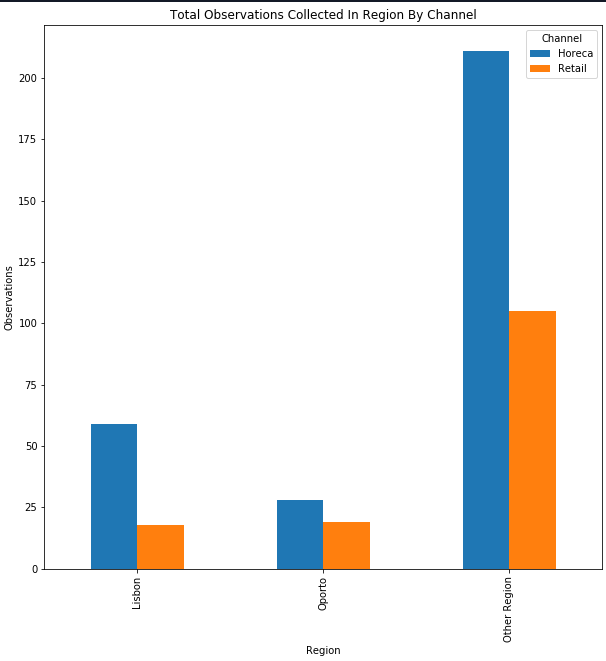
**Module B8IT105 Programming For Big Data  
Assignment CA4 – Pandas and Visualisations  
Student Code 10541255  
Student Name Alyosha Pulle**

# Introduction

The dataset for this CA is sourced from the UCI Machine Learning Repository and the dataset selected is a dataset from a wholesale distributor which gives a breakdown of annual spending by customers in different product categories by region.

The page in the repository relating to the dataset can be found here.  
<https://archive.ics.uci.edu/ml/datasets/Wholesale+customers>

The data is categorised by region and channel (Horeca and Retail). Horeca is a combination of Hotel, Restaurant and Catering. Below is a bar chart of the observations in the dataset.  


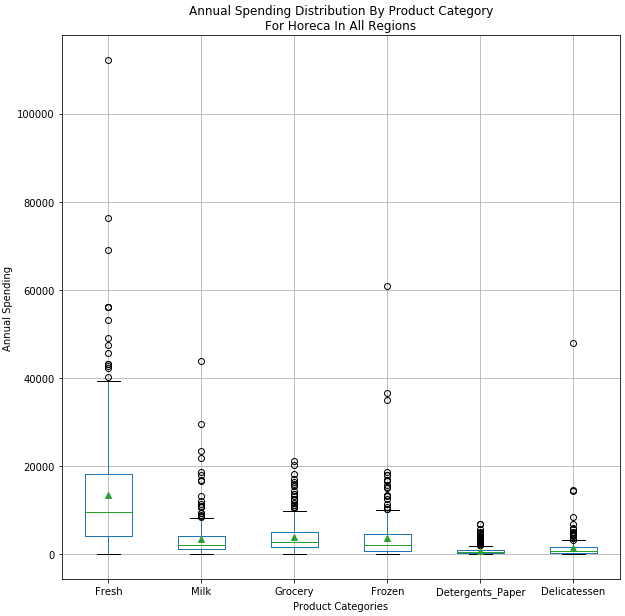
The regions are grouped by the two largest cities in Portugal and everywhere else.  
Horeca spending activity is significantly higher outside of the big cities.

# Total Spending In Regions

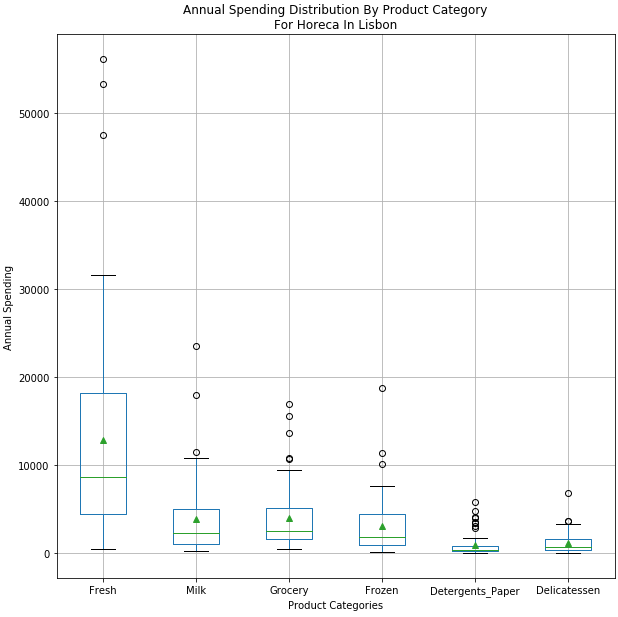
# 

From this we can see that for Lisbon, spending is significantly higher in Horeca than Retail.  
In Oporto, Horeca and Retail annual spending is very close.  
Spending in Other Region far outstrips the big cities, which can be explained by the higher level of spending activity as shown in the first bar chart.

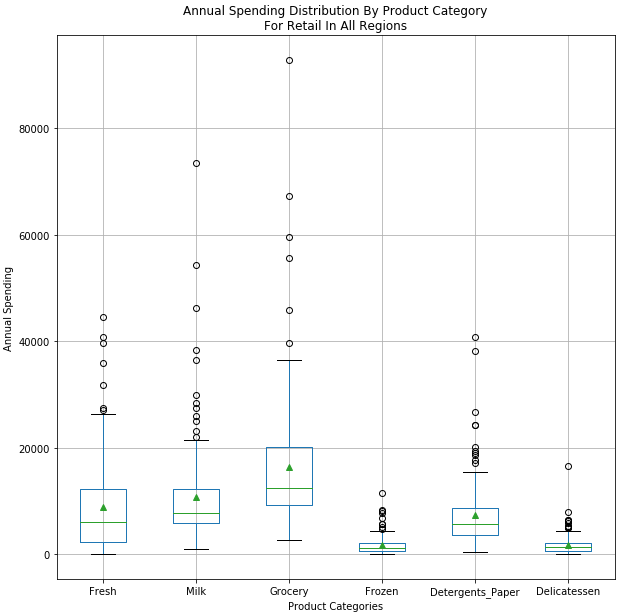
# Comparing Horeca Overall Annual Spending vs Lisbon



From this we can see that for the Hoteca channel, Fresh products category has the highest variability of spending. The outliers in this category may need investigation as to whether they are incorrect data or actual high spending customers.   
Looking at spending in Lisbon (see chart below) it is more or less in line with the overall spending trend across the product categories.



# Comparing Retail Overall Annual Spending vs Lisbon



From this we can see that for the Retail channel, Grocery product categories is where there is the highest variability in annual spending and Lisbon reflects the same trend.

