Learning Progress Review Week-13

OPTIMISTIC

DigitalSkola

ALDIVA LUTFIA ASPRIZAL MILLENIA GILANG

DigitalSkola

SESSION 37
EVALUATION METRICS AND MODEL SELECTIONS

METRICS FOR PERFORMANCE EVALUATION

- Confusion matrix adalah sebuah tabel yang sering digunakan untuk mengukur kinerja dari model klasifikasi di machine learning. Tabel ini menggambarkan lebih detail tentang jumlah data yang diklasifikasikan dengan benar maupun salah.

- True Positive (TP) : Jumlah data yang bernilai Positif dan diprediksi benar sebagai Positif.
- False Positive (FP) : Jumlah data yang bernilai Negatif tetapi diprediksi sebagai Positif.
- False Negative (FN) : Jumlah data yang bernilai Positif tetapi diprediksi sebagai Negatif.
- True Negative (TN) : Jumlah data yang bernilai Negatif dan diprediksi benar sebagai Negatif.

		Predicted	
		Positive	**Negative**
Actual	**Positive**	True Positive	False Negative
	Negative	False Positive	True Negative

DigitalSkola

METRICS FOR PERFORMANCE EVALUATION

- Accuracy adalah proporsi data yang diprediksi tepat baik yg bernilai positif atau negatif.terhadap seluruh data

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

- Precision adalah proporsi data bernilai positif yang diprediksi tepat dari perspektif data yang diprediksi

$$Precision = \frac{TP}{TP + FP}$$

- Recall adalah proporsi data bernilai positif yang diprediksi tepat dari perspektif data aktual

$$Recall = \frac{TP}{TP + FN}$$

- F-score adalah skor tunggal yang menyeimbangkan precision dan recall

$$F - score = 2\ x\ \frac{Precision\ x\ Recall}{Precision + Recall}$$

DigitalSkola

METRICS EVALUATION FOR REGRESSION

- Mean Square Error(MSE)/Root Mean Square Error(RMSE)

- MSE dihitung dengan jumlah kuadrat kesalahan prediksi yang merupakan output target dikurangi output yang diprediksi dan kemudian dibagi dengan jumlah titik data. Ini menunjukkan nilai absolut seberapa jauh hasil prediksi Anda menyimpang dari angka sebenarnya.

- Root Mean Square Error (RMSE) adalah akar kuadrat dari MSE. Ini digunakan lebih umum daripada MSE karena pertama nilai MSE bisa terlalu besar. Kedua, MSE dihitung dengan kuadrat kesalahan, dan dengan demikian akar kuadrat mengembalikannya ke tingkat kesalahan prediksi yang sama dan memudahkan interpretasi.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

DigitalSkola

METRICS EVALUATION FOR REGRESSION

- Mean Absolute Error(MAE)

- Mean Absolute Error (MAE) mirip dengan Mean Square Error (MSE). Namun, MAE mengambil jumlah nilai absolut kesalahan.

- Dibandingkan dengan MSE atau RMSE, MAE adalah representasi yang lebih langsung dari nilai error. MSE menghasilkan nilai error lebih besar untuk kesalahan prediksi besar dengan mengkuadratkannya sementara MAE memperlakukan semua kesalahan sama.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

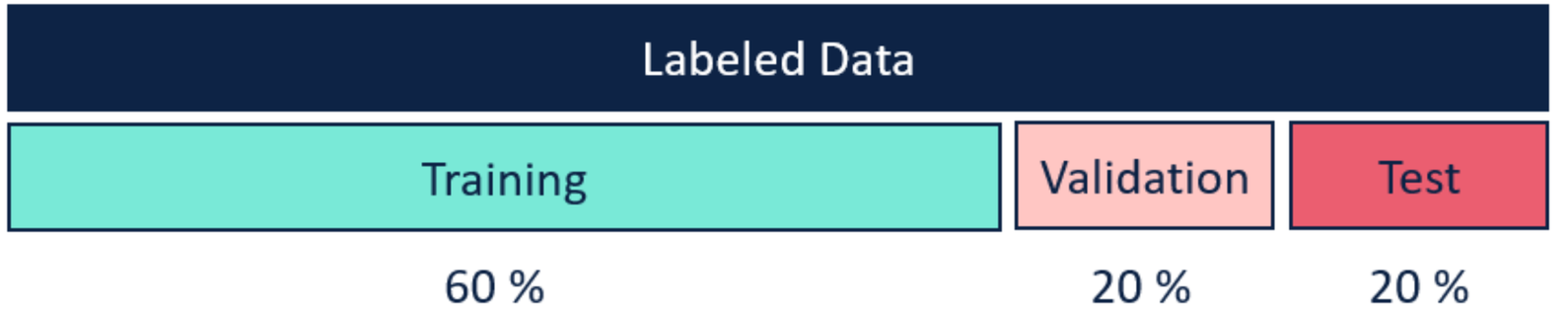
VALIDATION TECHNIQUES

HOLDOUT

- Holdout adalah teknik validasi model paling simpel dengan membagi data menjadi data latih dan data tes. Contoh proporsinya 80:20, 75:25, atau 70:30.

- Teknik holdout memiliki keterbatasan untuk mengatasi data dengan kondisi berikut:
  -Data yang terbatas
  -Data input memiliki sebaran yang berbeda dengan data sampel.

- Untuk kondisi tersebut bisa diatasi dengan teknik validasi lain yaitu Cross-Validation.

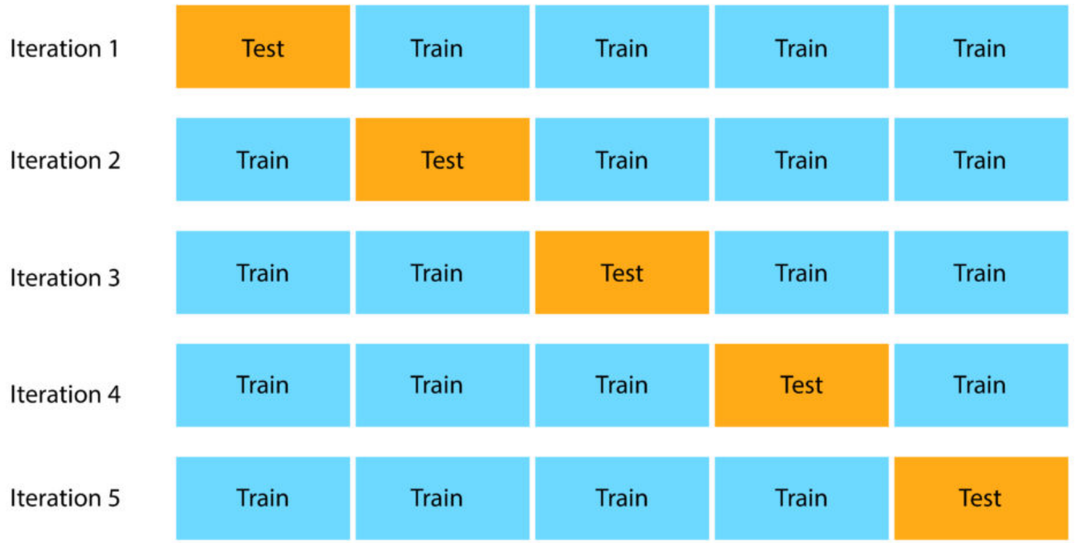


VALIDATION TECHNIQUES

CROSS-VALIDATION

- Cross-Validation adalah teknik validasi model dengan cara membagi dataset menjadi K jumlah fold atau bagian. Setiap fold memiliki kesempatan untuk menjadi data tes. Beberapa teknik cross-validation yang populer antara lain 5-k CV, 10-k CV, LOOCV

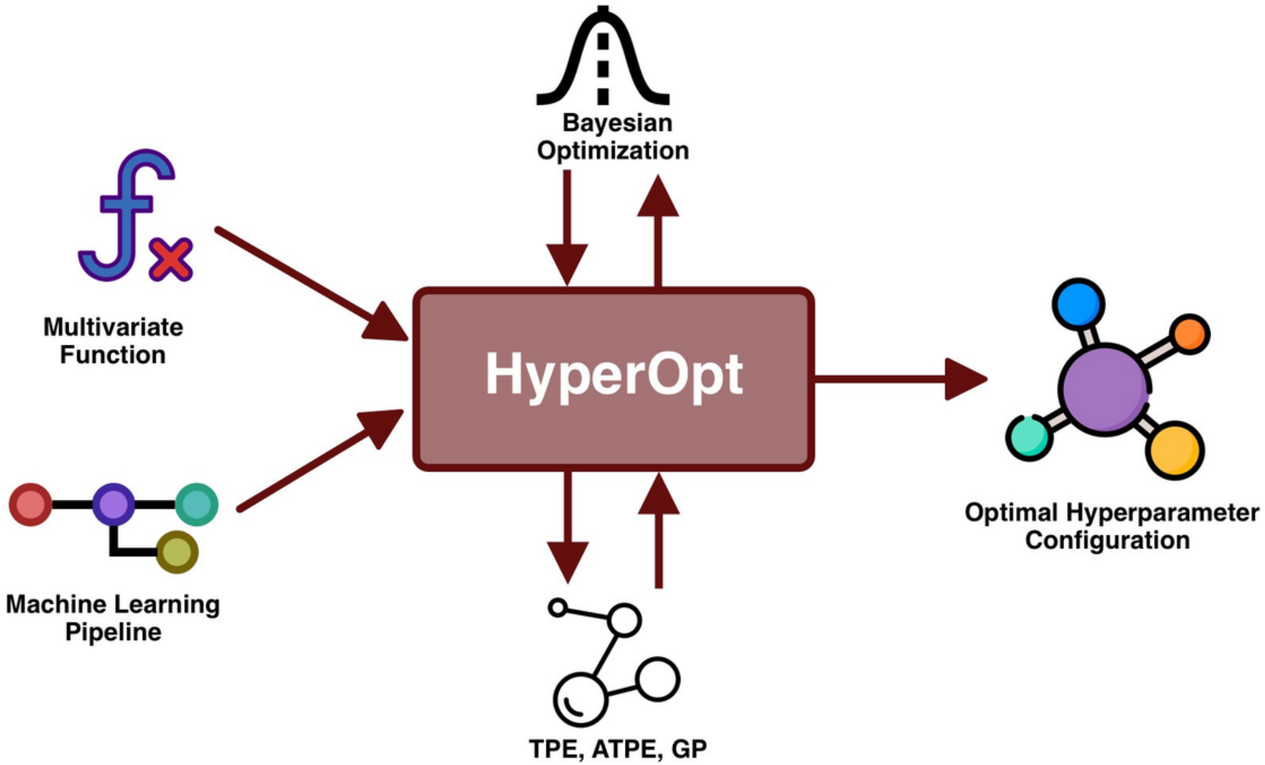
- Cara evaluasinya adalah membandingkan nilai evaluasi matriks dari masing-masing fold. Jika ada yang berbeda jauh, berarti ada yang perlu dicurigai dari model atau data.



HYPERPARAMETER TURNING

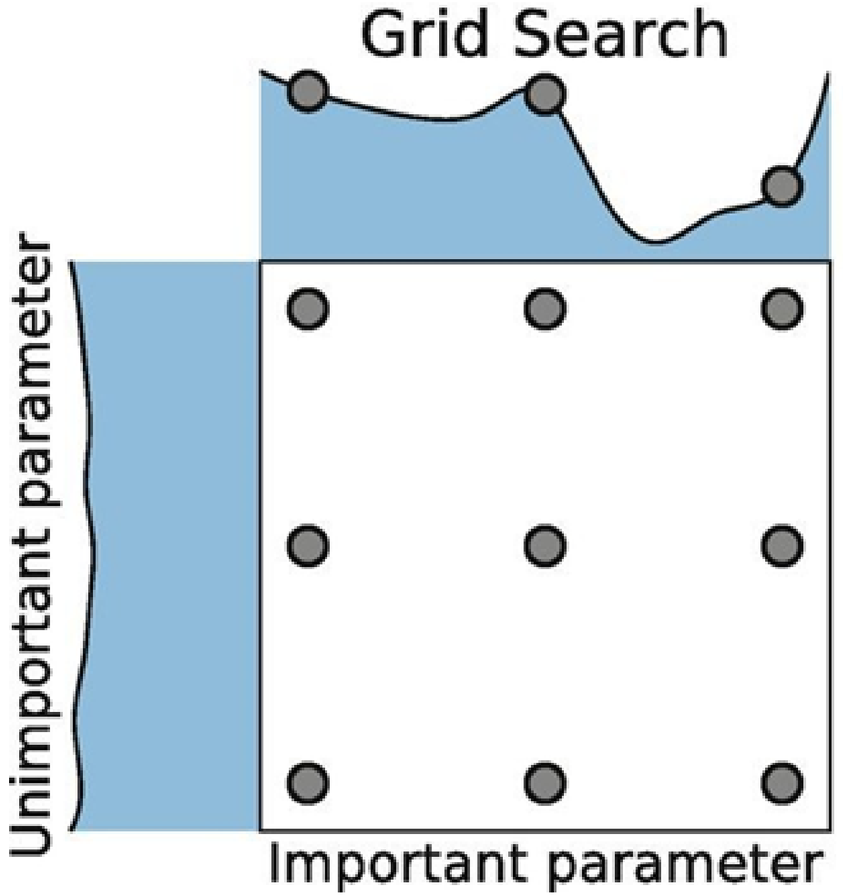
- Parameter model adalah variabel dari model yang dipilih yang dapat diperkirakan dengan menyesuaikan data yang diberikan ke model. Parameter adalah kunci algoritma dari machine learning, bagian dari model yang dipelajari dari data latih historis.

- Hyperparameter model adalah parameter yang nilainya ditetapkan sebelum model memulai pelatihan. Mereka tidak dapat dipelajari dengan menyesuaikan model dengan data..



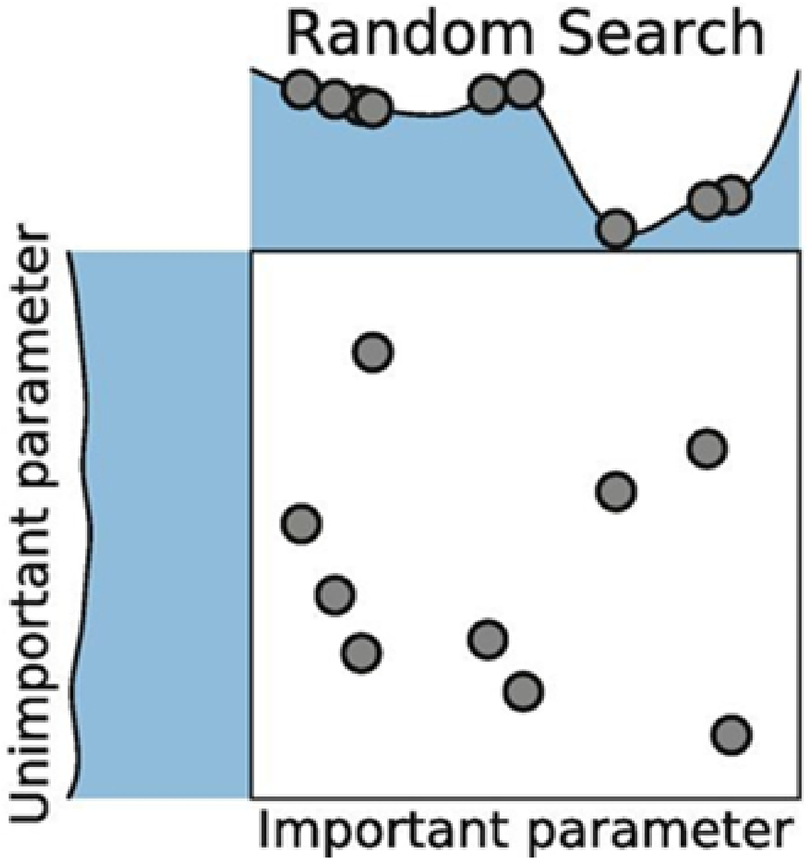
GRID SEARCH

- Bisa diartikan sebuah metode pemilihan kombinasi model dan hyperparameter dengan cara menguji coba satu persatu kombinasi dan melakukan validasi untuk setiap kombinasi.
- Tujuannya adalah menentukan kombinasi yang menghasilkan performa model terbaik yang dapat dipilih untuk dijadikan model untuk prediksi. GridSearch sangatbagus untuk kombinasi pemeriksaan spotyangdiketahuiberkinerja baik secara umum



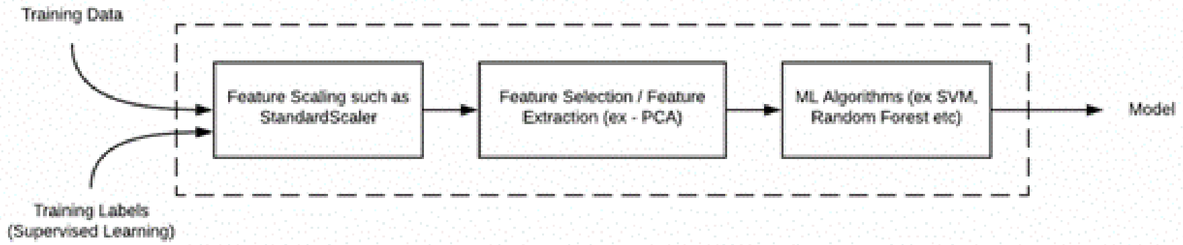
RANDOMIZED SEARCH

- **Random Search** sangat bagus untuk penemuan dan mendapatkan kombinasi hyperparameter yang tidak terduga secaraintuitif, meski butuh waktu lebih banyak. Random search adalah metode pencarian langsung yang tidak memerlukan derivatif untuk mencari domain continue.

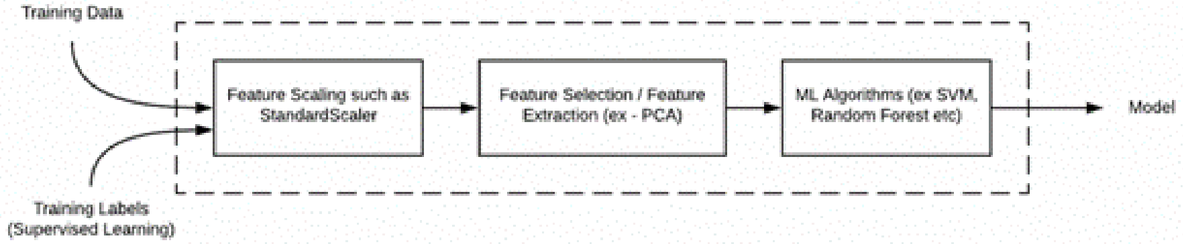


ML PIPELINE

- Machine Learning (ML) pipeline secara teoritis, mewakili langkah-langkah yang berbeda termasuk transformasi data dan prediksi yang dilalui data. Hasil dari pipeline adalah model terlatih yang dapat digunakan untuk membuat prediksi. Sklearn.pipeline adalah implementasi Python dari ML Pipeline . Daripada melalui langkah-langkah pemasangan model dan transformasi data untuk set data pelatihan dan pengujian secara terpisah, kita dapat menggunakan Sklearn.pipeline untuk mengotomatiskan langkah-langkah ini. Berikut adalah diagram yang mewakili alur untuk melatih machine learning model based on supervised learning :



Machine Learning Pipeline



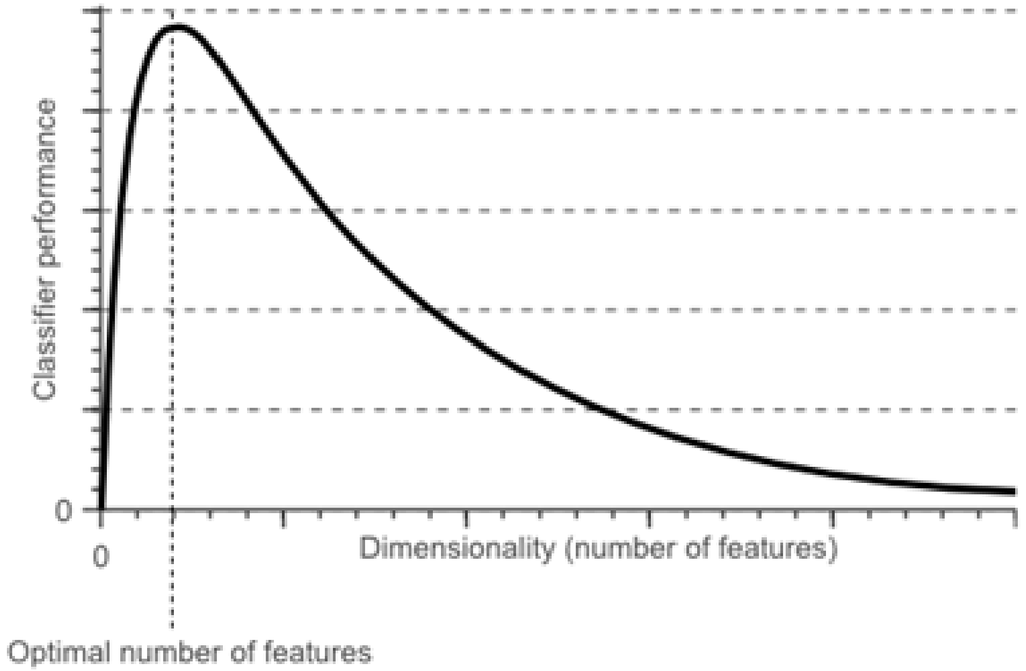
Machine Learning Pipeline

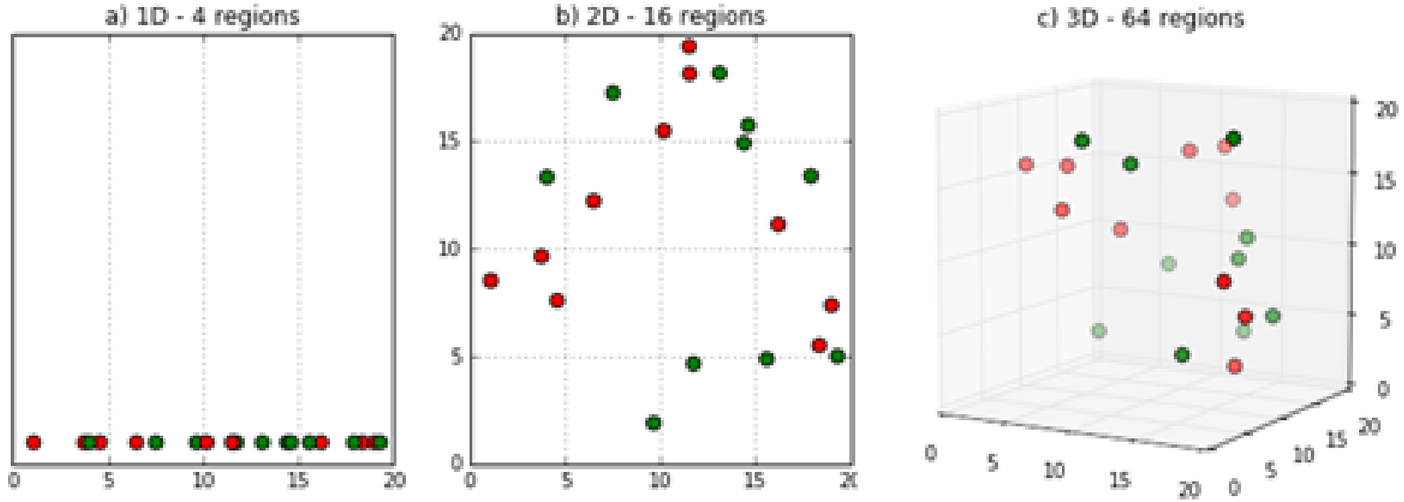
DigitalSkola

SESSION 38

Advanced ML Topics

Dimensionality Curse





Dalam pembuatan model machine learning terdapat dimensionality curse, dimana dimensionality curse adalah proses untuk mengurangi dimensi pada feature set atau dataset. Terdapat dua metode pada dimensionality reduction yaitu, feature selection dan feature extration.

Manfaat Dimensionality Reduction

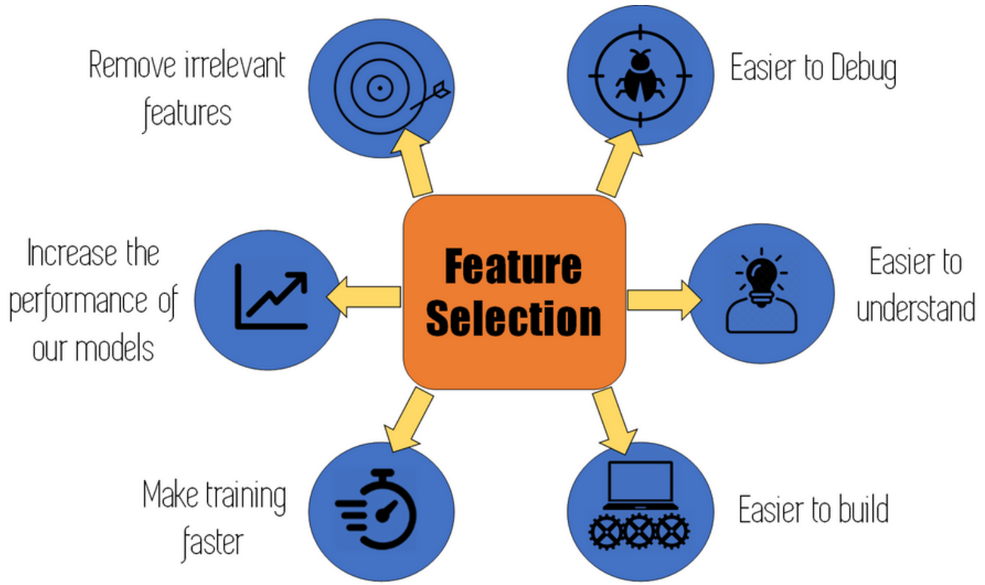
1. Mengurangi data yang misleading terhadap model.
2. Semakin sedikit kolom yang kita gunakan, maka semakin ringan beban untuk menghitung.
3. Mengurangi penyimpanan data.
4. Mengurangi data yang tidak penting.

Feature Selection

Feature selection adalah salah satu konsep pada machine learning yang mana konsepnya sangat berpengaruh besar terhadap performa model.

Feature Selection Methods

1. Univariate Selection
2. Correlation Matrix with
   - Pearson Correlation
   - Spearman Correlation
   - Chi-square
3. Feature Importance



Feature Importance

Setelah dilakukan training selanjutnya ialah tahap feature importance. Feature importance dapat dikatakan sebagai tolak ukur besaran kontribusi berbagai data feature yang dilatih kepada performa model prediksi. Manfaat dari feature importance

1.Dapat membantu memilih variabel yang penting.

2.Model mudah diterjemahkan dan dipahami oleh orang lain.

3.Dapat mengetahui bias pada model.

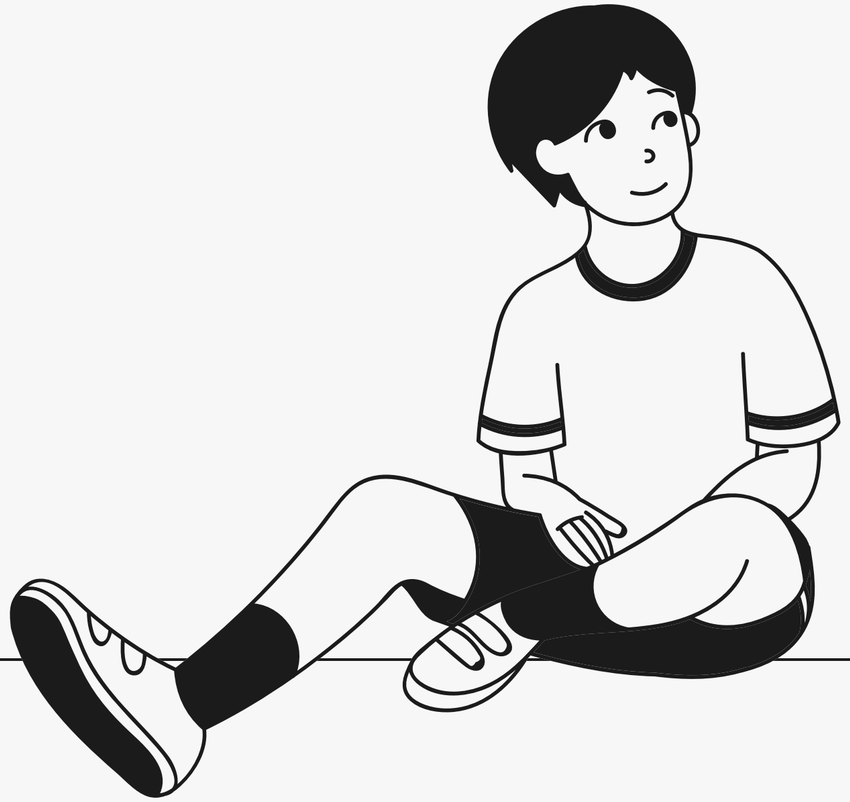
Principal Component Analysis

Principal Component Analysis atau PCA merupakan teknik yang digunakan untuk mengesktrak inti dari dataset. Kapan PCA digunakan, ketika data yang memiliki banyak kolom, visualisasi untuk multivariavble dataset. PCA tidak digunakan ketika ingin mengetahui feature importance, data dengan korelasi yang kecil, dan classification project.

SESSION 39
BUSINESS INTELLIGENCE

- Business Intelligence adalah sekumpulan teknik dan alat untuk mentransformasi dari data mentah menjadi informasi yang berguna dan bermakna untuk tujuan analisis bisnis.
- Business Intelligence adalah salah satu aspek yang akan membantu perusahaan dalam menentukan strategi pemasaran berdasarkan data pasar.
- Kumpulan data tersebut kemudian akan diolah oleh seorang BI (Business Intelligent) menggunakan metode, *tool* dan *software* yang sesuai.
- Seorang inteligen bisnis akan bertugas merencanakan, mengelola data dan memberikan hasil akhir berupa informasi yang mudah dipahami untuk seluruh stakeholder pada bisnis.

MENGAPA BI PENTING?



1. Membaca & menafsirkan data untuk membantu menentukan keputusan perusahaan.
2. Mendapatkan proyeksi yang lebih terencana untuk jangka panjang.
3. Pemilihan metode pemasaran yang sesuai dengan bisnis perusahaan.
4. Menganalisa halangan, keuntungan dan solusi yang akan terjadi.
5. Membantu perusahaan dalam menentukan perencanaan biaya.
6. Memudahkan pihak manajemen untuk mengerti kebutuhan pasar dan memperhitungkan dari segi bisnis.
7. Mengevaluasi data tren pasar secara real time.
8. Penentuan target atau KPI (Indikator Kinerja Utama) sesuai dengan data.

KEY PLAYERS IN BI

PROFESSIONAL DATA ANALYST
Bertugas untuk menganalisis data dan mendapatkan insights.

BUSINESS USERS
Bertugas untuk mengevaluasi atau monitoring.

IT TEAM
Bertugas untuk menjaga dan monitoring infrastruktur data/sistem.

HEAD OF THE COMPANY
Bertugas untuk mengambil keputusan dengan pandangan menyeluruh.

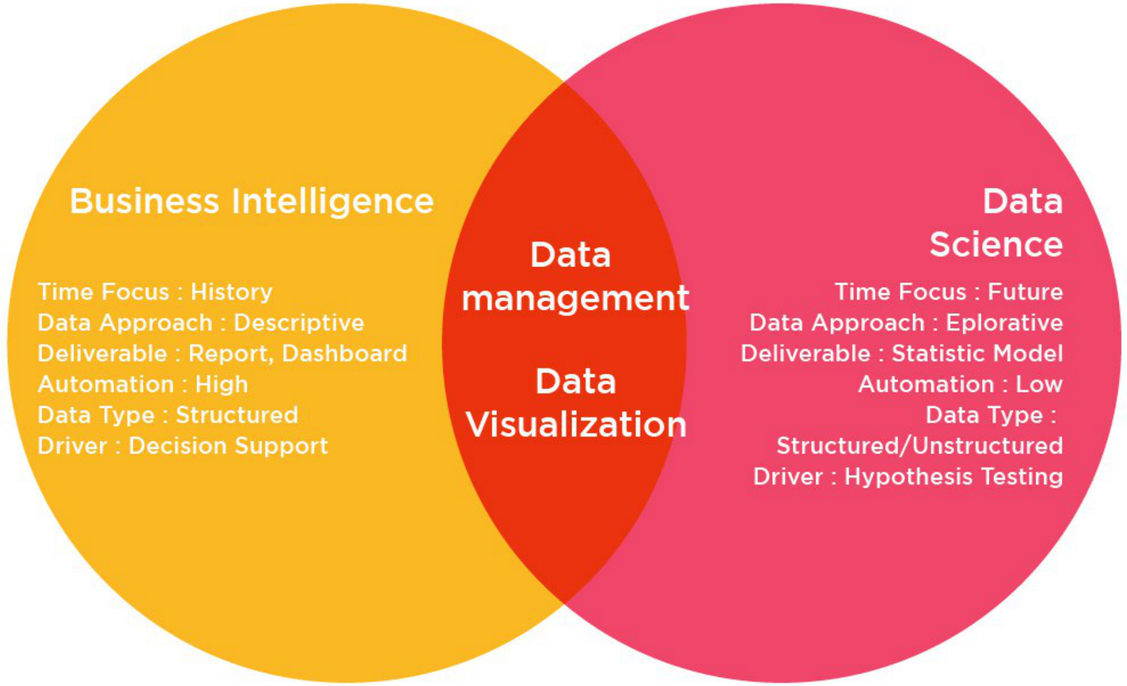
BUSINESS INTELLIGENCE VS DATA SCIENCE

PERSAMAAN :

- Menganalisis data
- Memvisualisasikan data
- Perhitungan statistik

PERBEDAAN :

- BI menganalisis dengan retrospeksi dan menampilkan kondisi masa kini atau masa lalu, sedangkan DS memprediksi suatu kondisi kedepannya.



DASHBOARD CONSTRUCTIONS

Sales Date
10/12/2021 - 11/11/2021

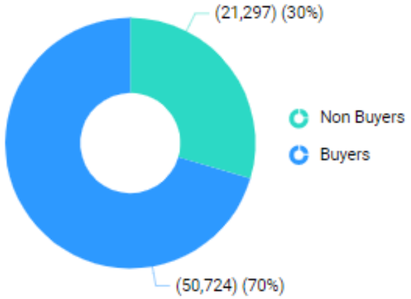
$56
Sales Per Sq Ft

$43
Average Unit In Retail (AUR)

6 PCS
Average Basket Size (ABS)

$271,196
Net Sales

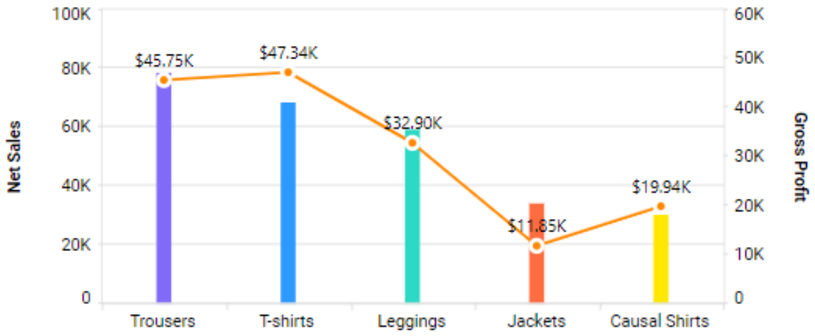
Total Visitor's Count



Average Basket Value
Last 7 days

There is no data to be displayed.

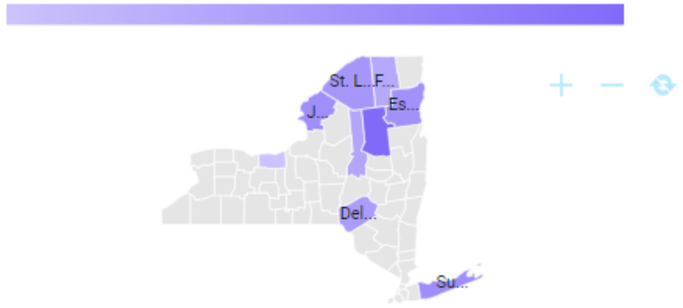
Sales and Gross profit by Category



Revenue table

Name	Total Sales	Discount	Net Sales	COGS	Gross Profit	Gross Margin
Store 1	$32,024	$1,316	$30,740	$9,412	$21,327	66.86%
Store 2	$36,323	$3,438	$32,927	$16,320	$16,612	66.64%
Store 3	$37,850	$3,659	$34,203	$17,911	$16,291	52.15%

Sales volume by Stores
Counties of New York state in USA



DigitalSkola

DASHBOARD CONSTRUCTIONS

- Isi dashboard harus memperjelasan tujuan apa yang akan dicapai.
- Dashboard hanya berisikan sesuatu yang penting .
- Menyesuaikan ukuran dan pisisi untuk menunjukkan hirarki dimana berguna untuk menunjukkan kepada audiens informasi mana yang paling penting.
- Menggunakan konteks angka sehingga audiens mampu memahami baik atau buruknya informasi yang disampaikan.
- mengelompokkan informasi yang memiliki metrik berhubungan agar lebih mudah ditemukan.
- Konsisten untuk menggunakan visualisasi dan tata letak yang sama dapat membuat perbandingan dilakukan dengan mudah.
- Menggunakan label yang jelas untuk membantu audiens memahami informasi dengan mudah.
- Membulatkan angka agar tidak menimbulkan kesan perbedaan yang besar.
- Selalu memperbaiki dan mengembangkan dashboard untuk mendapatkan hasil yang terbaik.

BI TOOLS

Database

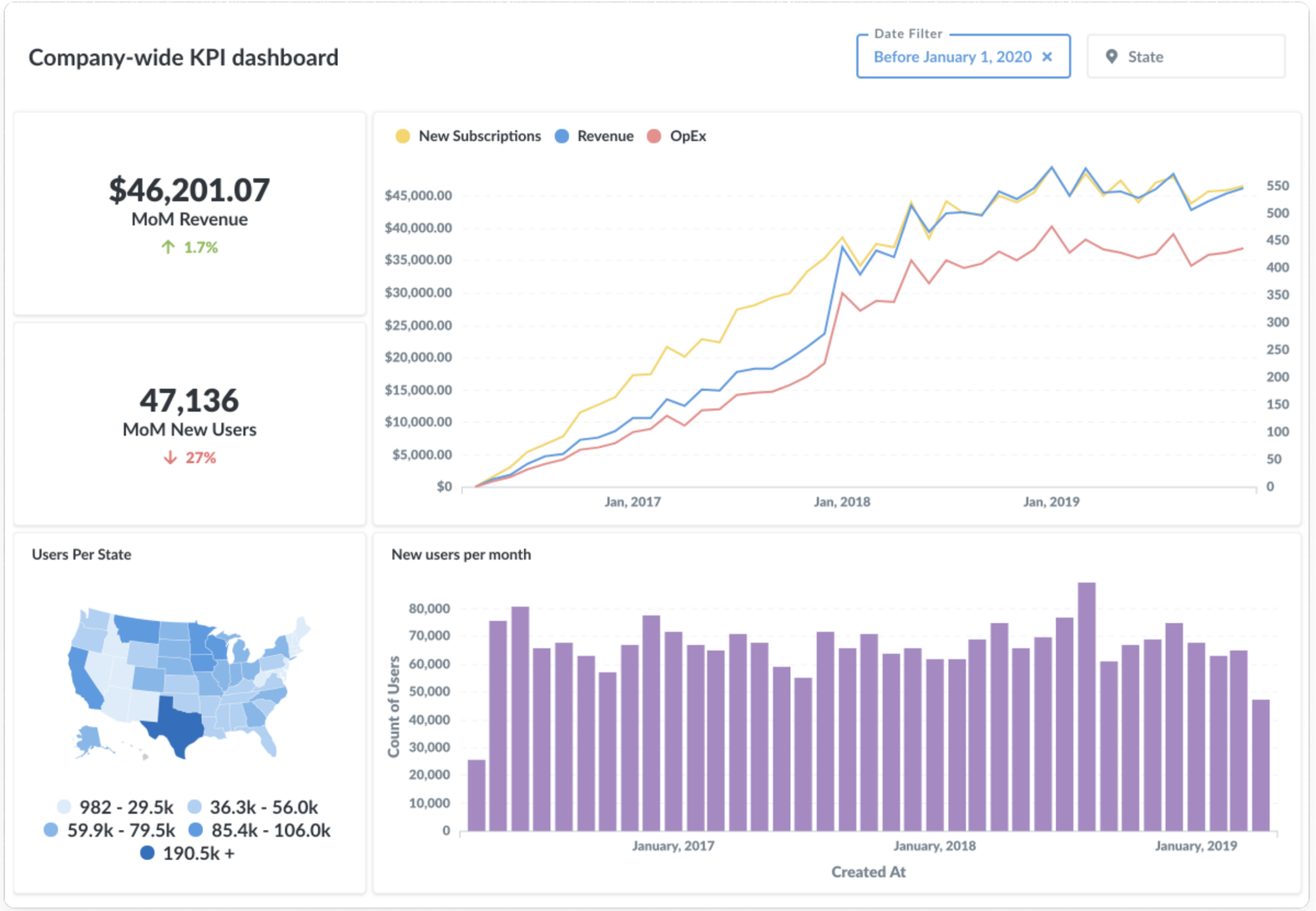
Power BI

Qlik

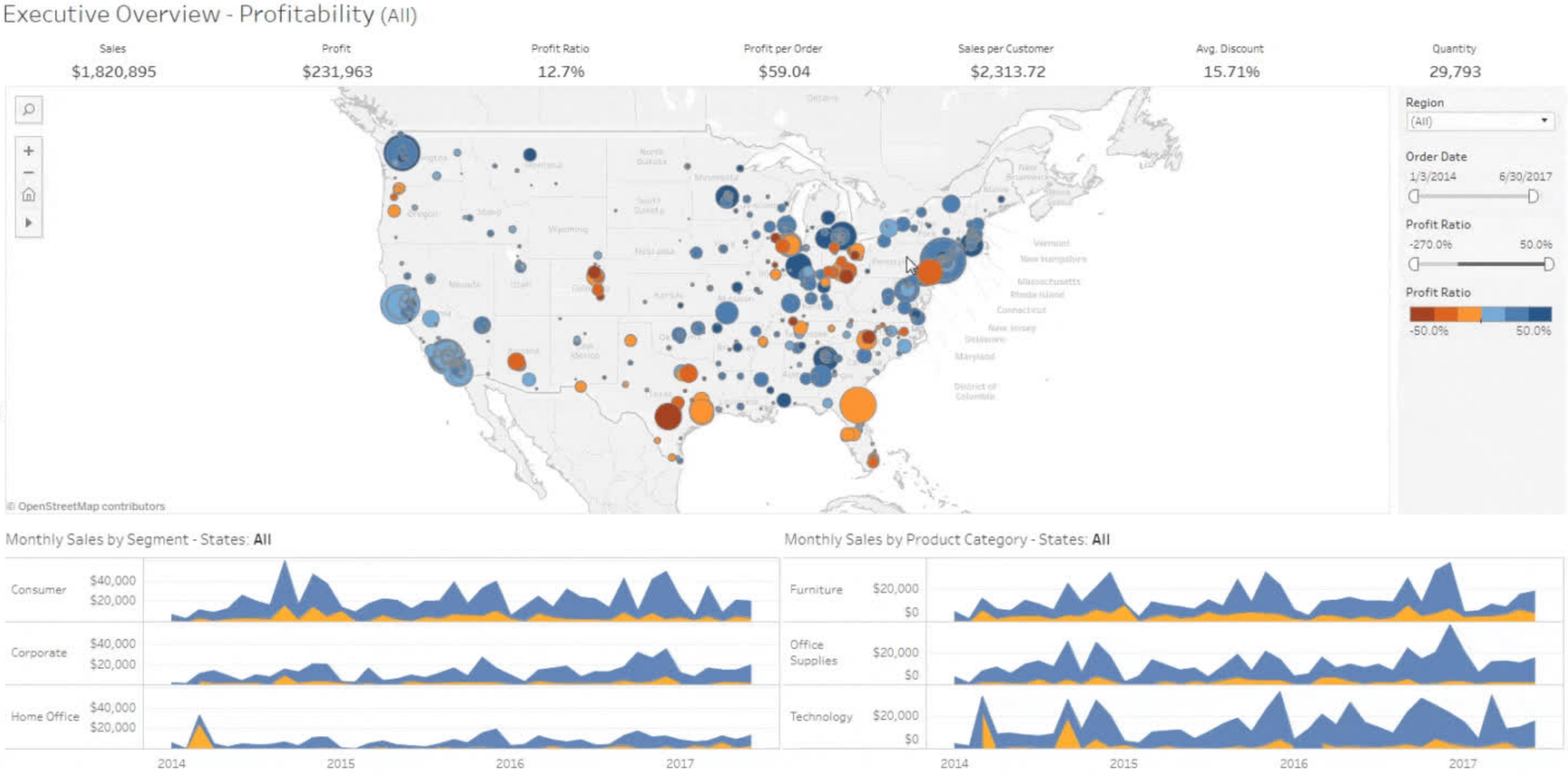
+ableau

Business Intelligence tools merupakan aplikasi untuk mengoleksi dan memproses data besar yang tidak terstruktur dari sistem internal dan eksternal serta dapat membantu menyiapkan data untuk analisis sehingga pengguna dapat membuat report, dashboard, dan visualisasi.

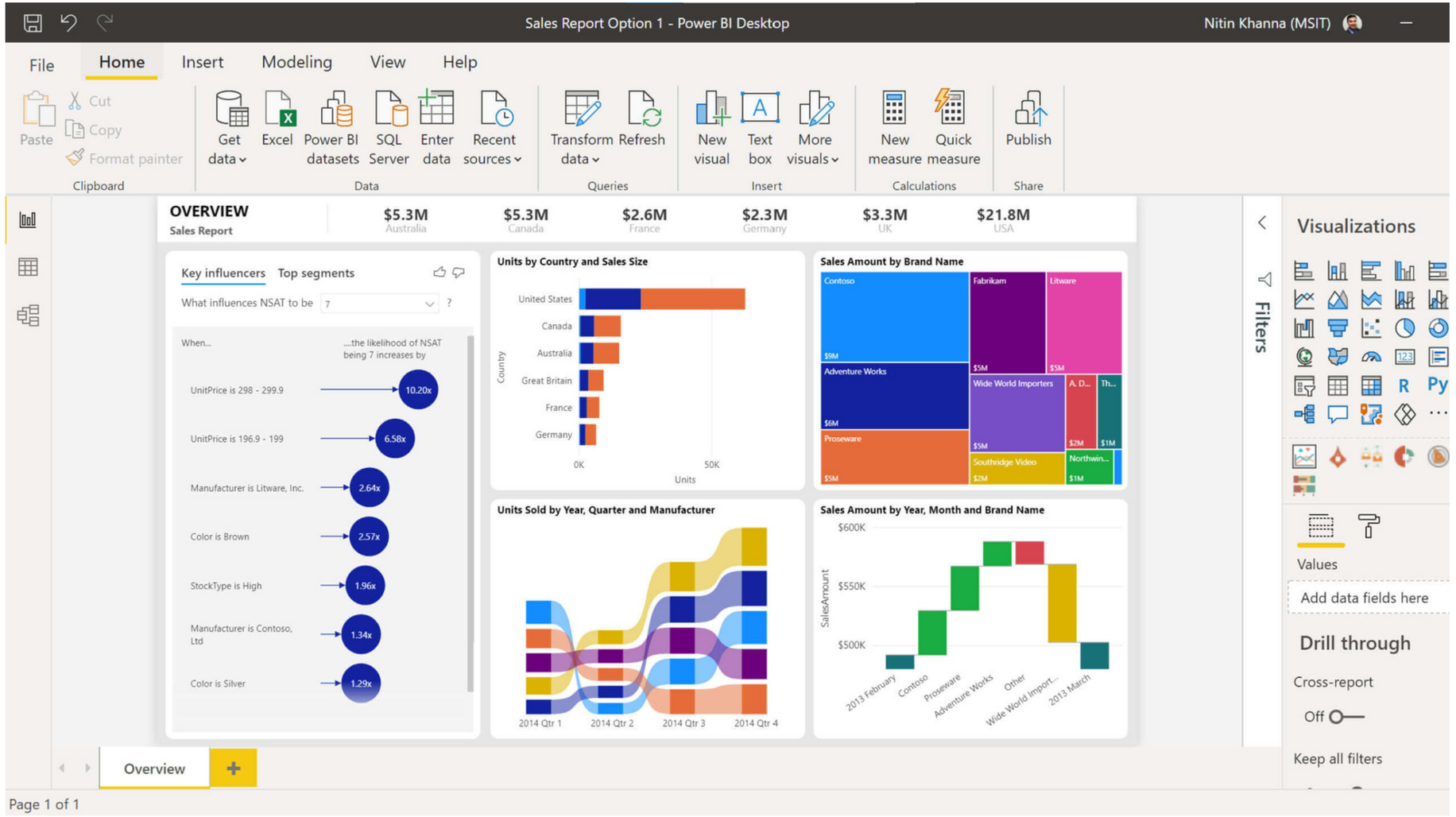
Metabase



DigitalSkola

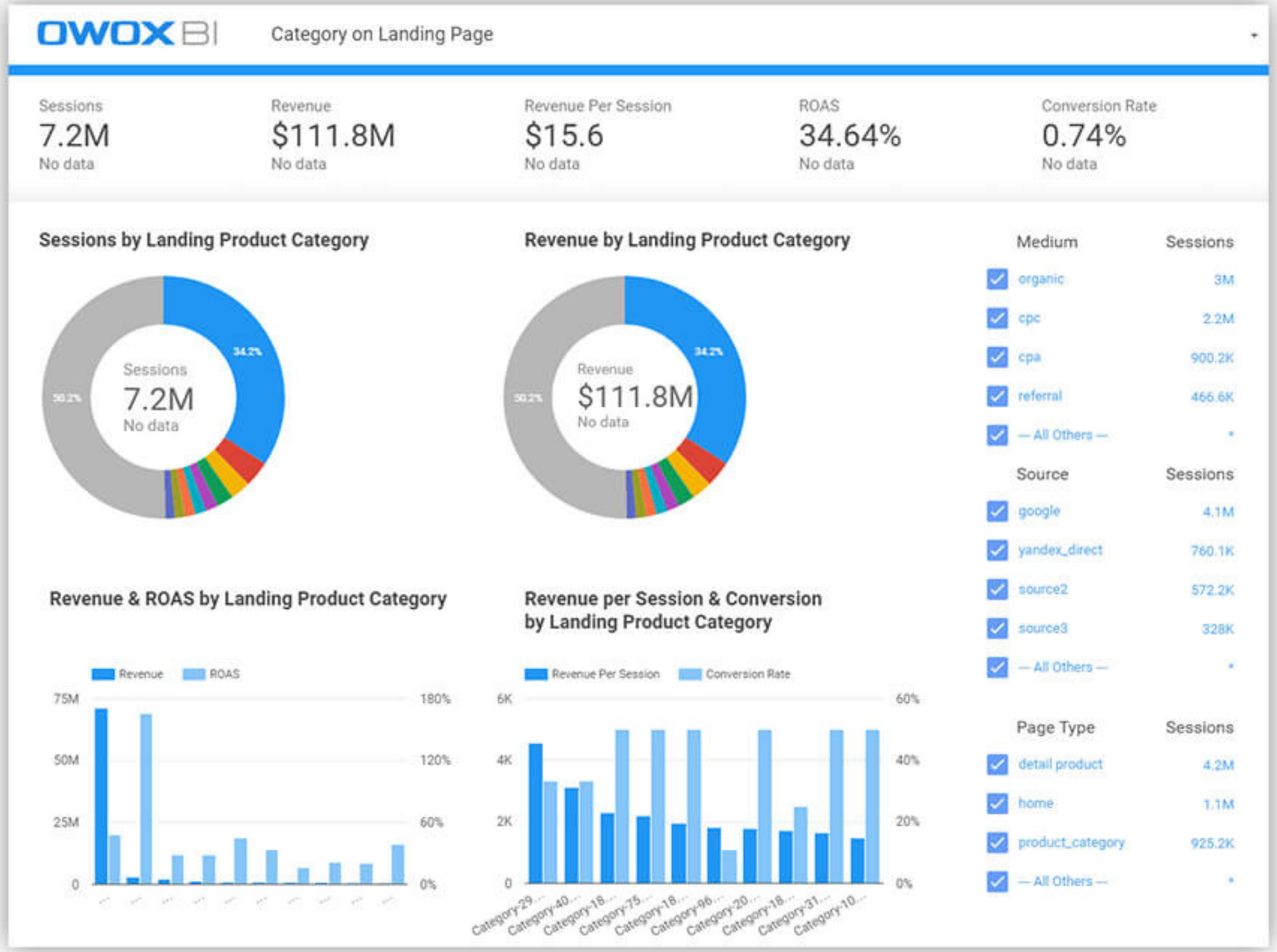

Executive Overview - Profitability (All)
Sales
$1,820,895
Profit
$231,963
Profit Ratio
12.7%
Profit per Order
$59.04
Sales per Customer
$2,313.72
Avg. Discount
15.71%
Quantity
29,793
Region
(All)
Order Date
1/3/2014
6/30/2017
Profit Ratio
-270.0%
50.0%
Profit Ratio
-50.0%
50.0%
© OpenStreetMap contributors
Monthly Sales by Segment - States: All
Consumer
Corporate
Home Office
$40,000
$20,000
Monthly Sales by Product Category - States: All
Furniture
Office Supplies
Technology
$20,000
$0
2014
2015
2016
2017

Power BI

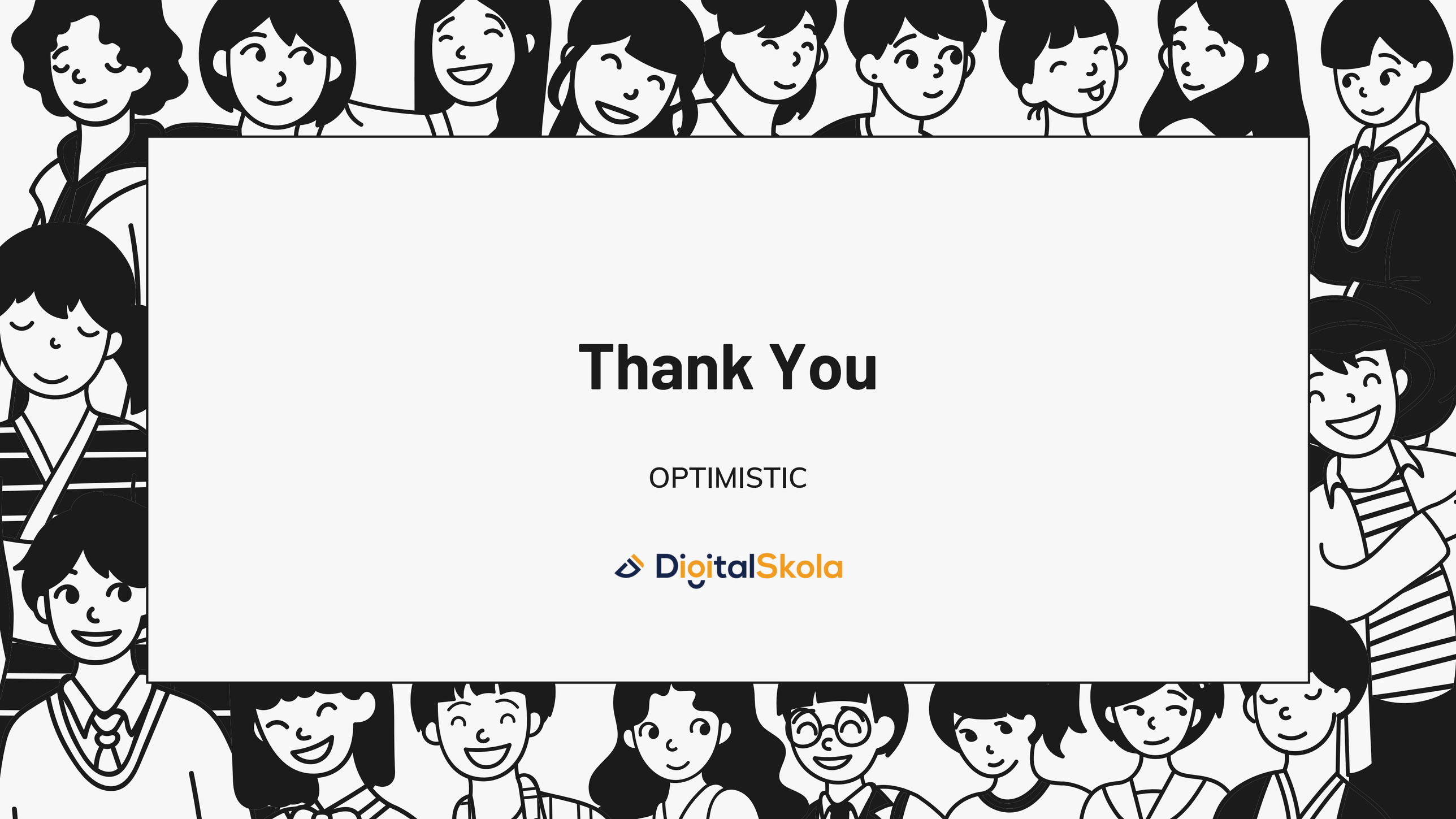


DigitalSkola

Google Data Studio



DigitalSkola



Thank You

OPTIMISTIC

DigitalSkola