CS 185C

Advanced Practical Computing Topics

Project Proposal

San Jose State University

1 Washington Sq, San Jose, CA 95192

*Project Members*

Malik Khalil

Anahit Sarao

Ahmed Ahmed

The project will consist of three members: Anahit Sarao, Malik Khalil, Ahmed Ahmed.

Anahit will handle the normalization of the data set listed below, and produce a working environment where a Kafka Broker will send a singular line of the csv file to the Kafka Consumer also known as the client. The data will then analyzed by Spark under the Hadoop ecosystem. Anahit will also provide scripted tools with the help of Ahmed. Malik will be responsible for the main algorithms used to analyze the data set, this will involve using the the Hadoop and Spark APIs. Ahmed will handle the needed transformations and visual analysis of the data. The whole team is committed to help in all parts, for example all group members would like work on the Kafka Producer and Consumer, this gives the opportunity for each member to experience everything related to the project while still being focused on their core tasks.

The data set which the group has chosen is in comma separated format. It provides a series of columns about house sale prices for King County, Seattle area. The data set provides a timeline of May 2014 to May 2015.

Link to dataset:

<https://www.kaggle.com/harlfoxem/housesalesprediction>

The team wants to be able to predict the fair market rate of a home going up for sale in the area based off the data provided. This will contain home values in all districts, the number of bedrooms, bathrooms and all any additional home information. The main algorithm used will be a Linear Regression. Modeling will be done using a two dimensional analysis, which will result in set of predicted numbers. The analysis will also generate margin of error for the data set. Examples of valued results are Predictor, Constant Coefficient, and Height.

In conclusion this project will utilize methods learned in class and be a valuable experience for all members. If additional time is permitted a more in depth analysis will be done and data will represented using advanced metrics.