Data Science

Burton, Lailynette D.

Cabanela, Charmie A.

Manuel, Meriel Necole T.

Balatong, Jayson

Table of Contents

6 Visualize

Advantage of good Data Visualization

Tools used in Data Visualization

Overview 1

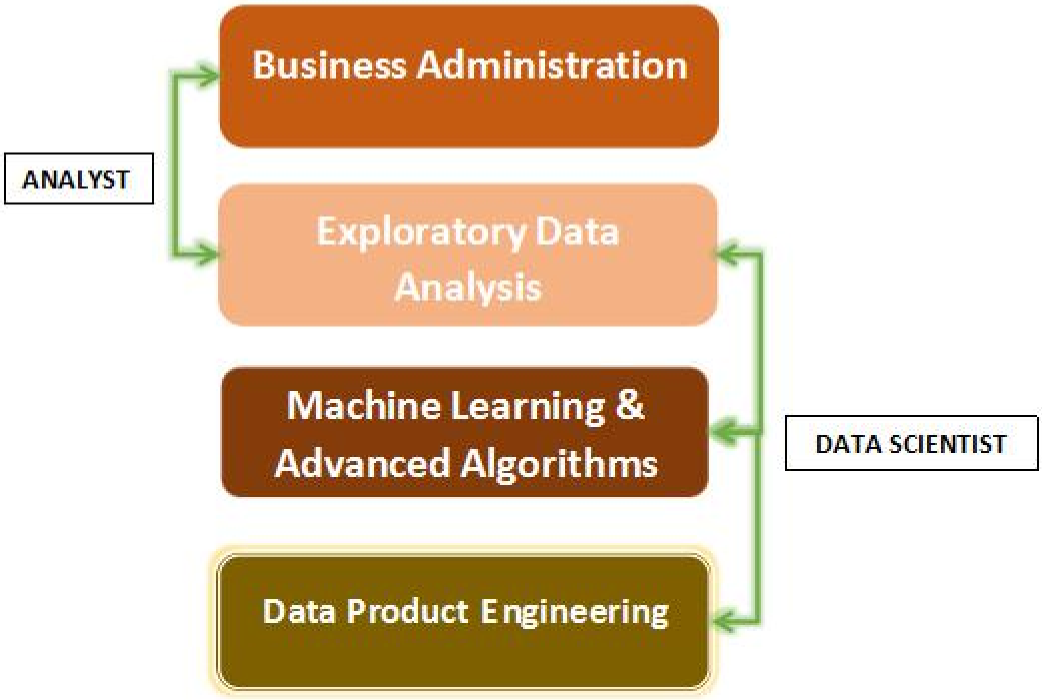
Data Science, to define it precisely, is the study of data. It is a process of understanding a certain data and make it applicable and understandable in many different things of the world. For example, you have a proposed explanation for a certain problem. In order to prove and convince the people to believe in you proposition, you have to support your explanation with various data. It is a skill of unveiling the underlying meaning in a data. Data science involves using different methods to analyze large amounts of data, it will gather the knowledge and information. We are going to convert data into a story - to a more understandable propositions. We use storytelling to explain a perspective or to express an insight, these insights will eventually help the decision making and come up to a strategic plan or preference for a company or institute.

The term 'data science' and its definition came up when some professors and IT Professionals, scientist were looking into a statistics curriculum. They decided to better call it as data science, later on, also called as data analytics. It is a field that extracts data of various forms and from different resources.

Why we need Data Science

How are decision and predictions are finalized using Data Science? Data science are composed of various tools, algorithms, and machine learning principles. Its main goal is to disclose a specific figure or produce a very useful information from a raw data.

Now, what was the difference of data science from those who does statistics?



The image above explains that, Data Analyst are the one who demonstrate and simplify what is going on by collecting and processing the history of data. While the Data Scientist, doesn't just regulate an exploratory analysis, but also uses advanced machines and algorithms to explain undoubtedly the data that has been collected for present and future decision makings. A Data Scientist can look at different angles for the data some can even distinguish new information that was unknown earlier.

Today, many companies are already using data science to gain insights from their business. Many companies use data science to improve the quality of their business and to produce a better offer to their customers, as well as their staffs and business partners. Aside from that, governmental organizations also rely on internal data scientists to discover substantial information, that can definitely share to the public. Non - government organization and Universities are also not foreign of using data. They are able to raise money for a cause and enhance study experience by meticulously taking into considerations of all the data that they gather.

Who is a Data Scientist?

There are lots of available words to describe a Data Scientist. But, to make it precise and simple, a Data Scientist is the one who has the knowledge and practices the art of Data Science. The term was coined by DJ Patil and Jeff Hammerbacher. Data Scientist are scientists who extracts and dissects every available data that they are able to gather. From that, they will convert it to a very useful information with their certain scientific disciplines.

The term 'Data Scientist' has been a subsistence after considering that they are collecting a massive amount of information from the scientific fields and applications regardless if the information is mathematical, computer science or statistical. They exercise using latest strategies, tools and technologies in finding solutions and encompassing conclusions that would be valuable for the success of an organization. Data Scientists interprets data in much more precise and functional way, compared to the available raw data that they have.

Further reading suggestions

Beginner's Guide to Data Science,

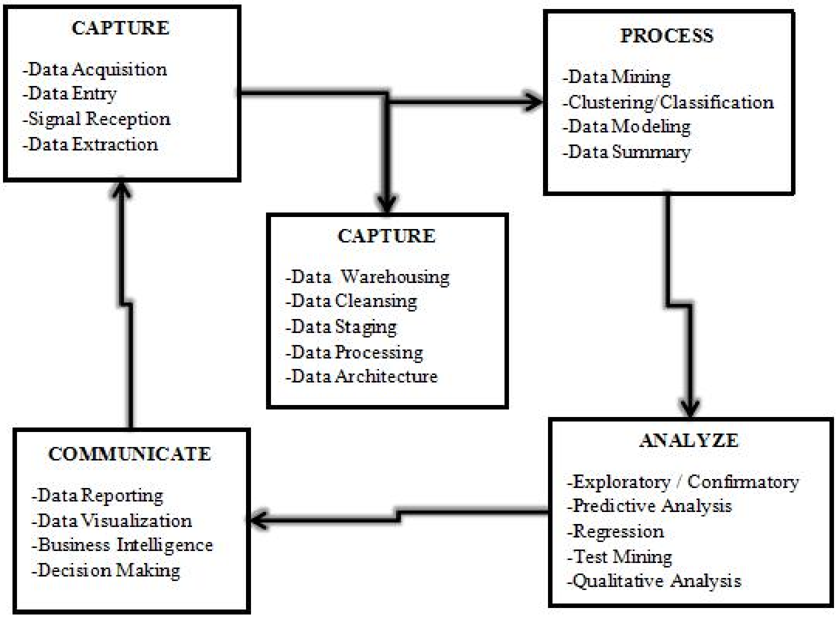
https://towardsdatascience.com/a-beginners-guide-to-data-science-55edd0288973

Cielen, Meysman, Ali, ***Introducing Data Science,***
https://books.google.com.ph/books?hl=en&lr=&id=EZAtAAAAQBAJ&oi=fnd&pg=PP1&dq=data+science&ots=yk1PNs4SFV&sig=AqS7KZ7kb6Hsz0HSFYbHUdSsFH8&redir_esc=y&authuser=1#v=twopage&q&f=false

Getting 2

Data science involves various skills that make and defines its profession in a constant struggle. As time goes by, data science also continues to evolve. It became one of the most promising and in-demand career paths for skilled profession. In order to unveil the underlying information and qualitative data in an organization, data scientists must master the full spectrum of the data science life cycle. They must posses a level of versatility to maximize returns to each phase of the process.

Five stages of Data Science Life Cycle



CAPTURE

-Data Acquisition
-Data Entry
-Signal Reception
-Data Extraction

PROCESS

-Data Mining
-Clustering/Classification
-Data Modeling
-Data Summary

CAPTURE

-Data Warehousing
-Data Cleansing
-Data Staging
-Data Processing
-Data Architecture

ANALYZE

-Exploratory / Confirmatory
-Predictive Analysis
-Regression
-Test Mining
-Qualitative Analysis

COMMUNICATE

-Data Reporting
-Data Visualization
-Business Intelligence
-Decision Making

Job that fits you in Data Science

Data, a term that can be use anywhere and everywhere. There are variety of terms that are related into mining, cleaning, analyzing, and interpreting data that are often used are interchangeably but can involves different skills and complexity in data. Here are the following jobs that are related in analyzing and organizing data in different industry.

Data Scientists

Data scientists examines questions that are answerable and finds way to solve the related data. They have business skills and analytical skills as well as the ability to mine, clean, and present data of the company. Businesses use data scientists to source, manage, and analyze large amounts of unstructured data. Results are then synthesized and communicated to key stakeholders to drive strategic decision-making in the organization.

Skills needed: Programming skills (SAS, R, Python), statistical and mathematical skills, storytelling and data visualization, Hadoop, SQL, machine learning.

Data Analysts

Data analysts bridge the gap between data scientists and business analysts. They provides the questions that needs to solve and answer from an organization and then organize and analyze the data to find results that aligns with the high-level business strategies. Data analysts are responsible for translating technical analysis to qualitative action items and effectively communicating their findings to diverse stakeholders.

Skills needed: Programming skills (SAS, R, Python), statistical and mathematical skills, data wrangling, data visualization.

Further reading suggestions

***Coursera's Getting and Cleaning Data**, https://www.coursera.org//learn/d*

Machine Learning walk through course,
https//www.dataquest.io./blog/machine-learning-preparing-data/

Cleaning 3

Data Cleaning means the process of replacing, modifying or deleting the necessity of the data from the identified incorrect, inaccurate, incomplete, irrelevant or missing part in a data. Data cleaning is considered as one of the foundation element of the basic data science.

For further details, suppose that you are a general manager of a company. Your company needs to collect data of different customers who buys product in which produced by the company. Now, as the general manager, you need to know which products people are interested mostly and according to the data, your company wants to increase the production of the product. But if the data is corrupted or contains missing values then you will be misguided and needs to remake the correct decision and that's a trouble in your company.

To be able to manage the trouble, Machine Learning can be used since it is a data-driven AI. In machine learning, if the data is irrelevant or error-prone then it leads to incorrect model building.

Different ways of Data Cleaning

1. Remove all unwanted observations.

First step to data cleaning is to remove the unwanted and excess observations. This includes the irrelevant and the duplication of each observations.

Duplicate observations

Duplicate observations mostly begin during data collection, such as when:

- Combining sets of data from multiple places
- Scraping data
- Received data from clients/other departments

Irrelevant observations

Irrelevant observations are those that doesn’t actually fit the **specific problem** that you’re trying to solve.

- For example, if you were building a model for home of those single-family only, you wouldn't want observations for Apartments in there.
- Reviewing the charts from Exploratory Analysis. You can look at the distribution charts for categorical features to see if there are any classes that shouldn’t be there.
- Checking for irrelevant observations **before engineering features** can save you many headaches down the road.

2. Fixing Structural Errors

Structural errors are those that arise during measurement, data transfer, or other types of **"poor housekeeping."**

For instance, you can check for **typos** or **inconsistent capitalization.** This is mostly a concern for categorical features, and you can look at your bar plots to check.

After we replace the typos and inconsistent capitalization, the class distribution becomes much cleaner.

Finally, check for **mislabeled classes**, i.e. separate classes that should really be the same.

- If ’N/A’ and ’Not Applicable’ appear as two separate classes, you should combine them.
- ’IT’ and ’information_technology’ should be a single class.

3. Filtering Unwanted Outliers

Outliers can cause problems with certain types of models. For example, linear regression models are less robust to outliers than decision tree models. In general, if you have a legitimate reason to remove an outlier, it will help your model’s performance. However, outliers are innocent until proven guilty. You should never remove an outlier just because it’s a "big number." That big number could be very informative for your model.

4. Handling Missing Data

Missing data is a deceptively tricky issue in applied machine learning.

First, just to be clear, **you cannot simply ignore missing values in your dataset.** You must handle them in some way for the very practical reason that most algorithms do not accept missing values.

The 2 most commonly recommended ways of dealing with missing data actually hard to solve.

1. **Dropping** observations that have missing values
2. **Imputing** the missing values based on other observations

Dropping missing values is sub-optimal because when you drop an observation, you are also dropping an information.**.**

Missing categorical data

Label the missing data as “Missing” to be able to identify and categorize them. It is the best way to handle missing data in categorical features.

- It is necessary and essential to add a new class for the categorical feature.
- It tells and identify the algorithm that the value that were given was missing.
- It gets around the technical requirements for no missing values.

Missing numeric data

In any missing numeric data, it must have flag and fill the values.

1. Flag the observation with its indicator variable of missing.
2. Then, fill the original missing value with 0 just to meet the technical requirement of no missing values.

By using this technique of flagging and filling, you are essentially **allowing the algorithm to estimate the optimal constant for missing**, instead of just filling it in with the mean.

Further reading suggestions

Coursera’s Getting and Cleaning Data*, https://www.coursera.org//learn/d*

Advanced Data Cleaning course*,*
https://www.dataquest.io/course/python-datacleaning

Organize 4

Why is it important to organize your project?

- ✓ Increases productivity. If a certain project is well organized, and with everything placed in one directory, it makes it avoid wasting time searching for project files such as datasets, codes, output files, and etc.
- ✓ A project that is well-organized helps to keep and maintain a record of ongoing and completed data science projects.
- ✓ A Completed data science projects can be used for building future models.
- ✓ A well-organized project can easily be understood by other data science professionals when shared on platforms such as Github.

Its always recommended that a project must maintain two versions, one locally, and the other on Github. The advantage is you can access the Github version of your project from anywhere and anytime in the world, as long as you have internet connection. And also if something happen with your local computer that could impact your computer adversely, such as viruses, then you can always be confident that you still have your project files on Github that can serve as a backup.

Once you have finish what you create, gather, or start manipulating data and files, they can easy become disorganized. To save time and prevent errors, the whole team should decide how you will name and structure the files and folders. Which includes documentation (or 'metadata') that will allow you to add context to your data so that you and others can understand it in the short, medium, and long-term.

Naming and Organizing files

Choosing logical and a consistent way to name and organise your files allows you and others to easily locate and use them. the best time to think on how to name and structure the documents and directories created is at the start of a project. a naming convention will help you to provide consistency, it will be easier to find and correctly identify your files, Organising your files will save you time and by helping you and your team find what you need when you need it.

Things that are needed to consider in naming a file:

- consistent
- meaningful to you and to your colleagues
- allows you to find the file easily.
-

It is useful if your project agrees on the following elements:

- Vocabulary – it must be a standard vocabulary for file names, so that everyone uses a common language
- Punctuation – decide on conventions for when you will be using punctuation symbols, capitals, hyphens and spaces
- Dates – agree on a logical use of dates (chronologically i.e. YYYY-MM-DD)
- Order - decide which element should go first, so that the files on the same theme are listed together and can be found easily
- Numbers – specify the amount of digits that will be used (listed numerically e.g. 01, 002, etc.)

Renaming example:

- You can rename variables programmatically or interactively.
- Rename interactively fix(mydata) # results are saved on close
- Rename programmatically library(reshape)
- mydata <- rename(mydata, c(oldname="newname"))

Documentation and Metadata

To ensure that you understand your own data, use and cite your data properly, it helps you to add documentation and metadata (data about data) to the documents and datasets you create.

When and how do I include documentation/metadata?

Embedded documentation

Informations about a file or datasets can be included within the data or the document . For digital datasets, this means that the documentation can sit in separate files (example: text files) or be integrated into the data file(s), as a header or at specified locations in the file.

Supporting documentation

This is a information in separate files that accompanies data in order to provide context, explanation, or instructions on confidentiality and data use or reuse.

Sorting

To sort a dataframe in R, use the order( ) function. By default, sorting is ASCENDING. Prepend the sorting variable by a minus sign to indicate DESCENDING order.

Sort by Net_Sales_Unit

```
newdata <- sdata[order(sdata$Net_Sales_Unit),]
```

Merging

merge() function recognizes that each DataFrame has an "employee" column, and automatically joins using this column as a key. The result of the merge is a new DataFrame that combines the information from the two inputs.

To merge the two dataframes (datasets) horizontally, use the merge function. In most cases, you join two dataframes by one or more common key variables (i.e., an inner join).

Merge two dataframes by ID

total <- merge(dataframeA,dataframeB,by="ID")

Merge two dataframes by ID and Country

total <- merge(dataframeA,dataframeB,by=c("ID","Country"))

Aggregating

A process where raw data is gathered and expressed in a summary form for statistical analysis

It is relatively easy to collapse data in R using one or more BY variables and a defined function. Using the sdata data set, let's aggregate Units_Billed by Customer.

aggdata <- aggregate(sdata["Units_Billed"],by=sdata[c("Customer")], FUN=sum)

Further reading suggestions

Tools for metadata tracking and data standards (ISA Tools)

Organizing Data Science Teams for strong ROI - insideBIGDATA,
insidebigdata.com/2016/01/16/organizing-data-science-teams-for-strong-roi/

A quick guide to Orgzanizing (Data Science) Projects, *https://medium.com/outlier-bio-blog/a-quick-guide-to-organizing-data-science-projects-updated-2016-4cbb1e6dac71*

Model 5

Key Phases in Building Data Science Model

1. **Data Extraction**

   To start with, you need to have an idea about the problem at hand, while the collection of data follows next. Not any data, but the collected unstructured data should be relevant to the business problem you are about to solve. And not all data is relevant and updated.

2. **Data Cleaning**

   This holds significance where you require to clean the data while you are collecting it.

   - Here are some common sources of data errors:
     - ○ Duplicated entries gathered from across many databases.
     - ○ The error with the input data in regard with accuracy.
     - ○ The data entries were changed/updated/deleted.
     - ○ Missing values in variables across databases.
   - Tricks to Eliminate the common sources of errors:
     - ○ Filter out duplicates by referring to the common IDs.
     - ○ Sort out the data by referring to the date it was updated, i.e. giving preference to the most recent entry.
     - ○ Fill in the missing data entries with mean value.

3. **Diving Deep into the Data**

   Now that every source of data is ready, you can start with analyzing the essential patterns involved. Deploying interesting tools such as Tableau or Micro Strategy can help a ton. All you have to do is build an interactive dashboard and see how your data becomes a mirror to important insights.

4. **Identifying the Critical Features**

   When seeking to get hold of key patterns in business, feature engineering can be deployed. This step can't be ignored as it forms the prerequisite for finalizing a suitable machine learning algorithm. In short, if the features are strong, the machine learning algorithm would produce good results. But better data beats fancier algorithms.

   - There are two categories of features that need to be taken care of:
     - Constant features that are less likely to change
     - Variable features whose values fluctuate from time to time

5. **Exploring the World of Machine Learning**

   This makes for one of the most important steps as the machine learning algorithm helps build a workable data model. There are many algorithms to choose from, but no worries as the data scientist would make it easy for you.

   In the words of data scientist, machine learning is the process of deploying machines for understanding a system or an underlying process and making changes for its improvement. And, an algorithm can be termed as a set of instructions to the computer system to drive a particular task.

6. **Evaluate and Deploy the Model**

Techniques such as cross-validating or even ROC (Receiver Operating Characteristics) curve, work well for generalizing the model output for new data. If the model appears to be producing satisfying results, you can now implement the model to see your business making a difference.

The engineers are given the power to deploy the model into the corresponding production phase. Here, the experts translate the model into a production stack language to facilitate a fine implementation.

Secondly, infrastructure is set up that further makes data scientists independent enough to deploy the data model all on their own. This is possible with APIs that are gaining momentum at a good pace. These APIs work on eliminating the lags between the data science and the teams involved in the project.

Machine Learning Methods

- Supervised Learning

  It is based on the outcomes of a similar process in the past. Supervised Learning helps in predicting an outcome based on historical patterns.

  Are models where there is a clear distinction between explanatory and dependent variables. The models are trained to explain dependent variables using explanatory variables. In other words, the model output attributes are known beforehand.

  The tried and tested algorithms for supervised learning:

    - Linear Regression
    - Random Forest
    - Support Vector Machines

- Unsupervised Learning

This learning method remains devoid of an existing outcome or pattern. Instead, it focuses on analyzing the connections and the relationships between the data elements. The model outputs are unknown or there are no target attributes: there is no distinction between explanatory and dependent variables. The models are created to find out the intrinsic structure of data.

The tried and tested algorithms for supervised learning:

- K-Mean Clustering
- Apriori Algorithm

Further reading suggestions

40 Techniques used by Data Scientist,
https://www.datasciencecentral.com/profiles/blogs/40-techniques-used-by-data-scientists

Predictive Modeling: The Only Guide You'll Need,
https://www.microstrategy.com/us/resources/introductory-guides/predictive-modeling-the-only-guide-you-need

Visualization 6

In Data Visualization we use computer graphics - create graphical representation which aid in understanding and representation of massive data and information. Using visual elements such as maps, graphs, and charts, we are able to interpret and understand trends, outlines, and patterns of data in a more accessible way. Data Visualization is an essential tool to analyze large amounts of information that can guarantee a data-driven decision making.

Data Scientists can choose several ways to represent their insights to the end user. Take a look at the examples below:

- ***One-time representation:*** research question will take only one time. The decision that will be made will bind the organization for long period of time. When the decision is made, the results will be delivered with a representation.
- ***New viewport on your data:*** best example for this is the customer segmentation. The segments will be communicated through reports and representation and it will form a tool. A clear and relevant customer segmentation will then be unveiled. From then on, people can produce their own reports and insights. A new discovered segmentation can be fed back to the database as the new angle on the data.
- ***Real-time dashboard:*** as a reliable data scientist, the task is not just discovering a new information, send it back to the database and then it's done. As a data scientist who discovered new information you must create an example, because for some instance, others might misunderstood it. It is a must that a data scientist will make the first reports, and others will just follow the steps. Make the first dashboard to shorten the representation and delivery of an insight will make good use of the information and its application.

Advantage of good data visualization

Data Visualization is a form of visual art. It's has the characteristic to grab people's interest and keeps the information more accessible and understandable. For instance, we see a chart, from that we can quickly see the information it beheld, the current trends and outliers. Creating and seeing graphs has a big difference when looking at a massive spreadsheet full of different data, people couldn't easily distinguish the the information they we're looking for.

Data Visualization has its way of story telling. It helps in curating the data in an easier way, make it easier to understand and highlight the information needed. A good visualization helps in removing the unnecessary information, like removing a noise from a story and highlighting the necessary information. However, creating a visual representation isn't just about making it stunning. Effective data visualization involves a delicate balance between the form and function. The data and the visuals must work together, taking a lot of consideration to come up with a great combination of analysis and storytelling.

Importance of data visualization in any career

Lots of fields benefit from making data more accurate. Government, finance, history, sport, education, marketing, consumer goods, service industries, and so on. Data Visualization is undeniably, a practical, real-life application. Since data visualization provides a high standard of understanding a data, it is also a good and most useful professional skill that one may develop.

Creating good use of a data, making it relevant for decision making, is increasingly valuable skill for a professional to develop. Using visuals to deliver stories of when data informs the who, what, where, when, and how is a great deal for a business industry. We are thought of a great difference between creative storytelling and technical analysis, but to be able to cope up with the changes of this modern world, having the quality to bind the two - having the skill in data visualization will make one prolific.

Tools used in Data Visualization

When we come across with data visualization, the first thing that will probably come in our mind are bar graphs or pie charts. While these were the common way of visualizing a data, one must realize that a certain information must be paired with the right visualization tool. There are things to consider, methods to present, to be able to show information in effective and interesting way.

Common tool used for data visualization:

- ***Charts:*** used to emphasize differences in proportion among few entities. Its common display type are: bar charts, line charts, pie charts, bubble charts, stacked charts, and scatterplots.
- ***Tables:*** have multiple data sets with different units of measure. User need to be precise in inputting value to precisely compare related values.
- ***Graphs:*** show general trends. Can interpret large data sets. It is used when wanted to reveal relationships among multiple values. Common display type of graphs are: line graph, bar graph, scatter plot, etc.
- ***Maps:*** Treemaps, a common type used for data visualization. It displays data using rectangles. Depending on choice the rectangle representing the leaf node is colored, sized or both according to chose attributes.
- ***Infographics:*** graphic visual representation of data or information which aims to accurately and quickly present the information. It utilized images and graphics that will improve human visual and cognitive to be able to see trends and sets of information.
- ***Dashboards:*** it simplify a complex data that will allow users to see information, view the current trends and performance in just a glance.

Software choices to represent data

In today's time, there are a lot of proven software packages that helped in interpreting our data. Best examples are: Tableau, MicroStrategy, Qlik, SAP, IBM, Microsoft, Spotfire, and so on. These companies offers dashboard tools. They can offer a free trial version, for the user to have the chance to text their product's credibility. In the end, they are going to offer and pay to acquire a full version, these applications might be worth, specially for those who are working with bigger companies that handles bigger data.

However, in you're looking for a free data visualization, you can quickly end up in using HTML, JavaScript libraries are free to use, one can easily plot data that they want. Example landscapes are the following:

- ***HighCharts:*** a browser-based graphing libraries
- ***Charkick:*** a JavaScript charting library for Ruby on Rails user.
- ***Google Charts:*** a free charting library of Google. It is free to use and offers a wide range or graphs.
- ***D3.js:*** not a graphing library but a data visualization library. It is the most versatile JavaScript data visualization library available.

Further reading suggestions

Stefanowsky, Informatiky, ***Data Visualization****,*
http://www.cs.put.poznan.pl/jstefanowski/sed/DM14-visualisation.pdf

Top Visualization tools for data science*,*
https://www.tableau.com/learn/articles/data-visualization

Quiz

Exercise

Assignments

Quiz / Seatwork

I.

1. Explain Data Science in your own words.

2. Why do you think Data Science plays an essential part for a project's success?

3. Explain the difference between a Data Scientist and a Statistician.

4. Give at least (3) misconceptions about Data Science.

5. In your own words, describe the importance of Data Science in a business world.

II.

1. What does Data Scientist performed?

a) Define the question
b) Create reproducible code
c) Challenge results
d) All of the mentioned

2. Which of the following is the most important language used for Data Science?

a) Java
b) Ruby
c) R
d) None of the mentioned

3. Which of the following is the correct statement?

a) Data has only qualitative value
b) Data has only quantitative value

c) Data has both qualitative and quantitative value
d) None of the mentioned

4. Data that summarize all observations in a category are called __________ data.

a) frequency
b) summarized
c) raw
d) none of the mentioned

5. Which of the following is another name for raw data in Data Science?

a) destination data
b) eggy data
c) secondary
d) machine learning

III.

1. Importance of organizing your project? Own opinion.

2. What are the four useful elements for your project.

3. Two ways to rename variables.

4. What is documentation and metadata in your own words

5. To merge two dataframes what function will you use?

IV.

1. Give at least 3 key phases in building a Data Science Model.
2. Give at least 2 common sources of data errors.
3. Give 1 tricks to eliminate common sources of data errors.
4. What is Supervised Learning Algorithm?
5. What is Unsupervised Learning Algorithm?

V.

1. What is Data Visualization? Discuss base on your own understanding.

2. Explain a situation where there will be no Data Visualization applied in a certain presentation. How do you think the presentation will turn out?

3. Discuss the impact of Data Visualization in business success.

4. Explain "*Data Visualization has its way of story telling*".

5. Give at least (3) reasons why software packages like Tableau, Microsoft, QlikView is a great help to handle data for bigger companies.

Exercises

I.

1. Create a Venn diagram to show the difference and similarities of Data Scientist and Data Analyst.

2. List down the character that a Data Scientist should have

3. Discuss at least (2) issues that a Data Scientist can encounter in the Business world

II.

1. Differentiate Data Scientist and Data Analyst.
2. Give the 5 Stages of Data Science Life Cycle. Explain it in your own idea.
3. In your own idea, define what Machine Learning is.

III.

Organize a sample project using:

- Renaming
- Merging
- Sorting and,

- Aggregating

IV.

1. Elaborate the tools / strategies used for Data Visualization

2. Give at least (3) indicators that a Data Visualization is properly applied in data presentation

3. Choose one example of a data visualization tool and illustrate it. Choose any form of data to represent.

Assignment

I.

Create list of at least (5) famous Data Scientists. Write down their experiences in the field they worked on.

II.

The purpose of this assignment is to let you learn and demonstrate your ability to collect, work with, and clean a data set. The goal is to prepare tidy data that can be used for analyzing the data.

You may submit:

1. A tidy data set as described below,
2. A link to a Github repository with your script for performing the analysis,
3. A code book that describes the variables, the data, and any transformations or work that you performed to clean up the data called CodeBook.md. You should also include a README.md in the repo with your scripts. This repo explains how all of the scripts work and how they are connected.

III.

You should create one R script called run_analysis.R that does the following.

1. Merges the training and the test sets to create one data set.
2. Extracts only the measurements on the mean and standard deviation for each measurement.
3. Uses descriptive activity names to name the activities in the data set
4. Appropriately labels the data set with descriptive variable names.

5. From the data set in step 4, creates a second, independent tidy data set with the average of each variable for each activity and each subject.

IV.

Create a simple model using a data that you can collect inside the classroom.Choose an example model (1) from Supervised Learning Algorithm, (1) from Unsupervised Learning Algorithm

V.

Create a data representation using any software tool/application used for Data Visualization. You can get any form of data or free example data that are available in web.

References

What is Data Science, *https://www.edureka.co/blog/what-is-data-science/*

Overview of Data Science, *https://www.geeksforgeeks.org/overview-of-data-science/*

Data visualization beginner's guide: a definition, examples, and learning resources,

https://www.tableau.com/learn/articles/data-visualization

https://geteducated.com/careers/how-to-become-a-data-scientist

https://datascience.berkeley.edu/about/what-is-data-science/

https://Chapter/203_/20Data/0Cleaning/20Steps/0and/20Techniques20-20Data20Science20l

https://towardsdatascience.com/what-is-data-cleaning-how-to-process-data-for-analytics-and-machine-learning-modeling-c2afcf4fbf45

https://rpubs.com/ninjazzle/DS-JHU-3-4-Final

https://www.sanfoundry.com/data-science-questions-answers-pandas-data-structure/

https://www.sanfoundry.com/data-science-questions-answers-raw-processed-data/

https://www.proprofs.com/quiz-school/topic/data-science

Seven Major Steps for Building a Data Science Model, *https://towardsdatascience.com/seven-major-steps-for-building-a-data-science-model-c1761408dd17*

10 Machine Learning Algorithm every Data Scientist should know*,* *https://analyticsindiamag.com/10-machine-learning-algorithms-every-data-scientist-know/*

https://www.data.cam.ac.uk/data-management-guide/organising-your-data

How to Organize Your Data Science Project, *https://towardsdatascience.com/how-to-organize-your-data-science-project-dd6599cf000a*