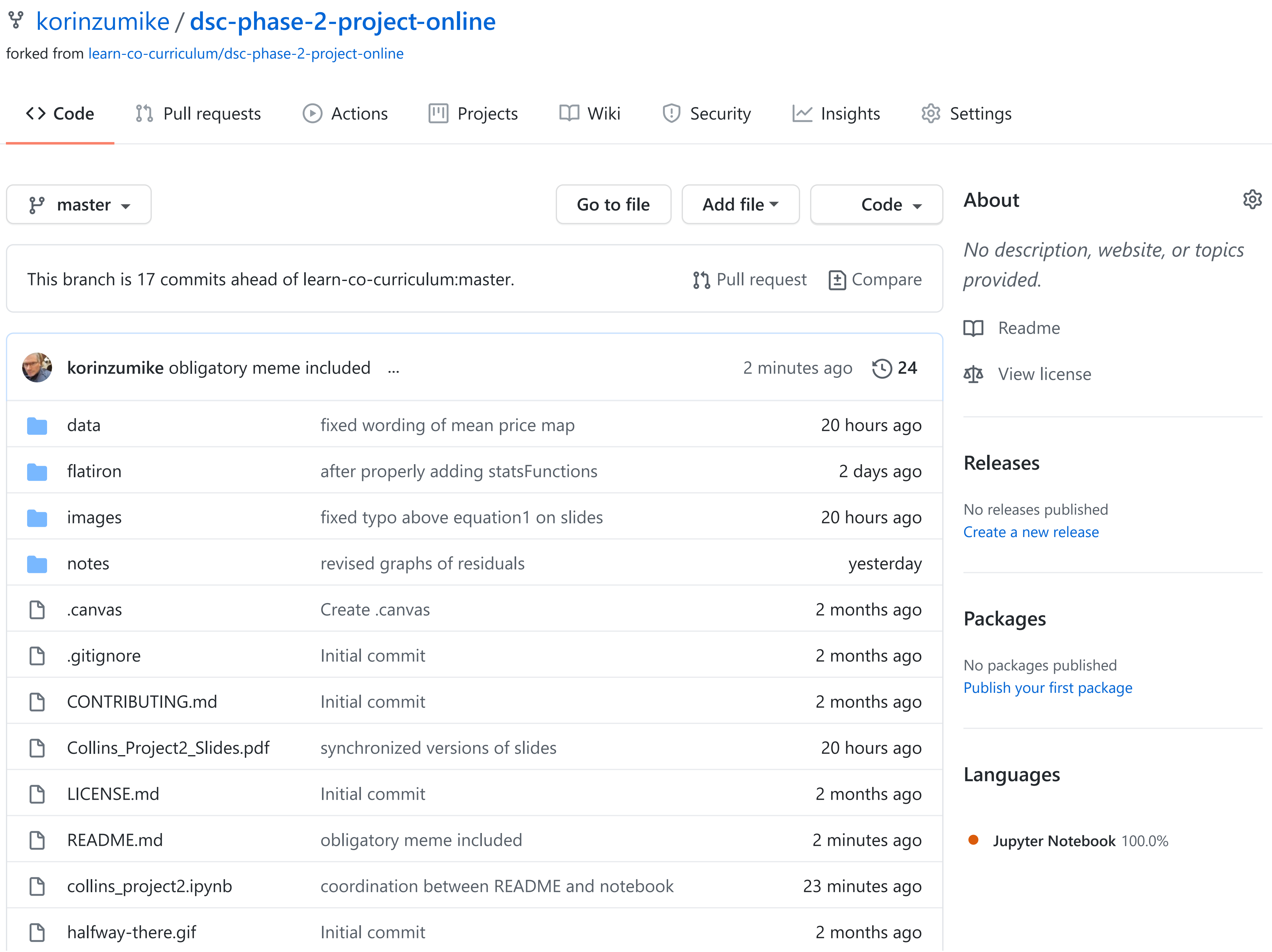

korinzumike / dsc-phase-2-project-online

forked from learn-co-curriculum/dsc-phase-2-project-online

This branch is 17 commits ahead of learn-co-curriculum:master. Pull request Compare

korinzumike obligatory meme included … 2 minutes ago 24

data	fixed wording of mean price map	20 hours ago
flatiron	after properly adding statsFunctions	2 days ago
images	fixed typo above equation1 on slides	20 hours ago
notes	revised graphs of residuals	yesterday
.canvas	Create .canvas	2 months ago
.gitignore	Initial commit	2 months ago
CONTRIBUTING.md	Initial commit	2 months ago
Collins_Project2_Slides.pdf	synchronized versions of slides	20 hours ago
LICENSE.md	Initial commit	2 months ago
README.md	obligatory meme included	2 minutes ago
collins_project2.ipynb	coordination between README and notebook	23 minutes ago
halfway-there.gif	Initial commit	2 months ago

About

No description, website, or topics provided.

Readme

View license

Releases

No releases published

Create a new release

Packages

No packages published

Publish your first package

Languages

● **Jupyter Notebook** 100.0%

halloween.jpg added obligatory halloween meme 8 minutes ago

README.md

Phase 2 Project: “Residential Real Estate Sales in King County, WA”

Michael Collins

Student of Data Science at Flatiron School

October 30, 2020

ONE DOES NOT SIMPLY
CARVE A DETAILED DESIGN IN A
PUMPKIN WITH CRUDE CAVEMAN UTENSILS
imgflip.com

In this Git Repository, you will find the following Jupyter Notebook file:

https://github.com/korinzumike/dsc-phase-2-project-online/blob/master/collins_project2.ipynb

This Jupyter Notebook has several cells, which need to be executed in order. Execution of all the cells can be achieved **by selecting "Run All"** from the "Cell" menu of the Jupyter Notebook interface.

- The first cell is labeled "**Importation of Libraries**". It imports python libraries that are used in subsequent cells.
- The second cell is labeled "**K-Nearest-Neighbor Evaluation**". It contains highly specialized code that I wrote specifically for this project. The code in this cell does an efficient job of listing the event indices for the "K nearest neighbors" of a particular sales event.

The metric used (for judging the proximity between two events P and Q) is the **geodesic distance** between (the location of event P) and (the location of event Q). This is analogous to the “great circle distance” between point P and point Q, except that Earth is assumed to be the shape and size of the **WGS84 Reference Ellipsoid **at sea level, rather than spherical.

When determining the K nearest events to “Event P at Point P”, Event P is EXCLUDED from the list. In other words, an event is NEVER its own neighbor. Furthermore, there are multiple examples (in the provided Dataset) of events that share the same exact (latitude, longitude) coordinates. When finding the K nearest events to “Event P at Point P”, EVERY event whose location is identical to Point P is also EXCLUDED from the list. This practice (of **excluding ALL events at Point P from the nearest K neighbors list** of Point P) prevents “known price(s) of property at Point P” from being included in the calculation of the “empirical price of a property at Point P”.

- The third cell is labeled **“IMPORT DATA”**. The code in this cell looks for a file of “clean” data called “kc_house_data_KNN.csv”, and imports data from that CSV file, if that file exists.

  If the file “kc_house_data_KNN.csv” is not available, the original “kc_house_data.csv” file is imported instead. As the original “kc_house_data.csv” file is being read, raw data from that file is filtered to “clean” the data. Clean data is then stored in a pandas dataframe.

  If the IMPORT DATA cell did not find previously evaluated nearest neighbor lists (a list of the event indices for the K nearest sales events, with one such list per sales event), the K-nearest neighbors lists will be calculated from scratch. For the data that gets tabulated in “kc_house_data_KNN.csv”, a value of K=50 was used in the nearest-neighbor evaluations. The process of identifying the event indices of the 50 nearest sales events (for each of the ~21000 sales events, in turn) required about 45 minutes of wall-clock time on the laptop computer of the author.

The number of nearest neighbors to use in computation of “Local Mean Price”, “Local Mean Number of Bedrooms”, “Local Mean Number of Bathrooms”, etc. was stipulated to be K=15 in this Jupyter Notebook, and the analyses therein. Other values of K were considered. K=15 was chosen, initially, because the provided dataset included columns labeled “sqft_living15” and “sqft_lot15”, which correspond to square footages of the NEAREST 15 properties that were sold in the vicinity of each sale.

A crucial task performed by the “IMPORT DATA” cell is the real-time computation of the local mean value and local median value for VARIOUS QUANTITIES associated with each property sale. These quantities include price, bathrooms, bedrooms, etc. All such quantities are added to the pandas dataframe “df”. A copy of the fully populated dataframe is written as a date-stamped CSV file in the “./data” folder each time the IMPORT DATA cell is executed. This backup copy of all data in dataframe “df” is intended to facilitate detailed scrutiny of any or all derived quantities, should those quantities become subjects of later interest.

- The next several cells (those that precede the “OLS Regression Model” cell) are related to feature selection.
- The “OLS Regression Model” cell produces the results of Ordinary Least Squares multivariate Linear Regression. These results are described in the slide show, linked below.
- Cells that appear AFTER “OLS Regression Model” are related to the production of graphs and maps that are used as visualizations in the slide show, linked below.

In this Git Repository, you will find the following Jupyter Notebook file:

https://github.com/korinzumike/dsc-phase-2-project-online/blob/master/images/Collins_Project2_Slides.pdf

This slide presentation (slide deck) contains 13 color slides, in which the graphs generated by collins_project2_workbook.ipynb are considered in detail.

In this Git Repository, you will find the following folder that contains Images (including graphs and maps) :

https://github.com/korinzumike/dsc-phase-2-project-online/tree/master/images

In this folder are previously-generated versions of the graphs that would otherwise be generated by collins_project1_workbook.ipynb. These graphs are provided as a courtesy to those who are unable to run the Jupyter Notebook and generate the graphs directly.