Data Incubator Project Proposal

Students graduate all the time, they get new jobs and the become more financially stable as they move forward with their career. Once young adults feel they are financially secure, they start hitting the market searching for houses to buy. In the process they scan the market, and get in contact with real estate agents to help them in the process.

There is a lot of publicly available data related to housing that can help a person with their decision. The problem with the data is that they are represented in a way that does not make sense to most people who are not in the data science field. Ordinary and busy buyers need something informative and maybe visual to assist in making a decision of where to buy a house, and what implications are made according to the location and size of the house.

I am looking forward to analyze the housing data and build a useful, simple and informative application people can rely on. The first step is to gather data, analyze it and look at different perspectives and how different attributes are related to each other. Discovering the relationships between attributes can assist in building predictive models that can help buyers with make a smarter decision of where to put their money to gain the best value. The mentioned relationships can be extracted by developing some mathematical models to help people address their financial concerns and help them predict where the ideal house can be located, its price, mortgage and installments predictions. The project might be a little too much for a six weeks' fellowship, but it definitely will be something that I personally will benefit from. So for this proposal I'm just generating a couple of graphs to see if things are correlated in some sense or not.

For the data set I used the 2013 housing data that can be found at the below link: https://www.huduser.gov/portal/datasets/hads/hads.html. The attributes in the data set are outlined with brief explanation of their meaning on the last page of a document that can be found here: https://www.huduser.gov/portal/datasets/hads/HADS_doc.pdf. I did not use the whole dataset to generate my figures, I extracted some columns that I saw interesting for the purpose I am addressing.

I will be showing two figures:

Figure 1, which the link for it is provided in the application shows the relationship between the cost of utility and the number of rooms in a house. It is seen that the data point exhibits a normal distribution. This can be explained in numerous ways, one of which that the utility cost is different from one region to another. Therefore, the cost of the utility cost in some regions is so low that leads to houses with large number of rooms to yield a utility cost that is not as high as houses with fewer number of rooms but located in locations in which the utility costs more.

Figure 2, shows the relationship between the market rent value of a house and the adjusted income with respect to number of rooms. The figure shows that there is a, increasing linear relationship between both of them. As the adjusted income increases the rent value of the house increase. To support this I used linear regression concept to find the line that represents the data. f