﻿This course aims to teach students data science concepts and machine learning methods for solving problems and gaining insights in real-world settings. In order to apply the concepts to practice, the students will be working on the following midterm project as one large component of this class. Students will work in groups of two to four students to practice completing the project.

This home dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015. The dataset consists of following variables

**id** - Unique ID for each home sold   
  
**date** - Date of the home sale   
  
**price** - Price of each home sold   
  
**bedrooms** - Number of bedrooms   
  
**bathrooms** - Number of bathrooms, where .5 accounts for a room with a toilet but no shower   
  
**sqft\_living** - Square footage of the apartments interior living space   
  
**sqft\_lot** - Square footage of the land space   
  
**floors** - Number of floors   
  
**waterfront** - A dummy variable for whether the apartment was overlooking the waterfront or not   
  
**view** - An index from 0 to 4 of how good the view of the property was   
  
**condition** - An index from 1 to 5 on the condition of the apartment,   
  
**grade** - An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level of construction and design.   
  
**sqft\_above** - The square footage of the interior housing space that is above ground level   
  
**sqft\_basement** - The square footage of the interior housing space that is below ground level   
  
**yr\_built** - The year the house was initially built   
  
**yr\_renovated** - The year of the house’s last renovation   
  
**zipcode** - What zipcode area the house is in   
  
**lat** - Lattitude   
  
**long** - Longitude   
  
**sqft\_living15** - The square footage of interior housing living space for the nearest 15 neighbors   
  
**sqft\_lot15** - The square footage of the land lots of the nearest 15 neighbors

The goal is to predict the sales price for each house based on the given features. For each Id in the test set, you must predict the value of the price variable and submit your results on Canvas by 7PM on Friday 11/04. You need to submit a csv file with two columns: ID in the test set and the predicted sales price.

**Accuracy Metric:** RMSE between the logarithm of the predicted value and the logarithm of the observed sales price will be used to evaluate the prediction accuracy. (Taking logs means that errors in predicting expensive houses and cheap houses will affect the result equally.)

Each team will present their findings and results in the class on Wednesday 11/03. Please make sure to cover the following topics in your presentation:

* A summary statistic of the data
* Preprocessing and data cleaning (if needed)
* Feature transformation (e.g., log transformation) (if needed)
* Encoding categorical features and in particular features with high cardinality
* Feature Selection and Model Selection
* Best model and best features
* Results, findings, and Learnings

Please submit your code and presentation on Canvas by 11/04.