Data Science Fundamentals

With Anthony Mipawa

About Me

- Software engineer @ Neurotech Africa
- Ambassador @ Zindi Africa

- LinkedIn: Anthony Mipawa
- Twitter: @LoytTony

What is data Science?

About Data Science

Turning data into informations

Analyzing data to get insights

Identifying trends, patterns and Correlations

Contextualizing, Applying and understanding them

> In data science we use tools from coding, statistics & math to work *creatively* with data.
>
> Ways may vary a lot.
>
> The Goal is to get insights.

What does a data scientist do?

“Data scientists use data to answer questions.”

- ❑ Get and process data to convert it from its raw format to a cleaner format
- ❑ Calculate and interpret statistical variables
- ❑ Create visualizations and draw conclusions for analysis
- ❑ Suggest applications from the information and develop machine learning implementations.

Essentials of data science

- ❑ Statistics
- ❑ Programming
- ❑ Domain Knowledge/Understanding

Statistics

- ❑ Understanding the different types of data you can encounter.
- ❑ Understanding key statistical terms.
  - ❑ Type of means
  - ❑ Fluctuations in data
- ❑ Splitting up, grouping, and segmenting data points.

Programming

- □ Python
- □ R
- □ SQL

Programming Cont..

Why knowing how to program makes your life so much easier:-

- Ease of automation
- Being able to customize,explore, prototype and test

Programming Cont..

Essential packages to use in python

- ❑ **Pandas** for data analysis
- ❑ **Numpy** for computational analysis
- ❑ **Matplotlib** and **seaborn** for data visualization
- ❑ **S-klearn** for data preprocessing and Modeling

Statistical Data Types

- Numerical/quantitative data
  - discrete
  - Continuous
- Categorical/qualitative data
  - eg gender, nationality, ethnicity
  - Can't be compared
- Ordinal- A mixture of numerical and categorical data
  - -eg hotel ratings

Three Types of Average

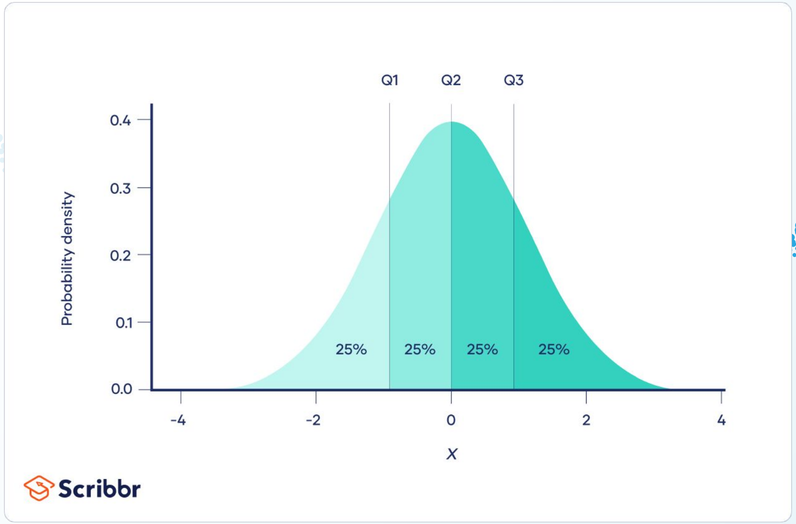
- Mean
  - Add up all the values and divide this total by the number of values.
- Median
  - Places all your values in order from smallest to highest and finds the one in the middle.
- Mode
  - Most commonly occurring value.

Three Types of Spread

- ❑ Range + Domain
  - ❑ Range = Maximum - Minimum
  - ❑ Domain is the value that your data points can take on.
- ❑ Variance + Standard Deviation
  - ❑ Variance tells how much the values of your data differ from the mean value.
  - ❑ Standard Deviation is a square root of variance.
- ❑ Covariance + Correlation
  - ❑ Covariance tells how much one value varies when the other varies
  - ❑ Correlation : Covariance divided by S.D of each variable

Quantiles + Percentiles

- Quantiles:
  - Splitting your data into a certain number of regions that each have the same probability.
  - Splitting data into a certain regions so that each contains equal number of data points.
  - e.g quartile(4 parts)



Quantiles + Percentiles Cont..

- ❑ Percentiles:
  - ❑ Splitting the data into 100 equal segments.
  - ❑ Examples:

    Getting a test score of 93% places you in the 99th percentile, meaning your score is high than 99% of the people that took this test.

    This is a good for normalization, because it lets you judge someone's performance by having it relatives to the performance of everyone else.

    A test score of 60% that puts you in 95th percentile means the test was very difficult and you did much better than most other people on it.

Data Visualization

- ❑ Roles of the computer
- ❑ Roles of Human Being
- ❑ Presenting Data
- ❑ Interpreting Data

Data Visualization Cont..

- ❑ Roles of the computer
  - ❑ Much faster at calculating than a human
  - ❑ Great for crunching numbers
  - ❑ Great for doing many repetitive tasks
  - ❑ Carrying out tasks that we gave based on logical thinking

Data Visualization Cont..

- ❑ Roles of the Human
  - ❑ We've developed to identify patterns.
  - ❑ Creativity.
  - ❑ bring in or remembering outside knowledge.
  - ❑ understanding summary values and images.

We are able to look at things and use our general understanding to recognise patterns.

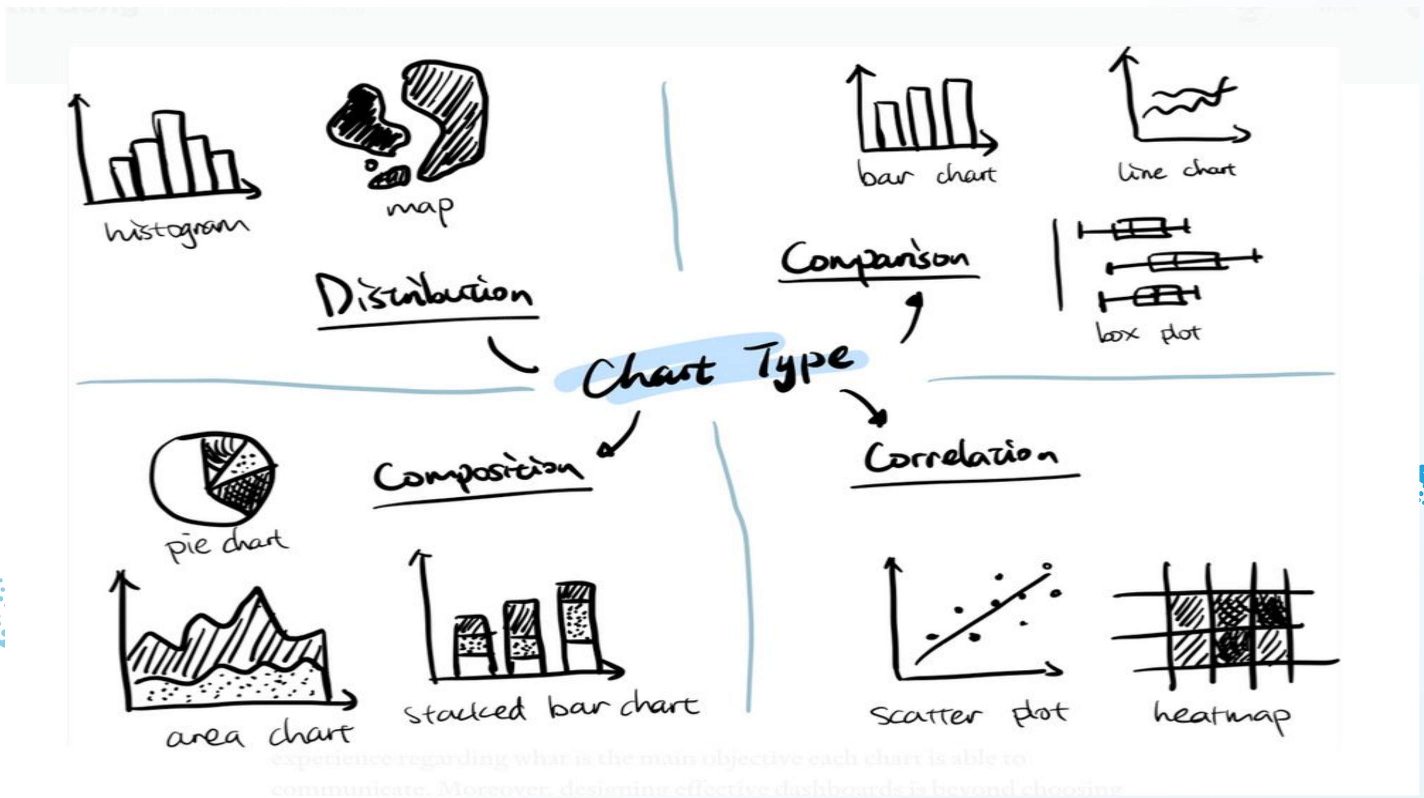
Data Visualization Cont..

- Presenting and Interpreting
  - Not always statistical summary can be useful to bring point across
  - Visualization allow us to communicate and understand the data
  - We use our domain knowledge to come up with findings
  - Considering the contextual of the data

Data Visualization Graphs

- ❑ Histogram
- ❑ Bar Plot
- ❑ Pie Chart
- ❑ Scatter Plot
- ❑ Line Plot
- ❑ Box and Whisker plots
- ❑ HeatMap

Data Visualization Graphs Cont..



experience regarding what is the main objective each chart is able to communicate. Moreover, designing effective dashboards is beyond choosing

The Roadmap

As a data professional the great work to do is even before the data and the algorithms that is all about decision making.

Where to start?

- Programming language
- Math basics
- Data science
- Machine Learning

Where to start? Cont..

- Deep Learning
- Natural Language Processing
- Business and communication
- Deployment

Resources & Tips

Resources

1. Coursera
2. Kaggle
3. Edx
4. WorldQuant University
5. Zindi Africa
6. Khan academy for math
7. Mathisfun.com

Tips

1. Curiosity
2. Mentor
3. Focus on acquiring skills rather than certifications
4. Find syllabus on the platforms like udacity etc to guide your journey
5. Community engagement

Question

Are you ready to be a data professional?

If yes?

Then start today, the world of data needs you!

Thanks!

Any questions?