**D204 MULTIPLE ATTEMPT/TEACH-BACK STUDY PLAN**

1. **Lesson 1: The Data Analytics Life Cycle – Review this table.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Data Analytics**  **Life cycle**  **(7 Phases)** | **Tools and**  **Techniques** | **Potential**  **Problems** | **Data Pathway Terms**  **(4 Phases)** | **Data Science**  **(6 Phases)** |
| 1. **Business Understanding/ Discovery phase**   An analyst defines the major **questions** of interest that need to be answered, determines the needs of the **stakeholders**, and assesses the **resource** constraints of the project. Defines project **outcomes** | **Tools:**  1.Scope Statement   1. Stakeholder Register 2. Gannt Chart 3. Network Diagram   **Techniques:**  1. Critical Path Method  2. KPI  3. Budget estimation techniques  4. Schedule estimation techniques. | Lack of clear focus on:   * stakeholders, * timeline, * limitations, and * budget   could potentially derail an analysis | 1.Planning:   * Define goals. * Organize resources (software, hardware, staff) * Coordinate people * Schedule the project | 1. Find a question |
| **2. Data Acquisition**  Collecting data phase. Data is **collected**, and **stored**, for easy **retrieval** from a database, perhaps a component of a data warehouse, by using a language like SQL. Web scraping and surveys to acquire data. | **Tools:**  1.SQL  2. Web Scrapping software  3. Survey  4. Input Data: Self-generated Data  5. NoSQL- Used to collect Unstructured Data.  **Techniques:**  1. ETL  2. API  3. Web scrapping | **Quality**: uniqueness, relevance, reliability, validity, and accuracy.  **Type of data:** structured, unstructured, semi-structure, quantitative, qualitative.   * May make access more difficult. | 2.Wrangling:   * Get Data | 1. Collect the Data |
| **3. Data Cleaning phase**. Also known as data **cleansing**, data **wrangling**, data **munging**, and **feature engineering.** Analyst will use SQL, Python, R, or Excel to perform data **modifications** and **transformations** | **Tools:**  1.Python  2. R  3. SQL  4. Excel  **Techniques:**  1. Dimensionality Reduction: reduce variables, deals with a single score not multiple scores.  2. Data Reduction: optimize storage capacity  3. Modification  4. Transformation  5. Anomaly Detection | 1. Some cleaning techniques could dramatically change data/outcomes.  2. Outliers not dealt with can cause problems with statistical models due to excessive variability. | 2.Wrangling:   * Clean Data | 1. Prepare the Data |
| **4. Data Exploration** phase. Analyst begins to understand the **basic nature of data**, the **relationships within it (btw data variables),** the **structure of the dataset**, the presence of **outliers,** and the **distribution** of data values. This phase uses data visualization tools and numerical summaries such as measures of central tendency and variability. | **Tools:**  1.Distributions: Normal or Skewed Curve  2. Visualization Tools: Tableau, R, Python, RStudio, and histogram.  3. Statistical Tools such as  mean, median, and mode.  **Techniques:**  1. Correlation Discovery  2. Pattern Discovery  3. Visualization Techniques: histogram, charts, tables, boxplot, etc.  4. Variability: STD, Quartile | Skipping this step could:  1. Enable faulty perceptions of the data which hurt advanced analytics.  2. The analyst will lack insight into the structure of the data set. | 2.Wrangling*:*   * Explore Data * Refine Data |
| 5**. The predictive**  **Modeling phase.** Allows the analyst to move beyond describing the data to **creating models** that enable **predictions of outcomes** of interest. Python and R are used in automating the training and use of models. | **Tools:**  1. Python  2. R  **Techniques:**  1. Data Modeling  2. Correlation Modeling  3. Regression Modeling  4. Time Series Modeling  5. Cross Validation  6. Regression Models  7. Classification Models  8. Clustering | 1. Too many input variables (predictors) can cause problems.  2. Correlation does not imply causation.  3. Time series models often need sufficient time data to offer precise trending.  4. Predictive model accuracy should be assessed using cross-validation. | 3.Modeling   * Create the model * Validate the model * Evaluate the model * Refine the model | 1. Create the Model 2. Evaluate the Model |
| 1. **The data mining phase.** Looks for **patterns in large** sets of data. Tools are Python and R. Also called Machine learning. A specialized segment of data mining techniques that continually update to improve modeling over time.   **Note.** Both Exploration and mining uncover patterns. The difference is that data exploration is an initial step to uncover initial patterns and using both manual and automated methods. While data mining is an in dept step to discover patters using automated methods like machine learning. | **Tools:**  1.Python  2. R  **Techniques:**  1. Training dataset to build models  2. Classification  3. Clustering  4. AI  5. Machine Learning  5. Deep Learning | 1. Running on entire data is problematic,  2. Needs to subset data into training and testing datasets to build models.  3. Training Data: machine learns on training data to improve models.  4. Testing Data: evaluate the model itself.  5. Too little sample could cause limited insight. |
| 1. **The data reporting phase.** Analyst tells the story of the data and uses graphs or interactive dashboards to inform others of the findings from the analyses. Tools such as Tableau is used to spot trends and patterns. Goal is to give actionable insight to stakeholders. | **Tools:**  1.Dashboards:  2. Tableau  3. Story telling: a feature of Tableau  3. Graphs, charts, images, histogram, etc.  **Techniques:**  1. Visualization  2. Stakeholder communication | 1. Due to potential large audience consumption, mistakes can cause bad business decisions and loss of revenue  2. Improper scales used in graphs could push for interpretations of the story that is inaccurate. | 4. Applying / Reporting and Visualization   * Present the model * Deploy the model | 1. Deploy the Model |

**Lesson 1: The Data Analytics Life Cycle**

### Define each component below and provide an example of each.

### Descriptive Analytics –

### Diagnostic Analytics –

### Predictive Analytics –

### Prescriptive Analytics –

### What is the relationship between predictive and prescriptive analytics?

### Lesson II: Data Analytics Tools and Techniques

1. **List and define the various data types and examples?** 
   1. **Structured –**
   2. **Semi – structured –**
   3. **Unstructured –**
   4. **Quantitative –**
   5. **Qualitative –**
2. **Define a relational database -**
3. **List the various data sources**.
4. **What are the various data analytics applications/processes?** 
   1. **Python** –
   2. **R** –
   3. **Tableau** –
   4. **API** -
   5. **XML**–
   6. **SQL –**
   7. **D3.js –**
   8. **Search engine** –
   9. **JSON –**
   10. **Boxplot –**
   11. **MLaaS**
   12. **ETL** –
   13. **ETLTL** –
   14. **Training Data** –
   15. **Test (Validation) data** –
   16. **Normal Distribution** –
   17. **Bell curve with a long tail end** –
   18. **Histogram** –
   19. **Qlik** –
   20. **Heatmap** –
   21. **Scatterplot** -

**Review this table.**

|  |  |
| --- | --- |
| **PYTHON VS R** | |
| **PYTHON** | **R** |
| Used by programmers that want to delve into data analysis or apply statistical techniques, and by developers and programmers that turn to data science. | Used primarily in academics and research and is great for exploratory data analysis. In recent years, enterprise usage has rapidly expanded. |
| A production-ready language with capacity to be a single tool that integrates with every part of your workflow! | Used by statisticians, engineers, and scientists without computer programming skills. It’s popular in academia, finance, pharmaceuticals, media, and marketing. |
| Easier for people with software engineering background. | Easier for people with no coding experience. |
| Coding and debugging is easy because of the simple syntax. | Statistical models can be written with only a few lines. |
| The indentation of code affects its meaning. | The indentation of code does not affect its meaning |
| Any piece of functionality is always written the same way with Python. | The same piece of functionality can be written in several ways. |

1. **What is the difference between scripting and programming used in data** **analytics?**
2. **Define/Describe the methods used to validate models.**

1. **Define Regression-**
2. **Define/Describe trend analysis.**
3. **Define/Describe time series –**
4. **Defined decomposition –**

1. **Spectral Density -**

1. **Define/Describe Machine learning and artificial intelligence.**

1. **Define/Describe the various data mining/machine learning methods and techniques**.
   1. **Classification** –
      1. **Discuss the general approach to classification**.
      2. **Classification vs clustering**.
   2. **Clustering** –
   3. **Baye’s theorem** –
   4. **Dimensionality reduction-**
   5. **Data Reduction**
   6. **Hierarchal Clustering** –
   7. **Anomaly Detection-**
   8. **Neural networks** –
   9. **Deep Learning** -
   10. **Decision trees** –
   11. **Optimization Analysis** -
   12. **Supervised Model** –
   13. **Unsupervised Model-**

**13. What is the most effective way of virtual communication**?

### Lesson IV: Organizational Needs and Ethical and Legal Issues

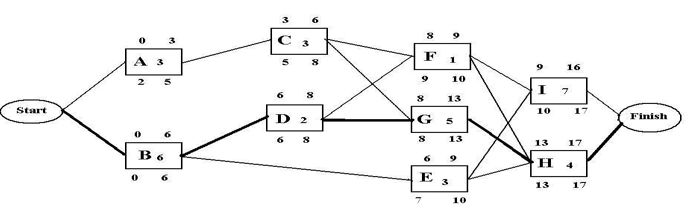
1. **What decisions are necessary to initiate a data analytics project?**
2. **What are the implications of undefined outcomes of potential data analytics projects?**
3. **How does one define research questions within an organization?**
4. **Summarize the legal frameworks for data governance.**

5**. Define Conflict of Interest in the context of data frameworks.**

6. **Define Democratization.**

### IV. Lesson 4 and Lesson 6: Teamwork and Collaboration, and Data Analytics Project

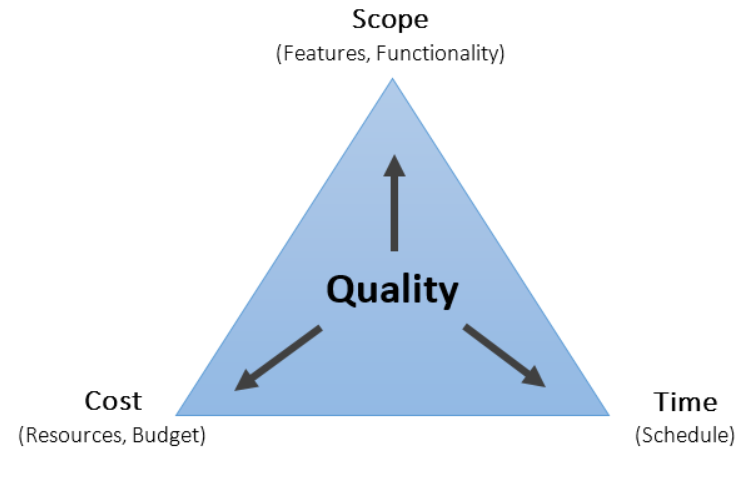
1. **Define the various roles in the workplace in a data analytic project** (provide examples if possible).
2. **Define the various roles of potential partners and stakeholders in data analytics projects (provide examples, if possible).** or groups.
3. **Explain and define the critical path and its relationship to project timeline**.



**Network Diagram showing critical path.**

**Longest Path = B, D, G, H; Minimum Time = 17 days**

1. **Explain the IRON triangle and the challenge of balancing resources in data analytics projects**.



1. **What are effective interpersonal communication skills**?
2. **What is Active listening?**
3. **Describe co-creation approaches and tools.**

**Additional Tips**

1. First attempt - Read everything, watch all the videos and answer the chapter quizzes.
2. 2nd attempt Focus mostly on the text in Lessons 1, 2, and 5 and use the videos as supplements.
3. 3rd attempt should meet with the CI for a teach back session with the 3rd attempt study guide.
4. Attend Cohort to have a live session with your instructor(s). The cohort is on the 1st and 3rd Sundays and Tuesdays of every month. See link below.

Cohort Student Link: <https://srm.my.salesforce.com/apex/CohortLiveEvent?cohortId=a2y3x000002ujwl>