**Analysis on IRS data and relations between AGI and Mortgages**

**Project 12 – SECTION 2**

**Sai Bharadwaja    Uppala  
Sundeep    Vuyyuru  
Venkateswara R    Kotha  
Srikanth** **Budi (from section 03)  
Tharun** **Madagouni** **(from section 30)  
Kesava Paruchuri (from section 29)**

**Project Documentation.**

After looking at the IRS data given we thought there would be some relation to mortgages paid and the location they are from.

For example, in states like California where housing is costly people would have to pay higher mortgages when compared to other

states like Arizona. We tried to extract this information from the data.

**Data Source**

The IRS returns data was used for this project.

**Cleaning & modelling Data**

* IRS data contained lot of data. We were interested in number of returns, state, agi, mortgages.
* Pandas handled null's and other issues in data.
* Data was remodeled to and a new column was added to data frame called ppm percapita mortgage interest payed.
* ppm = total mortgage paid/total returns
* Clustering
* The data was clustered into 3 clusters using k-means clustering algorithm
* 3 clusters were selected because the graphs of data showed a roughly into 3 groups

**Data Representation**

* Matplot lob was used to represent data.
* x- axis is Agi
* y axis is ppm
* Each cluster has a colour
* and each point in the graph is a state.

**Conclusion**

As mentioned California is in a cluster which has more ppm Arizona is in

the middle and North Dakota is part of the cluster which has least ppm