1. **Project 1: Supervised learning- regression problem**

Kaggle Dataset: <https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting/data>

consists of historical sales data of 45 Walmart stores located in different regions across the United States, it’s a good resource for data processing projects. Each store contains a number of departments. This dataset can be used to predict the department-wide sales for each store.

* Besides, Walmart runs several promotional markdown events throughout the year, mostly before four different holidays in a year (Super Bowl, Labor Day, Thanksgiving, and Christmas). The weeks including these holidays are weighted five times higher in the evaluation than non-holiday weeks. So, we are modeling to predict the effects of the markdowns on these holiday weeks in the absence of complete/ideal historical data.

**Data files:**

1. stores.csv: contains type and size of 45 stores
2. train.csv: historical data from 2010-02-05 to 2012-11-01; contains different features are store, department number, Date - the week, Weekly Sales - sales for the given department in the given store, IsHoliday - whether the week is a special holiday week or not
3. test.csv: contains all features as train.csv except weekly sales.

So, the task is to predict the sales for each triplet of store, department, and date in this file.

1. features.csv: contains additional data related to the store, department, and regional activity for the given dates and fields are store number, date - the week, temperature - average temperature in the region, fuel\_price - cost of fuel in the region, MarkDown1-5 - anonymized data related to promotional markdowns that Walmart is running. MarkDown data is only available after Nov 2011, and is not available for all stores all the time. Any missing value is marked with an NA. CPI - the consumer price index, Unemployment - the unemployment rate, IsHoliday - whether the week is a special holiday week

four holidays fall within the following weeks in the dataset (not all holidays are in the data):

* + Super Bowl: 12-Feb-10, 11-Feb-11, 10-Feb-12, 8-Feb-13
  + Labor Day: 10-Sep-10, 9-Sep-11, 7-Sep-12, 6-Sep-13
  + Thanksgiving: 26-Nov-10, 25-Nov-11, 23-Nov-12, 29-Nov-13
  + Christmas: 31-Dec-10, 30-Dec-11, 28-Dec-12, 27-Dec-13

1. **Project-2: Supervised Learning binary classification problem**

**Kaggle dataset:** [**https://www.kaggle.com/c/titanic**](https://www.kaggle.com/c/titanic)

On April 15, 1912, during her maiden voyage, the widely considered “unsinkable” RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren’t enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew. With luck, it seems some groups of people were more likely to survive than others. So, goal is to build a predictive model that answers the question: “what sorts of people were more likely to survive?” using passenger data (ie name, age, gender, socio-economic class, etc). The data has been split into two groups:

Data files

1. training set (train.csv), contains features PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare,Cabin, Embarked
2. test set (test.csv), contains all features as training set except Survived.
3. gender\_information.csv: features are passenger\_id, survived (0 – not survived, 1- survived).

Use training set to build machine learning models, provided the outcome (also known as the “ground truth”) for each passenger. Model will be based on “features” like passengers’ gender and class. The test set should be used to see how well your model performs on unseen data. For the test set, ground truth for each passenger is not provided. So, predict these outcomes. For each passenger in the test set, use the model trained to predict whether or not they survived the sinking of the Titanic.

1. **Project 3: Supervised Machine Learning- Regression problem**

**Kaggle dataset:** [**https://www.kaggle.com/c/house-prices-advanced-regression-techniques**](https://www.kaggle.com/c/house-prices-advanced-regression-techniques)

Goal is to predict price of a particular house using information like house area, number of bedrooms, number of bathrooms, and utilities

1. train.csv - the training set: contains 81 columns
2. test.csv - the test set: contains all columns as train.csv except SalePrice column.
3. data\_description.txt - full description of each column listed in train.csv, test.csv
4. sample\_submission.csv - a benchmark submission from a linear regression on year and month of sale, lot square footage, and number of bedrooms