Comp9321 Project plan

Property Price Predictor
Team **Simba**

Members:

Michelle Leman z5076202
Ethan Xu z5108944
You Zhou z515002
Qingchen Zhong z5119641
Sean Nguyen z5055824

Motivation:

The motivation behind this project is simple - it's a realistic product that can reach a large number of users. When planning our project we had other ideas such as a movie gross prediction service for movie investors but realised two things. The first being that the connection within the dataset between the columns we chose wasn't strong enough and produced a fairly low R^2(0.6). And the second being that the product itself really wasn't that useful.

The topic we ended up choosing was interesting to us personally as it gave us an insight into the Australian real estate market and could be beneficial to us in the long run when we start searching for properties to purchase.

Description:

Our website will be simple in design - the main page asking users to select from a drop down list the suburb, number of bedrooms, bathrooms, car spaces and the type of house they're after. We have chosen drop down lists to control the input to reduce the risk of incorrect input. The result page will state the predicted price of buying a home based on those requirements and things in the suburb they've chosen. This includes restaurants, supermarkets, schools and hospitals. We then list some other suggested suburbs which are in a $20,000 price bracket ($20K more, $20K less)

What is the aim of the service?
- Predicts a price for people looking to buy a home with specifications based on properties with similar features in their area

What are the datasets?
- A dataset containing Melbourne real estate property prices in 2016 off Kaggle with the address information (suburb, address, district etc) and house information (number of bedrooms, bathrooms, car spaces etc)
- https://www.kaggle.com/anthonypino/melbourne-housing-market#Melbourne_housing_FULL.csv
- We focused on the following columns: "Bedroom2", "Bathroom", "Car", "Type", "Suburb", "Price", "Longitude" and "Latitude" for the prediction and suggestions. We chose these columns in particular as when we factored in other columns such as "Landsize", we found that they didn't greatly impact the result. We also tried using the column "YearBuilt" as we realised this impacted the price - older houses with the same amount of bedrooms/bathrooms/car spaces compared to newer houses sold at a cheaper price. Although, many of the rows in this column were N/A and were too unpredictable to use.

Team Management

Our team has chosen to use Facebook Messenger to communicate and to use Github for our repository (https://github.com/seannguyn/9321_asg3). Our tech stack includes Python3, Scikit-learn, Flask, React and mLab. We have given each member a role in the project which is as follows:
- REST API & Machine Learning: John, Sean, You
- Doc, Design Planning & Frontend: Michelle
- REST API Doc & Frontend: Ethan

Mock-Ups:
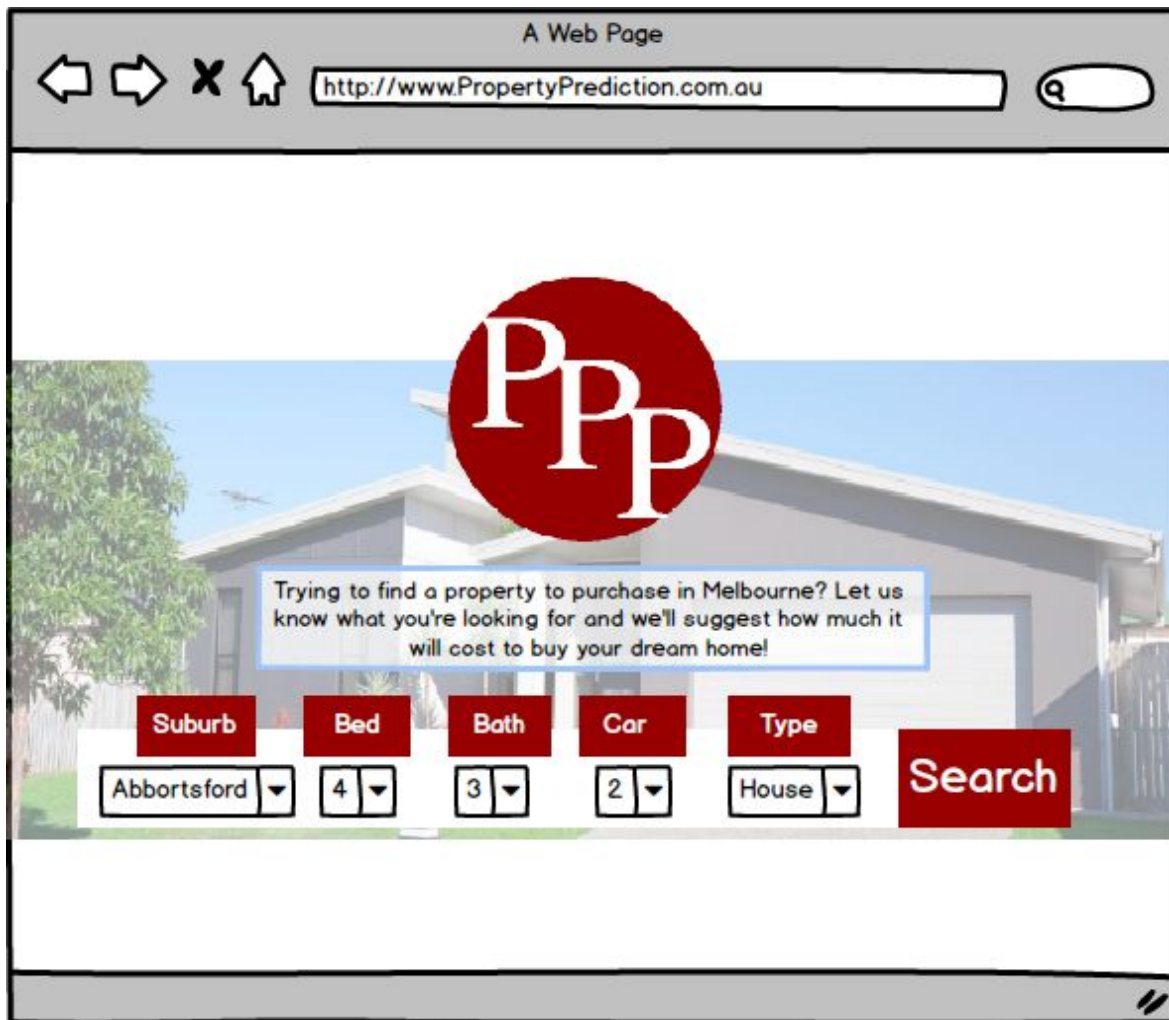
The homepage was designed to be clean and simple. We wanted to make it very clear what the service was offering and wanted to minimise the risk of invalid input so we chose to use drop down lists. Each of the input fields are the columns we have chosen to use from the database for the calculation of our prediction.

http://www.PropertyPrediction.com.au

PPP

Trying to find a property to purchase in Melbourne? Let us know what you're looking for and we'll suggest how much it will cost to buy your dream home!

Price Prediction

$750,000
For a 4 bed, 3 bath, 2 car house in Abbortsford

Abbortsford

St Therese Catholic School

The Cookhouse

Coles

Del Punto

Prince Of Wales

Eastlakes Primary School

Moroccan Feast

Aldi

Nearby Restaurants

The Cookhouse
141 Belmore Rd
★★★

Del Punto
40 Pauls Street
★★★★

Moroccan Feast
12 Avoca Street
★★★

Nearby Schools

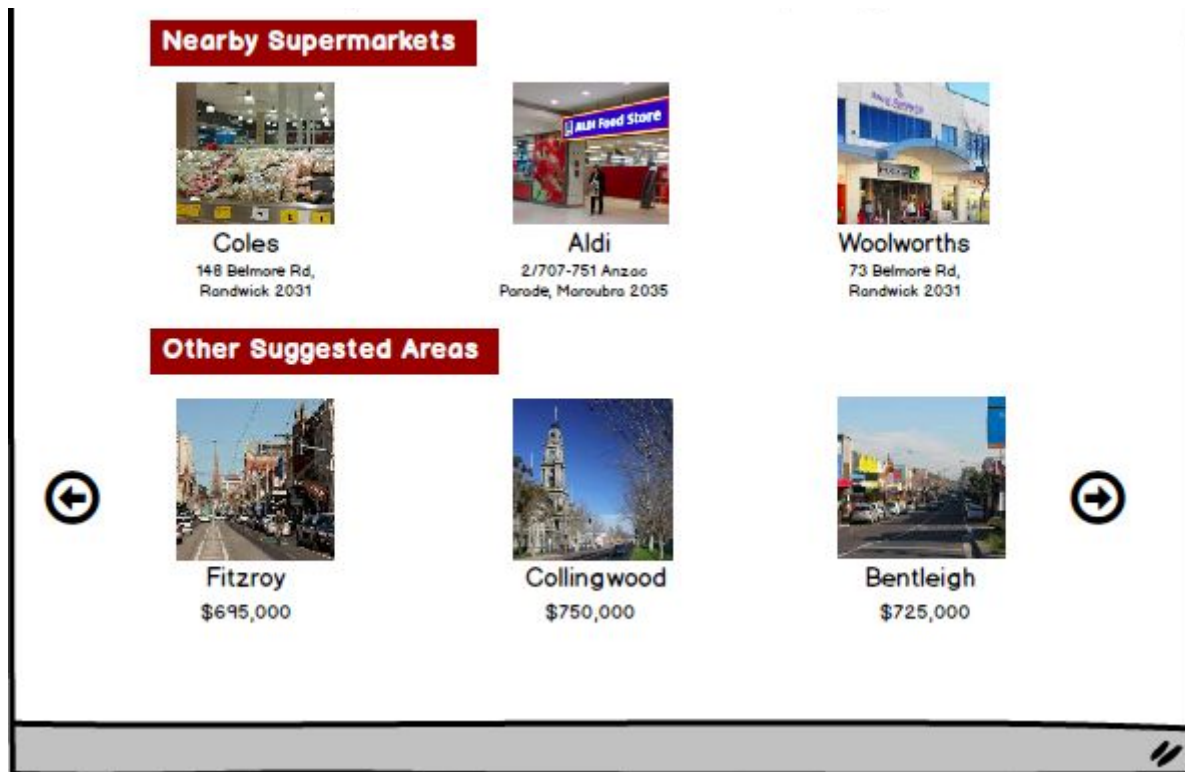|		
St Therese Catholic Primary School	2.8/5	43 Sutherland St, Mascot 2020
Eastlakes Public School	4.5/5	7 Florence Ave, Eastlakes 2018

Nearby Hospitals

|	
Prince of Wales	320-346 Barker St, Randwick NSW 2031
St Vincent's Hospital	390 Victoria St, Darlinghurst NSW 2010

Our results page is shown above. We wanted to go beyond just providing the prediction of the property's price. We decided to use the Google Maps API to display the suburb and all of the relevant information we provide. We then display useful things we thought homeowners might like to know - the best local restaurants, closest schools, hospitals and supermarkets. We then have a "Suggested Areas" section which provides the areas in which are in a similar price bracket (give or take $20,000)

System Workflow
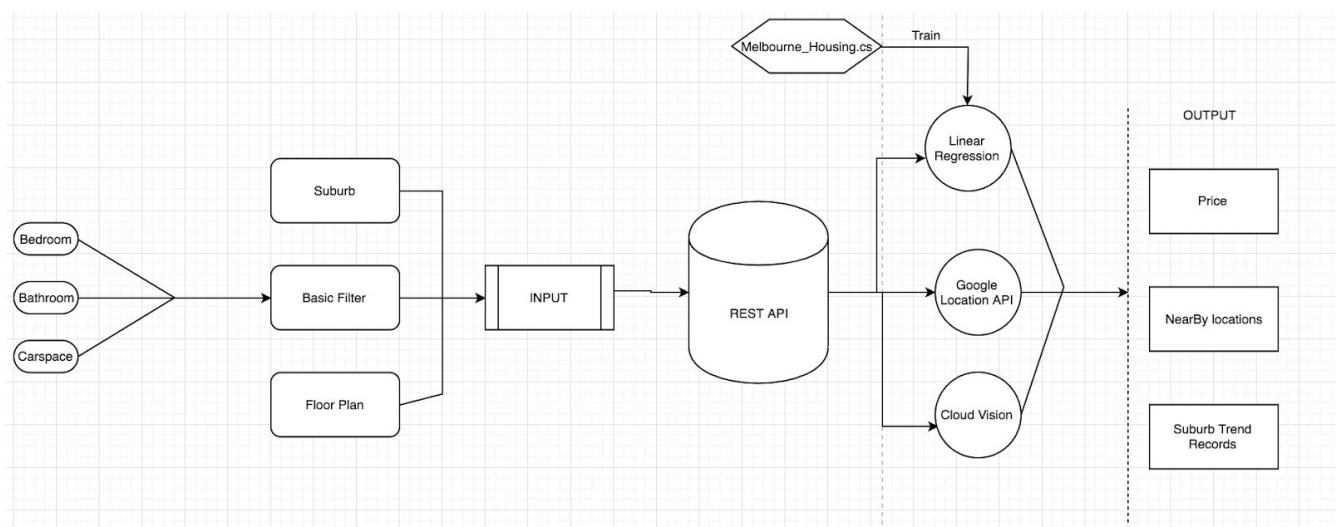
Our plan

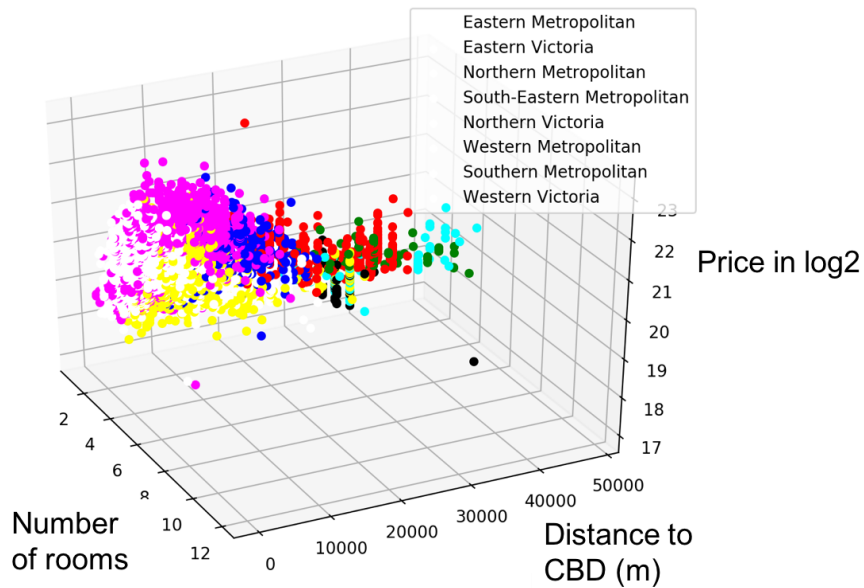| Category	Task	Staff related	Hour needed	Status
Plan/Analysis	Plot the data frame and choose a suitable algorithm for price prediction and make recommendations for suburbs.	John, Ethan,	3	Finished
Backend/mLab	Sort the dataset to mlab and allow users to retrieve them	John	2	Finished
Backend/Machine learning	Predict price for a property with a certain number of bedrooms, bathrooms, and car space specified by the user.	John	3	Finished
Backend/Machine learning	Recommend some suburbs where users are able to purchase a certain type of property within a specified price range.	You Zhou	5	Finished
Backend/Restful API	Build a Restful API for returning predictions of the price of properties and some recommendations.	Sean	6	Finished
Frontend	Design and build up the webpage which allows users to specify what kind of property they are looking for.	Michelle, Ethan	5	Finished
Frontend	Design and build up the webpage to show the returned prediction of price and relevant recommendations.	Michelle, Ethan	5	Finished
Frontend	Display the suggested suburb at the google map.	Sean	3	Finished
Frontend	Display a pie chart for the distribution of sales in all suburbs.	Sean	3	Finished

Why we have chosen linear regression to do the analysis:

The difference between the real price and predicted price
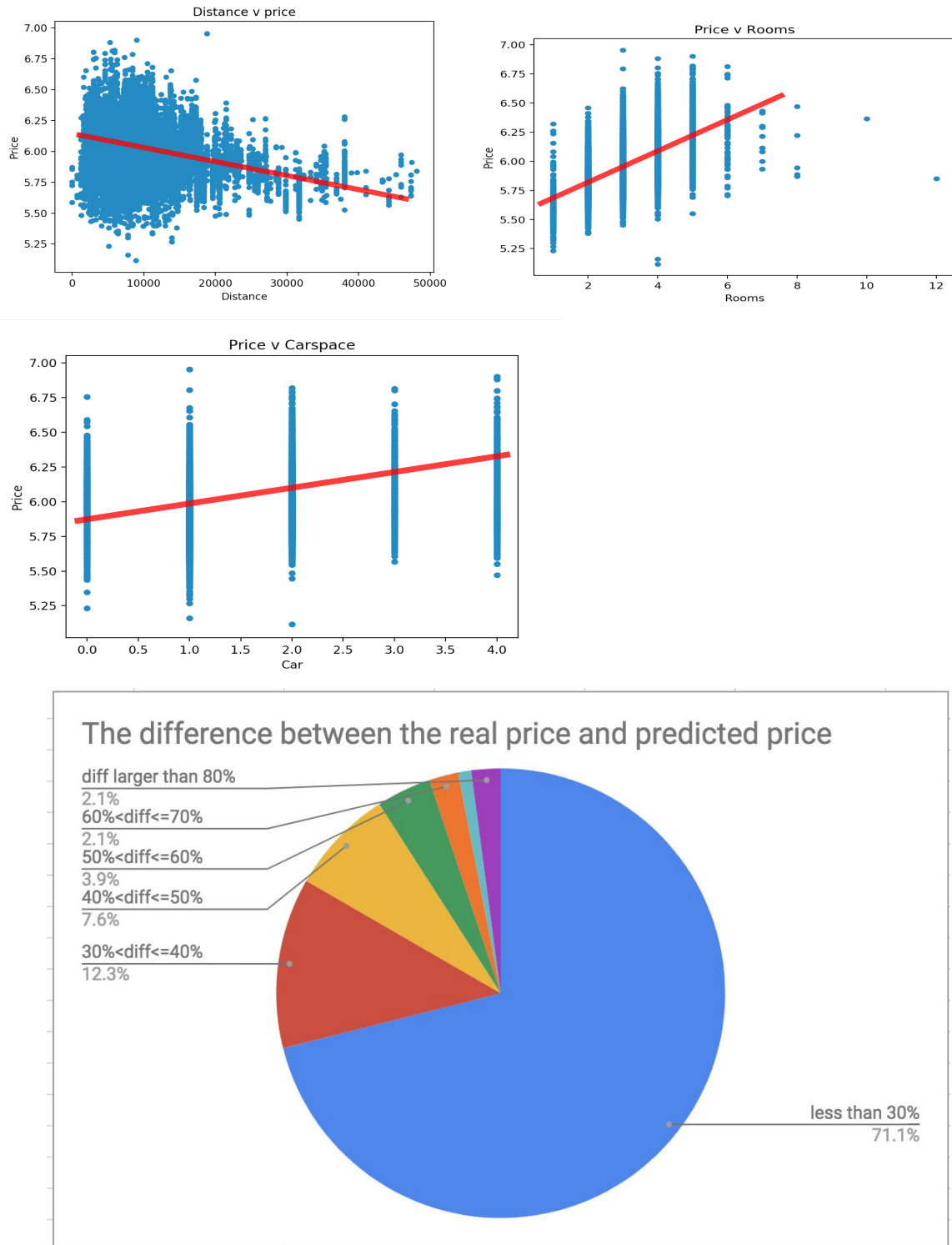
After plotting the data, we discovered that the distance between the property and the CBD, the number of car spaces and the number of rooms of a property have clear linear relations with the price. By using the linear regression model suggested above to predict the price, it produces a result with 71% of the 2561 samples in the test-set having a difference less than 30% between the predicted price and the real price of the test.

Why we did not use a simple KNN to return the prediction of suburbs?

As shown below, there is a strong overlap in the dataset and was not as useful in our prediction