## Introduction

Considering the datasets available from the London DataStore and the Health Profiles, I have chosen the following:

- Health Profiles: Adult smoking, Deprivation and unemployment indicators
- London DataStore: Income from 2010-2012 both at the borough and ward levels

The aim is to explore the relationship between adult smoking (dependant variable) and its relationship with income, deprivation and long term unemployment (independent variables). I expect that there should be a positive correlation between these factors. Also it presents an interesting combination of factors to analyse for potential correlations.

## Methodology

**Data Selection & Exploratory Analysis**

- Pick variables and look at pairwise relationships to find any interesting patterns
- Generate some basic statistics and clean outliers

**Data Cleaning and preparation for merging**

- Ensure relevant columns for merging are there in all data
- Rename columns to ensure data from different sources are identifiable
- Drop redundant data which does not add value and merge data

**Modelling and Analysis of merged data**

- Look at merged data and ensure integrity of product
- Drop redundant columns in data
- Look at pairwise relationships in merged data and plot some regression trends

## Exploratory Analysis, Initial Insights and Data Pre-processing

**Fig 1, 2 and 3** in the Appendix section show the exploratory stage of the data from the health profiles. We see that for most of the data, it is only available for a year so aggregate columns in the data set are not available which have been dropped. Also, the data is visualised by Significance of the indicator values as a hue in the plots; this is a categorical variable defined as significantly better, worse or not much different from the average for that health indicator value for England. We already note some positive trends in the data which will be investigated further after data merging.

**Fig 4 and 5** show violin plots for the median income at ward and borough level. I am looking at both because I want to explore the wider trends in incomes and also because the health indicator data is at borough level so I want my income dataset to be equivalent. The violin plot is chosen to visualise the income data as it shows an estimate for the distribution of the data. From our first look we see that incomes at the top end have risen for the time period of the data, while the median has stayed fairly static. This could possibly be showing the rise in income inequality and an example of the rich getting richer principle. But also it could be the extreme values skewing the data. This highlights the need for data scaling for conducting meaningful analysis. I consider log, z-standardisation and min-max standardisation of the income data. Since the health profile data chosen is in the time period of 2010-2012, I have kept the income data for this period and dropped the rest.

The scaling of the data ward income data is shown in **Fig 6, 7 and 8**. I have chosen the min-max scaling as I feel that it does not skew the relationships in the data as much as the z-standardisation

but is more interpretable than the log normalisation. Prior to merging, the health indicator data is scaled the same way to ensure consistency of processing. **Fig 9, 10 and 11** show the same plots but for borough level data included for completeness.

The health data is merged using the 'ONS Code (new)' field the equivalent in the income data is the 'LAD Code'. Also the income data can be merged with borough names, which does not result in data loss. The all data sets are scaled prior to merging and columns are renamed with which indicator they represent to make it easier to identify post merging. The challenges faced in merging the data are that they are at different resolutions which results in data loss when merged. One approach was to merge the income data at ward and borough level with each other, drop the duplicates and blank items and then merge this with the health data. This led to a much more complete data set, which was used for final analysis.

## Analysis, Modelling and Conclusions from merged data

To gain insight into the data, a pair plot of all variables with a linear regression model fitted was generated and interesting variables identified for further exploration. The Year on Year Income increase at Ward and Borough level was plotted against the Indicator values for the health profiles, with a linear regression model fitted see **Fig 12**.

We observe a positive correlation between Year on Year increase in income deprivation, unemployment and smoking. It appears as one would expect that lower income increases year on year are likely to be in areas that have higher levels of deprivation and long term unemployment which may be leading to higher rates of smoking. It is probably because such areas have few economic opportunities. We also see that the kernel density estimate plots for YoY data at the ward level that there is a sharp spike, implying that a few wards have great concentrations of rising incomes. The KDE plot at the borough level shows a distinct peak, with a few smaller peaks probably due to a few high income wards being spread across a few boroughs. We see an almost linear relationship between deprivation and smoking data which may imply that tackling deprivation could bring down smoking levels which could lead to better health outcomes.

A Gradient Boosting Regression Model is fitted to the data and used to predict adult smoking rates shown in in **Fig 13**. We see that the model performs quite well in minimizing deviance over 10 iterations to predict adult smoking rates. Also this model is used to derive feature importance shown in **Fig 14.** This shows that the most important variables in determining smoking in adults is year on year increase or decrease in incomes. We observed a positive correlation between these variables before but it is surprising to note its importance as determined by this model. Thus we have achieved a deeper level of insight that just using a simple model this can now be used in practice to inform policy. Rise or fall in incomes are an indication of economic opportunities which is linked to levels of deprivation. This can be used as a simple metric to screen for areas that undergo massive disruption in incomes and can be used to model potential health outcomes as a result.

In conclusion it can be said that authorities should focus on increasing the economic opportunities in most deprived areas as increasing income could have a positive impact in reducing levels of deprivation, unemployment which could aid in improving health through lower smoking rates. Although there may be confounding variables not considered in this analysis which could affect these conclusions.

## Appendix of Figures

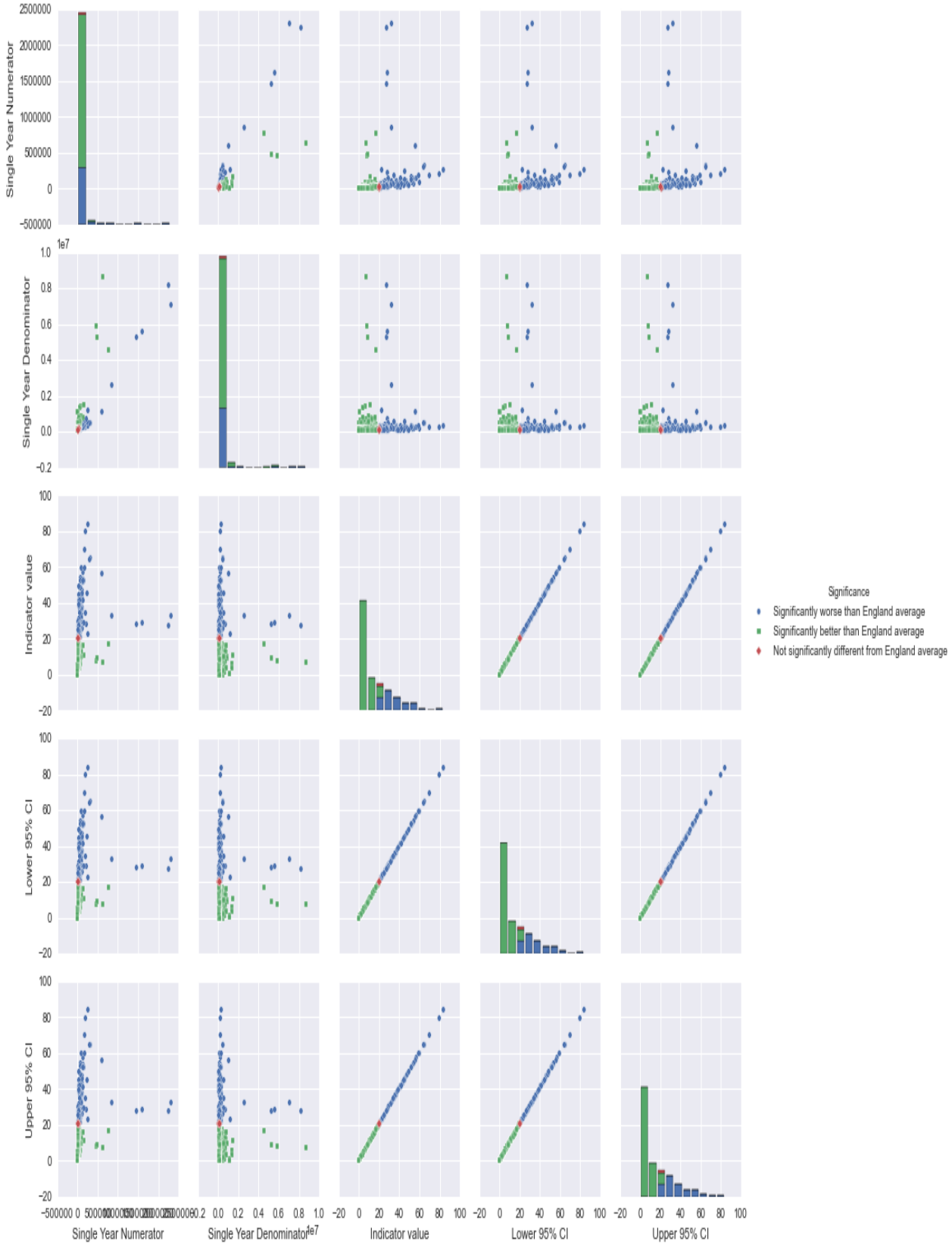


**Figure 1:** Shows pair plot of Adult Smoking Data coloured by significance. Outlier associated with the aggregate value for all of England has been removed.

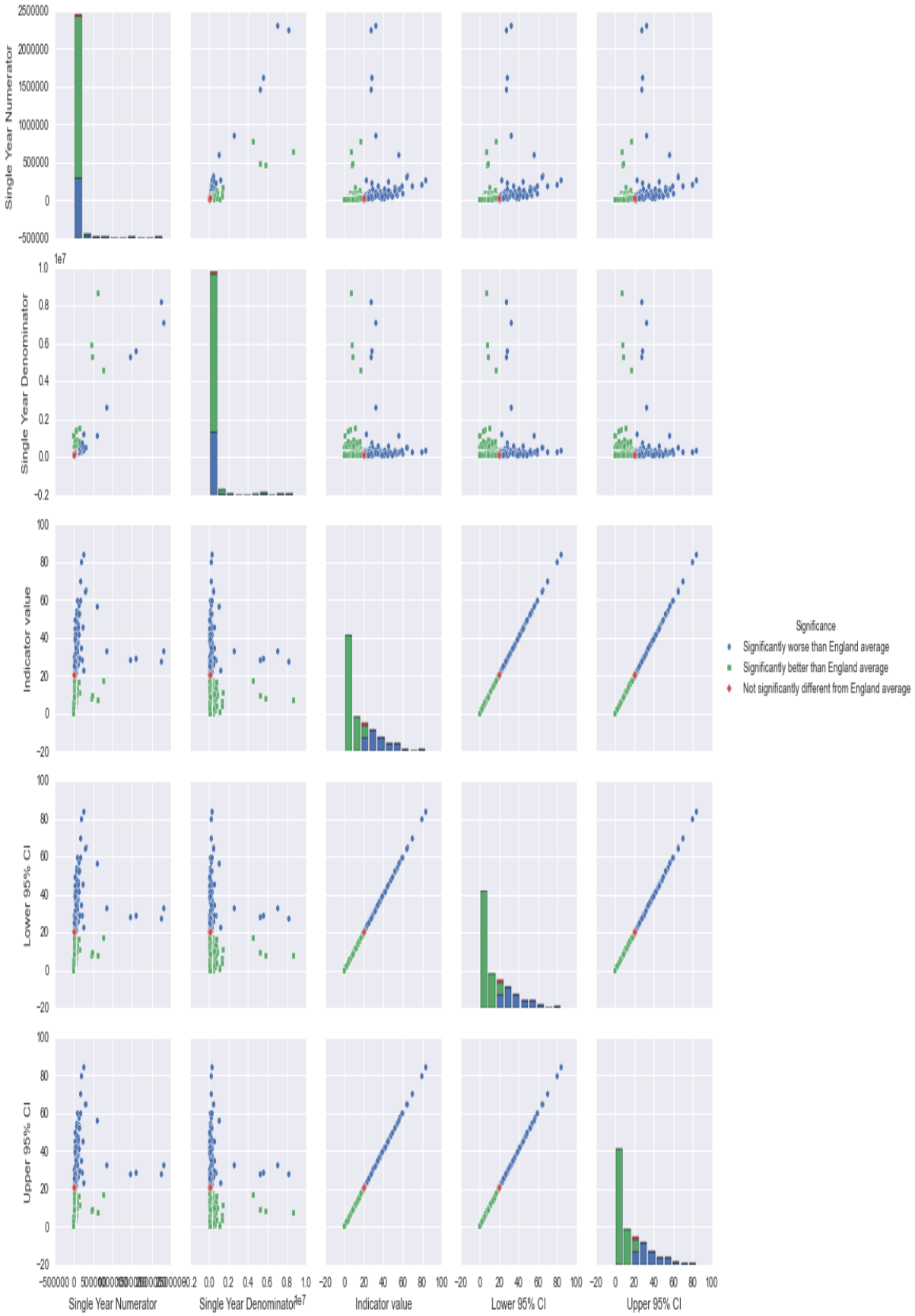


**Figure 2:** Shows pair plot of Deprivation Data coloured by significance. Outlier associated with the aggregate value for all of England has been removed.

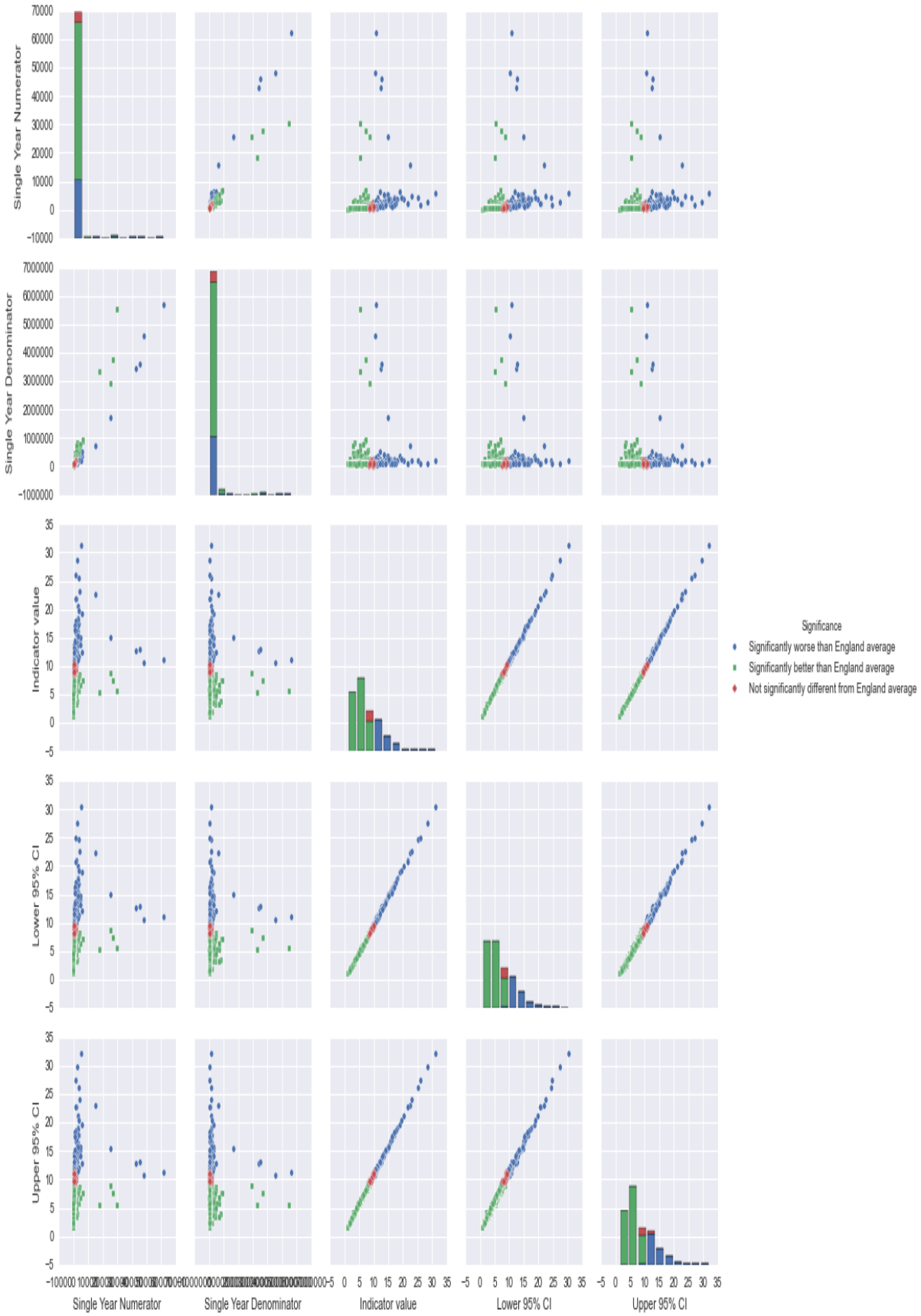


**Figure 3:** Shows pair plot of Long term unemployment data coloured by significance. Outlier associated with the aggregate value for all of England has been removed.

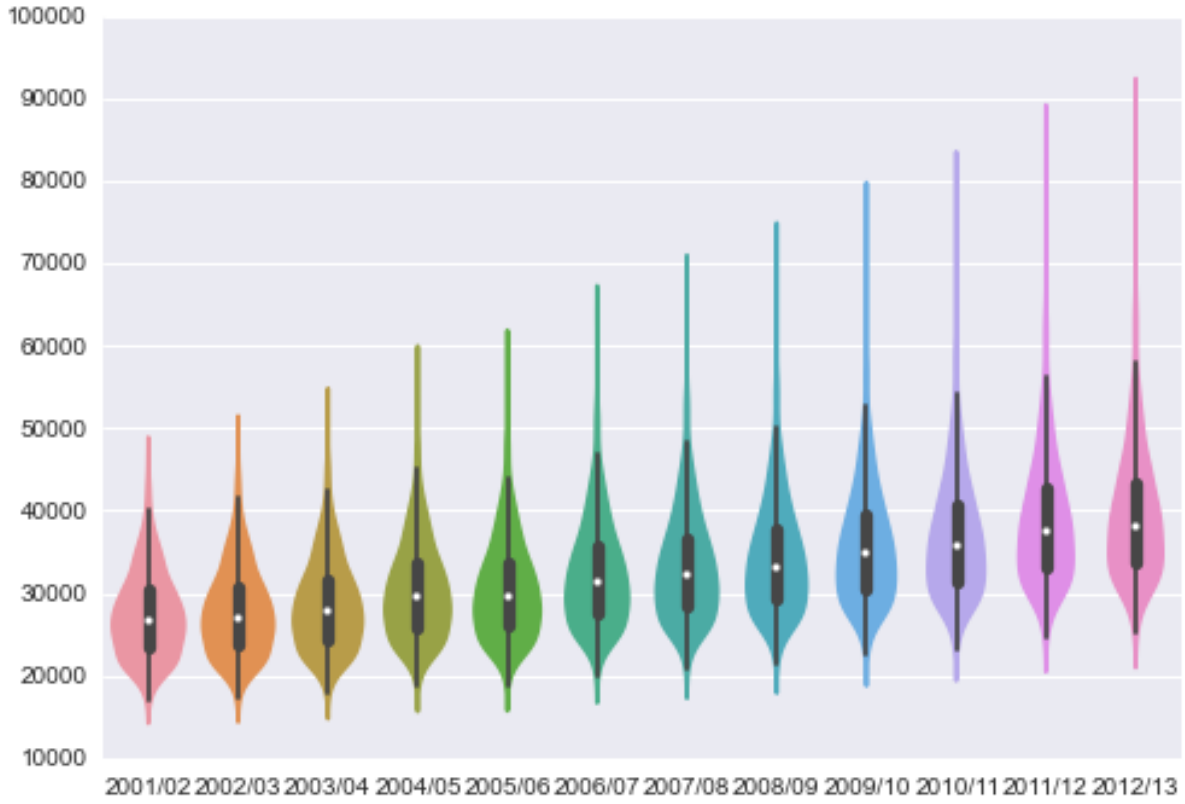


**Figure 4:** Shows of violin plot of the median data at ward level from the London Data Store. Note no scaling has been applied to this data. The violin plot shows an estimate of the probability distribution of the data so more informative than a boxplot.

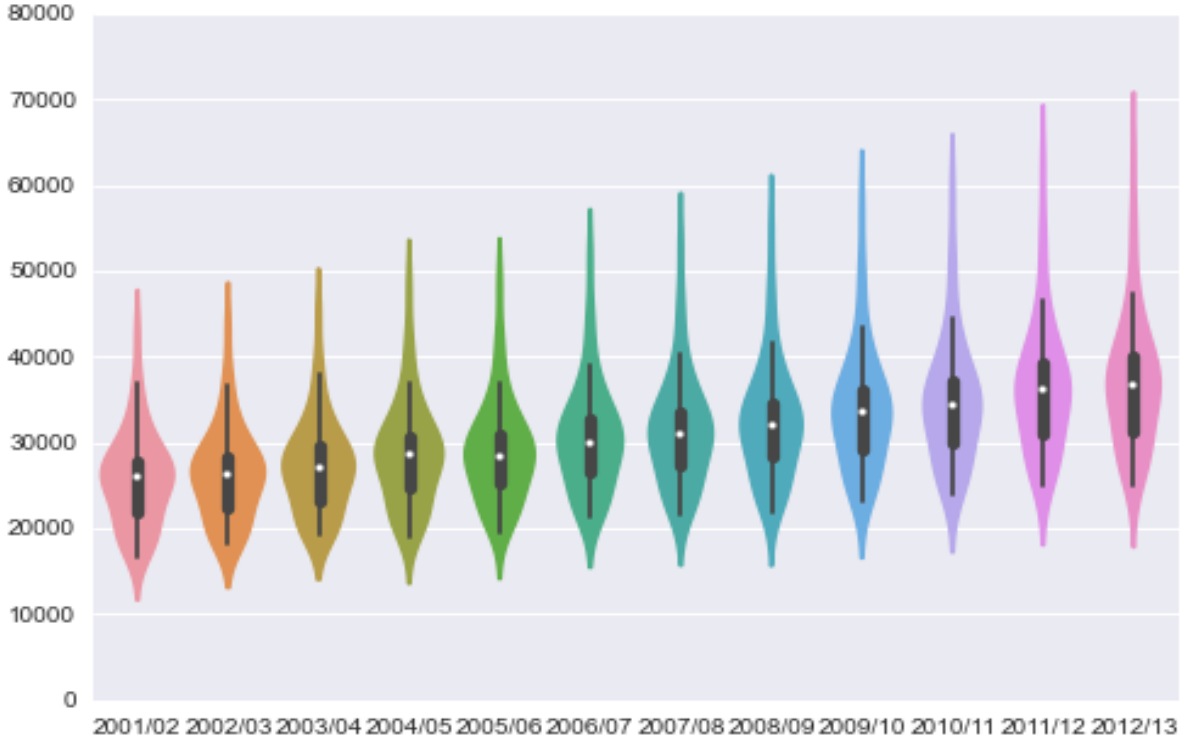


**Figure 5:** Shows of violin plot of the median data at borough level from the London Data Store. Note no scaling has been applied to this data

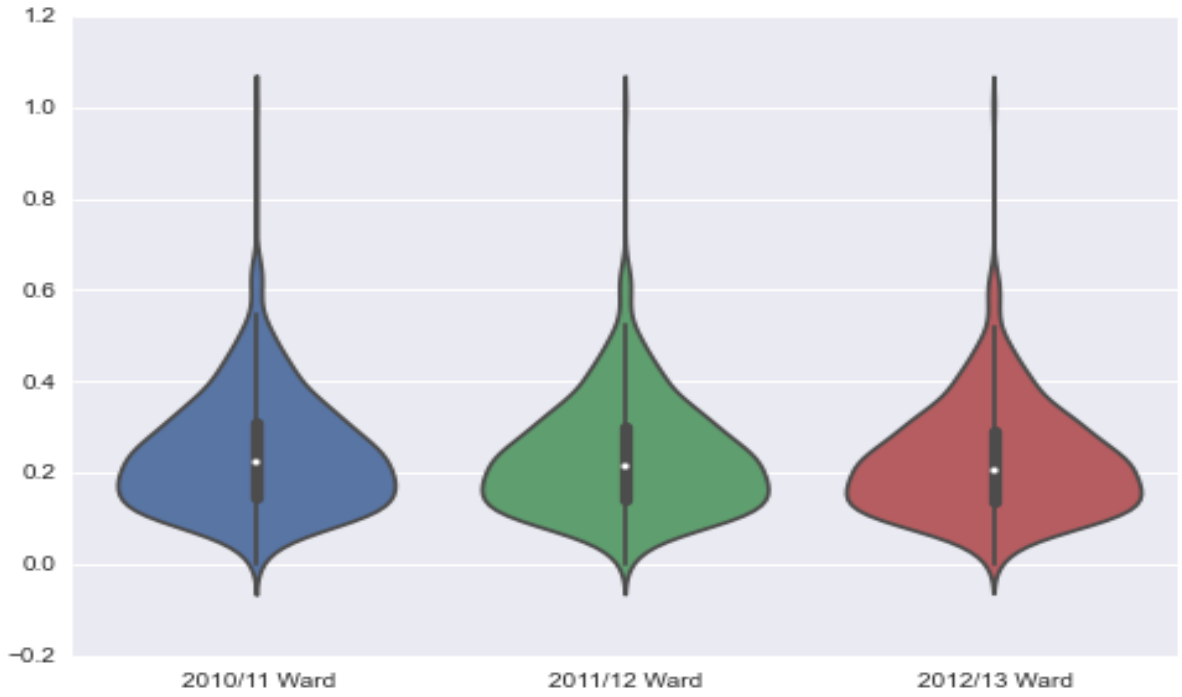


**Figure 6**: Median incomes at ward level with min-max scaling. We see top end is fairly constant with the bulge of the distribution being around the average.

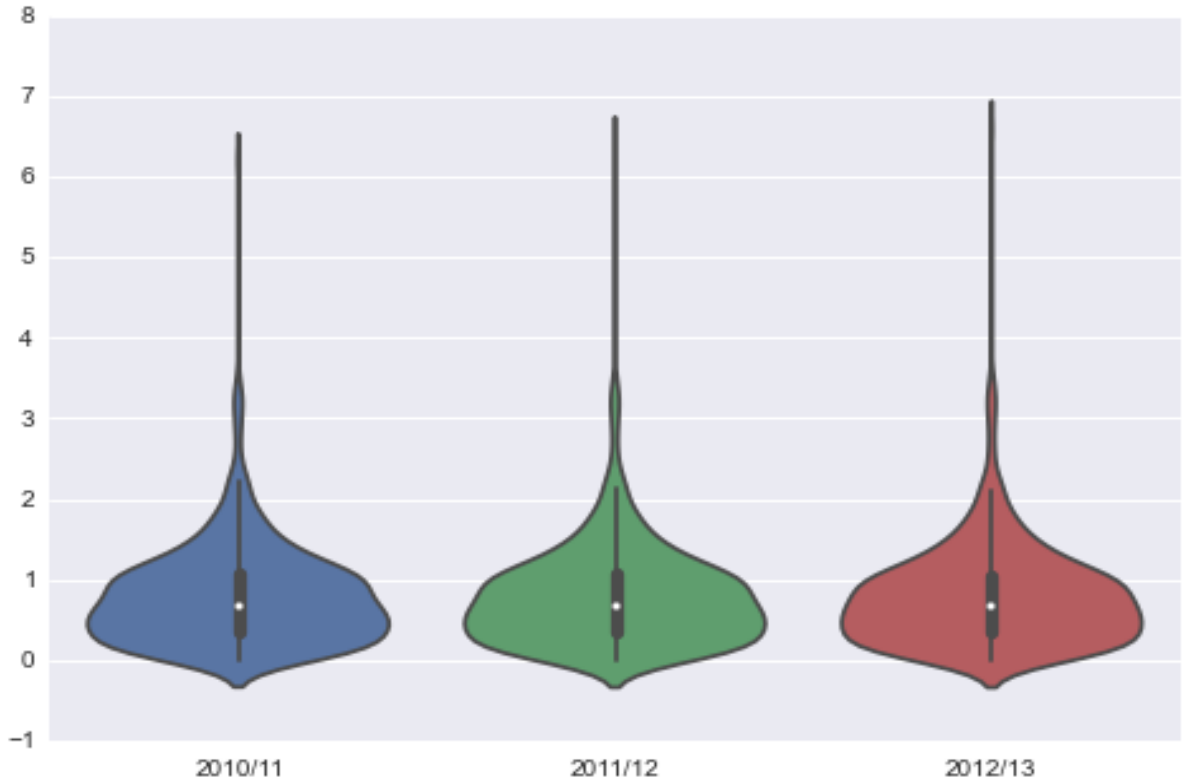


**Figure 7:** Median incomes at ward level with z-standardization. We see top end is more prominent here with the bulge of the data being squashed around the average this scaling seems to skew the relationship in the data around the mean. Less symmetric than above.

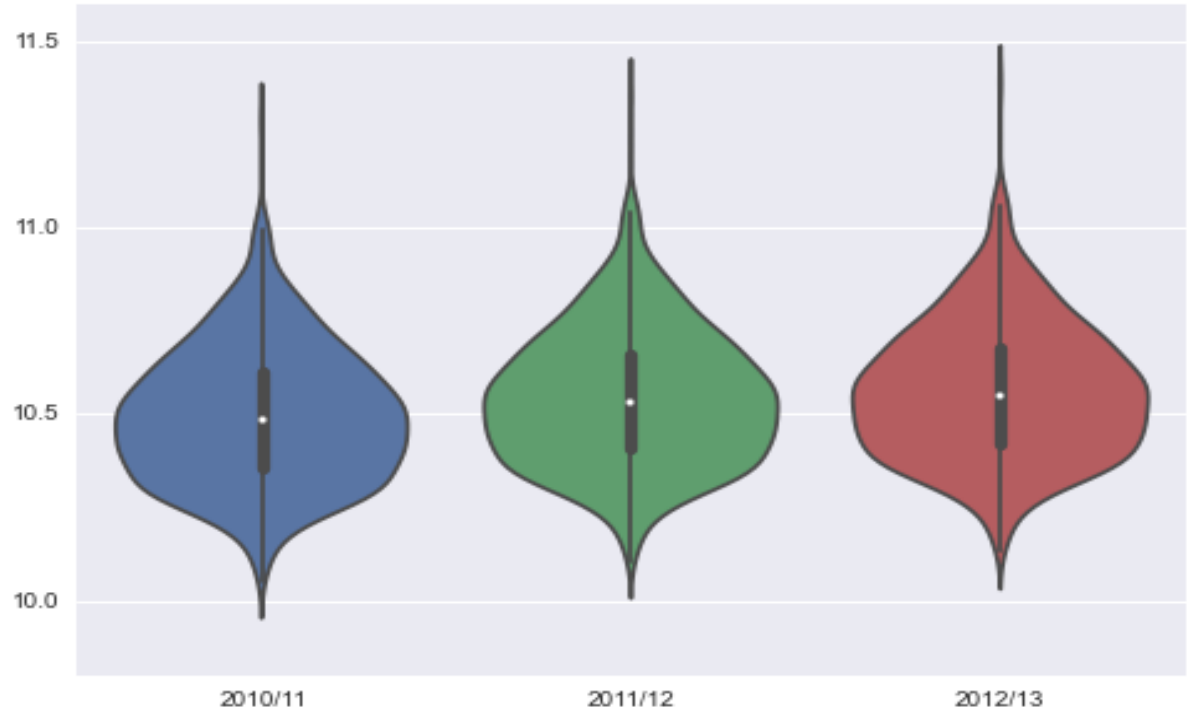


**Figure 8:** Median incomes at ward level with log normalisation. This produces a very symmetric scaling of the data however is difficult to interpret.


**Figure 9:** Median incomes at borough level with min-max standardization.

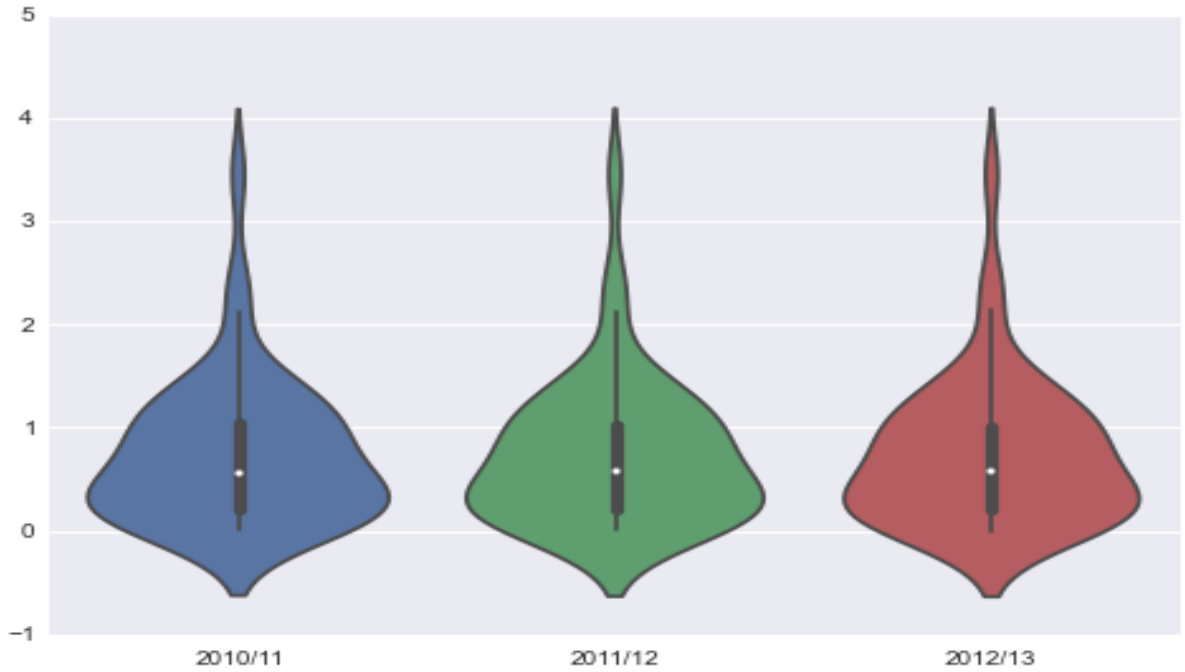


**Figure 10:** Median incomes at borough level with z-standardization.

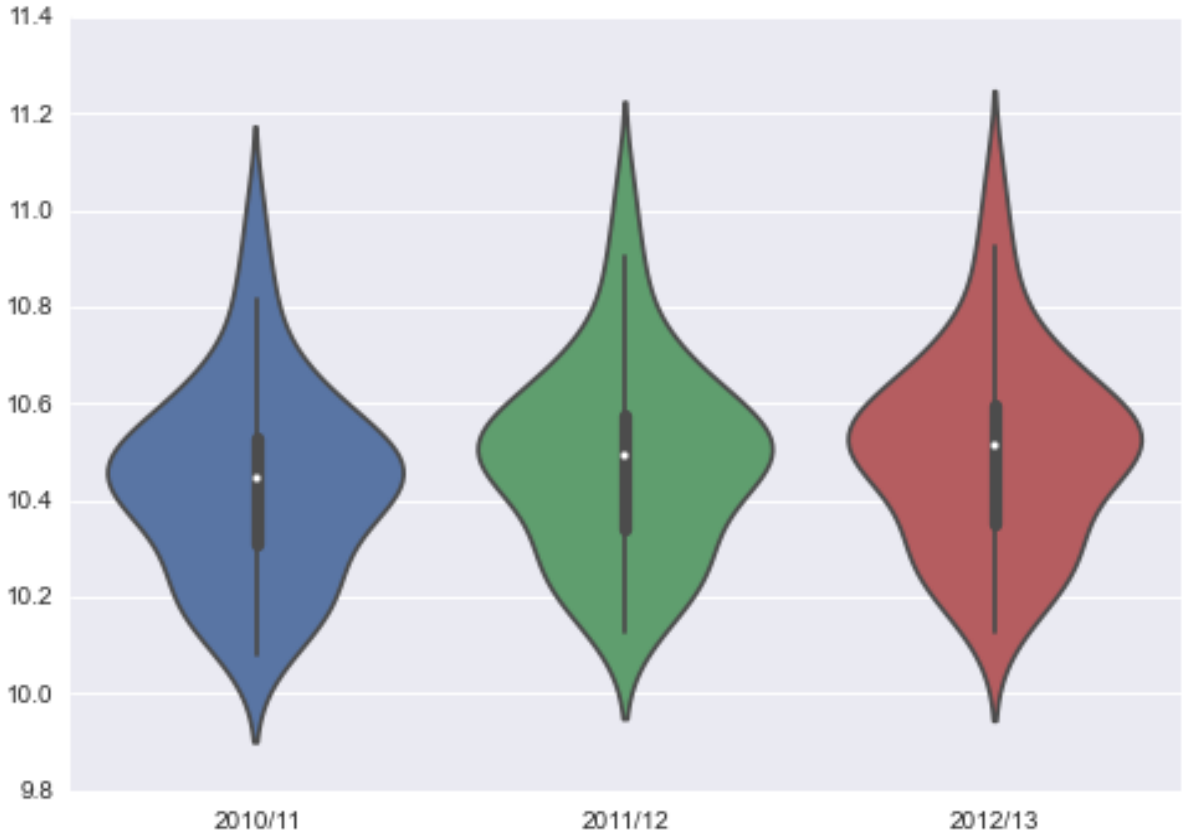


**Figure 11:** Median incomes at borough level with log normalisation.

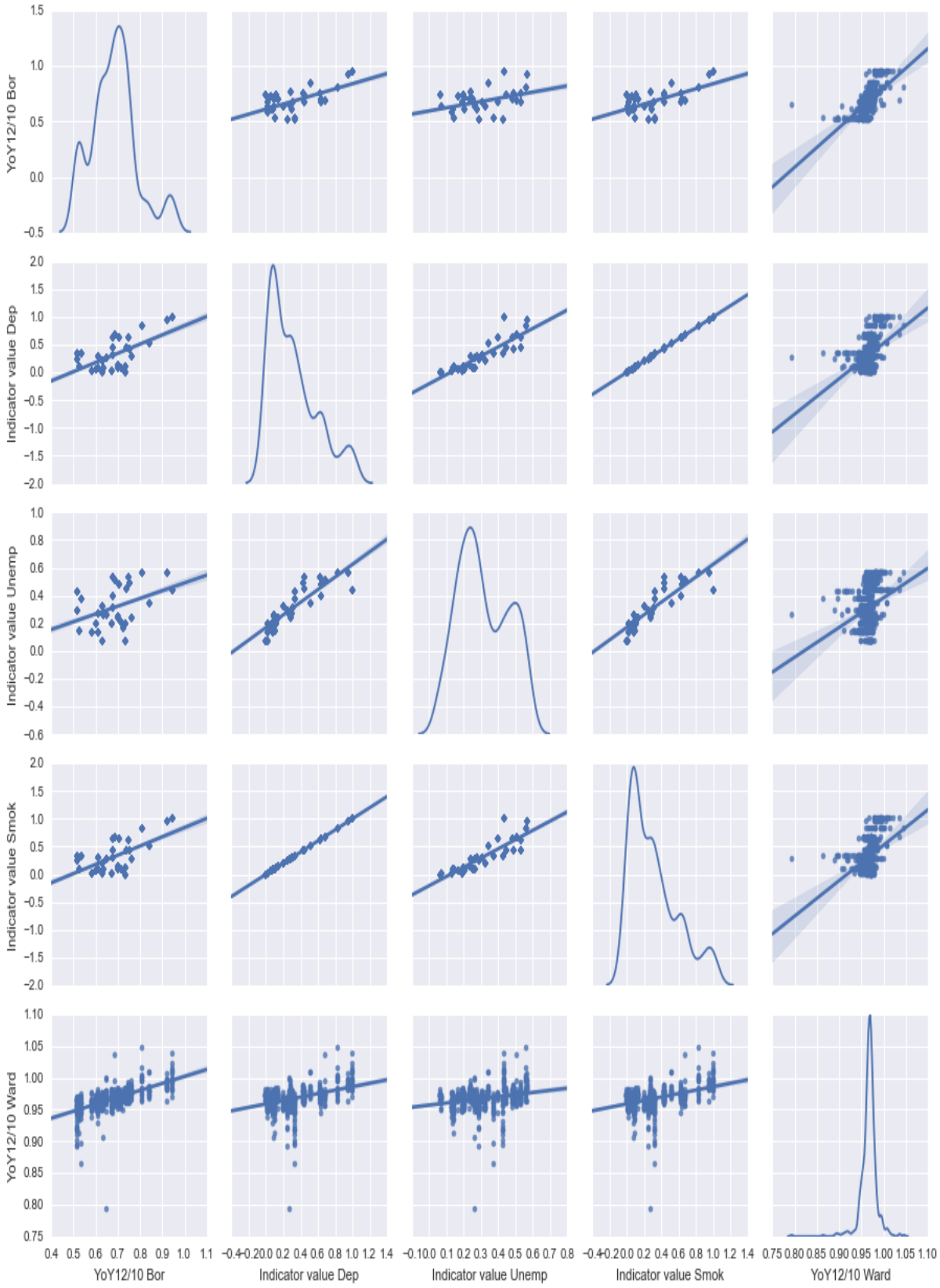


**Figure 12:** Shows highlights from final merged data with scaling. Here YoY represents year on year increase or decrease in income at both borough and ward level. The indicator values are for each health indicator with linear regression model fitted to data.

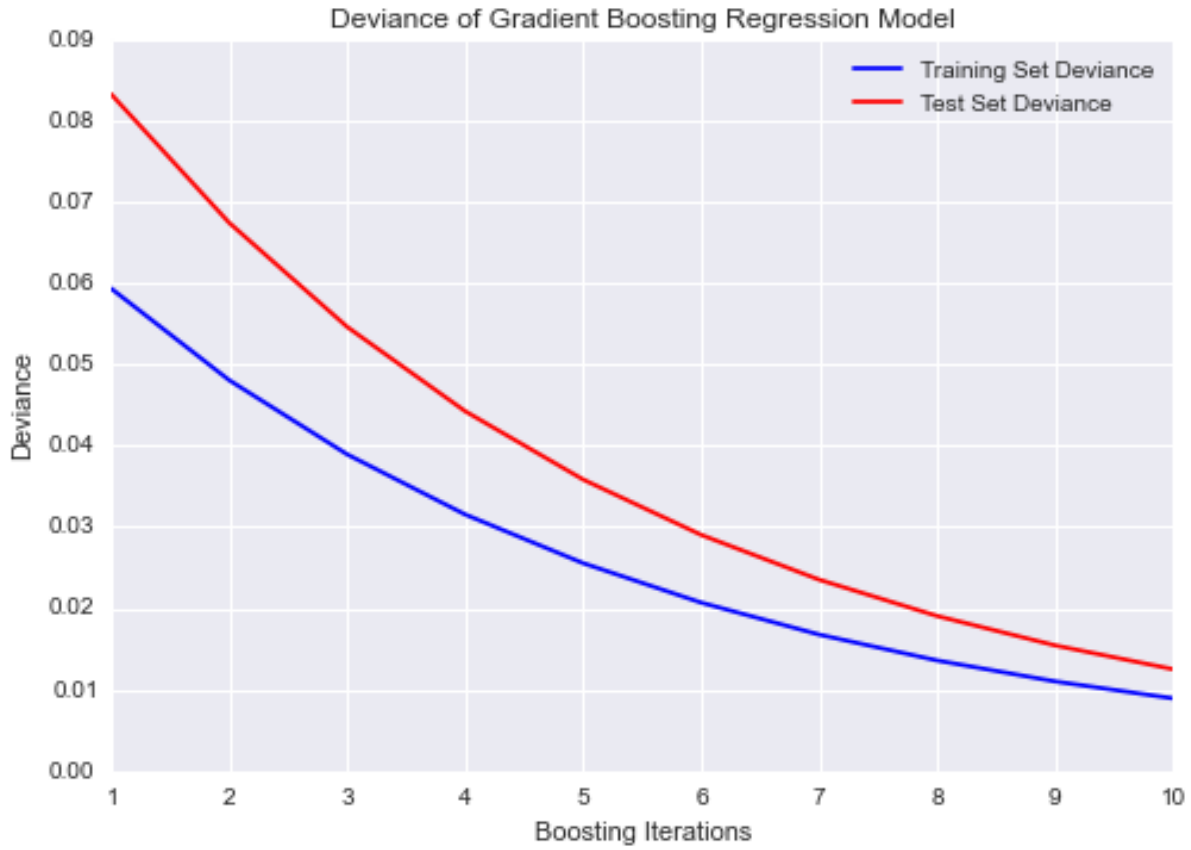


**Figure 13:** Plot of Deviance on training and test data over 10 iterations on Gradient Boosting Regression Model of final data.

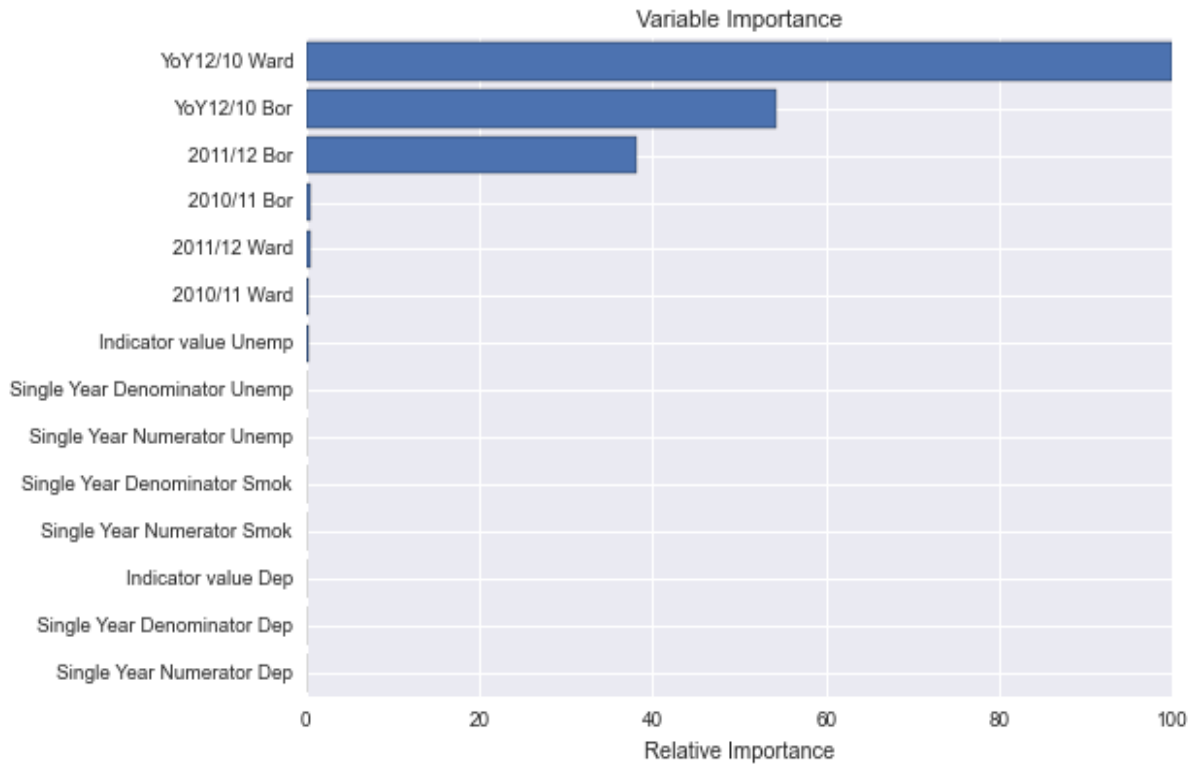


**Figure 14:** Plot of relative importance of variables in predicting smoking among adults derived from final data.