Big Data & Predictive Analytics
Lecture Notes on Data munging - Exploration and Visualisation

School of Informatics
University of Leicester

Data rarely arrives in the form needed for analysis. Data munging is typically the most time consuming part of a data science project. It is an iterative process often discovered using appropriate visualisation tools in order to fix modeling problems.

Let us first review what is meant by data analytics. Data analytics consists in transforming data into insights so as to make business decisions. Data represents unprocessed facts about the world. By processing data, we gain information about the world. The gained information forms a set of beliefs or knowledge about the world over time. The set of beliefs resulting from this process can guide decision-making on what actions to take for specific objectives. Now that we have the setting, let us explore the tools and methods that a data analyst can employ in order to gain these insights. In these notes, we will focus on exploratory analysis. To give a simple illustration of this type of analysis, consider a time series. Finding the average or a period is a descriptive task; plotting the data is an exploratory task; and calculating the value of a future data point is a predictive task.

1 Exploratory Analysis

Exploratory Data Analysis (EDA) consists in using graphics in order to inspect and visualise data. The motivation comes from the idea that graphs enables us to identify patterns that may be unusual and to view the overall picture of a given dataset. In general, EDA should follow from the descriptive analysis of the dataset.

1.1 Covariance and correlation

Correlation is a key tool into predictive modeling as it exposes relationships between two variables of interests x, y in the form

$$y = f(x).$$

The function f can be linear ($y = ax + b$)) or exponential ($y = \exp(ax) + b$). A correlation between x and y means that a change on the value of x impacts the value of y. For example,

when x increases, y increases as well. The degree of correlation between the two variables x, y, for a given set of n couples of data points, is calculated as follows:

$$corr(x, y) = \frac{cov(x, y)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

wherein

$$cov(x, y) = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

is the covariance between x and y and $\bar{x}, \bar{y}$ represent respectively the mean of the sample data points for x, y. The correlation can be positive or negative. A higher correlation indicates a stronger relationship between the two variables.

If we use the above formula to calculate the correlation between the variables `TV` and `Sales`, we get the coefficient 0.78. However the correlation coefficient between `Newspaper` and `Sales` gives 0.23. We therefore see a strong correlation between advertising on TV and sales. Figure 1 exposes the correlations between TV & Sales and Newspaper & Sales.

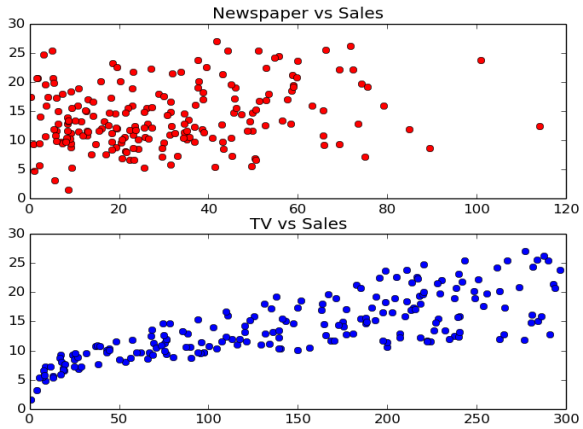

Figure 1: An example of strong and weak correlation between two variables

1.2 Big data and spurious correlation

In the era of big data, spurious correlation on variables for some snapshot data at some given time interval can be observed. If variables are correlated, they should be naturally related all the time by common sense. Figure 2 (taken from a slide of Bill Howe) illustrates this problem.

2 Data cleaning and Transformation

As we have seen in previous lectures, data can have missing or repeated values in various rows and need cleaning. To perform such a task, we need to be able to access specific rows or columns of the dataset. Such an operation is called *subsetting*. We can also create new columns; refer the cheat sheets 1 and 2 for the Python functions to carry out these tasks.

Figure 2: An example of spurious correlation

2.1 Grouping the data – aggregation, filtering, and transformation

In this section, you will learn how to aggregate data over categorical variables. This is a very common practice when the data consists of categorical variables. This analysis enables us to conduct a category-wise analysis and take further decisions regarding the modeling

2.2 Cleaning outliers and errors

Outliers are the points out of the league of the other points in the dataset. They can be spotted using different plots since the outliers can be easily will lay significantly away from the other data points. The outliers, see Figure 3, need to be removed before using the dataset for modeling as they can distort the model and reduce its efficacy even if they are less in number. For example, a 1% outlier data is capable of distorting a model.

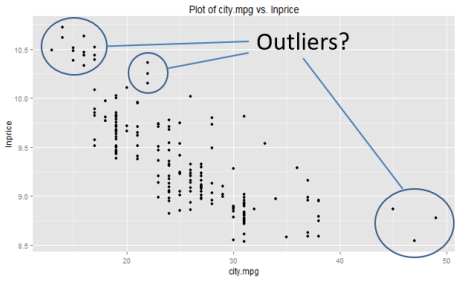

Figure 3: Spotting Outliers

The following are ways in which one can spot outliers:

- Plotting scatter plots of the concerned variables to visualise their relationships
- Boxplots are potent tools to spot outliers in a distribution. Any value 1.5*IQR below the 1st quartile and 1.5*IQR above the 1st quartile can be classified as an outlier.

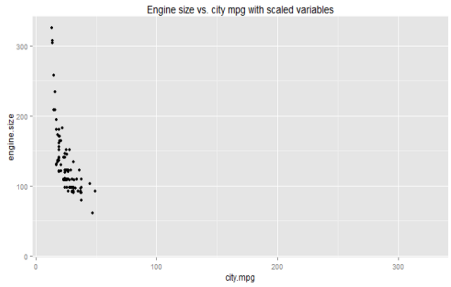

Figure 4: An example of scaling

- We can also deal with outliers a posteriori by calculating the error (the difference between the actual value and the value predicted from the model) and set a cut-off for the error. Anything outside this cut-off will be an outlier.

2.3 Scaling data

The plots of relationships between variables can sometimes be misleading if the variables are not in the same numerical range. This is due to the fact that the variables with larger numerical values can introduce bias, see for example Figure 4, by showing a bigger impact than the variables with smaller numerical values.

In such a case, we need to *scale* the values of the variables in order to bring them in the same numerical range. The scaling entails the *numerical transformation* of each value so that they come in the same numerical range. Note that scaling data has to happen after removing outliers. *Normalization* is an example of scaling wherein values are scaled down to be between 0 and 1. For example the distribution of a random variable can be scaled to zero mean and unit variance. Other scaling methods includes min-max as follows:

```
df_norm = (df-df.min())/(df.max()-df.min())
```

In the above code, `df` is a Python data frame and the transformation is normalizing the whole dataset by scaling down each column value between 0 and 1.

2.4 Data Visualisation

Descriptive analysis is only a first step in exploring a dataset since different samples of different shapes and distributions can have the same summary statistics. To further illustrate the importance of EDA, let us consider the Anscombe's quartet as shown in Figure 5 taken from `https://en.wikipedia.org/wiki/Anscombe's_quartet`.

Anscombe's argument on using exploring data by using graphs was that statistical calculations rely upon assumptions that need be checked and validated in order to gain confidence on the results of those calculations and graphs are valuable tools for that purpose. One of the main task in data analytics is *data cleaning*. Data cleaning is a data preparation process wherein we identify missing values, outliers, or other anomalies in a way that if data rest unlean, it may lead to one or more of the followings:

- incorrect analysis of the relationships between variables,

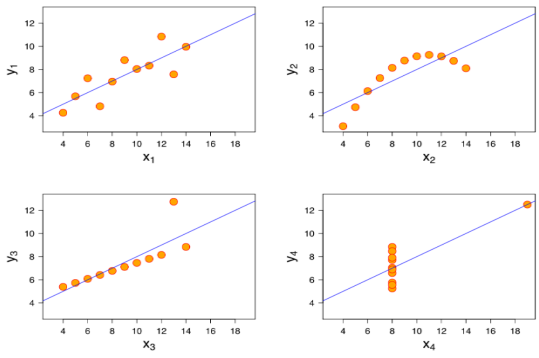

Figure 5: All four sets have the same mean, variance, correlation and regression line, but vary considerably when graphed

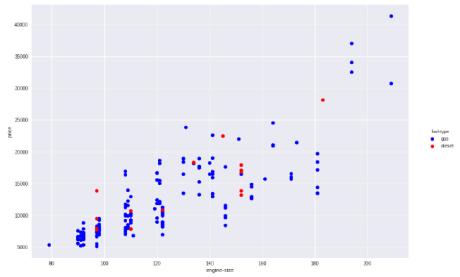

Figure 6: An plot using aesthetics to hihlight auto prices for various categories

- incorrect interpretation of the data,
- incorrect inference from the data.

These anomalies on data must be treated before carrying out further analysis and that is why EDA is an important step of big data analytics. Moreover, graphs can help us to match our expectations of what clean data or 'good' behaviour between two variables might look like. For example, plotting data may lead us to observe a Gaussian distribution of a variable.

To visualise a dataset, we can start from basic plots such as a **scatter plot, pie chart, histogram, box-plot**, etc. These plots will help us to identify complex relationships in the dataset. Data exploration requires multiple views, which reveal different aspects of the relationships. Often, we need different plots using various aesthetics such as color, shape, or size to better highlight the relationships between variables in the dataset. For example, we can use color to highlight price differences for various categories of automobiles as shown in Figure 6.

A scatter plot matrix, as shown in Figure 7, enables us to highlight possible correlations between a set of numerical variables. This is a useful tool to explore how different input variables impact an output variable.

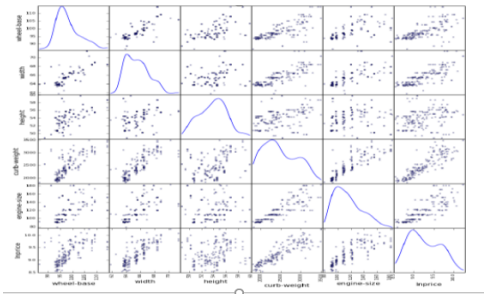

Figure 7: A scatter plot matrix

- Many features show significant correlation, such as `wheel-base, width and curb-weight, or curb-weight and engine-size and width`.
- A number of features show a strong relationship with the variable `lnprice`, such as `wheel-base, engine-size, curb-weight, width`.
- The column height does not show correlation with `lnprice`; we can say that `wheel-base, engine-size, curb-weight, width` can be good predictors for the price.

Last Updated January 11, 2021 by Emmanuel Tadjouddine