**COMP 4993/6993 Project**

**Due date: TBA**

# Goal

The goal of this project is to build and evaluate models of home prices. Although you can borrow techniques in the California Homes dataset, the project is more extensive and comprehensive.

You will have to form a team of two persons, formulate the questions and problems you want to address, collect data, build different models & evaluate them in various ways, and present your work to the instructor and your classmates.

There is much room for creativity in this project. You are not restricted to any method, although linear regression should be a core component of your project. You are not restricted to any strategy or heuristic. You are free to employ any strategy, heuristic, trick that will result in accurate modeling and prediction.

# Team Formation

You will have a week to form a team of two persons. After a week since the project starts, you will have one chance of forming a new team. So, form your team and choose your teammate carefully.

Both team members will work on all aspects of the project. There are two scenarios in terms of collaboration between the team members. First, a task requires lots of effort (e.g. data collection), it can be split between the two members. Second, a task requires testing (e.g. implementing a specific procedure), team members can alternate between code creation, code testing, etc. There might other scenarios, but both team members must be involved in all aspects of the project.

# Problem Description

The main goal is to build the best models you can possibly build. These models should as accurately as possible relate the home prices to other attributes of a home. These models should be as accurately as possible predict recent home sales. It is likely the case that you will need more than one models.

You should first start with formulating the problems your team will solve. You might want to answer very specific questions such as:

* Do I want model(s) that works anywhere in the USA? Or do I want model(s) that work in specific areas (cities, zip codes, etc.)?
* What information or methods do I need to know to accomplish this project?
* How do I collect data?
* What hypotheses do I have about good models? (These hypotheses might be mutually conflicting). How do I select features to build these models?
* What strategies that are helpful in terms of preprocessing, postprocessing in making my predictions more accurate?
* How do I evaluate my models and strategies? What do I compare my predictions to?

The first part of your report should be dedicated to answering these questions.

# Data collection

Start with the Ames Housing dataset.

For extra credit, you can additionally collect real and recent data, e.g. home sold in various areas of Memphis. In this case, you must identify the sources, the method of collection, and tasks that preprocess your data. For example, you might need to figure out how to translate addresses to longitudinal and latitudinal coordinates.

All of these must be described in your report.

# Modeling

You should study different ways of modeling the relationship between home prices and other attributes.

Starting from the basic linear regression, you can try other methods in scikit-learn.

You can also incorporate heuristics that you think can help your modeling and/or prediction.

You must describe the modeling strategies, methods and heuristics in detail, concisely and precisely.

# Evaluation

You must evaluate your models and strategies. You must describe how you evaluate your models and predictive power of your models. The more meaningful and comprehensive you evaluate your methods, the more valuable your product is.

# Grading

Your project will be graded based on:

1. The weekly effort on your project as reported in your weekly project update.
2. How much you help other teams on technical issues related to the project.
3. The clarity of your report and presentation.
4. The thoughtfulness of your work.
5. The comprehensiveness of proving your models. You can have multiple models.
6. How much work you spend on creating the components of your project. For example, you can have a very impressive project, but 90% of it is done by other sources (which you must explicitly cite), then you will get only 10% of the grade.
7. How your models help answering realistic questions. For example, if someone needs to spend $200,000 to buy a home, what are your recommendations?

If you registered for COMP 6993, you are graded based on a higher standard. You should be proactive in leading your team and contribute more work to finish the tasks.

# Build an interface for your model

Your model is useful only if people can use it. There are various ways to make your model easy to use. I recommend you to use Plotly Dash to build the interface to your app. Each team member should be involved in designing/implementing the app.

# Academic Honesty

If you plagiarize, you will be reported to the department Chair and the College.

If you use material that you do not create, you must cite it. This means your report should have a list of references.

If you share material with other teams, you must report it. If you do not and the other team does not report it, you are potentially responsible for a dishonest act.

# Project Updates, Report and Presentation

On Wednesday each week, we expect you to send in a project update, which is essentially an on-going draft of your project report. You should date and comment on what you’ve done in the past week on the project.

You will turn in one report that describes all the components of your work, including:

* Team member contribution,
* Problem formulation and scope of your project,
* Features of your dataset (using Pandas),
* How you collected and processed data (in case you collect your own data)
* Model building, strategies, and heuristics,
* Evaluation of your models,
* Conclusion which reflects on the process and the quality of your work.

Name your report COMP4993\_Project\_Lastname1\_Lastname2

You will have 15 minutes to present your project to the class.