**Data Analyst**

**The funfamental of data analyst 1**

**Definisi**

Merupakan segala sesuatu yang di proses dengan cara

* Proses data collection = Mengumpulkan data baik itu melalui survey, device, wawancara, dsb.
* Data cleaning and transformation = data yang tidak sesuai dapat di cleaning dan ditransformasikan terlebih dahulu
* Data visualization = bagaimana data tersebut dapat diolah menjadi bentuk visual yang dapat kelola orang banyak orang
* Data modelling = menentukan model yang akan digunakan agar lebih mudah untuk dipahami oleh orang lain.

Tugas dari seorang data analyst ialah menyampaikan informasi penting dan Menarik kesimpulan dari data yang telah diolah.

4 jenis data analisis :

* Deskriptif = Menggambarkan suatu peristiwa
* Diagnostik = Menjelaskan penyebab atau latar belakang dari suatu pristiwa
* Prediktif = Memperkirakan apa yang akan terjadi di masa yang akan datang
* Preskriptif = Merekomendasikan tindakan

**Data Analisis dalam Bisnis**

4V of Challenges in Big Data

* Volume = Data semangkin hari semangkin besar 2-3 triliun gigabyte data dihasilkan setiap hari
* Velocity (kecepatan) = 50.000 pencarian di google, 125000 vidio youtube di tonton, dan 2 juta email terkirim setiap detik
* Variety (keseragaman data) = data terbagi menjadi structured data dan unstructured data audio, visual, dokumen, informasi geografis, dll.
* Veracity (kebenaran dari data) = bias informasi, fitnah, adu domba, hoax.

Manfaat data analis untuk bisnis

* Menghasilkan keputusan yang lebih baik
* Memperkecil risiko bisnis
* Mengurangi biaya
* Menghasilkan informasi yang transparan dan objektif
* Memberikan kendali yang lebih luas

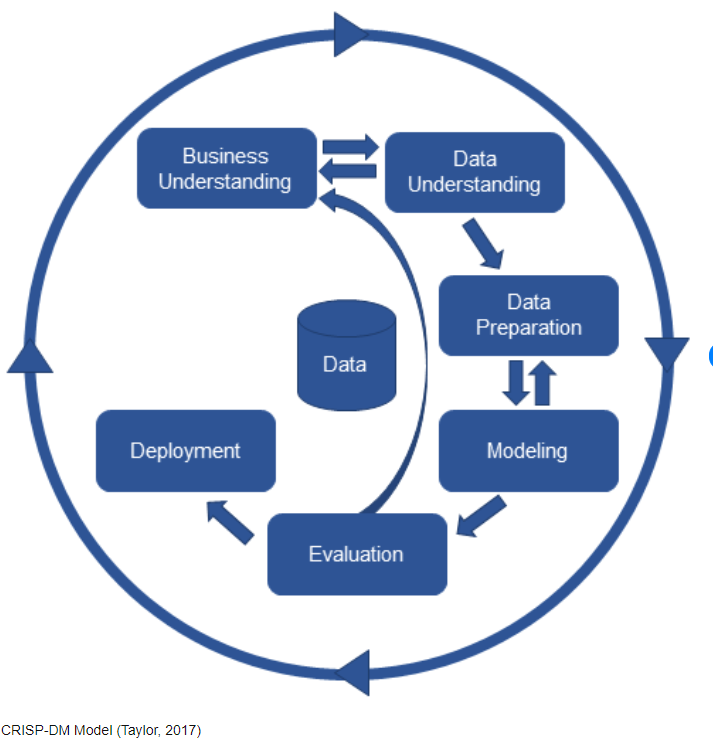
Contoh kasus

Perusahaan **Target**

* **Target** merupakan salah satu perusahaan ritel yang berbasis di amerika serikat
* untuk meningkatkan penjualan, target menggunakan analisis prediktif untuk membuat sistem rekomendasi
* Sistem rekomendasi adalah sistem yang memberikan penawaran terpersonalisasi berdasarkan data historis transaksi konsumen

**Proses Data Analis**

Framework CRISP-DM (Cross Industry Standard Process for Data Mining)



* Business Understanding bertujuan untuk mengetahui apa yang sedang dibutuhkan oleh bisnis ?
* Data Understanding bertujuan untuk mengetahui apa saja yang kita miliki dan butuhkan ?
* Data Preparation bertujuan untuk mengetahui bagaimana data itu disiapkan untuk analisis ?
* Modeling bertujuan untuk mengetahui teknik model apa saja yang bisa diterapkan ?
* Evaluation bertujuan untuk mengetahui model mana yang paling sesuai ?
* Deployment bertujuan untuk mengetahui bagaimanan stakeholder mengakses hasil analisis ?

Contoh penggunaan Framework CRISP-DM pada Dealer mobil bekas

Business Understanding

* Berapa harga yang layak dibayarkan untuk mobil bekas dengan spesifikasi tertentu ?
* Fakor apa saja yang mempengaruhinya ?

Data Understanding

* Dari hasil survey, kamu mengumpulkan data-data mobil bekas dengan informasi yang meliputi : harga, merk, tahun pembuatan, mesin, dll.
* Exploratory data analis, eksplorasi untuk lebih memahami data

Data Preparation

* Membersihkan data dari inkonsistensi, kesalahan input, missing value contohnya seperti terdapat informasi kosong seperti kilo meter yang tidak diketahui.
* Mempersiapkan data ke dalam format yang dapat diolah oleh komputer

Modeling

* Membangun model regresi untuk memprediksi harga

Evaluation

* Memilih model dengan metric RMSE (Root mean squared error) terbaik
* RMSE (Root mean squared error) merupakan cara standard and popular untuk mengukur kesalahan suatu model dalam memprediksi data kuantitatif yang menunjukan seberapa tersebar data di sekitar garis yang paling cocok.

Deployment

* Menerapkan model untuk digunakan sebagai referensi pakar dalam menentukan harga.

**Tollset data analist**

SQL

* Digunakan setiap hari oleh data analyst
* Berfungsi untuk berinteraksi dengan data di database

Python

* Digunakan untuk mengolah data berskala besar
* Membangun model statistic atau machine learning

Tableau / BI

* Menyederhanakan proses data analisis
* Sistem reporting yang terintegrasi
* Membuat dashboard monitoring dan infrografis

**Understanding Business Problems**

Menentukan atau memecahkan masalah menjadi skill yang sangat dibutuhkan dalam melakukan berbagai jenis analis. Pemecahan masalah biasanya dapat memakan waktu 60-70% dari pengerjanan analisis, hal ini dikarenakan sulitnya mencari akan masalah yang tepat. Terkadang kita perlu melakukan beberapa kali percobaan untuk menentukan akar masalah yang tepat. Akar masalah juga sangat mempengaruhi kecepatan hasil dalam menjawab persoalan yang ada

Langkah-Langkah Problem Solving

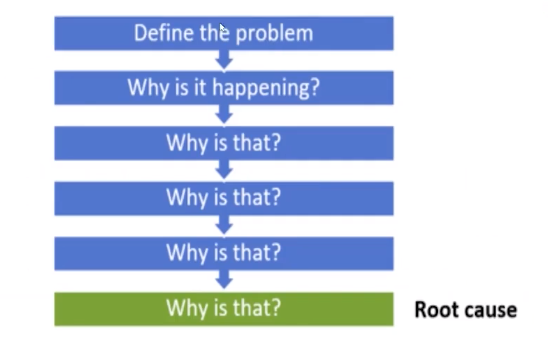
* Memahami latar belakang dan konteks.
* Menentukan stakeholders dan bertanya untuk menambah informasi.
* Membuat kerangka berpikir untuk menentukan akar masalah.

Problem solving tools

* 5 Why = Metode mencari Problem Solving denga cara menanyakan apa penyebab permasalahan muncul dan mencari jawabannya. Hal ini dilakukan berkali-kali sampai permasaahan dapat tergambar dengan jelas.

Langkah tools 5 why :

1. Tentukan masalah secara umum
2. Buat dan jawab pertanyaan kenapa masalah itu bisa terjadi
3. Ulangi langkah ke 2 sampai menemukan akar masalah yang cukup jelas, biasanya 5 pertanyaan sudah cukup untuk mendapatkan akar masalah yang didapat
4. Harus kritis, detail dan spesifik untuk setiap iterasi



* Action Priotity Chart = Diagram yang dapat digunakan untuk mencari prioritas masalah berdasarkan impact atau benefit yang bisa diberikan. Dengan diagram ini kita dapat memilih masalah yang impactnya mayoritas dari permasalahan yang ada.

Istilah

1. Effort = Usaha yang diperlukan untuk menyelesaikan masalah
2. Impact = seberapa besar hasil yang bisa didapat
3. Quick Wins = Membutuhkan usaha yang minim tapi hasil yang besar
4. Major Project = Membutuhkan usaha yang besar walaupun hasilnya besar juga
5. Fill-ins = Membutuhkan usaha minim dan menghasilkan hasil yang minim juga
6. Thankless Task = Sebisa mungkin hindari karena butuh usaha besar tapi hasil tidak terlalu berpengaruh.

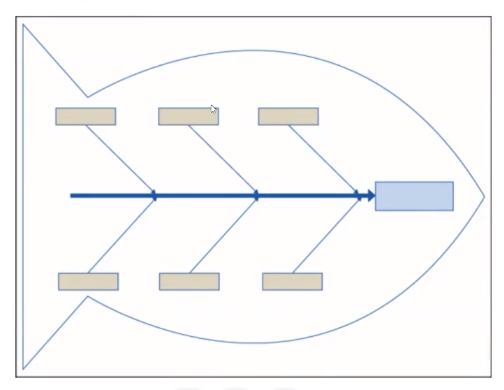
Langkah membuat Pareto Chart :

1. Tentukan masalah atau akar masalah yang ingin dianalisis
2. Tentukan impact dari setiap masalah dan buat kuantitas dari impact tersebut.
3. Tentukan effort yang diperlukan untuk menyelesaikan masalah tersebut.
4. Petakan setiap masalah pada diagram disamping, prioritaskan menyelesaikan masalah yang termasuk quick wins



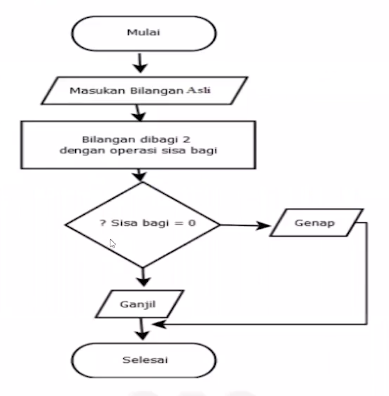
* Fishbone diagram = diagram yang berbentu tulang ikan. Diagram ini digunakan untuk merinci masalah dan membantu mencari akar masalah yang dapat dilihat dari beberapa point of view.

Langkah fishbone diagram :



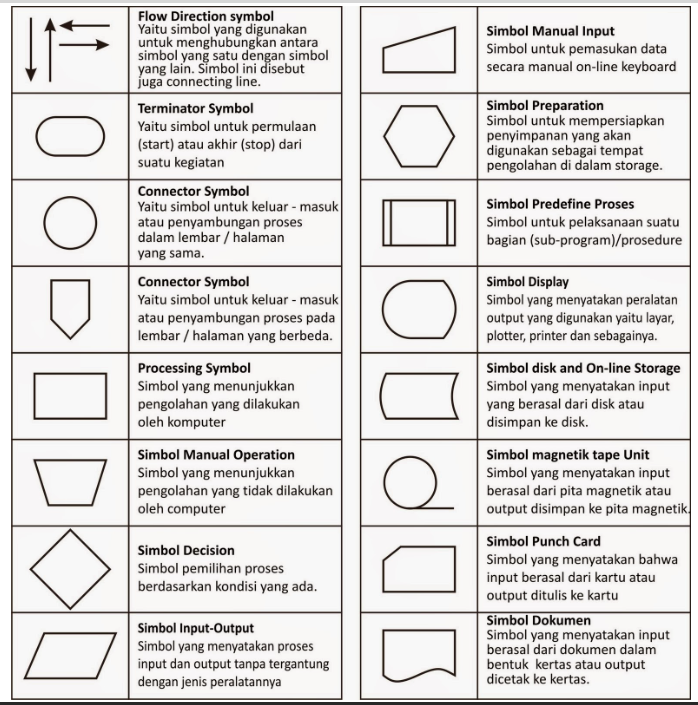
1. Kotak warna biru = sysmptoms atau masalah yang dirasakan atau diutarakan oleh stakeholders.
2. Kotak Warna Abu-abu = kategori dari point of view untuk mencari akar masalah.
3. Pada garis kecil diantara garis biru tebal dan kotak abu-abu kita bisa menuliskan akar masalah berdasarkan point of view nomor 2.

* Flowchart / Algo = Metode pemecahan masalah dengan cara memetakan flow atau membuat pseudo algoritma untuk menentukan masalah dan cara untuk mencari solusinya secara sistematis.



Langkah Flowchart / Algo :

1. Tentukan titik mulai dari chart
2. Lalu tentukan variable input
3. Tentukan proses yang akan dilakukan sampai titik akhir.



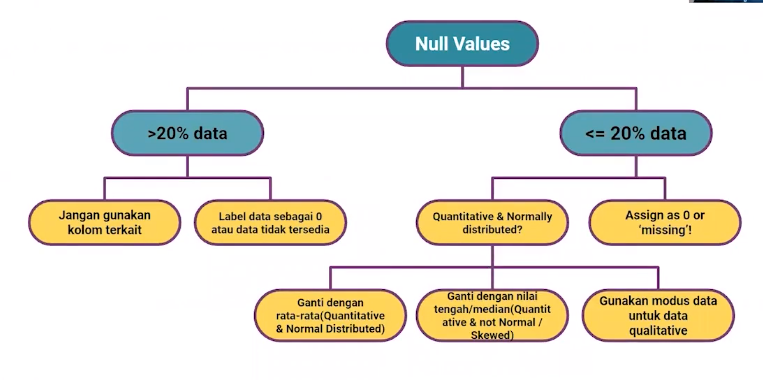
**Data Cleaning and Preparation**

Dalam proses pengolahan data sering kali data yang diolah belum tentu sudah dapat langsung digunakan dengan baik. Adapun masalah yang biasanya kita temui ketika mengolah data adalah :

1. Null Value (data kosong)
2. Kesalahan strukturan (kesalahan ketik / input)
3. Duplikasi data
4. Outlier (nilai dari data terlalu besar / kecil)

Null Value (Data kosong). Data yang tidak baik dapat menghasilkan hasil yang buru. Hal ini disebut sebagai GIGO (Garbage in garbage out). Kita dapat melakukan process pembersihan data (data cleaning) dan pengubahan data (data transformation).

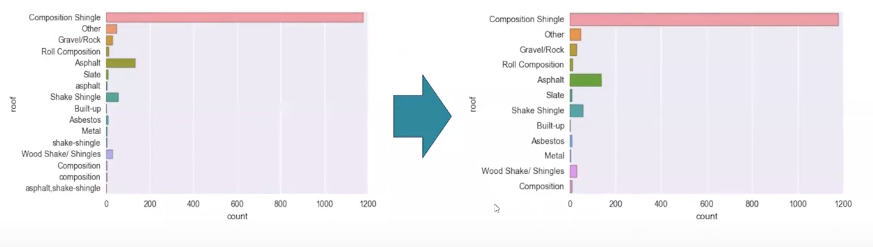
Data Cleaning (Handling Null Values)



Data Cleaning (Structural Errors)

Kesalahan strukturan merupakan kesalahan yang muncul saat pengambilan data, pemindahan data atau buruknya penyimpanan data.

Sebagai contoh, kesalahan dalam typo atau tidak konsistensinya penggunaan huruf besar atau kecil. Hal ini sering menjadi masalah pada kategorisasi. Kita dapat megeceknya dengan membuat diagram bar dan melihat frekuensi kemunculannya. Untuk menyelesaikan masalah ini kita dapat merubah typo atau menghapusnya saja jika frekuensi kemunculannya kecil dan tidak terkait hal ini.



Data Cleaning (Duplikat dan Outlier)

Untuk memeriksa duplikasi data kita dapat menghitung dan mengelompokan data lalu mengeluarkan data yang muncul lebih dari sekali. Selanjutnya hapus duplikasi data jika memang data tersebut benar-benar sama. Akan tetapi lebih baik kita pastikan terlebih dahulu apakah ada perbedaan antar data tersebut.

Outlier adalah data yang sangat jauh dari titik tengah suatu distribusi atau muncul diluar ekspektasi. Biasanya outlier adalah data yang nilainya lebih dari persentil 99% ataupun kurang dari persentil 1%. Tetapi sebenarnya tergantung dari konteks. Data outlier ini sebaiknya kita keluarkan dan tidak kita gunakan dalam pengolahan data.

**Data Transformation**

* **Formating**

seringkai kita menemukan data dimana tipe data atau nilainya tidak sesuai dengan bentuk yang kita inginkan. Sebagai contoh terkadang kita menemukan suatu kolom yang berisi nilai angka dimana angka tersebut sebenarnya mewakili suatu nilai lain. Memudahkan pengolahan data selanjutnya kita perlu mengubah nilai tersebut.

* **Ubah data ke matriks yang lebih relevan**

Rerkadang data yang tersedia belum dapat menggambarkan nilai yang kita inginkan. Tetapi dari data yang ada kita dapat membuat perhitungan klausa logika sehingga kita dapat menghasilkan nilai lain yang lebih relevan. Contohnya, kita perlu menghitung total jumlah biaya dari tabel komponen biaya, kita perlu membuat rumus dan menghitung total biaya dari komponen tersebut.

* **Normalisasi**

Normalisasi adalah proses perubahaan nilai data dengan cara membagi tiap data dengan total data atau dengan nilai terbesar dari yang ada. Biasanya ini dilakukan untuk menghitung proporsi atau untuk membandingkan nilai yang ada

**Exploratory Data Analysis & Tools**