Data Science and Big Data
2018 Post Graduate Masters Program

Delivered exclusively at:

Data Science Academy
Matrix TRC
Technical information: Mr. Walid Semaan
Phone: 961 3 636 208
Website: www.matrixtrc.com

Center for Learning & Development
American University of Science & Technology
Beirut – Ashrafieh – Alfred Naccache Avenue
Registration information: Mrs. Zeina Nsouli
Phone: 961 1 218716/7 Ext. 333
Website: www.aust.edu.lb
Data Science Diploma

1. Data Analysis Foundations
   1.1. Data Collection  1.2. Descriptive Stat  1.3. Estimations

2. Profiling Techniques: Basic & Intermediate
   2.1. One Group  2.2. Two Groups  2.3. Dep. & N.P

3. Profiling Techniques: Advanced & Expert
   3.1. Multiple Groups  3.2. Single Regressions  3.3. Dep. & N.P

4. Machine Learning (Unsupervised)

5. Machine Learning (Supervised)
   5.1. Multiple & Logistic Regressions  5.2. Predictive Models  5.3. E - Miner

6. Q – F – E
   6.1. QC Measurements  6.2. Forecasting  6.3. Epidemiology

7. Data Management
   7.1. Business Intelligence  7.2. Oracle  7.2. SAS Data Flux

8. Statistical Programming
   8.1. R  8.2. Python  8.3. SAS Studio

9. IoT and Data Information Security
   9.1. Internet of Things  9.2 – 9.3 Data and Cybersecurity

10. Big Data
    10.1. Technologies  10.2. Analytics  10.3. Project
Matrix TRC Data Science Academy, in partnership with SAS technologies, offers a complete flexible post graduate “masters” diploma program (450h into 10 courses / 30 credits) that encompasses exhaustive Data Science and Big Data topics, that:

- Covers most needed expertise in the emerging strategic numerical economy
- Puts all Machine Learning algorithms into practical applications
- Helps companies implement a vital internal Data Center for scientific information
- Brings a synergy between the main information system stakeholders
- Joins a group of local and international experts in each field
- Supports participants with latest analytical and complete IT technologies
- Tackles comprehensively the whole IoT and Big Data components
- Brings highest ROI with participants applying 6 projects from within their businesses.
- SAS complete package set of more than 20 analytical tools ($40,000) ... free of charge!!
He is the founder and president of Matrix Training Research & Consulting, a company specialized in Data Science. He graduated in Engineering from Ecole Supérieure d’Ingénieurs à Beyrouth. He holds a degree in finance and marketing from the Ecole Supérieure de Commerce de Paris (ESCP) and an MBA from the University Paris-Dauphine-Sorbonne in Paris. He is the creator and architect of the automated artificial intelligence behind “Triple One Analytics”, winner of the Best Innovative ICT Project at the Arab Golden Chip Award 2011, and architect of Matrix TRC “Data Science Academy”. Walid is the main instructor in data analytics and machine learning.

Murray de Villiers is a Senior Manager for SAS’ Global Education Practice Academic Program. He is responsible for the establishment and development of undergraduate and post-graduate SAS academic data science and advanced analytics programs at universities in Southern- and Eastern Europe, Middle East and Africa, as well as in other emerging economies. This includes the establishment of SAS Academic Flagship Universities, SAS Academic Thought Leaders and the establishment of various inter-university and industry-university collaboration programs, empowering academia, students and learners with appropriate skills for lucrative careers. Prior to taking this position, Murray has held various roles in SAS, including leading Marketing, Education, IT, Sales and SAS Partner enterprise system sales efforts in South Africa.
Anthony Franklin is an expert statistician with experience in many software languages including open source technologies. Anthony conducted his doctoral studies at North Carolina State University and completed his graduate assistantship with the NCSU Institute of Advanced Analytics. Anthony resides in North Carolina and currently is a Sr. Analytics Architect at SAS Institute with experience in risk management, hospitality and sports analytics.

Riaan de Jongh is the Director of the Centre for Business Mathematics and Informatics® at the North-West University (NWU). He has received awards for the best published research in a particular year from the Operational Research Society of South Africa (ORSSA), the South African Statistical Association (SASA), the Actuarial Society of South Africa (ASSA) and Risk.net. He is a Past President and Fellow of SASA, received a Thought Leader Award from SASA and a National Recognition award from ORSSA. He also served on the boards of several local and global associations. He is the Vice-Chairperson of the NWU Pension Fund and served on various committees for the Global Association of Risk Professionals, the Professional Risk Managers’ International Association, the International Society for Business and Industrial Statistics, and the Global Certified Enterprise Risk Actuary (CERA) Review Panel.
Courses & Schedule

C1. Data Analysis Foundations
C2. Profiling Techniques (Basic – Int.)
C4. Machine Learning 1 (Unsupervised)
C5. Machine Learning 2 (Supervised)
C6. Quality / Forecasting / Epidemiology
C7. Business Intelligence
C8. Statistical Programming: R - Python
C9. IoT and Data & Cybersecurity
C10. Big Data

Project 1
Project 2
Project 3
Project 4
Project 5
Final Project
Concerned Industries

Data Science is the rising prerequisite knowledge for all companies desiring to extract most of the information from their data. It is no longer valid to make decisions based on shallow reading of the increasing amount of data pouring from everywhere. A complete and exhaustive scientific approach is becoming a must for successful institutions.

Below is a list of the main industries that are involved in meeting with the data market requests.
Who Should Attend?

This program is primarily aimed to serious people who aspire to learn state-of-the art Data Science techniques, and how it can be applied to solve data and business issues. It specifically suites practitioners involved mainly, but not exclusively, to the below fields. It is also addressed to anyone who wants to "get connected" with latest data management and analytical solutions, how different tools can be used, and what factors should be considered when planning for any “Big Data” project.

Prerequisites
Previous knowledge or experience in statistics, data mining, or artificial intelligence algorithms is not required at all. However, a serious interest and continuous active participation is a must for a successful completion of the diploma.
Clients and Participants
Some Testimonials

Abed Al Rahman Itani (Azadea)
IT Business Solution Specialist
The diploma was very inspirational, energizing, bringing lots of ideas. I learnt a lot. I am inspired to rethink the way I plan and present my work. A very high pedagogical standard. Well done!

Ali Zaiter (Edumotion)
Senior Data Analyst/Methodologist
The program has widened and sharpened my analytical skills. It was complete and very easy to master through very knowledgeable instructors. DS & BD program has opened doors for to attend data science challenge weeks in top prestigious universities worldwide and supported fully my advanced post graduate studies. It is the first program of its kind in the MENA region that provides the necessary statistical and technical tools for big data mining and data analysis in a very well-structured manner. I highly recommend this program for those who are willing to gain experience with complex data analysis and machine learning algorithms.

Radwan M. Hadwan (Bank Audi SAL)
Systems, Policies and Regulatory Body Relations Unit
Matrix TRC Data Science program opens new scopes for participants to new job opportunities, widens the critical thinking, and enhances the analytical skills of those who are mostly implicated in data.

Ghassan Zein (Ideatolife)
Senior System Admin.
The Data Science program provided me with the complete needed knowledge to develop a product for Data Analysis and Predictions to help our customers taking the right decision at the right time.

Sabine Oussi (Sabis)
Data Analyst – IT
Attending Data Science & Big Data post-graduate diploma has made me discover data from a different perspective. It is not only about what the business intelligence tools can do nowadays but mainly about the real analysis that lies behind their logic. DS & BD is a full program from basic to Big Data analytics going through all the methodologies and applications needed as well as other contemporary topics.

Kris Khairallah (BOB)
IT Developer - Data architecture & BI
The course exceeded my expectations in many regards, especially in the depth of information supplied and the access to the instructor for feedback on work in progress. In a very non-threatening environment, I learned key principles of design that I can implement immediately. A diploma I highly recommend!
Some Testimonials

Marc Lati (Ethos Interactive)  
Co-founder

I joined the Data Science program, unsure of the benefit I might gain. But results exceeded my expectations by far, with rarely encountered experts as competent and enthusiastic about their courses. Some of them master their knowledge, reaching professional skills that made complex information look so easy to understand and put into practice. The diploma reveals great professionalism that brought me efficient added value. I would advice it to any person who wishes to grasp a solid understanding of Data Science and specifically Big Data analytics.

Mario Gergi Sarkis (Bank Audi S.A.L.)  
Business intelligence officer

Data Science is a trend that has become a crucial element for every major Enterprise. Learning data analysis smartly based on correct gathered information, affected dramatically my business development skills in various aspects. Furthermore, predictive and classification modules has helped me, as a professional daily data modeler, to reach better insight for all key elements affecting clients, communication, business management and product assessment.

Jean Abi Saab (A.U.B)  
Senior Research Analyst

The Data Science program has given me the right exposure not only to a broad range of data models but also to various statistical tools that helped me in my research projects. After the training, I was directly able to analyze data using statistical methods and new approaches without becoming a slave of data. The rich curriculum in data analysis and data mining allowed me to think beyond needed reports to develop new insights and find hidden patterns.

Pierre Nehman (B.O.B)  
Senior HRIS & Payroll Officer

The program joins both levels for experienced professionals as well as for beginners. The complete diploma enhanced my skills on how to deal even with complex data through machine learning algorithms and predictive analytics that were all supported with cutting edge technologies. I would highly advise anyone to be part of it!

Madonna Beyrouthy (Sabis)  
Associate Manager  
Book Orders

A very interesting, useful, and rich diploma program! I am certain that it will be of a benefit to both myself and my team, especially when it comes down to the different approaches in reading and analyzing data. The material and topics covered are very well presented and the instructors are very professional. I believe that this new topic will be an important one during the coming years, and I highly recommend everyone that deals with data to enroll in it.

Ramzi Fanous  
Statistics Assistant  
UN-ESCW

As an experienced statistician, with big excitement and enthusiasm I enrolled in the post graduate diploma in Data Science at AUST. The program covered a wide range of statistical tests, techniques and models. The courses greatly enriched my analytical knowledge especially in the fields of machine learning and big data. What I greatly liked is that, although many of the subjects were very technical and rigid, they were smoothly tailored and delivered by professionals. In addition, every topic we learned was applied in projects that helped me in understanding better and linking several analytical techniques together. The initial big excitement and enthusiasm were indeed highly met.”
C1.1. Data Analysis Foundations

Data Collection and Reports

- Data collection
- Types of variables
- Sampling
- Questionnaire design
- Qualitative Research
- Quantitative Research

Ethics and Reliability of Research

- Ethics
  - Rules and Regulations
  - Values
  - Research
  - Ethical Principles
  - Moral Principles

Data Science Academy
Descriptive Statistics

TV Time

<table>
<thead>
<tr>
<th></th>
<th>Count</th>
<th>Mean</th>
<th>Variance</th>
<th>St. Dev.</th>
<th>Coef. of variation</th>
<th>Standard Error</th>
<th>CI 95%</th>
<th>Shapiro test</th>
<th>Minimum</th>
<th>First quartile</th>
<th>Median</th>
<th>Third quartile</th>
<th>Maximum</th>
<th>Range</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>850</td>
<td>548.79</td>
<td>9,944.17</td>
<td>99.72</td>
<td>0.18</td>
<td>3.42</td>
<td>[542.07 - 555.5]</td>
<td>91.8%</td>
<td>220</td>
<td>483.75</td>
<td>547</td>
<td>613.25</td>
<td>662</td>
<td>0.03</td>
<td>0.00</td>
<td>466,468</td>
<td></td>
</tr>
</tbody>
</table>

Descriptive Table

Charting and Visualization

Summary Tables

<table>
<thead>
<tr>
<th>N.O.I</th>
<th>Gender / Invoice (Sum)</th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$2,827</td>
<td>$3,056</td>
<td>$5,883</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>$20,366</td>
<td>$28,597</td>
<td>$48,963</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>$23,304</td>
<td>$56,730</td>
<td>$80,034</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>$8,276</td>
<td>$20,283</td>
<td>$28,559</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>$2,139</td>
<td>$8,691</td>
<td>$10,830</td>
<td></td>
</tr>
<tr>
<td>TOT.</td>
<td>$56,912</td>
<td>$117,357</td>
<td>$174,269</td>
<td></td>
</tr>
</tbody>
</table>
C1.3. Data Analysis Foundations

Estimations

Central Limit Theorem

Probability Laws

Data Science Academy

Probability Laws

Central Limit Theorem

Estimations

Applied technologies:
C2.1. Profiling Techniques *(Basic & Intermediate)*

One Group

Benchmarking Variables

Analysis Process

Statistical Tables

**Hypothesis Tests**

\[-1.96 = -Z_{0.025} \]

\[Z_{0.025} = 1.96 \]

**One-Means’ Deviation Sorting** table

<table>
<thead>
<tr>
<th>X: Heart beat / Y: Type of illness</th>
<th>Average</th>
<th>R.V</th>
<th>Raw deviation</th>
<th>Std. Dev.</th>
<th>Std. Error</th>
<th>Stat</th>
<th>d.f</th>
<th>$H_0$ limits</th>
<th>P-Value</th>
<th>Test Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pulse</td>
<td>83.34</td>
<td>84</td>
<td>-0.66</td>
<td>17.7</td>
<td>0.694</td>
<td>-0.95</td>
<td>646</td>
<td>[-1.96; 1.96]</td>
<td>0.343</td>
<td>t-student</td>
</tr>
<tr>
<td>Upper Tension</td>
<td>12.80</td>
<td>13.5</td>
<td>-0.70</td>
<td>2.4</td>
<td>0.095</td>
<td>-7.38</td>
<td>646</td>
<td>[-1.96; 1.96]</td>
<td>4.8x10^-13</td>
<td>t-student</td>
</tr>
<tr>
<td>Lower Tension</td>
<td>7.29</td>
<td>7.5</td>
<td>-0.21</td>
<td>1.5</td>
<td>0.058</td>
<td>-3.67</td>
<td>646</td>
<td>[-1.96; 1.96]</td>
<td>2.6x10^-4</td>
<td>t-student</td>
</tr>
<tr>
<td>Temp</td>
<td>37.23</td>
<td>37</td>
<td>0.23</td>
<td>0.8</td>
<td>0.032</td>
<td>7.03</td>
<td>646</td>
<td>[-1.96; 1.96]</td>
<td>5.2x10^-12</td>
<td>t-student</td>
</tr>
<tr>
<td>Resp. Frequency</td>
<td>19.99</td>
<td>19.0</td>
<td>0.99</td>
<td>3.3</td>
<td>0.131</td>
<td>7.61</td>
<td>646</td>
<td>[-1.96; 1.96]</td>
<td>9.7x10^-14</td>
<td>t-student</td>
</tr>
</tbody>
</table>
C2.2. Profiling Techniques (Basic & Intermediate)  
**Two Groups**

### Statistical Tables

#### Comparison of means between two independent groups

<table>
<thead>
<tr>
<th></th>
<th>X: Treatment (two modalities)</th>
<th>Y: LDL cholesterol level (continuous)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Mean</td>
<td>139</td>
<td>136</td>
</tr>
<tr>
<td>Variance</td>
<td>31.3</td>
<td>19.1</td>
</tr>
<tr>
<td>Count</td>
<td>15</td>
<td>10</td>
</tr>
</tbody>
</table>

#### Profiling Variables

- GPA
- IQ
- EQ
- Satisfaction
- Gender
- Baccalaureate type

#### Profiling Solutions

- Public
- Private

#### Chance Analysis
C2.3. Profiling Techniques *(Basic & Intermediate)*

**Dependent and NP Tests**

### Dependent Samples

**CONTROL**
- \( n_{1}(+) = 150 \)
- \( n_{1}(-) = 50 \)

**TREATED**
- \( n_{2}(+) = 160 \)
- \( n_{2}(-) = 40 \)

### Power Tests

<table>
<thead>
<tr>
<th>Pair</th>
<th>Before</th>
<th>After</th>
<th>Difference</th>
<th>Sign</th>
<th>Rank of A.V</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>83</td>
<td>93</td>
<td>10</td>
<td>+</td>
<td>18</td>
</tr>
<tr>
<td>2</td>
<td>79</td>
<td>87</td>
<td>8</td>
<td>+</td>
<td>15.5</td>
</tr>
<tr>
<td>3</td>
<td>68</td>
<td>64</td>
<td>-4</td>
<td>-</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>81</td>
<td>75</td>
<td>-6</td>
<td>-</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>84</td>
<td>92</td>
<td>8</td>
<td>+</td>
<td>15.5</td>
</tr>
<tr>
<td>6</td>
<td>82</td>
<td>89</td>
<td>7</td>
<td>+</td>
<td>13</td>
</tr>
<tr>
<td>...</td>
<td>73</td>
<td>71</td>
<td>-2</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>86</td>
<td>92</td>
<td>6</td>
<td>+</td>
<td>10</td>
</tr>
<tr>
<td>23</td>
<td>84</td>
<td>80</td>
<td>-4</td>
<td>-</td>
<td>6</td>
</tr>
<tr>
<td>24</td>
<td>87</td>
<td>94</td>
<td>7</td>
<td>+</td>
<td>13</td>
</tr>
<tr>
<td>25</td>
<td>85</td>
<td>81</td>
<td>-4</td>
<td>-</td>
<td>6</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>78.12</td>
<td>81.36</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Non Parametric Tests

<table>
<thead>
<tr>
<th>Pair</th>
<th>Before</th>
<th>After</th>
<th>Difference</th>
<th>Sign</th>
<th>Rank of A.V</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>83</td>
<td>93</td>
<td>10</td>
<td>+</td>
<td>18</td>
</tr>
<tr>
<td>2</td>
<td>79</td>
<td>87</td>
<td>8</td>
<td>+</td>
<td>15.5</td>
</tr>
<tr>
<td>3</td>
<td>68</td>
<td>64</td>
<td>-4</td>
<td>-</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>81</td>
<td>75</td>
<td>-6</td>
<td>-</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>84</td>
<td>92</td>
<td>8</td>
<td>+</td>
<td>15.5</td>
</tr>
<tr>
<td>6</td>
<td>82</td>
<td>89</td>
<td>7</td>
<td>+</td>
<td>13</td>
</tr>
<tr>
<td>...</td>
<td>73</td>
<td>71</td>
<td>-2</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>86</td>
<td>92</td>
<td>6</td>
<td>+</td>
<td>10</td>
</tr>
<tr>
<td>23</td>
<td>84</td>
<td>80</td>
<td>-4</td>
<td>-</td>
<td>6</td>
</tr>
<tr>
<td>24</td>
<td>87</td>
<td>94</td>
<td>7</td>
<td>+</td>
<td>13</td>
</tr>
<tr>
<td>25</td>
<td>85</td>
<td>81</td>
<td>-4</td>
<td>-</td>
<td>6</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>78.12</td>
<td>81.36</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Applied technologies:

- **SAS**
- **STATISTICA**
C3.1. Profiling Techniques *(Advanced & Expert)*

**Multiple Groups**

### Data

\[
j = 1, k = 1, n_j = 1
\]

\[
\sum \sum (x_{ij} \ - \ \bar{X})^2
\]

<table>
<thead>
<tr>
<th>Groups</th>
<th>Count</th>
<th>Sum</th>
<th>Average</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>South</td>
<td>6</td>
<td>3,625</td>
<td>604.17</td>
<td>37,104</td>
</tr>
<tr>
<td>North</td>
<td>8</td>
<td>6,300</td>
<td>787.50</td>
<td>44,642</td>
</tr>
<tr>
<td>Beirut</td>
<td>7</td>
<td>7,250</td>
<td>1,035.71</td>
<td>20,595</td>
</tr>
<tr>
<td>Mount Lebanon</td>
<td>8</td>
<td>7,925</td>
<td>990.63</td>
<td>29,453</td>
</tr>
</tbody>
</table>

**ANOVA Table**

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>P - value</th>
<th>F_{0.05}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>786,503.1</td>
<td>3</td>
<td>262,167.7</td>
<td>7.92</td>
<td>0.0007</td>
<td>2.99</td>
</tr>
<tr>
<td>Within Groups</td>
<td>827,764.1</td>
<td>25</td>
<td>33,110.57</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1,614,267.2</td>
<td>28</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Multiple Qualitative**

- Benchmarking sub groups
- Pair Comparisons
- Multiple Means / ANOVA Table

**Multiple Means / ANOVA Table**

### Applied technologies

- SAS
- STATISTICA
- SPSS
C3.2. Profiling Techniques *(Advanced & Expert)*

**Single Regressions**

### Linear Regressions

- **$\hat{Y}_{obs} = 72.9$ kg**
- **$\hat{Y}_{ass} = 171.2$ cm**

### Non-Linear Regressions

- **Polynomial**
- **Exponential**
- **Power**
- **Logarithmic**

### Web visits

- **$L(Y) = N(72.9, 12.5)$ kg**
- **$L(e) = N(0, 6.52)$ kg**

### Applied technologies:

- *SAS*
- *STATISTICA*
- *IBM SPSS*
C3.3. Profiling Techniques (Advanced & Expert)
Dependent and N.P Tests

| Dependent and N.P Tests | Dependent Samples |

**Pulse by DAY per DRUG.**

<table>
<thead>
<tr>
<th>Dichotomous</th>
<th>Nominal / Ordinal</th>
<th>Quantitative</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Independent</strong></td>
<td><strong>Dependent</strong></td>
<td><strong>Independent</strong></td>
</tr>
<tr>
<td>Two Means</td>
<td>Two Means</td>
<td>Multiple Means</td>
</tr>
<tr>
<td>$H_0: \mu_1 = \mu_2$</td>
<td>$H_0: \mu_1 = \mu_2$</td>
<td>$H_0: \mu_1 = \mu_2 = \mu_k$</td>
</tr>
<tr>
<td>F test</td>
<td>F test</td>
<td>One-way ANOVA</td>
</tr>
</tbody>
</table>

| Non Parametric | | | |
| Two Means | Two Means | Multiple Means | Multiple Means | Correlation |
| $H_0: \mu_1 = \mu_2$ | $H_0: \mu_1 = \mu_2$ | $H_0: \mu_1 = \mu_2 = \mu_k$ | $H_0: \mu_1 = \mu_2 = \mu_k$ | $H_0: \beta_1 = 0$ |
| Wilcoxon / Mann Whitney rank-sum test | Wilcoxon SR test | Kruskal Wallis test | Friedman test | Spearman rank-correlation test |

**Explained**

$$\sum_{k=1}^{5} \sum_{i=1}^{9} (\bar{x}_{ik} - \bar{x})^2 = 11,152.6$$

**Residual**

$$\sum_{k=1}^{5} n_k (\bar{x}_k - \bar{x})^2 = 2,196.3$$

$$\sum_{k=1}^{5} \sum_{i=1}^{9} (x_{ik} - \bar{x}_k)^2 = 6,285.46$$

**Total**

$$\sum_{k=1}^{5} \sum_{i=1}^{9} (x_{ik} - \bar{x})^2 = 11,152.6$$

**Spearman Correlation**

- Beirut
- Big cities
- Villages

Applied technologies:

- SAS
- STATISTICA
- IBM SPSS
C4.1. Machine Learning (Unsupervised)
P.C.A and M.D.S

**Introduction to Algebra**

\[
Au_1 = \begin{bmatrix}
a_{11} & \ldots & a_{1p} \\
\vdots & \ddots & \vdots \\
a_{n1} & \ldots & a_{np}
\end{bmatrix}
\begin{bmatrix}
u_{11} \\
\vdots \\
u_{1p}
\end{bmatrix} = \begin{bmatrix}
\ldots \\
\ldots \\
\ldots 
\end{bmatrix} = \sum_{j=1}^{p} a_{ij}u_{1j}
\]

**Applied technologies**:

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>J</th>
<th>B</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beirut</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jounieh</td>
<td>45</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Byblos</td>
<td>85</td>
<td>35</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Tripoli</td>
<td>95</td>
<td>65</td>
<td>43</td>
<td>0</td>
</tr>
</tbody>
</table>
C4.2. Machine Learning (Unsupervised)
Clustering and Correspondence Analysis

Histogram of agglomeration indices

Simple & Multiple Correspondence Analysis

Cluster 1
Cluster 2
Cluster 3
Cluster 4

Pareto’s association

British
BBC
CNN

French

Arabs
Jazeera

F2 – 19.8%
F1 – 80.2%

-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2

-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2

-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2

-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2

0%
20%
40%
60%
80%
100%
120%

Applied technologies:

SAS
STATISTICA
IBM
SPSS
C4.3. Machine Learning (Unsupervised)

Factor Analysis and Canonical Correlations

**Factor Analysis**

**Quadrant Analysis**

**Canonical Correlations**

**Text Mining**

**Applied technologies:**
- SAS
- STATISTICA
- IBM SPSS
Gradient Descent Algorithms

Logistic Regressions

\[ P(Z) = \frac{1}{1 + e^{-z}} \]

Multiple and Stepwise Regressions

C5.1. Machine Learning (Supervised)

Multiple Regression Models

\[ Y = 12.51 + 2.3X_1 \]

\[ Y = 5 + 2X_1 + 3X_2 \]

\[ Y = 19.93 + 2.49X_1 \]
C5.2. Machine Learning (Supervised)
Predictive Models

Discriminant Analysis

Decision Trees

Neural Network
Deep Learning

Applied technologies:

- SAS
- STATISTICA
C5.3. Machine Learning (Supervised)
SAS – E Miner
### C6.1. Q – F – E

**Quality Control Measurements**

#### SPC for Measures

![SPC Diagram](image)

#### Process Capability Analysis

$$C_{pk} = C_p$$

- **$0 < C_{pk} < C_p$**
- **$C_{pk} \leq 0$**

**Fishbone Diagram**

**D.O.E**

**R&R Table**

<table>
<thead>
<tr>
<th></th>
<th>Est. S.D</th>
<th>Est. Variance</th>
<th>% R&amp;R</th>
<th>Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repeatability</td>
<td>$A$</td>
<td>$A^2 = s_{Rept.}^2$</td>
<td>$s_{Rept.}^2 / s_{R&amp;R}^2$</td>
<td>$A^2 / D^2$</td>
</tr>
<tr>
<td>Reproducibility</td>
<td>$B$</td>
<td>$B^2 = s_{Repr.}^2$</td>
<td>$s_{Repr.}^2 / s_{R&amp;R}^2$</td>
<td>$B^2 / D^2$</td>
</tr>
<tr>
<td>Measures (R&amp;R)</td>
<td>$A^2 + B^2 = s_{R&amp;R}^2$</td>
<td>100%</td>
<td>$(A^2 + B^2) / D^2$</td>
<td></td>
</tr>
<tr>
<td>Production</td>
<td>$C$</td>
<td>$C^2 = s_{Prod.}^2$</td>
<td>100%</td>
<td>$C^2 / D^2$</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>$D^2 = A^2 + B^2 = s_{Total}^2$</td>
<td>100%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Applied technologies**: SAS, STATISTICA

- $C_p$
- $C_{pk}$
- $D.P.M$
- 6σ Performance
C6.3. Q – F – E
Epidemiology

Measures in Epidemiology

Types of epidemiological indicators

Static
- Prevalence
- Proportional mortality
- Lethality or fatality

Dynamic
- Incidence rate
- Mortality rate

Morbidity

Mortality

Diagnostic tests

Types of Studies

False Positives
False Negatives

Survival Curves

Types of Studies:
- Etiological
- Descriptive
- Observational
- Interventional

Applied technologies:
Sas
Statistica
C7.1. Business Intelligence
Data Management from A to Z

**L1**
Contains source data, typically stored in **Transaction Data Bases**.

**L2**
Extracts information, and transforms source data into Multi-key & Time-dependent data.

**L3**
Stores such transformed information.

**L4**
Processes transformed information according to various purposes.

**Data Bases**
- POS
- ERP
- CRM
- Legacy
- OLTP
- Web Doc.
- Flat Files

**O.S/Data**
- Open Source Data Entry

**Data Warehouse**
- Meta Data Repository
  - **Extraction**
  - **Transformation**
  - **Loading**

**Data Mart (Mark.)**
**Data Mart (Mgt.)**
**Data Mart (Finance)**

**Analysis & Visualization**
- DSS
- Graphs
- OLAP
- Business Analytics
- Alert
- Forecast
- Data Mining
- Web browser

**BI T OOLS**
- Intranet
- Web browser
C7.1. Business Intelligence
Oracle BI Solutions Workshop
C7.3. Business Intelligence
SAS Data Flux Studio Workshop
2 million users worldwide

40% yearly increase of users

Leading programming language in data analytics and data science.
Module oriented for data scientists to start and develop their proper analysis.

Worldwide use of Python to come up with insights from data and gain competitive edge.

Most powerful tool to store and manipulate data.

C8.2. Statistical Programming
Python

```python
# Example code snippet
from scipy import stats
import pandas as pd

# Import Data
df = pd.read_csv('C:/Users/trafl/Downloads/Matrix/Example/Data.csv')

# Show the first 5 rows
df.head()
```

```
<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Size</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>df</td>
<td>DataFrame</td>
<td>(125, 3)</td>
<td>Column names: unnamed 0, City, SQM price</td>
</tr>
<tr>
<td>df_S</td>
<td>DataFrame</td>
<td>(16, 3)</td>
<td>Column names: unnamed 0, City, SQM price</td>
</tr>
<tr>
<td>y_S</td>
<td>Series</td>
<td>(125,)</td>
<td>class 'pandas.core.series.Series'</td>
</tr>
</tbody>
</table>
```

In [22]: final_stats = stats.ttest_1samp(y_S, 1000)

In [23]: final_stats.pvalue
```
```
C8.3. Statistical Programming

SAS Studio

Applied technologies:
IaaS: reducing operational and infrastructure costs

Virtualization and Containerization

Virtual Machines
- MySQL
- MySQL
- App
- Guest OS
- Guest OS
- Guest OS
- Hypervisor
- Host OS
- Server

Containers
- MySQL
- MySQL
- App
- App
- App
- Bins/Libs
- Bins/Libs
- Bins/Libs
- Container Engine
- Host OS
- Server

Connecting Everything to the Internet

IoT Value Add by 2020 - $1.9 Trillion

IoT market segments and value

Manufacturing
Healthcare
Insurance
Banking & Sec.
Retail & Wh.
Comp. Services
Others

IoT Communication Protocols

IoT Architecture
C9.2-3. IoT and Data Information Security
Data and Cybersecurity

System, Network Security

Attack Anatomy and cycle

Defense in Depth

Application and Database Security

Blockchain and Security Implications

Cybersecurity Damage
C10.1. Big Data

**BD Technologies**
C10.2. Big Data

BD Analytics
C10.3. Big Data BD Project and PMP Overview

1. Integration
2. Scope
3. Time
4. Cost
5. Quality
6. Human Resources
7. Communication
8. Risk
9. Procurement
10. Stakeholders

---

Project Charter

**Problem:** Big Data Project

**Objective:** To develop a comprehensive Big Data project plan

**Team:**

**Leader:**

**Manager:**

---

Cumulative values

- Expenditures
- Cost Baseline
- Funding requirements

**Completion date**

---

**ISO 9000**
Guest Speaker
Demography

6,184,701 (July 2015)

Total: 29.4 years
28.8 years
30 years

(Country comparison to the world: 118)
Professional Certificates

- **PC1 - Data Analysis**
  (90 hours)
  - Understand data structure
  - Quantify and illustrate information
  - Sort out characteristics of different groups in seconds!
  - C2 Profiling Techniques
    - Basic - Intermediate
  - C3 Profiling Techniques
    - Advanced - Expert

- **PC2 - Machine Learning**
  (90 hours)
  - Evaluate decisions error level
  - Optimize models to predict behaviors
  - Project complex data sets into maps
  - C4 Machine Learning
    - Unsupervised
  - C5 Machine Learning
    - Supervised

- **PC3 - Big Data**
  (90 hours)
  - Internet of Things and Security of information.
  - Real-time streaming analysis
  - Manage Cloud Data Platforms
  - Docker, Spark, ...
  - C9 Internet of Things Data and Cybersecurity
  - C10 Big Data Technologies
    - Big Data Analytics
The Specialties and DS Diploma

**Junior Data Analyst**
135 hours

- C1 Data Analysis Foundations
- C2 Profiling Techniques *(Basic – Intermediate)*
- C3 Profiling Techniques *(Advanced – Expert)*

**Senior Data Analyst**
270 hours

- C4 Machine Learning *(Unsupervised)*
- C5 Machine Learning *(Supervised)*
- C6 Quality / Forecasting / Epidemiology

**Data Scientist**
450 hours

- C7 Business Intelligence
- C8 Statistical Programming: R and Python
- C9 IoT - Data and Cybersecurity
- C10 Big Data: Technologies and Analytics

Benefit from the most flexible repayment plan. Settle part of your tuition fees gradually at 0% interest, and the remaining amount over 1 year after a grace period of 6 months.
Pierre EL HADDAD is currently a Hubert H. Humphrey Fellow at The Maxwell School for Citizenship and Public Policy, Syracuse University, New York, US. He is also a consultant, university teacher in management at University of Balamand-Lebanon, and has twenty years of experience in private business management. His present concerns relate to the participatory and collaborative institutional development towards sustainable organizations and communities. In parallel, he is developing a CSR decision-making model, and studying the role of the private sector in public administration reform.

Pierre has a doctorate in Business Administration from Jean Moulin University-Lyon III as well as a master in civil engineering from Saint-Joseph University and an MBA in economics and finance from Notre-Dame University in Lebanon.
Helgard Raubenheimer is an Associate Professor and Head of the SAS Lab and BMI professional masters programmes at the Centre for BMI, North-West University. He received a Ph.D. in Risk Analysis from the North-West University. His main research interest is in quantitative risk management. He has also supervised several master degree students on their industry research projects and is involved in research and consultation projects for financial institutions. He is also co-author of numerous peer-reviewed papers, which have received national and global awards.
Michel Sabbagh comes from a data science background having started his career as Oracle Database Administrator for Libatel Lebanon and Libatel Jordan and later on for Path Solutions. In 2002 he moved to SABIS as a project manager for an oracle laboratory where he produced many data projects that includes international data replication, and data warehouse. He spent 4 years in SABIS educational services before moving to BSynchro in 2006 as Head of Business intelligence department where he delivered many business intelligence projects including HAAD the data warehouse and business intelligence project for ministry of health in UAE. In 2009 he moved to BDL as Project Manager for one of the biggest and more advanced data warehouse in the ME region.
Christopher Ilmberger is the founder at Rayet Technologies, Inc. with over 17 years experience in data management technology, including projects focusing on data quality, data integration, modeling, MDM, and data governance. Christopher has worked in all verticals implementing data management practices, data warehousing, providing consulting, support, and pre-sales assistance. Deploying systems domestically and abroad building scalable solutions. He earned a bachelor's degree from North Carolina State University and was an early employee of DataFlux, which was purchased by SAS in 2000. Assisted in the integration of DataFlux into SAS.
The Experts

Karim Saikali

Karim earned a PhD degree in Software Engineering and is passionate about Service Oriented Architecture, Cloud Computing & Software methodology. He has a broad experience working with leading multi-national companies in Europe. Karim has also a sound academic experience and is currently assistant professor and coordinator of the business computing degree at St. Joseph University. He is actual a leading consultant at Element^n, expert in cloud development solutions.
Elie Zeidan is currently the Head of Cyber Intelligence Unit at Potech Consulting. Mr. Zeidan received his Engineering Degree in Communication and Computer from the Lebanese University and pursued his Masters’ Degree in Systems and Networks Security at Saint Joseph University. He has executed numerous missions in the information security field such as penetration tests, digital forensics, source code review, database and systems hardening for large networks in several countries such as Lebanon, France, UK, Cyprus, Jordan, KSA etc. He is currently an instructor at Saint Joseph University in Information Security, Cybercrime and Ethical hacking. He was part of several public talks such as the Anti-Cybercrime forum - the Cybercrime Bureau of the Internal Security Forces- Iktissad wal Amal Group, ICAR’17 – Antonine University, IoT Security – ISACA Lebanon, the Pre-Emptive Cybersecurity Training - USJ – Berytech.
Dr. Saadeh is an Instructor in the Department of Orthodontics at Lebanese University, Faculty of Dental Medicine and a Clinical Associate at the Division of Orthodontics and Dentofacial Orthopedics at AUBMC. She is the founder and director of “Charts- Research Design and Biostatistics”, a company that handles research designs, statistical analyses and epidemiology in the biomedical field. She is a current senior editor of the International Journal of Oral and Dental Sciences (IJODS). Her scholarly contributions include publications in peer-reviewed indexed journals and presentations at national and international meetings.

Dr. Maria Saadeh received her dental degree (DDS) from the Lebanese University School of Dentistry, and her postgraduate residency education in Orthodontics at the AUBMC. She holds a Master of Science in Human Morphology from the American University of Beirut Faculty of Medicine and a Master of Public Health (MPH) from the Faculty of Health Sciences at the same university. In addition, she earned her PhD degree in Odontological Sciences from the Doctoral School of Science and Technology at the Lebanese University in collaboration with the American University of Beirut.
Suzanne Menhem
(Demography Data)

Suzane holds a PhD from the Social Sciences Department of Sociology and Demography [École Doctorale Sociétés et Organisations at the University of Poitiers] in partnership with the Department of Social Sciences at the Lebanese University (Lebanon). She associated on many researchers, to name few: L’Institut Français du Porche Orient (IFPO), Lebanese Emigration Research Center (LERC), the Lebanese Emigration Research Center. She actually manages the demography program at the Lebanese University.

Christian Saab
(TV and Audience monitoring)

Christian Saab is Managing Director at AGB Stat Ipsos, where he manages operations and client relations and oversees business growth. With more than 23 years of experience in the TAM business, Media Research and Information Technology, Christian works alongside TV stations, the advertising agencies and advertisers to make decisions on allocating resources and budgets by providing insights about their target audiences. Christian is also part of the Joint Industry Committee of Media Research in Lebanon and a speaker on industry related topics at universities.