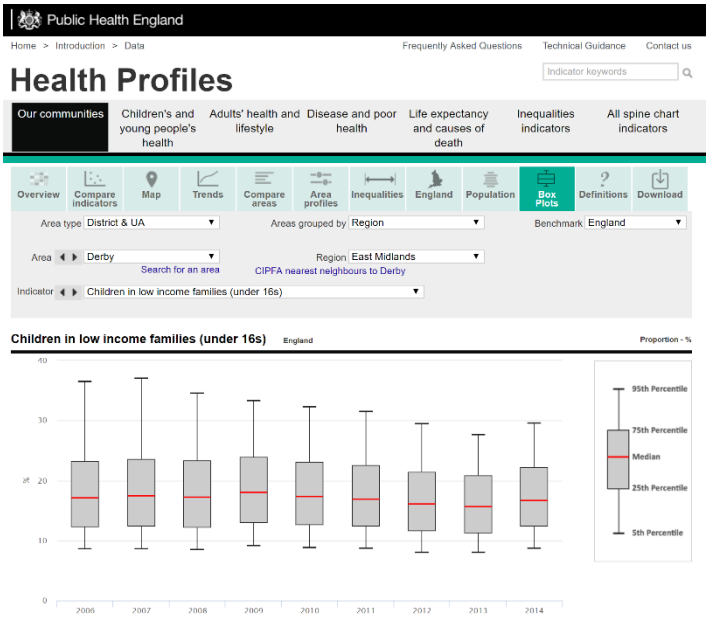
Emma Luk
Address: Bristol, Mobile: 07974 522 805 Email: emma_luk2@hotmail.com LinkedIn Data Science Projects

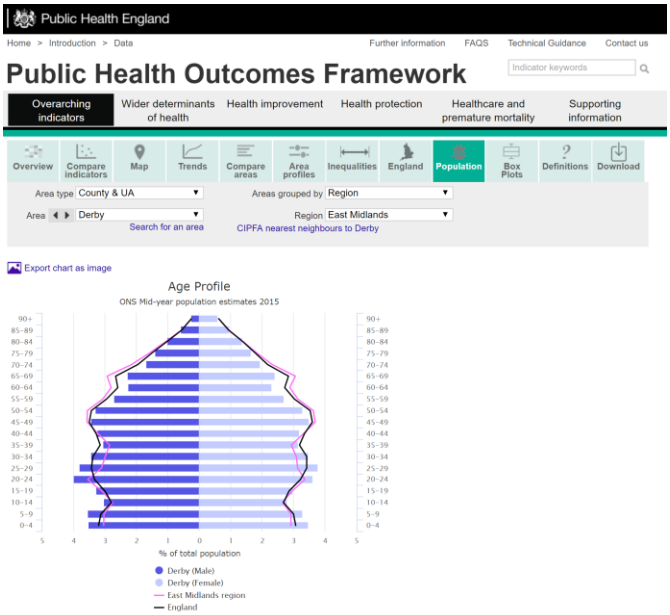
A selection of my data science work

Appendix: Emma Luk (Mobile: 07974 522 805):

1A. Examples of data visualisation from when I worked on the 'Fingertips' website for Public Health England (Government Agency):

On the left are boxplots depicting the percentage of children in low income families in the East Midlands between the years 2006 and 2014; on the right is a bar graph with negative stack depicting the proportion of males and females of different age groups in the East Midlands region.





1B: Data Visualisation in Microsoft Power BI & Tableau

- The aim was to create Executive Dashboard, tracked and reported on business metrics & the KPIs
- This dashboard included key performance, top ten products, top performing cities and top performing city, customer reviews and sales by Month.

Microsoft Power BI:

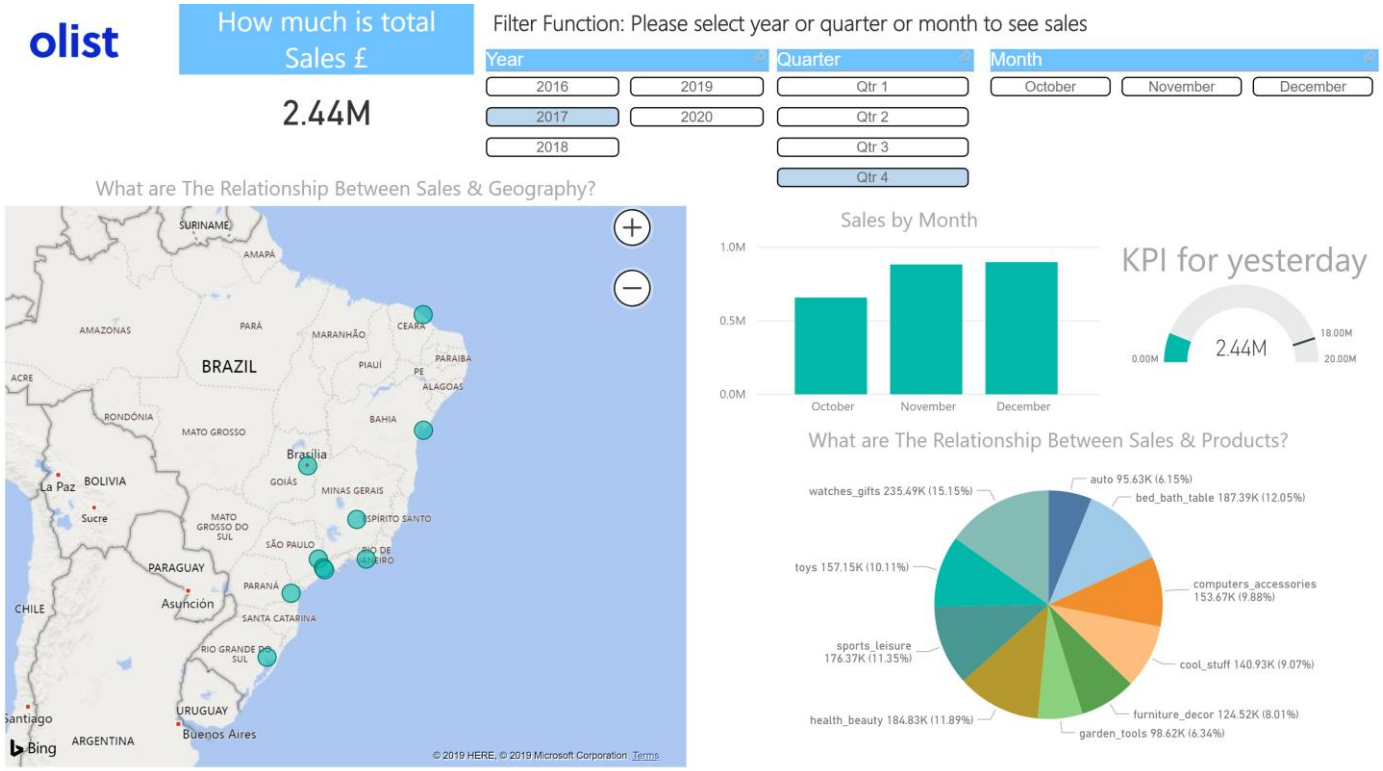
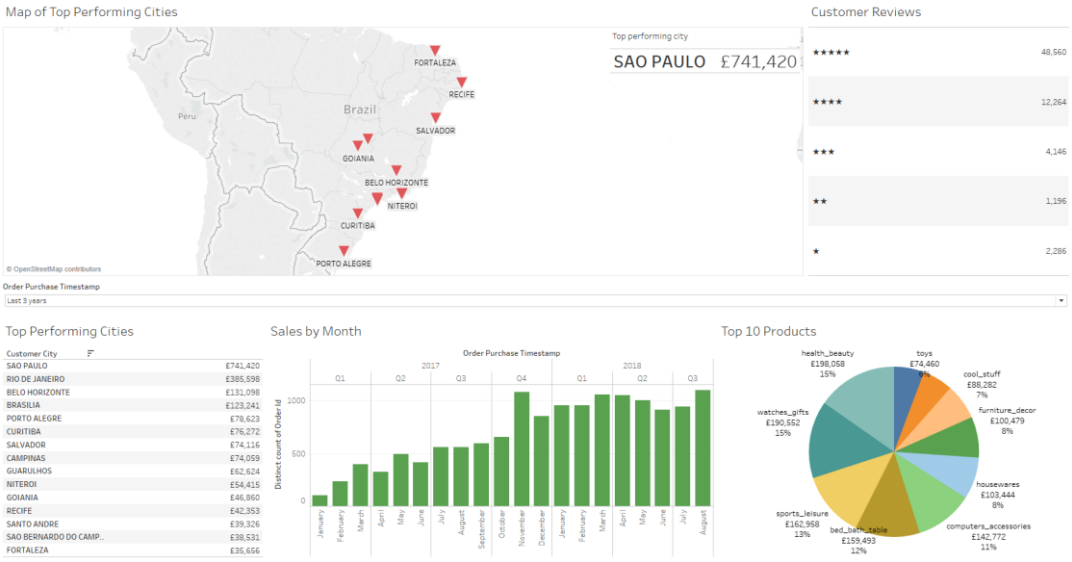
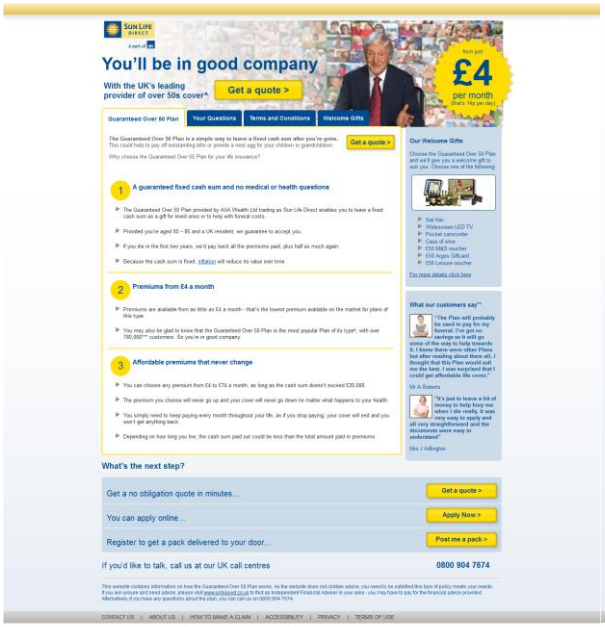


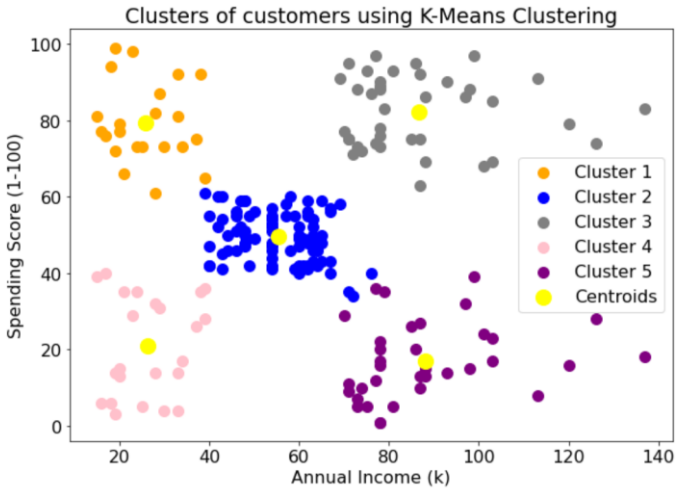
Tableau:



1C. Examples of A/B Testing and Multivariate Testing (MVT) from when I worked at AXA.

The two examples below depict different versions of the same webpage, which were used to provide insight to drive future strategies and identify business opportunities and problems.





Just looking at the graph tells us about the five different types of customers

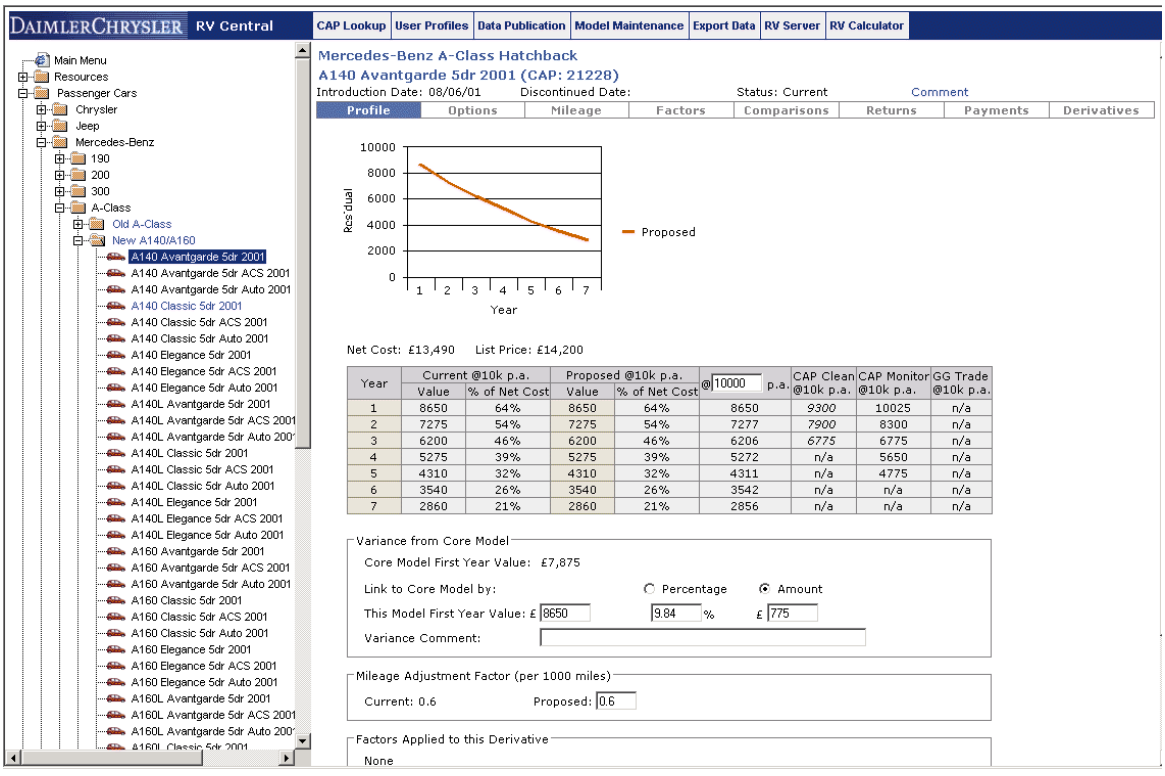
I. Cluster 1: Low income, High spenders
II. Cluster 2: Average income, Average expenditure
III. Cluster 3: High income, High spenders
IV. Cluster 4: Low income, Low Spenders
V. Cluster 5: High income, Low spenders

Machine learning is used to understand customers, drive personalisation, streamline processes and create convenient and memorable customer experiences.

Rather than relying on a marketer's intuition to separate customers into groups for campaigns, data scientists can use clustering and classification algorithms to group customers into personas based on specific variations among them with similar needs and behaviours. These personas account for customer differences across multiple dimensions such as demographics, browsing behavior, and affinity. Connecting these traits to patterns of purchasing behavior allows data-savvy companies to roll out highly personalised marketing campaigns that are more effective at boosting sales than generalised campaigns.

1D. An example from when I worked at Mercedes-Benz UK:

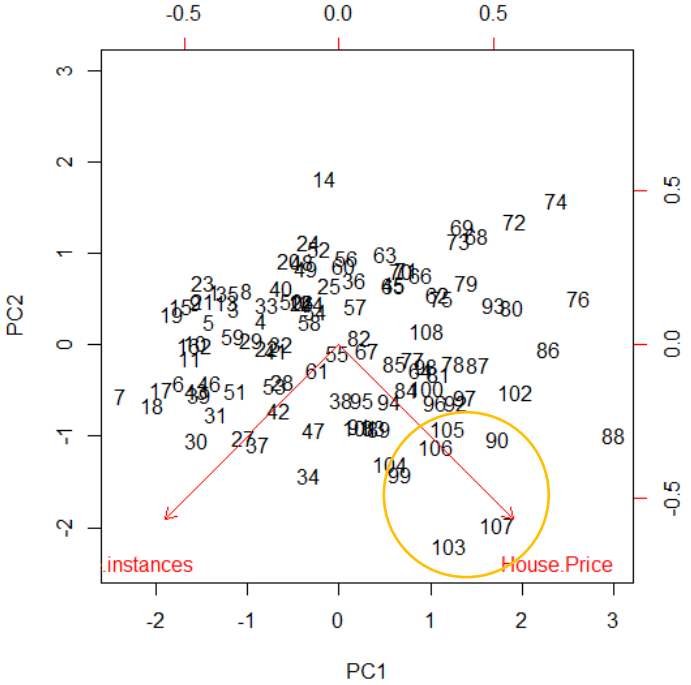
The first 'RV Central' in the UK, which provides the residual value of your used car. This system calculates the residual value for used cars and has since been implemented by Mercedes-Benz throughout all their UK dealerships. I also applied data science skills and new technology to improve business processes and business efficiency and reduced costs; this helped the company implement automated processes that would previously have been paper-based, resulting in increased efficiency and reduced environmental impact.



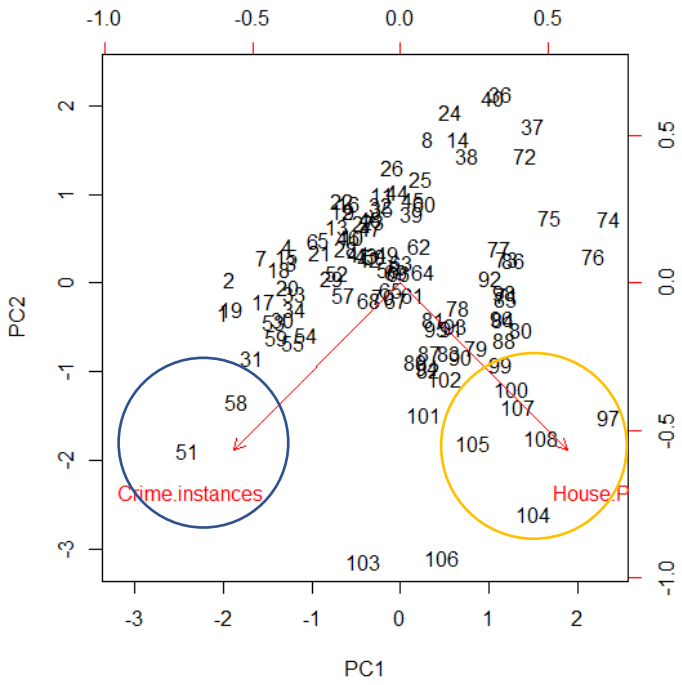
1E: Analysing House Prices and Crime in R

Unsupervised learning: Principal Component Analysis (PCA):

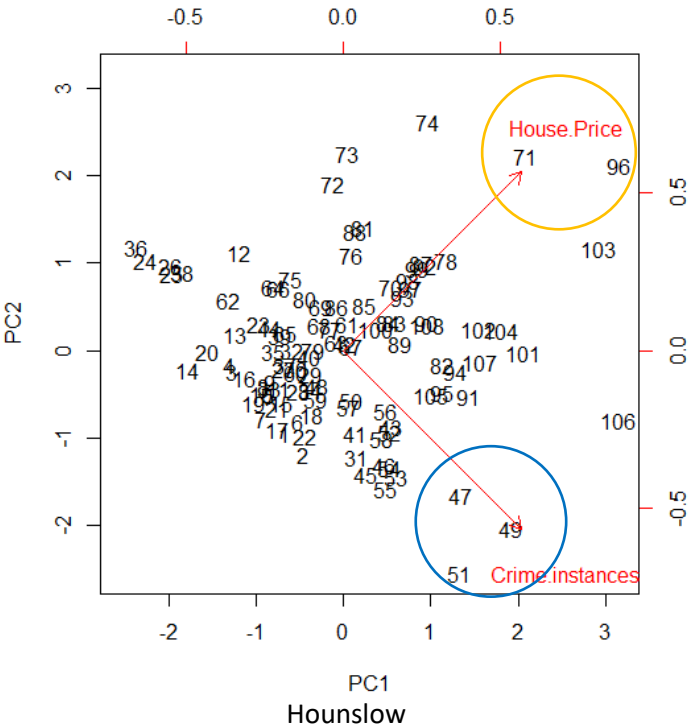
- The aim was to investigate how current house prices were affected by recent crime levels in London Boroughs
- Performed data cleaning, transformation, manipulation and conducted Principal Component Analysis (PCA): this was the process by which compute principal components and used them for **better understanding of the data**. PCA is considered an **unsupervised machine learning** method because it involves only a set of feature variables and no associated response variable. PCA also serves as a useful tool for **exploratory analysis** and **data visualisation**



Merton



Wandsworth



Hounslow

In Wandsworth:
House 51: £ 538,999.3
House 104: £ 973,938.4

In Hounslow:
House 71: £689,162.6
House 49: £396,623.0

For PCA:
House 104 & 71 data points are near House Price point. They are more expensive.

House 51 and 49 data points are near Crime point. They are less expensive. This is an indication that crime has some effect on house prices.

Figure 3: Plot of the first two principal components (PC1 & PC2) for Hounslow, Merton & Wandsworth

1F Industry Project: How Robots are making Farming Profitable
Weather Data Analytics Using Hadoop

- Leading the big data flow of the application starting from data ingestion from upstream to HDFS, processing and analysing the data in HDFS and data visualisation in R & JavaScript

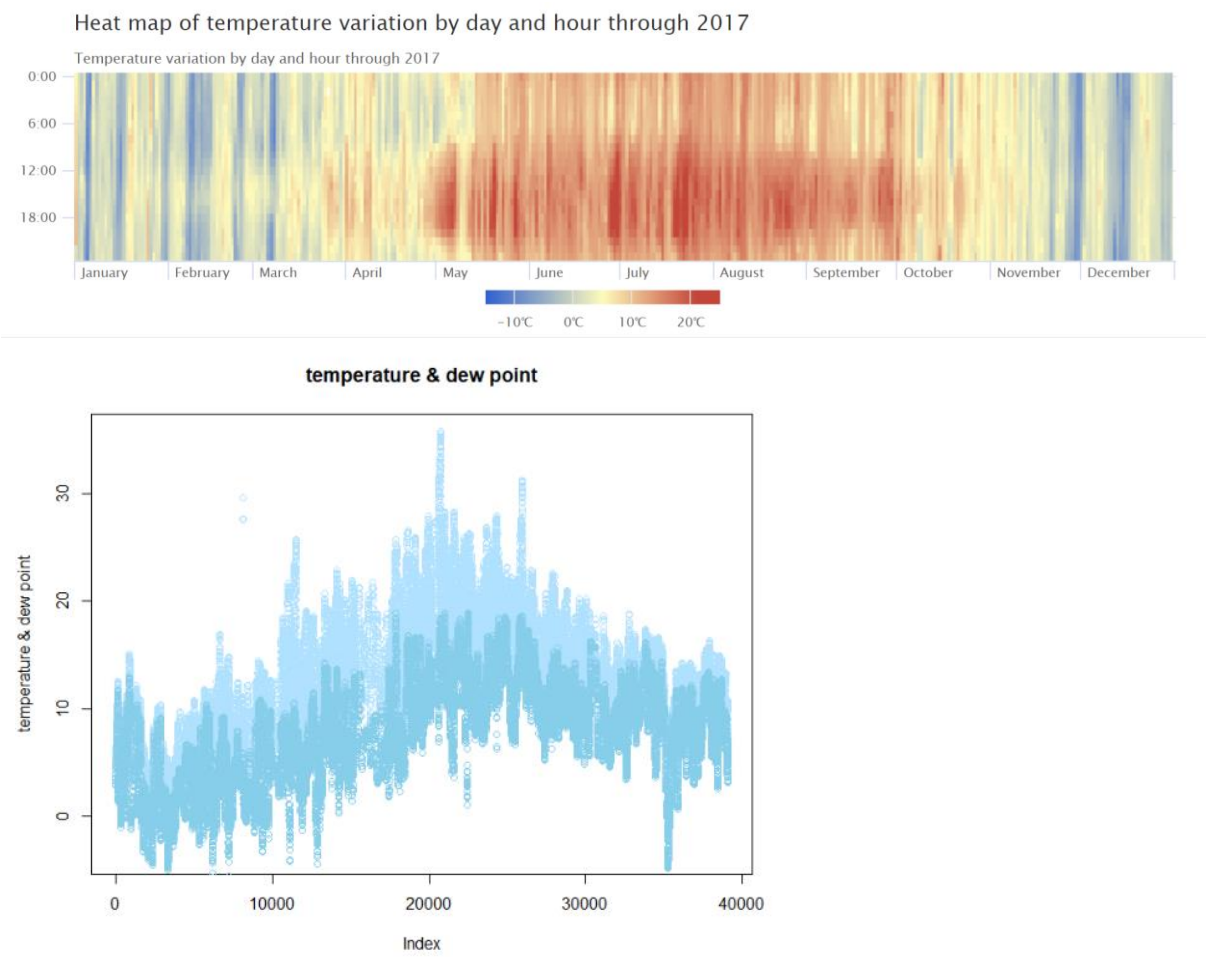


Figure 6: Scatter plots with two level
Added two level first level is dew point and second level is temperature in R

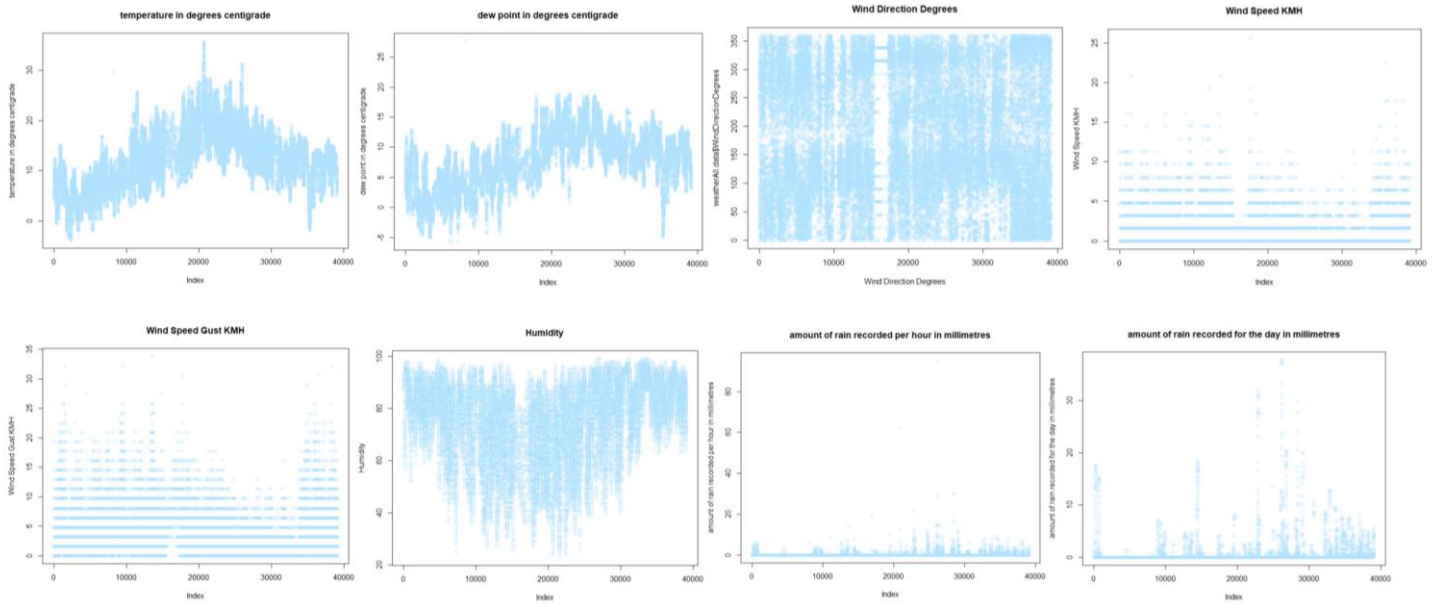
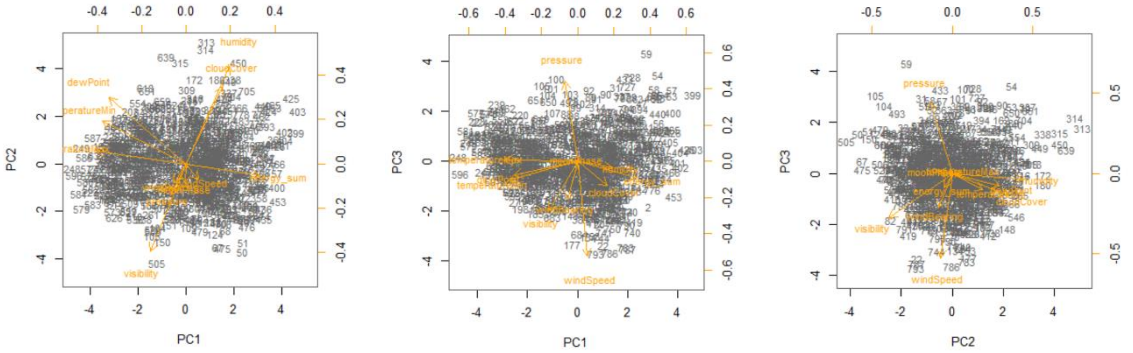


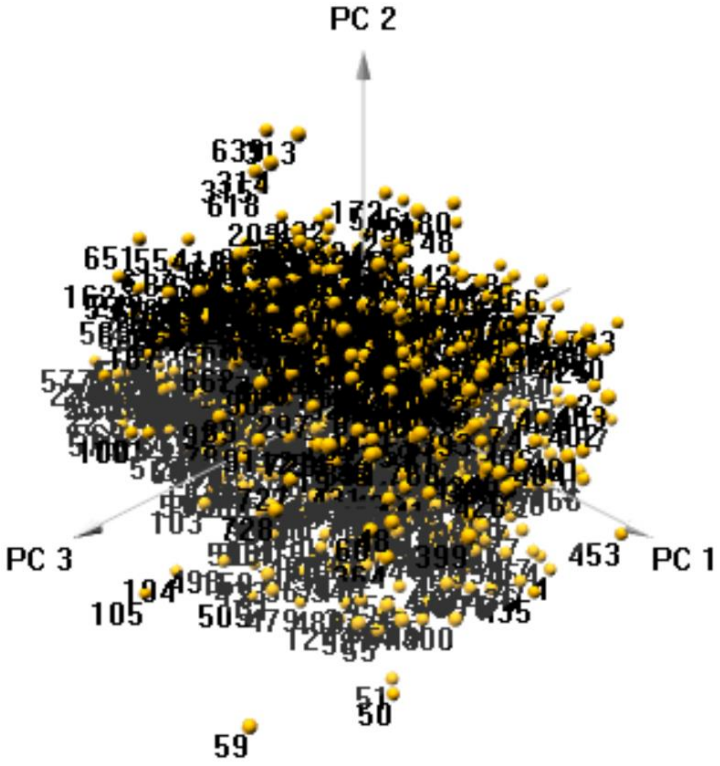
Figure 7: Scatter plots display the values of variables

1G: Explore relationships between weather conditions and energy consumption with R

- The goal was to draw a graph that shows how the samples are related (or not related) to each other.
- What was the relationships between weather conditions and energy consumption in London

Principal Component Analysis (PCA) Results with 2D & 3D graphs:





The project has over 11 features. PCA transformed variables into a new set of variables, which was a linear combination of the original variables.

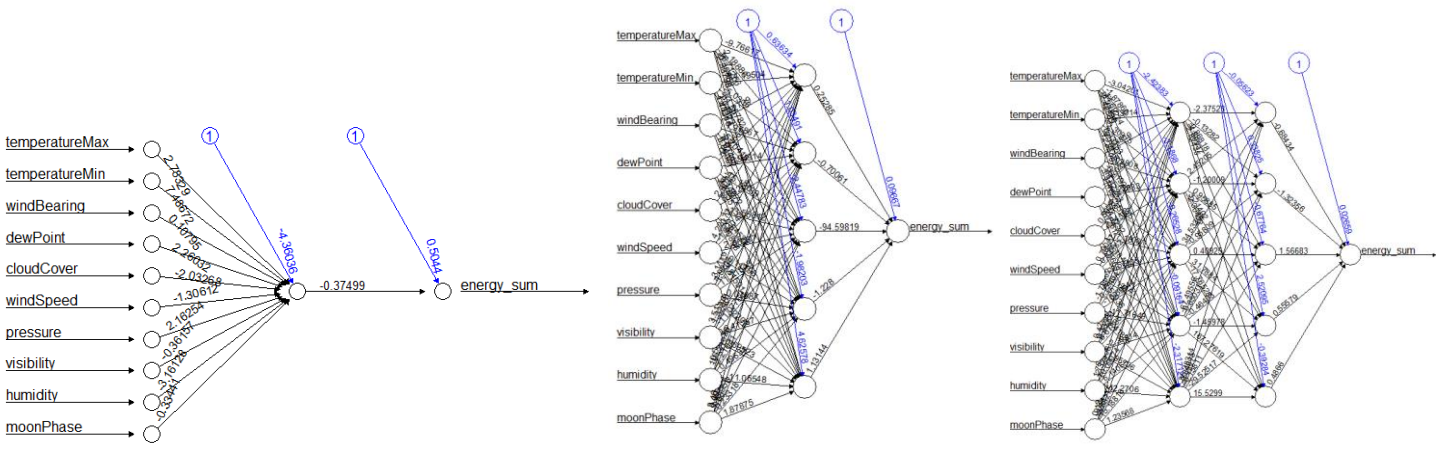
PCA is deterministic. So, the correct answer is guaranteed. It makes data plottable on a 2D graph.

PCA is a popular technique to transform a dataset onto a lower dimensional subspace for visualisation and further exploration. PC are Eigen-pairs. They describe the direction in the original feature space with the greatest variance.

House 59, 51 and 50: PCA sensitive to outliers and may cause wrong eigendirection.

Neural Networks

An artificial neural network was used. This kind of network has ten input layers, one output layer, and a number of hidden layers. The nodes in each layer are called neurons which perform non-linear calculations.



1H: Deep Learning with TensorFlow Long Short-Term Memory (LSTM) Neural Network for Stock Market Predictions with Python.

Deep learning is a subset of machine learning in artificial intelligence (AI) that is capable of learning from data.

Figure 1 shows Predicted Stock Prices (red) and Actual Stock Prices (blue)

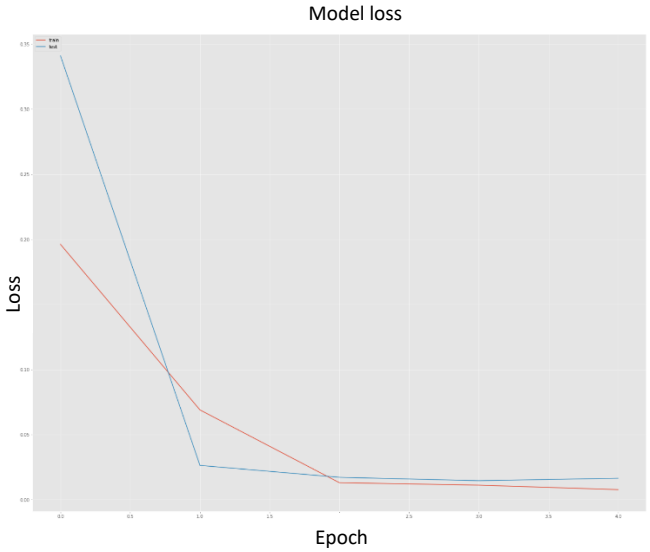


Figure 2 shows that a model is correct or just right training error (red line) slightly lower than test error (blue line).

PS: In deep learning, a loss function that quantifies the badness of our model, a model that is underfit will have high training and high testing error while an overfit model will have extremely low training error but a high testing error.

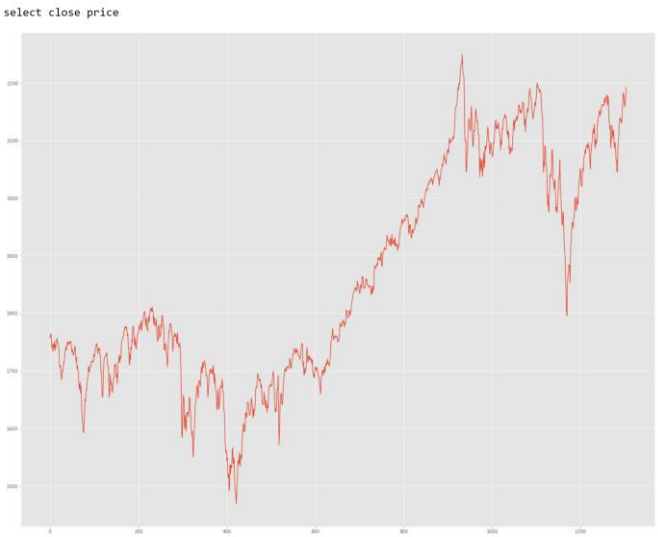


Figure 3: Plot of 'Close' for Global Equity Income sector price history