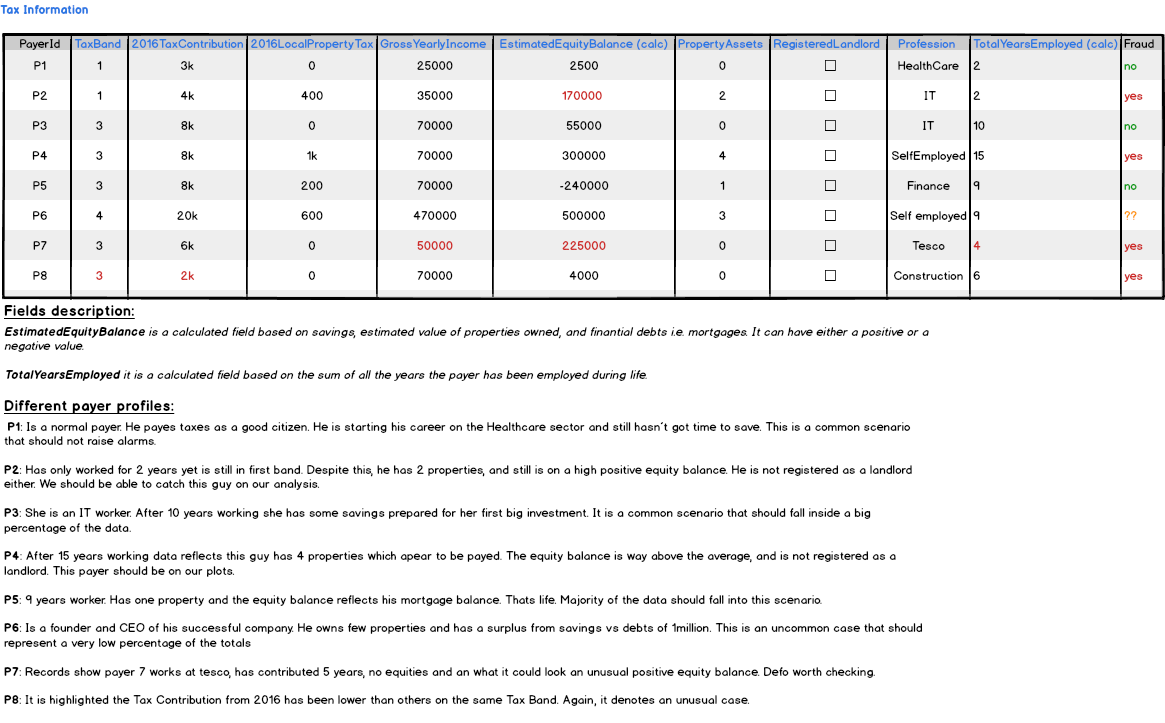
Data Analysis & Programming CA 1

Cesar Marrades Cortés – x00150376

# Design



# Part A – Fraud investigation

### First approach - Yearly Saving Ratio

What we can see on the table is each payer has contributed a different number of years. The more contribution, the more time the payer had to save. In order to get meaningful data with these values we could perform the following operations:

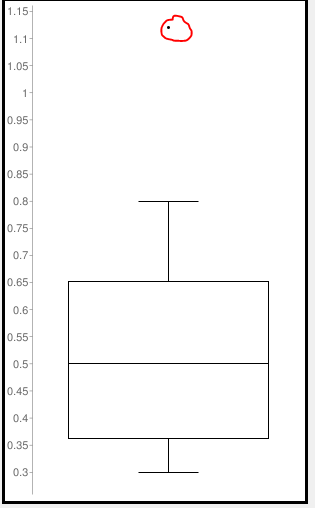
Estimated Equity Balance / Years of Contribution 🡪 Yearly savings

Yearly savings / GrossYearlyIncome 🡪 Percentage based on yearly savings vs current income.

This percentage will be called ***YearlySavingRatio***

Now we could either calculate this function at dabase level, or we could compute it on the fly by using appropriate functions on R.

If we were to boxplot this “*YearlySavingRatio*” **filtering out people without assets**, we would get a graphic where straight away we would see **Payer7**, where this ratio surpasses 100%. In the first table we can see payer 7 earns 50k a year, has only been working for 4 years, but it has a positive equity balance of 225k. So he saves yearly more than he earns. The boxplot would show an outlier which in this case points us to the payer in question that we are trying to detect.



As the upper quartile is on the 0.65, the Max value (0.8) could be an indicator of something worth investigating.

This scenario calculates the ratio based on “YearlySavings/GrossYearlyIncome”, so boxploting by different professions would not add valuable information given the ratio for a payer earning 500k and saving 250k would be the one from a payer earning 6k and saving 3k (50%).

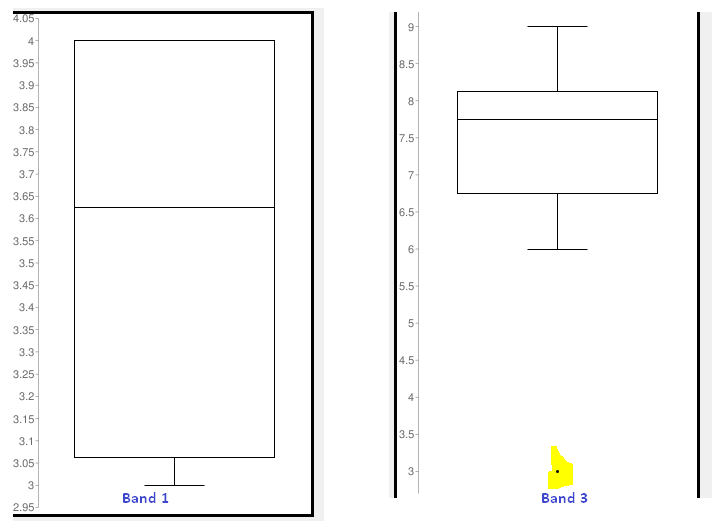
### Second approach – Tax Contribution

We could go further apply the boxplot to multiple fields that could show us relevant data. In this case, we could boxplot the 2016TaxContribution by taxbands:



Boxplotting *TaxContribution* would only have given us everybody on the same box, and there is not much information we could had read from it.

On this case it is useful to display *TaxContribution* based on the tax band, so they can easily be compared together. Different bands contribute at a different level, so boxplotting them by separate unmasks outliers for each band. For this approach, we detect an inconsistency worth checking for **Payer 8**, which is shown as an outlier for Band 3. It is important to remark we would have missed this outlier when boxplotting all bands together.

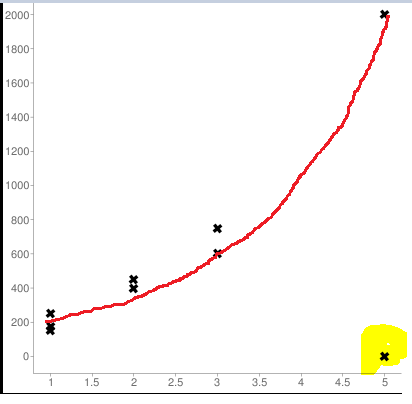


***Note:*** *the image shows 2 independent boxplots put together rather than a boxplot by TaxBands.*

### Third approach – Scatter Plotting Landlords

Another easy way to find abnormalities would be to scatter plot the local income tax with the number of properties. We could establish, the more houses you own, the more property tax you pay, or come up with some linear ratio based on square meters / local property tax. The example below shows the plot described, with few remarks:

* Top value represents a person owning 5 properties paying 2k.
* Lowest value represents a person owning 5 properties and **not paying property tax**.
* Although there is not enough data to support this evidence, supposing we would have thousands of records within this pattern, the shape of the curve corroborates our assumption: the more properties you own, the more you pay.



#### Investigating our data

1. **Correlation**

We could run tests to study our data and ensure our dataset is right.

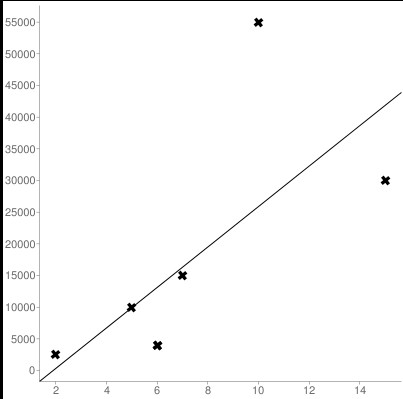
For example we could use a correlation test, to prove the assumption that for people with no assets, the more years they work, the more positive equity they are on. Running a quick test with the examples on the table, gives us a strong correlation coefficient **of 0.88** that proves this assumption.

1. **Linear Regression**

Once we find our assumption is right, we can also draw our linear regression model. Let´s say the independent variable is the years you have worked, and y, our dependent variable represents “equityBalance”.

We could get our linear model from our two variables, e.g. lm (equityBalance ~ totalYearsEmployed), and either :

1. Histogram our residuals, and see how our data is distributed.
2. Plot it, draw our straight and see how close our values are from it.



The graph above gives us visual information that can be used to support our findings. This model can also be used to predict “equityBalance” in x years.

#### Note

*We have covered different approaches that unmask Payers 7 and 8. We could run other operations that would leave other payers in an outlier position, making them target of fraud teams. E.g. P2 has only been working 2 years and has a positive equity of 170000eur. This would be obviously out of the norm and would be shown on a plot on years / equity balance as a point far away from the average.*

# Part B – Data Storage

As known, large amounts of data are difficult to process as they require big processing capacity.

Sometimes statistical operations cannot be performed over the total set of data, so when this happens, different techniques can be adopted.

Chunking is one of them. So for this scenario, we could have a big Database e.g. SQL, with all the data, and processing could be done by performing statistical operations on partitions of the whole data. After all the chunks would be processed, all the results would be computed and outputted if necessary.

This operation has some limitations:

* No dependency must exist between chunks.
* Chunks must have a manageable size.
* They must come from the same dataset

For this case, we could write a process chunking our data by manageable batches:

* “Manageable size” should be based on existing processing capacity (RAM, CPU).
* Processing should be performed on a separate server, so it wouldn´t add any load to the DB server, or slow down any applications running on an App Server.
* The information we are trying to process has no self-dependencies, so it fits in our rules.

See diagram below:

