A Detailed Study on Data Science Work Flow

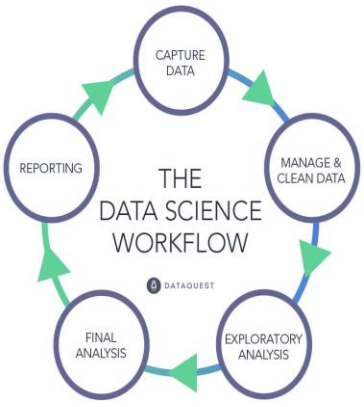
Ms R. Divya Sharon, Mr.Prem Saidhar
II CSE ,
Muthayammal Engineering College
Kakaveri, Rasipuram, Tamilnadu, INDIA

***Abstract*: Data science is a field of study and practice that's focused on obtaining insights from data. Practitioners of data science use programming skills, statistics knowledge, and machine learning techniques to mine large data sets for patterns that can be used to analyze the past or even predict the future. Analyzing these new digital data sets data required both the statistics knowledge of a statistician and the programming skills of a computer scientist. At the same time, computer processing power had advanced to the point that complex analyses of huge data sets was possible, and more advanced techniques like predictive analytics with machine learning were coming into reach. In this paper I analysis the detailed workflow of data science.**

I. INTRODUCTION

Data science

Both business and academia began to recognize the value of having experts with the programming skills required to collect, manipulate, and analyze digital data *and* the statistics skills required to select the type of analysis needed to accurately answer questions and gain meaningful insights. "Data Science," a term that had been around for decades by that point, became the mainstream phrase of choice to describe this confluence of skills. In day to day work, data scientists are often responsible for everything that happens to data, from collecting it all the way through analyzing it and reporting on the results. Although every data science job is different, here's one way to visualize the data science workflow, with some examples of typical tasks a data scientist might perform at each step.



It works like this:

1. **Capture data.** For example: pulling the data from a company database, scraping it from a website, accessing an API, etc.
2. **Manage data.** For example: properly storing the data, and will almost always involve cleaning the data.
3. **Exploratory Analysis.** For example: performing different analyses and visualizing the data in various ways to look for patterns, questions, and opportunities for deeper study.
4. **Final Analysis.** For example: digging deeper into the data to answer specific business questions, and fine-tuning predictive models for the most accurate results.
5. **Reporting.** For example: presenting the results of analysis to management, which might include writing a report, producing visualizations, and making suggestions based on the results of analysis. Reporting might also mean plugging the results of analysis into a data product or dashboard so that other team members or clients can easily access it.

II. BUSINESSES BENEFIT FROM HIRING DATA SCIENTISTS

At the highest level, data science is what allows companies to convert data into actual business value. Consider, for example, a specialty ecommerce retailer. Such a company might get tens of thousands of page views every day, and hundreds of orders. With each page view it can use automated tools to collect a large amount of data about who the visitors are and what actions they've taken on the site. And with each order, their sales system can easily collect a variety of data points about actual customers. Because there are many different types of data companies can collect, there are a wide variety of ways data scientists can add value. Here are just a few examples of how data science adds value at businesses across the globe:

- Improving decision-making — data science gives management actionable intelligence that leaders can use to shape short- and long-term strategies.
- Improving hiring — data science can help more objectively evaluate candidates and root out inefficiencies and biases.
- Predicting the future — using machine learning algorithms, data scientists can find patterns in data that humans would not be able to, and forecast future results with a higher level of accuracy.
- Improving targeting — data science can help companies find new target markets, better understand existing customers, and more accurately predict what customers want.
- Identifying new opportunities — by exploring data and looking for patterns, data scientists can identify new business opportunities that might not otherwise be apparent.

- Improving risk assessment — data science often makes it possible to "test" risky ideas by running the numbers before putting them into action, allowing companies to avoid potentially costly risks and mistakes.
- Fostering data-first culture — a data scientist or data science team can help facilitate data-based decision-making in every team across the company by providing them with data tools like dashboards and the training necessary to understand them.

With continuing path-breaking advancements in information technology, majority of data in today's world is stored and transferred in the form of document files, .PDF files, electronic-forms, codes, mails, web-content etc. However, a significant amount of data continues to accumulate in the form of manually filled forms, written documents, letters etc. In such a scenario, where there is no singularity of either form or sources of data, multiple methods of data capturing are required which can be used for data capture based upon the original source or form in which data is present.

Organizations and businesses need to determine the best way of carrying out data capture, as fits their purpose. Here, in this article, we identify some basic methods of data capture and highlights the significant differences among them.

III. EFFECTIVE WAYS TO DATA CAPTURE

Depending upon the procedure of collecting information, the data capture process can be divided into two segments:

Manual Data Capture: In manual data capture process, the data is entered manually by an operator using input devices like keyboard, touch screens, mouse etc. for keying in data in the form of figures or text into particular software such as Excel or any other data or word processing program. This method of data collection is labor intensive, time consuming and so businesses find it efficient to migrate to automated methods of data capture. Briefly, the methods of manual data capture include using:

- Mouse
- Graphics tablet
- Keyboard
- Touch-screen – e.g. PDA
- Tracker ball

Automated Data Capture: Automated data capture involves the use of computerized technology to capture data. This method has a high initial cost on account of the initial investment required as for instance, the purchase of technology but as the project proceeds, is found to lower the operating costs significantly on account of low manpower requirement. Automated data capture includes the use of different technologies such as OCR, ICR, OMR and others, which are individually described here.

Optical Character Recognition (OCR): OCR technology is used to convert different types of machine-printed documents including image files, PDF files or scanned paper documents, into searchable and editable data.

Intelligent Character Recognition (ICR): ICR technology helps to recognize and capture handwritten printed characters from image files. As handwritten text caries significantly, so ICR is less accurate and complicated as compared to other technologies. However, the technology evolves continuously by itself and as the number of samples processed increases, the accuracy increases.

Optical Mark Reading (OMR): OMR technology is used to capture human marked data from documents such as forms and surveys. The technology has the capacity to differentiate between marked and unmarked boxes and so, is used for capturing data through boxes that are checked manually on documents. **Magnetic Ink character Recognition (MICR):** :It is a data capture technology capable of recognizing characters. It involves the recognition of specially formatted characters that are printed in magnetic ink, by a machine. The technology is mainly used by banking industry to speed up the processing of cheques and other documents. . **Magnetic Stripe Cards:** :Magnetic stripe cards store data using magnetic properties of certain materials. They possess stripes of iron-based magnetic materials on the card. They are used for electronically storing particular numbers related to credit cards, identity cards and they enable automated data transfer when they reswiped in magnetic readers. **Smart-Cards:** :Smart cards are pocket-sized cards with embedded integrated circuits. They can function on contact or can be contactless. They contain more memory than magnetic cards and can be used for data related to personal identification, authentication, biometrics etc. Upon interaction with suitable reading devices they enable automated information transfer and data access.

Web-Data Capture: Data capture from web involves the capture of data from electronic forms through internet or intranet.

Voice-Recognition: Voice recognition is the process of converting speech into text. The text can be simple text or can be a set of commands. It finds application in dictation systems, small controlling systems and certain processes of data entry and word-processing environment. A suitable selection of a data capture tool is bound to make the business process resource- efficient, cost-effective and time saving. .

IV. CLEAN DATA USING DATA CLEANING TECHNIQUES

Data forms the backbone of any data analytics. To perform the data analytics properly we need various data cleaning techniques so that our data is ready for analysis. It's commonly said that,"Data scientists spend 80% of their time cleaning and manipulating data and only 20% of their time actually analyzing it." Thus, it is important to grow accustomed to the process of data cleaning techniques and all of the data cleansing tools that are related to data cleansing methods. This article covers the following data cleaning steps in Excel along with data cleansing examples:

1. Get Rid of Extra Spaces
2. Select and Treat All Blank Cells
3. Convert Numbers Stored as Text into Numbers
4. Remove Duplicates
5. Highlight Errors

6. Change Text to Lower/Upper/Proper Case
7. Spell Check
8. Delete all Formatting

Data Cleaning: Data cleansing or data cleaning is the process of identifying and removing inaccurate records from a dataset, table, or database and refers to recognizing unfinished, unreliable, inaccurate or non-relevant parts of the data and then restoring, remodelling, or removing the dirty or crude data. Data cleaning techniques may be performed as batch processing through scripting or interactively with data cleansing tools. After cleaning, a dataset should be uniform with other related datasets in the operation. The discrepancies identified or eliminated may have been basically caused by user entry mistakes, by corruption in storage or transmission, or by various data dictionary descriptions of similar items in various stores.

Are Data Cleaning Techniques Essential?

Data cleaning techniques are not only an essential part of the data science process – it's also the most time-consuming part. Without the data cleaning techniques, the neural networks and image identification modules will not be as efficient as we want them to be. With the rise of big data, data cleaning methods has become more important than ever before. Every industry – banking, healthcare, retail, hospitality, education – is now navigating in a large ocean of data.

And as the data pool is getting bigger, the variables of things going wrong too are getting larger. Each fault becomes difficult to find when you can't just look at the whole dataset in a spreadsheet on your computer. In fact, this could be true for a variety of reasons.

Data Cleansing Examples and Data Cleaning Methods in Excel

In this post, I will show you various ways to clean data in Excel with data cleansing examples & data cleansing techniques.

1. Data Cleaning Techniques-Get Rid of Extra Spaces

Here I have the text **Welcome To Digital Divya** written in four different ways.

welcome to digital divya

welcome to digital divya
welcome to digital divya
welcome to digital divya

Now, this could typically be the case if you get this data from a colleague or you get it from a text file or imported from a database. So to clean this data and get rid of these extra spaces you can use the function *trim*.
Syntax: =TRIM(Text)

Trim function takes one single argument and it would remove all the leading spaces and trailing spaces and extra spaces between words except one single space that is allowed.

DATA CLEANING TECHNIQUES-SELECT AND TREAT ALL BLANK CELLS

If you just need to use the text you can convert it into values by using *paste special* I have student names here and their marks in three subjects. You can see that there are gaps in this dataset which could be because the student could not appear in the exam. Now you may not want to leave this data set with blanks, you may want to type *not appear* in all these cells which are blank. So to do that you can either go and select each cell manually and type *not appear*. But if you have a huge data set that because this could be very tiresome. So to do it at one go,

- Select the entire data set,
- Go to *find and select* and select this option *Go to Special* this opens the go-to special dialog box. You can also use the keyboard shortcut *F5* and when you do this it opens the go-to *dialog box* here you have *special button*, click on it and it again opens it equal to special dialogue box

- Click *blanks* and click *okay*, this would select all the blank cells in your data set at the same time
- So now you have these cells in grey and the first cell is in white because this is the active cell so to type *not appear* in all these cells just start typing *not appear* and hit *ctrl+enter* and as soon as you hit *ctrl+enter* this gets entered in all the cells.

Data Cleaning Techniques-Convert Numbers Stored as Text into Numbers : Here I have this number entered in three different ways,

123
123
'123

In the first case, it is a number as you can see it is aligned to the right of the cell numbers are always aligned to the right while text gets aligned to the left of a cell and in the other two cases, you can see these are text format because these are aligned to the left. Now to convert all these three back into numbers.

4. Data Cleaning Techniques-Remove Duplicates

Here I have a data set of students and their marks in three subjects and there are duplicates in this data, there are two ways to do it first is using **conditional formatting**,

- So you can select the data set
- Go to home -> conditional formatting
- *Highlight cell rules -> duplicate values* and as soon as you select this it gives you the option to highlight

duplicates and the formatting. I will keep the formatting as a red fill with dark red text and when I hit OK you can see that this has been highlighted and all those numbers and names that appear more than once it highlighted in red.

5. Data Cleaning Techniques-Highlight Errors

Here I have a dataset for five companies. I have their revenue number for three years and net income numbers for three years and using these numbers I have calculated the net income margin which is net income by revenue. if you have a huge data set these errors could be difficult to spot so to do that you can use **conditional formatting,**

So select this entire dataset go-to *home -> conditional formatting* and select *new rule* within *new formatting rule dialog box* select *format only cells that contain* and from this drop-down select *errors* when you select errors you would get the option to format the cells which have error, in this case, let me select *red* and I click OK and as soon as I do this all the cells that have errors in it get highlighted in red

6.Data Cleaning Techniques-Change Text to Lower/Upper/Proper Case **:**Here I have names written in different ways you can see either it could be all caps, it could be all lowercase and in some cases, it's a mix-and-match of uppercase lowercase so to make it all consistent you can use one of these three formulas,

SYNTAX:
LOWER() – Converts all text into Lower Case.
ex. mary jane
UPPER() – Converts all text into Upper Case.
ex. MARY JANE
PROPER() – Converts all Text into Proper Case.
ex. Mary Jane

7. Data Cleaning Techniques-Spell Check: select the data and press **F7** and when you do that it runs the spellcheck for you and it is the same thing that you see in Microsoft Word or PowerPoint it will show you the text that it thinks is a spelling error and it will show you the suggestions as well so you can change these.

8. Data Cleaning Techniques-Delete all Formatting **:**If you have a worksheet where there is a lot of formatting and you need to clear all the formatting you can quickly do that by,

- Selecting the entire data
- Go to Home –> Clear –> Clear Formats
- You can also use *clear all* this would remove everything from your sheet including the content you can only clear the content would remain the formatting would remain intact you can clear the comments and the hyperlinks.

Data Cleansing Tools : Here are some interesting Data Cleansing tools relating to data cleaning techniques, analysis and modeling of data,

JASP – Open Source statistical software similar to SPSS with support of COS
Rattle – GUI for user-friendly machine learning with R
RapidMiner – Another point and click machine learning package
Orange – Open Source GUI for user-friendly machine learning with Python
Talend data preparation – Data cleaning, preparation tool with smarts
Trifacta Wranger – Data cleaning, preparation tool with the match by example feature
They are all open source or have free versions focusing on cleaning, analysing and modelling data.

Data cleaning is an inherent part of the data science process to get cleaned data. In simple terms, you might divide data cleaning techniques down into four stages: collecting the data, cleaning the data, analyzing/modelling the data, and publishing the results to the relevant audience.

V. EXPLORATORY DATA ANALYSIS

Exploratory data analysis is generally cross-classified in two ways. First, each method is either non-graphical or graphical. And second, each method is either univariate or multivariate (usually just bivariate).

Exploratory Data Analysis with Chartio

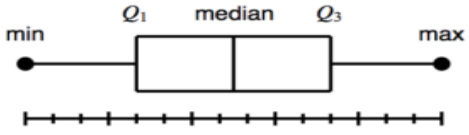
Let us perform exploratory data analysis on the **iris dataset** to familiarize ourselves with the EDA process. Let's look at the sample data:

epal_length	sepal_width	petal_length	petal_width	species
5	2	3.5	1	versicolor
6	2.2	4	1	versicolor
6	2.2	5	1.5	virginica
6.2	2.2	4.5	1.5	versicolor
4.5	2.3	1.3	0.3	setosa

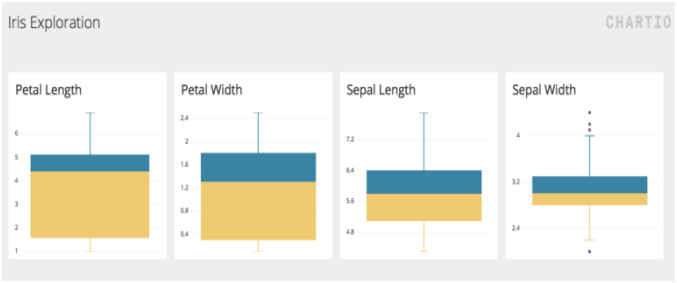
The dataset contains four features – **sepal length, sepal width, petal length, and petal width** for the different species (versicolor, virginica, setosa) of the flower, iris. Also, for each species there are 50 instances (rows of data).

Univariate Analysis: Univariate analysis is the simplest form of data analysis where the data being analyzed contains only one variable. Since it's a single variable it doesn't deal with causes or relationships. The main purpose of univariate analysis is to describe the data and find patterns that exist within it. Let us look at a few visualizations used for performing univariate analysis.

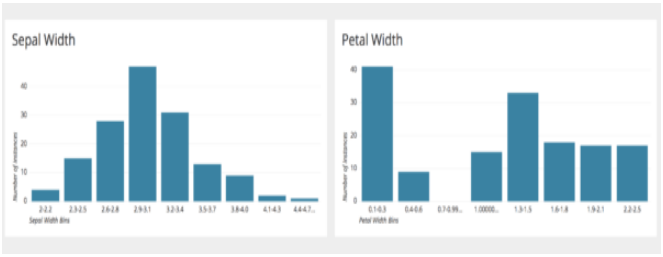
BOX PLOTS : A box and whisker plot – also called a box plot – displays the five-number summary of a set of data. The five-number summary is the minimum, first quartile, median, third quartile, and maximum.



The **box plots created in Chartio** provide us with the summary of the four numerical features in the dataset. We can observe that the distribution of petal length and width is more spread out, as exhibited by the bigger size of the boxes.



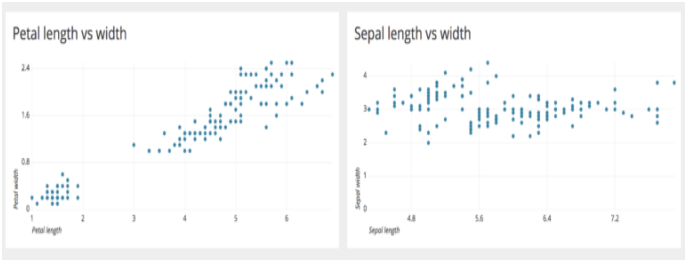
HISTOGRAM :A histogram is a plot that lets you discover, and show, the underlying frequency distribution (shape) of a set of continuous data. This allows the inspection of the data for its underlying distribution (e.g. normal distribution), outliers, skewness, etc.



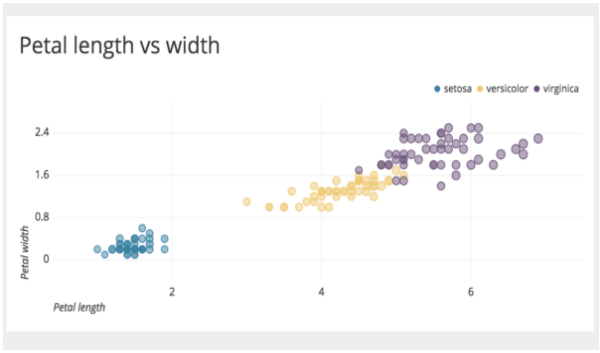
The above plots show the histogram of sepal and petal widths made in Chartio. From the charts it can be observed that the sepal width follows a **Gaussian distribution**. However, petal width is more skewed towards the right, and the majority of the flower samples have a petal width less than 0.4 cm.

Multivariate analysis :Multivariate data analysis refers to any statistical technique used to analyze data that arises from more than one variable. This essentially models reality where each situation, product, or decision involves more than a single variable. Let us look at a few visualizations used for performing multivariate analysis.

SCATTER PLOT: A scatter plot is a two-dimensional data visualization that uses dots to represent the values obtained for two different variables – one plotted along the x-axis and the other plotted along the y-axis.

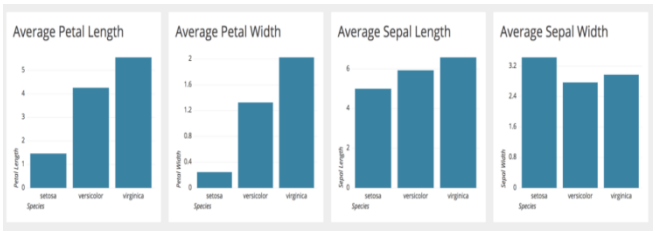


Above are examples of two scatter plots made using Chartio. We can observe that there is a linear relationship between petal length and width. However, with increase in sepal length, the sepal width does not increase proportionally – hence they do not have a linear relationship.



We can observe that 'setosa' species has the lowest petal length and width, 'virginica' has the highest, and 'versicolor' lies between them. By plotting more dimensions, deeper insights can be drawn from the data.

Bar Chart : A bar chart represents categorical data, with rectangular bars having lengths proportional to the values that they represent. For example, we can use the iris dataset to observe the average petal and sepal lengths/widths of all the different species.



Observing the bar charts, we can conclude that 'virginica' has the highest petal length, petal width and sepal length, followed by 'versicolor' and 'setosa'. Apart from the charts shown in our EDA example, we can use various other charts depending on the characteristics of our data:

1. **Line plot** to represent changes over time
2. **Pie charts** to show the relationship between a part to a whole
3. **Map charts** to visualize location data

VI. STRUCTURE OF A DATA ANALYSIS REPORT

A data analysis report is somewhat different from other types of professional writing that you may have done or seen, or will learn about in the future. It is related to but not the same as: Divide the body up into several sections at the same level as the Introduction, with names like: – Data – Methods – Analysis – Results This format is very familiar to those who have written psych research papers. It often works well for a data analysis paper as well, though one problem with it is that

the Methods section often sounds like a bit of a stretch: In a psych research paper the Methods section describes what you did to get your data.

VII. CONCLUSION

Such a way that , the data science is organized by the data scientist. Data Science is the study of data. It is about extracting, analyzing, visualizing, managing and storing data to create insights. These insights help the companies to make powerful data-driven decisions. Data Science requires the usage of both unstructured and structured data. It is a multidisciplinary field that has its roots in statistics, math and computer science. It is one of the most highly sought after jobs due to the abundance of data science position and a lucrative pay-scale. Being a less-saturated, high paying field that has revolutionized several walks of life, it also has its own backdrops when considering the immensity of the field and its cross-disciplinary nature. Data Science is an ever-evolving field that will take years to gain proficiency.

VIII. REFERENCES

[1] John Tukey-The Future of Data Analysis-July 1961
[2] Schutt, Rachel; O'Neil, Cathy (2013). Doing Data Science. O'Reilly Media. ISBN 978-1-449-35865-,
[3] Clean Data in CRM: The Key to Generate Sales- Ready Leads and Boost Your Revenue Pool Retrieved 29th July, 2016
[4] Microsoft Research. Retrieved 26 October 2013.Perceptual Edge-Jonathan Koomey-Best practices for understanding quantitative data-February 14, 2006
[5] Hellerstein, Joseph (27 February 2008). "Quantitative Data Cleaning for Large Databases" (PDF). EECS Computer Science Division: 3. Retrieved 26 October 2013.
[6] Stephen Few-Perceptual Edge-Selecting the Right Graph For Your Message-September 2004
[7] Behrens-Principles and Procedures of Exploratory Data Analysis-American Psychological Association-1997
[8] Grandjean, Martin (2014). "La connaissance est un réseau" (PDF). Les Cahiers du Numérique. 10 (3): 37–54. doi:10.3166/lcn.10.3.37-54.
[9] Stephen Few-Perceptual Edge-Selecting the Right Graph for Your Message-2004
[10] Stephen Few-Perceptual Edge-Graph Selection Matrix
[11] Robert Amar, James Eagan, and John Stasko (2005) "Low-Level Components of Analytic Activity in Information Visualization"