Final Project Proposal

Group 2
Glenda Decapia
Sarah Kachelmeier
Christopher Sadlo
Katrice Trahan

Github Repo:
https://github.com/csadlo/home-pricing-project

Topic Selected:
Analyzing home values wrt home characteristics, features, and amenities.

Methods Used/Goals:
We will use a Jupyter Notebook to clean our datasets and train our model. Then save the model and load it inside of a Flask app. Two models will be trained. The first will be a multi variable linear regression model. The second will be a Neural Network model.

A website will provide a static report of our findings in cleaning up the data, implementing the two models, and comparing the output and accuracy of the two models. It will also provide an interface for the user to design a fictional house of their choosing and receive an approximate home value.

The Flask app and website will be hosted either on Amazon AWS or Heroku.

Pieces to work on: Blue is group work

1. Load up the github repo for the Seattle, WA
2. Walk through the Jupyter Notebook together as a group
    a. Identify and familiarize with features of the dataset
    b. Train the model. Is it MLR or NN?
    c. Save the trained model.
3. Use the Jupyter Notebook to identify data that needs to be removed and clean it if the original author did not do so already. When doing this, keep in mind how to make the Notebook flexible for other housing datasets other than Seattle, WA. This doesn't need to be in the first draft, just don't make any decisions that would make using different datasets difficult. Is the provided modeling method MLR or NN? Add in the code to train the opposite model as well.
4. One person will modify the website search parameters to select features of homes (Identified from 2a) and send those parameters to the Flask app. Modify the website to switch the theme from hurricanes to housing prices.

5. Modify the Flask app to receive the provided parameters. Take the model saved in **2c** and load it into the Flask app. Work with the person doing Task **3** and add the second trained model from Task **3.**
6. Keep track of findings and regularly update the website's static report portion (req'd). Write the README file. Decide presentation outline/structure.
7. Setup the hosting of the website. (Chris - later)

Proposed Datasets:
https://www.kaggle.com/dansbecker/home-data-for-ml-course
https://www.kaggle.com/moezabid/zillow-all-homes-data
https://www.kaggle.com/tianhwu/brooklynhomes2003to2017
https://www.kaggle.com/prakharrathi25/home-prices-dataset
https://www.kaggle.com/new-york-city/nyc-property-sales

ORIGINAL SOURCES?
https://www.kaggle.com/harlfoxem/housesalesprediction
https://rstudio-pubs-static.s3.amazonaws.com/155304_cc51f448116744069664b35e7762999f.html

https://blue.kingcounty.com/Assessor/eRealProperty/default.aspx

https://www.kaggle.com/neuromusic/avocado-prices

Topic Suggestions:
- ~~Fake News Detection~~
  ~~https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset/tasks?taskId=832~~
- ~~Credit Card Fraud Detection~~
  ~~https://www.kaggle.com/mlg-ulb/creditcardfraud~~
- Home Price Detection
  https://www.kaggle.com/dansbecker/home-data-for-ml-course
  https://github.com/nickvega1989/Predicting-Housing-Prices

Overview:
- Scikit-Learn or ML???
- 2 of the following:
  Pandas, Matplotlib, HTML, Plotly, Leaflet, D3,
  PostgreSQL, SQLite, MongoDB,
  Amazon AWS, Google Cloud SQL, Tableau???

Proposed Layout:

- Reuse hurricane website layout?

Sources:

- Kaggle?
- Github?

Jupyter notebook to clean and train the model, save it
Load model into app.py

Tasks:

- Build the website
- Build machine learning model
- Build app.py
- Host website
- README
- Presentation
- Integration

Timeline:
(Please indicate your availability/unavailability outside class times/days below)

Sun	Mon	Tue	Wed	Thu	Fri	Sat
15	**16**	**17** **List ML proposal topics/links**	**18**	**19** **Finalize topic; Proposal Deadline**	**20** **S: After 5:30 pm** **G:any time**	**21** **Planning, analysis and cleanup** **G:2-4pm, 6pm onwards** **S: after 4 pm** **Takeaways: watch ML class videos, start the website layout**
22 **G:4-7pm**	**23** **Meeting @6** **S: After 5:30 pm** **G: 6pm onwards** **C:9am-6pm** **K: >6pm**	**24** **Train & save initial model** **G:8am-6pm** **C:9am-6pm** **K:>6pm / ping me for other time** **Takeaways:**	**25** **S: After 5:30 pm** **G: Anytime before 6pm** **C:9am-6pm** **K: vary-ping me**	**26** **Thanksgiving Day** **C: 3-8pm (very drunk tho)**	**27** **G:any time except 3-7pm** **C:9am-6pm** **K: vary: ping me** **S: possibly evening**	**28** **(Study Hall) Finalize model** **G:10:30am onwards** **C: 9-3pm**

		improve on the model				
29 G:any time	30 S: After 5:30 pm C:9am-6pm K: >6pm G:any time	1 MLR, correlation, html contents C:9am-6pm G:9am-6pm	2 Flask, input validation S: After 7:00 pm G:5pm onwards C:9am-6pm	3 Host website Model predict G:any time C:9am-6pm K:>6pm S: any	4 Improve website layout README Repo cleanup Submit project G:5-8pm C:9am-6pm S: any	5 Final Project Presentation