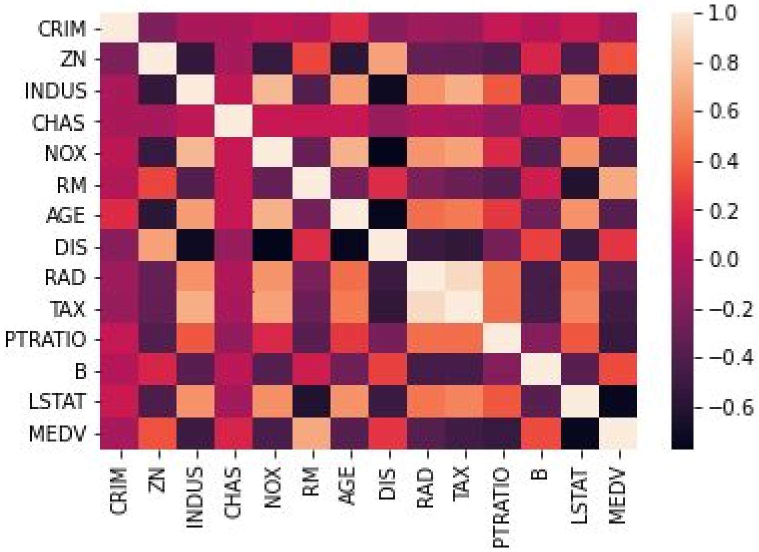
AIHack 2021 Report

MathBois

In order to get a good idea of the data and how the different variables interacted with each other we started by making a heatmap showing correlation between the variables:

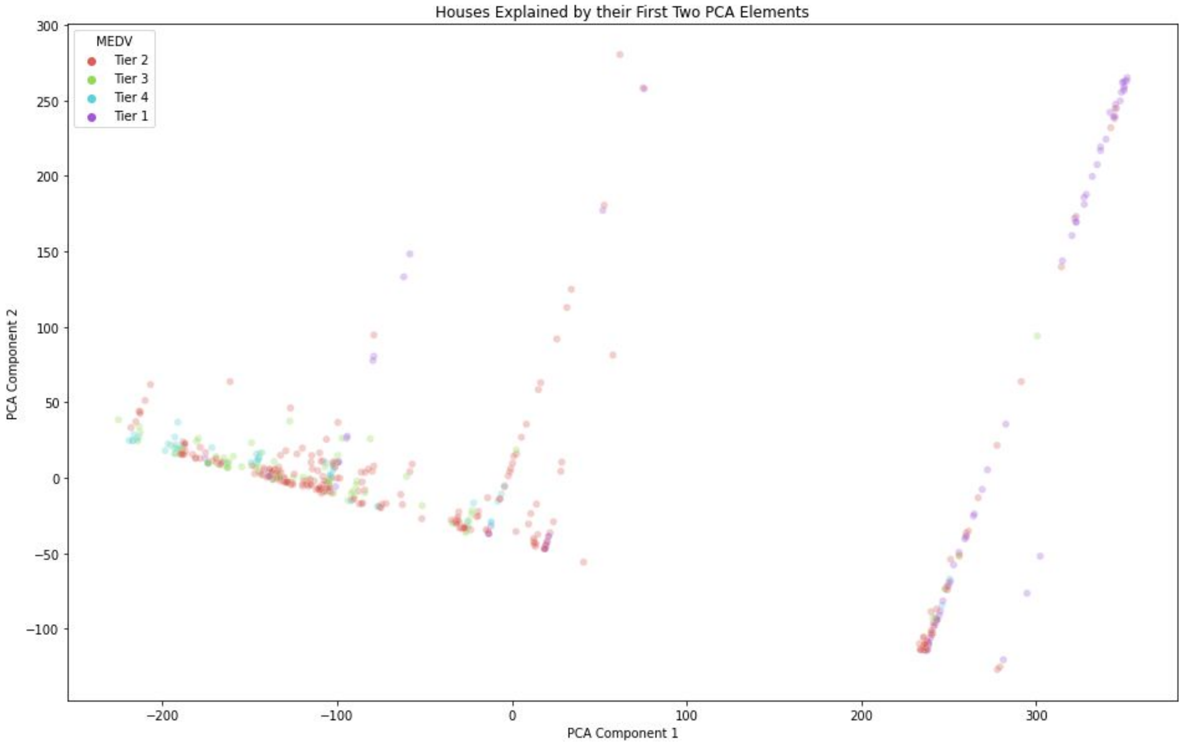


As you can see the only pair of variables with a strong positive correlation were RAD and TAX. Which means that there is a strong positive correlation between the accessibility of radial highways and the tax rate.

This graph also shows that MEDV (the median value of houses) strongly correlates negatively (i.e is negatively proportional) with the following variables:

- INDUS (number of non-retail business acres there are)
- NOX (nitric oxide concentration)
- AGE (proportion of buildings built before 1940)
- RAD (accessibility of radial highways)
- TAX (property tax rate)
- PTRATIO (ratio of pupils to teachers)
- LSTAT (percentage values of lower status population (whatever that means))

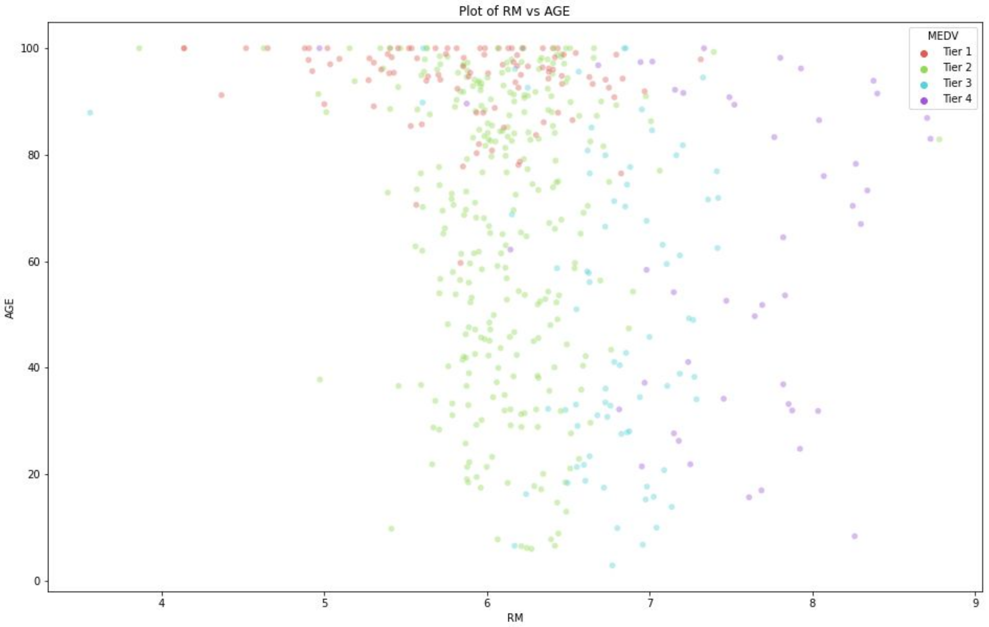
PCA Graph Analysis



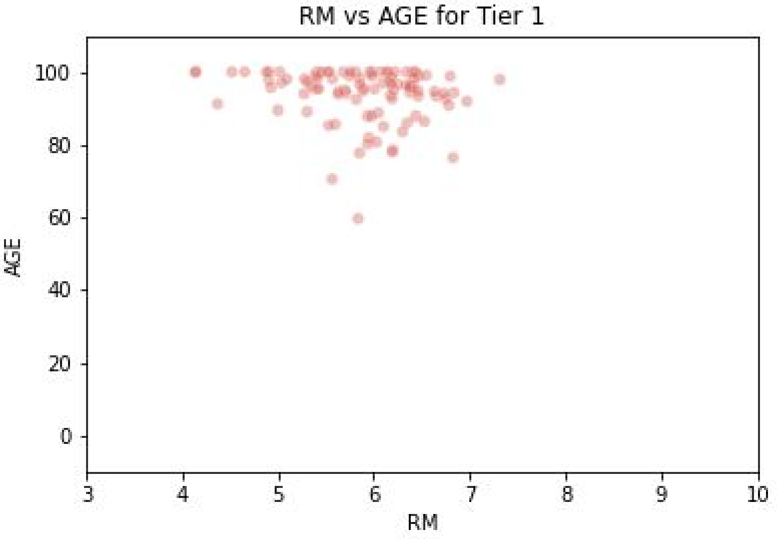
A first look at the PCA graph shows that the blue Tier 4 points are located mostly at the bottom left of the graph. PCA Component 1 has the largest variance in the data and stores the most information. PCA Component 1 also seems to be mostly composed of the property tax per USD 10,000. PCA Component 2 is also heavily weighted on the number of the black people in that area. The graph gives a negative correlation between the property tax and the median house prices. This is consistent to what is seen in the data. However, factors such as the pupil and tutor ratio do not have a heavy weightage in either of the components. One reason for this is because the PCA dimension reduction loses a high amount of data (around 40% in this instance), which can be seen as a setback of reducing 14 dimensions to 2 dimensions. Although the graph does seem to obey correct relationships as per the data with factors such as RAD (index of accessibility to radial highways per town) and INDUS (proportion of non retail businesses per acre per town). They have respective PCA values of 0.45 and 0.28.

AGE vs RM Analysis

Performing this analysis yielded the following graph:

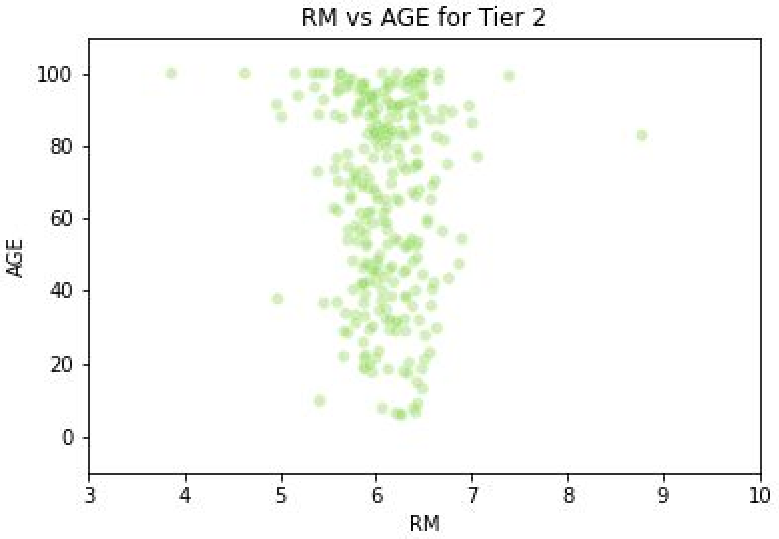


The graph shows many interesting properties. We can decompose this graph into the tiers and analyse them more profoundly. For tier 1 we produced the following graph:



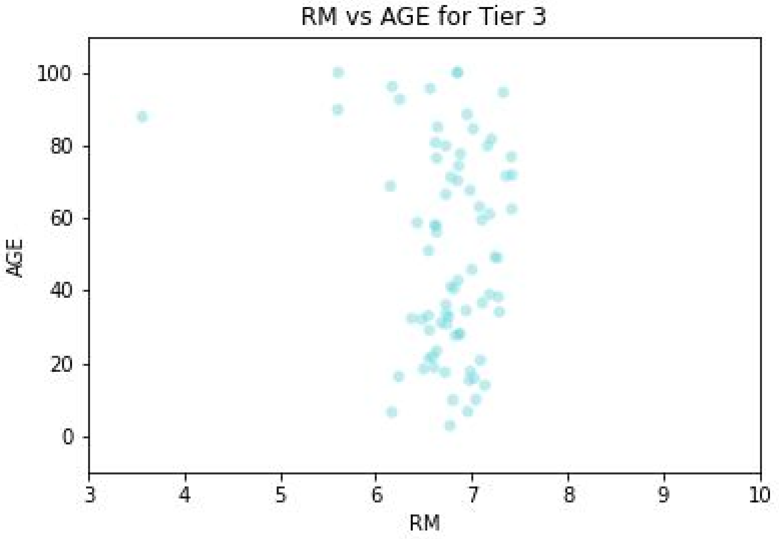
From this graph we can see that in general Tier 1 houses are quite old, with most of them having an AGE value between 80 and 100. Additionally, they all have less than 8 rooms on average, as the highest RM value is still less than 8. In this case, RM and AGE don't show an evident correlation. Hence I can conclude that, for Tier 1 houses, there is not enough evidence to suggest a correlation between RM and AGE. We can see that there is a dip between 5 and 6 RM, which suggests that newer houses tend to have between 5 and 7 rooms on average, but there is not enough data to support this as this could simply be due to outliers.

We now move onto analysing Tier 2:



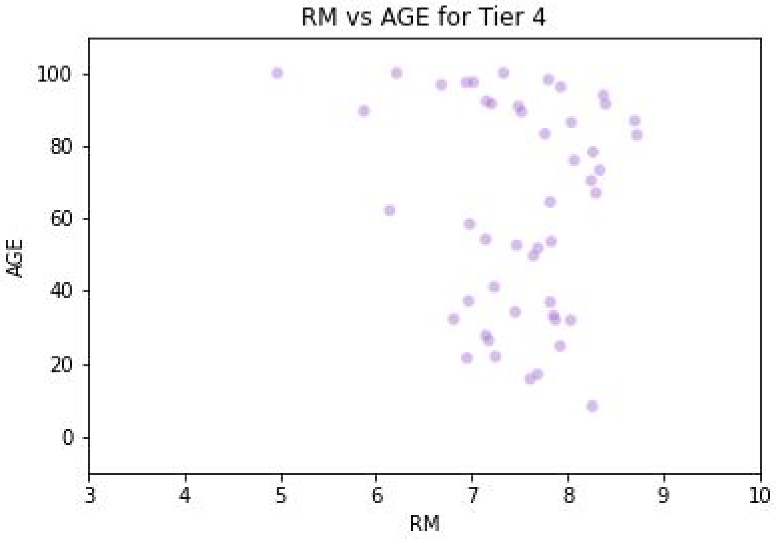
We can now see a very different picture: In this case we see that in general, Tier 2 houses clearly have between 5 and 7 rooms on average, with most of them having 6 as their RM value. Hence for tier 2 houses we can safely conclude that there is enough evidence to support the claim that Tier 2 houses have between 5 and 7 rooms on average. It is also important to note that the more recent houses have also had between 5 and 7 rooms.

For tier 3 we produced the following figure:



We can see a similar shape to that of Tier 2 houses, but slightly shifter to the right. The graph seems to suggest that most tier 3 houses have between 6 and 7 rooms, with most of them skewing on the latter saide. It is also worthwhile to note that there is an outlier with about 3.5 rooms, but it is an isolated event.

Lastly, for Tier 4, we obtained:



We can see a radically different shape here. We see that most houses have around 7-9 rooms, with the number of houses with more than 8 rooms on average tending to decrease throughout the years. It seems that there has been a recent RM value of more than 8, but we cannot determine if this is an isolated outlier or the beginning of a new trend.