Module 37 Data Analytics and Visualisation

Module title	Data Analytics and Visualisation
Module NFQ level (only if an NFQ level can be demonstrated)	8
Module number/reference	BSCH-DAV
Parent programme(s)	Bachelor of Science (Honours) in Computing Science
Stage of parent programme	Award stage
Semester (semester1/semester2 if applicable)	Semester 2
Module credit units (FET/HET/ECTS)	ECTS
Module credit number of units	5
List the teaching and learning modes	Direct, Blended
Entry requirements (statement of knowledge, skill and competence)	Learners must have achieved programme entry requirements.
Pre-requisite module titles	BSCH-FC, BSCH-LA, BSCH-PAS
Co-requisite module titles	None
Is this a capstone module? (Yes or No)	No
Specification of the qualifications (academic, pedagogical and professional/occupational) and experience required of staff (staff includes workplace personnel who are responsible for learners such as apprentices, trainees and learners in clinical placements)	Qualified to as least a Bachelor of Science (Honours) level in Computer Science or equivalent and with a Certificate in Training and Education (30 ECTS at level 9 on the NFQ) or equivalent.
Maximum number of learners per centre (or instance of the module)	60
Duration of the module	One Academic Semester, 12 weeks teaching
Average (over the duration of the module) of the contact hours per week	3
Module-specific physical resources and support required per centre (or instance of the module)	One class room with capacity for 60 learners along with one computer lab with capacity for 25 learners for each group of 25 learners

Analysis of required learning effort		
	Minimum ratio teacher / learner	**Hours**
Effort while in contact with staff		
Classroom and demonstrations	1:60	18
Monitoring and small-group teaching	1:25	18
Other (specify)		
Independent Learning		
Directed e-learning		
Independent Learning		54
Other hours (worksheets and assignments)		35
Work-based learning – learning effort		
Total Effort		125

Allocation of marks (within the module)					
	Continuous assessment	**Supervised project**	**Proctored practical examination**	**Proctored written examination**	**Total**
Percentage contribution	60%			40%	100%

Module aims and objectives

This module aims to provide learners with the foundations necessary for understanding and extending the current state of the art in data analytics and visualisation. In the recent years, the world has been flooded with ever-increasing amounts of data. We are required to possess ourselves with data analytics techniques to better understand this data and represent it meaningfully. Visualization provides one means of tackling data overload, as a well-designed visual representation of the data can help in improving comprehension, memory, and decision making. In this module the learners study techniques and algorithms for carrying basic data analytics (using the statistics) and creating effective visualizations based on well-established principles from graphic design, visual art, and perceptual psychology. The module is targeted both towards learners interested in data analytics and information visualisation.

Minimum intended module learning outcomes

On successful completion of this module, the learner will be able to:

1. Apply data cleansing and statistical operations on datasets to address the issues of data quality and dimensionality.
2. Select and apply suitable statistical methods and analyses techniques for data of various structure and content and present summary statistics.
3. Apply the key techniques and theory used in information visualisation, including data models, graphical perception and techniques for visual encoding and interaction.

4. Design, develop, and implement various techniques used for data visualization to effectively communicate information.

Rationale for inclusion of the module in the programme and its contribution to the overall MIPLOs

The module enables learners to understand concepts which form the basis of data analytics and visualisation techniques. It enables the learners to understand current techniques and develop effective data visualisations to communicate information.

Appendix 1 of the programme document maps MIPLOs to the modules through which they are delivered.

Information provided to learners about the module

Learners receive a programme handbook to include module descriptor, module learning outcomes (MIMLO), class plan, assignment briefs, assessment strategy, and reading materials.

Module content, organisation and structure

Data Analytics

- Overview of Data Analysis and Visualisation
    - Why do we need to analyse data?
    - Types of data,
    - Basics of information visualisation,
    - Opportunities and Challenges.
- Data Cleaning and Transformation
    - Handling noisy/dirty data
    - Dimensionality reductions (e.g. PCA, LDA)
- Basics of Statistics
    - Qualitative and Quantitative data summaries,
    - Statistical attributes (e.g. variance, standard deviation, etc.)
    - Normal distribution and Sampling
- Hypothesis Testing
    - Statistical Inference,
    - Stating Hypotheses,
    - Test Statistics and P-Values
    - Evaluating Hypotheses
    - Significance tests and Confidence Intervals,
    - Types of hypothesis tests and their applications (e.g. one sample, two sample.),
    - ANOVA

- Overview of Machine Learning algorithms
    - Pick any two/three basic algorithms,
    - Apply algorithms on sample datasets,
    - Perform cross-validation, etc.

Data Visualisation

- Exploratory Data Analysis
    - Types of charts/plots
    - Handling outliers, etc.
- Understanding association between two quantitative variables
    - Correlation,
    - Linear Regression,
    - Scatterplots
- Visualisation Techniques and Applications
    - Planning and designing visualisations,
    - Current tools for data visualisations (R, D3, etc.),
    - Techniques for:
        - Spatial data,
        - Geospatial data,
        - Multivariate data,
        - Trees,
        - Graphs,
        - Clusters, etc.

Module teaching and learning (including formative assessment) strategy

The module is taught using a combination of lectures, demonstrations, and tutorials. The demonstrations and tutorials focus on getting learners up to standard in practical application development. The lectures supply the necessary theoretical background. In a fast-changing technology field, learners are expected under guidance to engage in research in relation to the different technologies and products available.

The module has both a continuous assessment element and a final examination. It requires the learner to show an understanding of the current techniques in the area of data analytics and visualisation. The weekly lab work allows the learners to gain practical hands-on experience of cleaning, transforming, analysing, and visualising the datasets. The final examination assesses their understanding to the theoretical concepts of the module.

Timetabling, learner effort and credit

The module is timetabled as one 1.5-hour lecture and one 1.5-hour labs per week.

The number of 5 ECTS credits assigned to this module is our assessment of the amount of learner effort required. Continuous assessment spreads the learner effort to focus on weekly worksheets before integrating all steps into the overall process of data analysis and information visualisation.

There are 36 contact hours made up of 12 lectures delivered over 12 weeks with classes taking place in a classroom. There are also 12 lab sessions delivered over 12 weeks taking place in a fully equipped computer lab. The learner will need 54 hours of independent effort to further develop the skills and knowledge gained through the contact hours. An additional 35 hours are set aside for learners to work on worksheets and assignment that must be completed for the module as a part of the continuous assessment.

The team believes that 125 hours of learner effort are required by learners to achieve the MIMLOs and justify the award of 5 ECTS credits at this stage of the programme.

Work-based learning and practice-placement
There is no work based learning or practice placement involved in the module.

E-learning
The college VLE is used to disseminate notes, advice, and online resources to support the learners. The learners are also given access to Lynda.com as a resource for reference.

Module physical resource requirements
Requirements are for a classroom for 60 learners equipped with a projector, and a 25 seater computer lab for practical sessions with access to recommended data analysis and visualisation toolsets (this decision may be left on the lecturer to pick the current toolsets, possibly R/Python and D3).

Reading lists and other information resources
Recommended Text
Lander J. (2017) *R for Everyone: Advanced Analytics and Graphics*, Addison Wesley Data & Analytic Series Upper Saddle River: Pearson Education

Secondary Reading
Lakin S., (2011) *How to Use Statistics*, Harlow: Pearson Education

Chang, W., (2013) R graphics cookbook, Sebastopol: O'Reilly

Matthew W., Georges G., Daniel K., (2015) *Interactive Data Visualization: Foundations, Techniques, and Applications*. Boca Raton: CRC Press

Recent top-quality journal papers on Data Analytics and Visualisation Techniques as well as documentation from Toolsets

Specifications for module staffing requirements

For each instance of the module, one lecturer qualified to at least Bachelor of Science (Honours) in Computer Science or equivalent, and with a Certificate in Training and Education (30 ECTS at level 9 on the NFQ) or equivalent.. Industry experience would be a benefit but is not a requirement.

Learners also benefit from the support of the programme director, programme administrator, learner representative and the Student Union and Counselling Service.

Module Assessment Strategy

The assignments constitute the overall grade achieved, and are based on each individual learner's work. The continuous assessments provide for ongoing feedback to the learner and relates to the module curriculum.

No.	Description	MIMLOs	Weighting
1	Lab worksheets Learners must be submitting lab worksheet solutions on weekly bases (of which at least 5 must be graded). This work involves the learners to apply data analytics and visualisation techniques they learn during the course of the module.	1-6	40%
2	Assignment This is a major assignment which should be handed over to learners between after week 8 of the module. Learners will be given a dataset to analyse and build a visualisation to communicate insights.	1-6	20%
2	Written exam that tests the theoretical aspects of the module.	2 - 4	40%

All repeat work is capped at 40%.

Sample assessment materials

Note: All assignment briefs are subject to change in order to maintain current content.

1 Description

In this lab we are going to explore the way of visualising and summarising data in R.

2 Data set

This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica. iris is a data frame with 150 cases (rows) and 5 variables (columns) named `Sepal.Length`, `Sepal.Width`, `Petal.Length`, `Petal.Width`, and `Species`. Data set can be loaded in R by running:

```
library(datasets)
```

Then it is available under the name `iris` (type it and run to see what it contains). Check the help to get more information about the data set.

Previously we have seen some data types, especially different types of vectors. Here, the data set is returned as a data frame structure which is a structure that enables you to store different variables of the same data, together. If you would like to store all students information, you could have a data frame with columns such as name, date of birth, address etc, and in each row we would store student's records (more on this during the lecture). You can check if Iris data set is a data frame:

```
is.data.frame(iris)
```

You can access data frames in the same way you would access an array - like in the first lab. You can just call `iris[1, 2]` to get the second column of the first row. To retrieve the whole column or row, just leave the index empty, like `iris[, 2]` to get the whole second column. It is also possible to access a column by the name - `iris$Sepal.Length` will return column containing mpg. This one:

```
iris[1:10,]$Sepal.Length
```

takes first 10 rows and then it returns only the column `Sepal.Length`.

3 Visualisation

Before we start exploring different ways of visualising data, there is a useful R command that aggregates categorical data:

```
table(iris$Species)
```

3.1 Pie Chart

Pie charts are very useful when you would like to present some summaries of categorical data. In R you create them in the following way:

```
slices <- c(10, 12, 4, 16, 8)
lbls <- c("US", "UK", "Australia", "Germany", "France")
pie(slices, labels = lbls, main="Pie Chart of Countries")
```

3.2 Bar Chart

Similar to pie charts - we use them if we have categorical data and we would like to visualise how many of observations we have in each category.

```
counts <- table(iris$Species)
barplot(counts, main="Species Distribution", xlab="Species")
```

3.3 Histogram

Useful to summarise numerical data by aggregating them into buckets and counting.

```
hist(iris$Sepal.Width, main="Histogram of sepal width", xlab="Sepal Width")
```

3.4 Scatter Plot

Scatter plot displays values for typically two variables for a set of data. By colouring the points or changing the shape it is possible to increase the number of variables.

```
plot(iris$Sepal.Length, iris$Sepal.Width,
     xlab="Sepal Length", ylab="Sepal Width", pch=19)
```

4 Summaries

We also have very easy and useful functions to summarise the data. Consider the following data (discussed in last lecture).

```
x <- c(414, 123, 72, 79, 66, 84, 169, 144, 102, 110, 162) # 11 elements
y <- c(414, 123, 72, 79, 66, 84, 169, 144, 102, 110)      # 10 elements
```

4.1 Median

Median is the centre value of a data set. It splits the data set into two halves. Check with the results from the lecture slides.

```
median(x)
median(y)
```

4.2 Mean

Mean is an arithmetic average of the data set:

```
mean(x)
mean(y)
```

4.3 Quantiles

To get the quantiles of the data set run:

```
quantile(x, type=1)
quantile(y, type=1)
```

Check different types.

4.4 Head

Probably viewing the whole data set is not sufficient. Instead it is possible to view only top `n` rows of the data set:

```
head(iris)
```

There is also `tail` function. Try it to see what it does.

4.5 Summary

Instead of checking min, max, median, mean, etc. separately for each column, it can be done all together:

```
summary(iris)
```

4.6 Standard Deviation

Standard deviation is a measure that is used to quantify the amount of dispersion of dataset values:

```
library(stats)
sd(x)
sd(y)
```

For calculating variance type `var(x)` and `var(y)`.

Lab worksheet 2

1 Description

Your task for today is to practice the concepts of correlation and regression on different datasets - both manually and using the R functions.

2 Tasks

2.1 Task 1

Explore `Boston` dataset presented in the lecture. After you load `datasets` package, you can find it in the variable `Boston`. Observe the relationships between different variables. Try this:

- scatter plot of two variables - function `plot(...)`,
- you can add a regression line after you draw a plot - function `abline(lm(y x))`,
- correlation of different variables - function `cor(...)`.

2.2 Task 2

You have the following data set, manually calculate:

- correlation coefficient,
- regression function.

Compare your results with R results. You can get the regression line formula with the `lm(...)` function.

x	4	7	5	6	1	5	9	10	10	3
y	33	37	34	32	32	38	43	37	40	33

2.3 Task 4

The following data set gives the average heights and weights for American women aged 30-39 (source: The World Almanac and Book of Facts, 1975).

x	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72
y	115	117	120	123	126	129	132	135	139	142	146	150	154	159	164

What is the estimated regression line? Compute Coefficient of Determination. What else can you tell about this data set?

2.4 Task 5

Explore the data set `mtcars`. Perform the analysis of basic statistics for selected variables - these should be numerical variables. Analyse the relationship between the variables.

Lab worksheet 3

1 Description

Your task for today is to practice the concepts of Geoms and Aesthetics in ggplot library. All graphics must be drawn using ggplot2 package.

2 Tasks

You will be using the `Boston` dataset. After you load `MASS` package, you can find it in the variable `Boston`. Now draw graphics as follows:

1. Using the Boston dataset generate a scatter plot of rm versus medv.
2. Using the Boston dataset generate a scatter plot of crim versus medv.
3. Extending the plot from previous task, add a colour scale to the scatterplot based on the age variable.
4. Using an additional geom, add an extra layer of a fit line to the solution from task 3.
5. Change the method of geom_smooth() in previous task to lm. Observe the difference.
6. Following from task 3, a grid of plots (using facets) by using appropriate variables.
7. Generate a histogram of medv with each bar coloured grey.
8. Practice other geoms and aesthetics, use ggplot2's reference documentation available at `ggplot2.org`.

Lab worksheet 4

Preparing Dataset

WHO.csv (uploaded on moodle) is a data set of tuberculosis (TB) reported between 1995 and 2013 sorted by country, age and gender. The data comes from 2013 WHO Global Tuberculosis Report.

The data set is messy, especially the columns. The most unique feature of this data set is its coding system. Columns five through sixty encode four separate pieces of information in their column names:

1. The first three letters of each column denote whether the column contains new or old cases of TB. In this data set, each column contains new cases.

2. The next two letters describe the type of case being counted. We will treat each of these as a separate variable:

   - `rel` - stands for cases of relapse
   - `ep` - stands for cases of extra-pulmonary TB
   - `sn` - stands for cases of pulmonary TB that could not be diagnosed by a pulmonary smear
   - `sp` - stands for cases of pulmonary TB that could be diagnosed by a pulmonary smear

3. The sixth letter describes the sex of TB patients: `m` for males and `f` for females.

4. The remaining numbers describe the age group of patients:

- `014` - 0 to 14
- `1524` - 15 to 24
- `2534` - 25 to 34
- `3544` - 35 to 44
- `4554` - 45 to 54
- `5564` - 55 to 64
- `65` - 65 or older

Read the `tidy data article` uploaded on moodle. Make the WHO.csv dataset tidy in order to prepare it for performing analysis.

Lab worksheet 5

For this assignment please create a visualisation based on multi-dimensional data using a parallel coordinate visualisation. Please submit a .zip file to Moodle including all your Processing Sketch code and data sufficient to run your assignment. In addition please submit 3 samples screen grabs from your visualisation along with a document about your visualisation. This document (minimum 1,500 words) should describe the end to end process in terms of the data you acquired, parsing, filtering, mining, representation, refinement and interaction.

You must select a data source which relates to people. For example, details about members of a football team (or all teams), all players of international cricket, all CEOs in the USA, all your friends in facebook, people over 55 working in IBM etc.

However, also describe the conceptual process you undertook to identify interesting user tasks prior to deciding on which data sources to use etc. Please ensure you select a data source with at least 7 dimensions and with at least 40 cases (entities). Next, detail the design process you undertook to get the data into the correct format etc. For this assignment it's particularly important to highlight if your visualization reveals any interesting patterns in the data.

To complete this assignment basics (0 - 55%)

- Identify an interesting set of user tasks which rely on a multi-dimensional data source to answer
- Find or identify an interesting source of data with at least 7 dimensions and 40 cases
- Devise a process (whether manual or computational) which allows your processing sketch to access the data source (eg. collect the data and place it into .tsv files OR write processing code to access the required data source (or sources) along with normalizing the data so each dimension fits on a uniform 0..1 range
- Draw one vertical line per dimension (min 7 lines) with labels for each dimension showing the min and max values for each dimension
- Draw a poly-line per data case
- By using animation, colour, icons, transparency etc. develop an interaction technique to allow a user to select a range from within the data (eg. date, data type etc.)

Extended (55% - 70%)

- Ensure you support mouse overs so that a user can inspect the individual cases (eg. select a sin- gle line to see just one poly-line (data case) highlighted)
- Encode non-quantitative dimensions without use of a vertical line. eg. gender is a nominal value, in a PCV all polylines will either intercept a male or female point on a gender line. In- stead encode this dimension with colour, texture, dotted line etc.

Exceptional (70% - 100%)

- By using animation, colour, icons, transparency etc. develop an interaction technique to allow a user to select a range from within the cases (a mouse sweep over or other)

- Allow the user to select only a range of values within a particular dimension which are in scope (those in scope are shown in colour, those out of range are "greyed-out")
- Use an aspect of pre-attentive features to highlight an aspect of your visualisation upon first viewing it or changing a range of values.

Module Assignment

Overview:

The diamonds dataset contains the prices and other attributes of almost 54,000 diamonds. Your task is to perform an EDA and hypotheses testing on the dataset. Submit a report summarising your findings together with the source code. This assignment is part of the continuous assessment and worth **20%** of the module grade.

Tasks:

Your task is to perform EDA and calculate the strength of relationships between the variables of the dataset. Consider below as a guideline:

(a) Focus your analysis on the price variable:
   a) Show the histogram of the price variable. Describe it briefly. Include summary statistics like mean, median, and variance.
   b) Group diamonds by some price ranges (like low, medium, high, etc.) and summarize those groups separately.
   c) Explore prices for different cut types. You might want to use the boxplot.
   d) How different attributes are correlated with the price? Which 2 are correlated the most?

(b) List the frequencies of diamonds for various cuts and clarity levels. Create 2 scatter plots and colour the diamonds price by clarity and cuts.

(c) Now focus your analysis on the carat, depth, table and dimensions (x, y, z) variables:
   a) Compute a volume variable from x, y, z - add it to the dataset. Plot it against the price. Describe your findings.
   b) Are the carat and volume attributes correlated? Is that a strong relationship? Draw a plot with regression line.
   c) Explore the relationships between table and depth variables. Now explore relationships between table and rest of other variables. Compute correlations and describe your findings.

(d) You should now state at least 4 different hypotheses, each to test different data (so not all hypotheses should be checking the same statement just on different variables). Remember that there are different types of tests and you should use as many as you can (given if they are valid and make sense). Your ultimate goal is to report some findings. You should also prove that these findings are statistically correct. Take the below points as hints but do not limit yourself to these:
   - Look at different plots you have created during exploratory analysis. What conclusions can be drawn based on these? These could become your hypotheses.
   - If you focus on one attribute, what is your intuition about the distribution that could explain such results? You can check and measure how well the data fits some distribution.

GRIFFFITH COLLEGE DUBLIN

QUALITY AND QUALIFICATIONS IRELAND
EXAMINATION

DATA ANALYTICS & VISUALIZATION

Lecturer(s): …
External Examiner(s): …

Date: … **Time: … – …**

THIS PAPER CONSISTS OF FOUR QUESTIONS
ALL FOUR QUESTIONS TO BE ATTEMPTED
ALL QUESTIONS CARRY EQUAL MARKS

Question 1

a) What is correlation coefficient? How do you interpret the following correlation measures for two variables x and y?

   i. 0.85

   ii. -0.91

(10 marks)

b) Describe the difference between correlation and regression.

(5 marks)

c) Provide a list of the classes of objects in R.

(5 marks)

d) What is the purpose of data frames in R? How do they differ from matrices?

(5 marks)

Total (25 marks)

Question 2

a) Provide the description of Central Limit Theorem.

(5 marks)

b) What is standard error of sample means?

(5 marks)

c) Two diets (Low-Fat & Other) were compared for their weight loss (or gain) over 3 weeks. Two independent samples of 100 overweight people were randomly selected. One group was put on the low-fat diet and the second group on the other diet. Summary results were as follows:

	Low-Fat Diet	Other Diet
Sample size	100	100
Sample mean	9.3	7.4
Sample variance	22.4	16.3

Using a 0.05 significance level, is there evidence that the two diets differ in their mean weight loss over 3 weeks?

(15 marks)

Total (25 marks)

Question 3

a) Discuss, in your own words, the significance and importance of Data Cleansing to the processes of Data Mining and Data Visualisation. In your answer, you should highlight why Data Cleansing is needed in most real-world applications, explain some of the steps in a Data Cleansing process, cross-reference it to relevant data management strategies, and critically assess the advantages and disadvantages of specific Data Cleansing approaches. Total (25 marks)

Question 4

a) Discuss the psychological aspects of visual perception as they pertain to the design requirements for data visualisation with reference to concepts which may include, but are not limited to the following aspects of perception:

1. Perceived hierarchy of features
2. Preattentive and attentive processing
3. Colour choices
4. Visual inferences

(15 marks)

b) Describe the visualization techniques used for Spatial and multivariate datasets.

(10 marks)