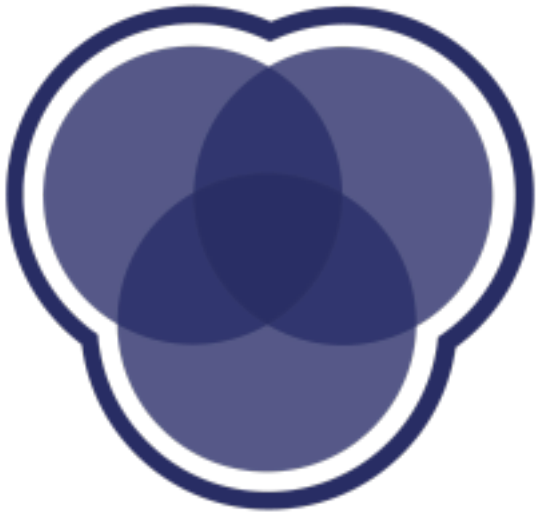


Milestone 3: Regression & Web Scraping
Onboarding Team

Welcome to the third stage of the data science pipeline! We will start modelling with linear regression, and also learn how to scrape data (in case you want to make a dataset).

Notion & Slack Profiles
Please add a photo to your Slack profile, and add a profile to the team directory if you haven't already, or else.

Coffee Chats
Coordinate and do your 2-on-1 coffee chat!

INFO 1998
Attend the INFO 1998 lecture on Monday 7:30-8:20 PM! Importantly, all new members must **attend the Web Scraping workshop this Monday @ 8:20 Gates G01**!

PM Meeting (30 minutes)
Meet with your PM for 30 minutes this week! Come prepared with questions and take notes on any insights you found useful.

Some Resources
- Use `scikit-learn`, `matplotlib`, `beautifulsoup`, and/or `selenium`!

Regression & Web Scraping

We recommend using a jupyter notebook in addition to regular scripts!

1. **Regression**: Create function(s) for each of the following:
    a. [Add 2-3 tests!] Split your dataset into a train, validation and test set. If the dataset is too large, feel free to use a smaller partition.
    b. Choose a continuous numerical column as your target.
    c. Create and save a correlation matrix between features and the target. Consider using a one-hot encoding for categorical features if it makes sense.
    d. Create and save scatterplots of the features with highest correlation.
        i. Bonus/Optional: see if you can apply any transformations that make the relationships more linear.
    e. **Train at least 3 linear regressors** on different inputs based on your findings above.
    f. Save plots of each regressor over the training data, and create a table of training and validation errors. For the regressor with lowest validation loss, report the test error.
    g. What did you learn about your dataset? What did you learn about linear regression?
2. **Preparing a hypothesis**: informed by your data visualizations from MS2 and the linear regression above, come up with at least one question you would like to answer about your dataset for the remainder of the project.
3. **Web scraping**: (optionally, you may choose a different website to scrape data from if it's feasible and related to your dataset.)
    a. Choose a wikipedia article related to your dataset. e.g. If my dataset is about data scientist roles, I might choose data science.
    b. Create a script that
        i. scrapes the content from this wikipedia article and cleans it so it is readable
        ii. scrapes the "See also" links.
        iii. saves both as a text or json file in the repo. You may push this to the git repo if you desire.
        iv. create a test that the wikipedia data is stored in the right directory and that you can extract the text and see-also links.
    c. Create a script that loads and prints the wikipedia page and see also links.

4. **Make sure all tests pass**
    a. Have a command that runs all tests. Make sure all tests pass.
5. **Add necessary packages to the project's virtual environment**.
6. **Add to the README**:
    a. Update the structure of your repo.
        i. Tell viewer what each important folder/script is for.
    b. Add instructions about installing a virtual environment if you haven't.
    c. Add instructions to run tests.

Submission

All of the above should be completed and submitted by **9:30 AM on Saturday, March 15th.** Code should be merged and pushed into `main`, and a report `milestones/ms3.pdf` should also be pushed with the following info.

● What, if any, insights did you get from your PM meeting?

From Jake, we gained insight as to how we can use ML in our actual project. Our original idea was simple arbitrage, but usage of ML can help us calculate an uncertainty boud

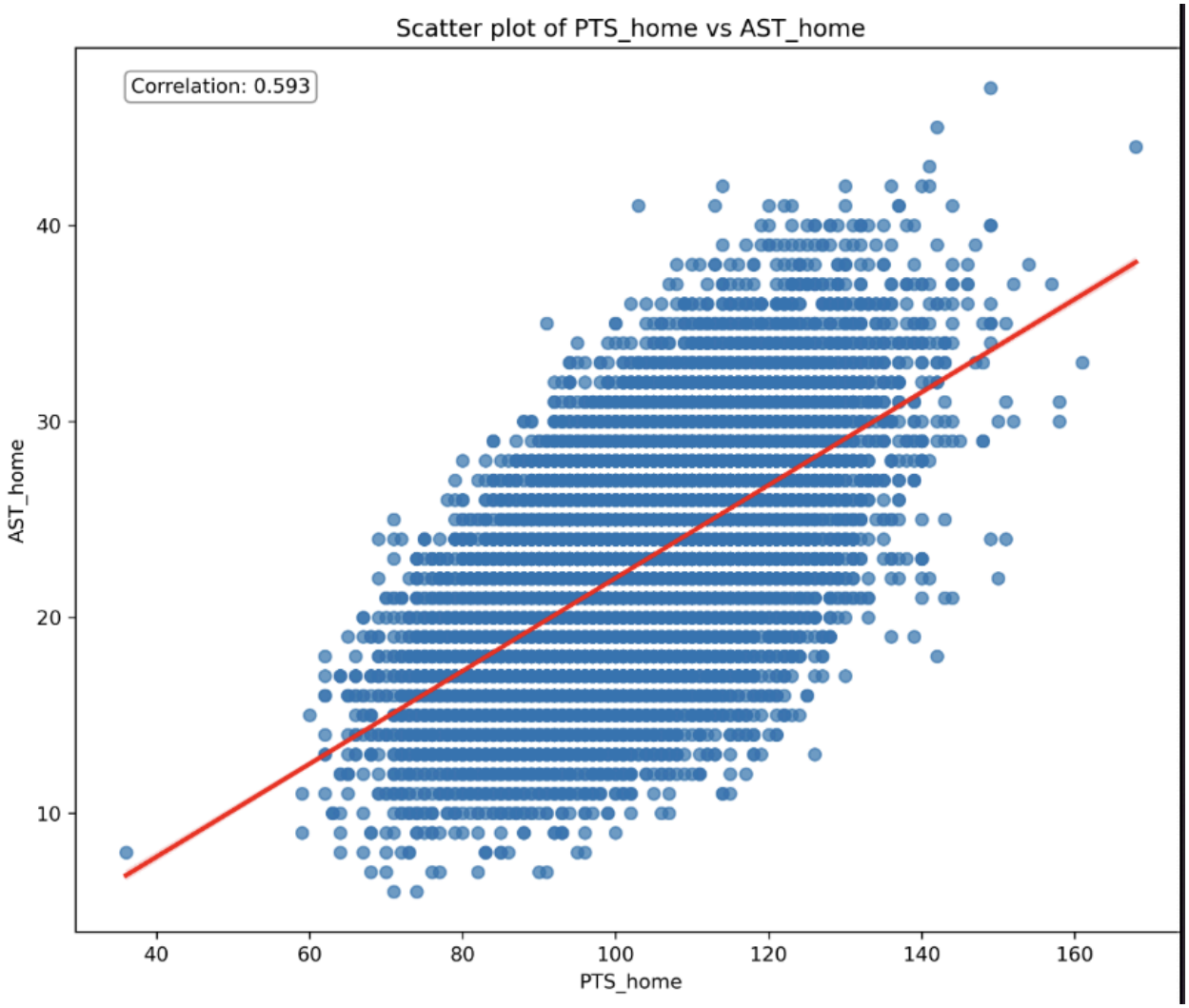
● Include all plots and tables mentioned above. Please neatly format everything.
● Include the following:
    ○ What did you learn from linear regression?
    We learned that there is a statistically significant correlation between home statistics and away statistics.
    ○ What did you learn about your dataset?
    We learned that there is a high amount of variance within the dataset even at each level set.
    ○ What is the question you are interested in exploring? What do you need to do to start answering it?
    Eventually, our goal is to combine this dataset with raw data from the market in order to build a probabilistic model.
● What challenges arose this week?
Using API keys with Kalshi in order to retrieve raw data from the internet.
● What went well?
Using regression on the dataset itself.

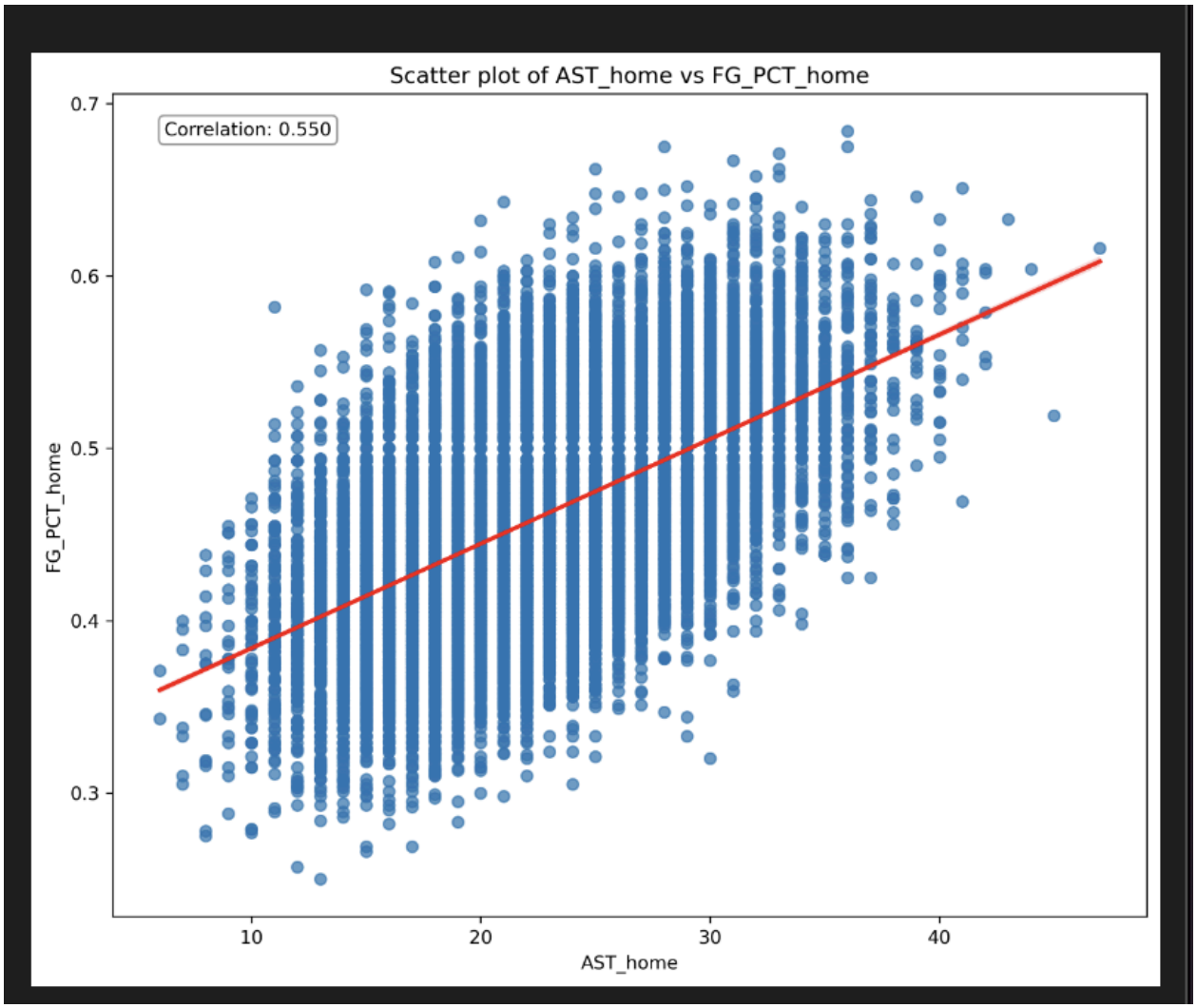
At the end of the report, each team member should type their name (and optionally insert a signature) to indicate that they have read the final version of the report and assent to everything in it, including the teamwork contract.

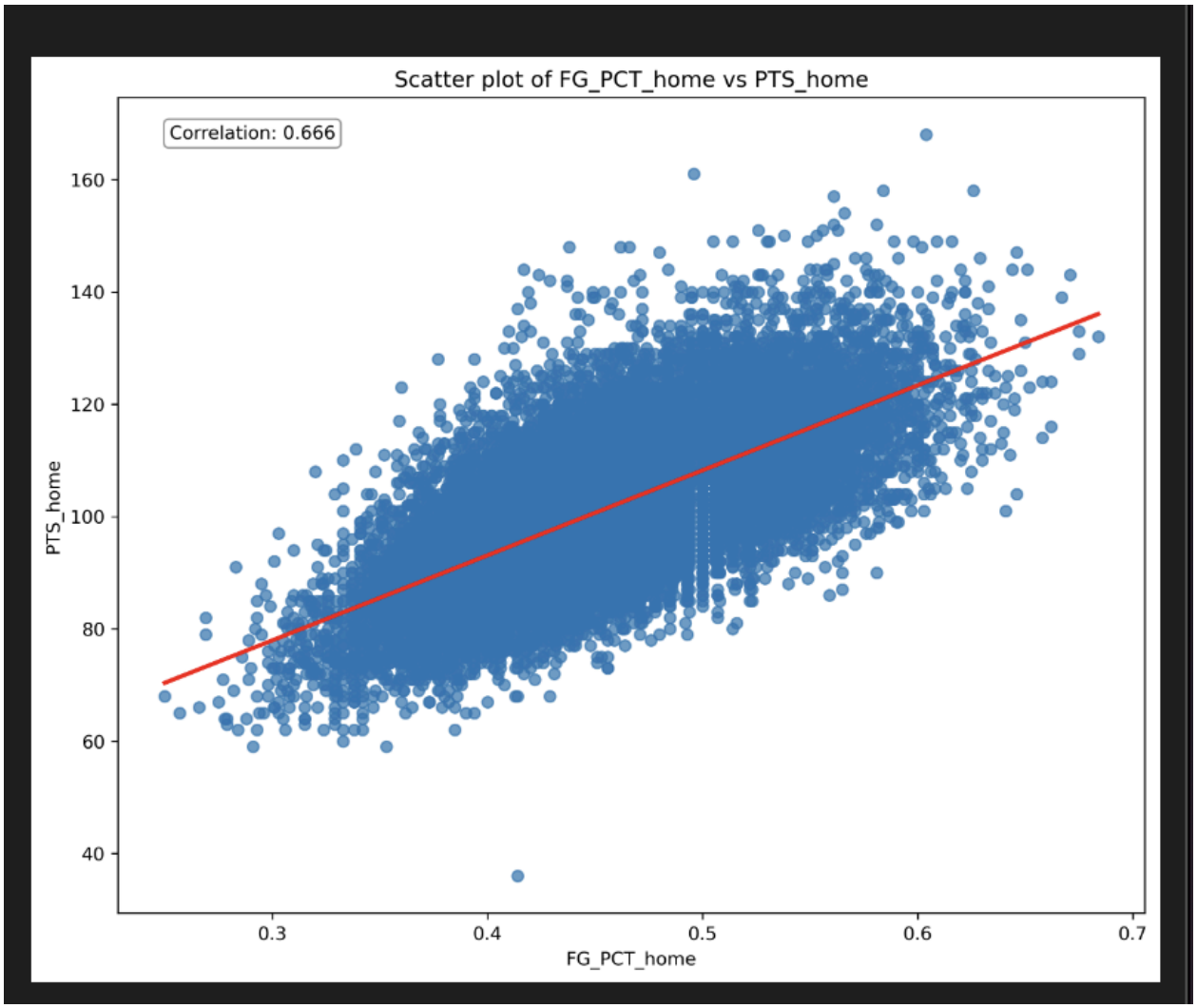
Grading

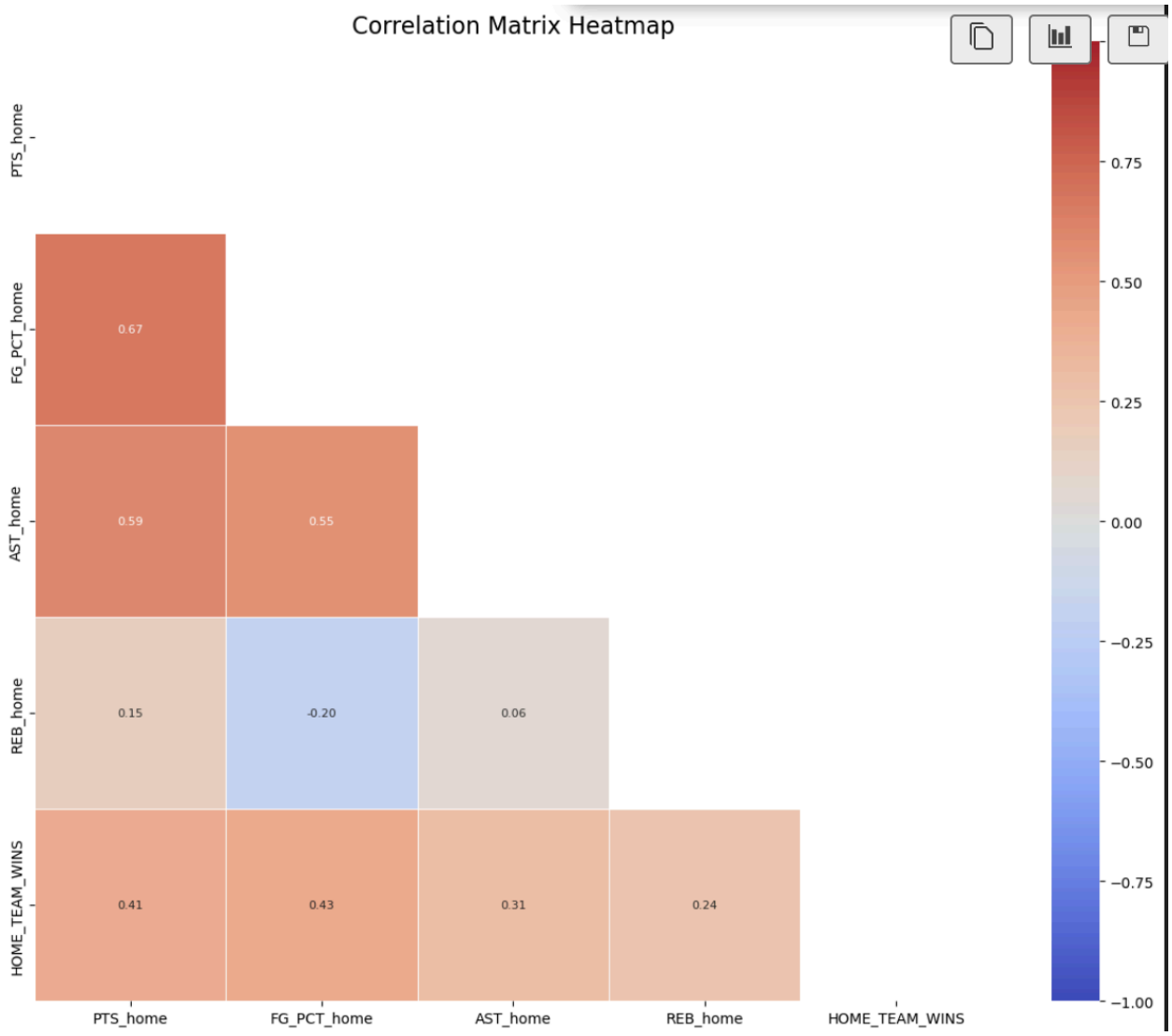
You will be graded on a 5 star scale based on:

- How much of the report and code fulfills the milestone.
- The report is neat and readable.
- The code is neat and readable.
- The Git commits are descriptive.
- The milestone is submitted in a timely manner.




Scatter plot of AST_home vs FG_PCT_home
Correlation: 0.550
FG_PCT_home
AST_home
0.7
0.6
0.5
0.4
0.3
10
20
30
40


Scatter plot of FG_PCT_home vs PTS_home
Correlation: 0.666
160
140
120
100
80
60
40
PTS_home
0.3
0.4
0.5
0.6
0.7
FG_PCT_home



Sucheer Maddury
Ronald Feng
Leo Qian
Aydan Gerber