Final Project Presentation

Nomor Kelompok: 3
Nama Mentor: Aditya Bariq
Nama:

- Muhammad Rizqiansyah
- Nadia Rizky Hairunnisa

Machine Learning Class

Program Studi Independen Bersertifikat
Zenius Bersama Kampus Merdeka

Latar Belakang

Latar Belakang Project

Sumber Data: Walmart Dataset
https://www.kaggle.com/datasets/yasserh/walmart-dataset

Problem: **Regression**

Tujuan:

- Memprediksi penjualan mingguan di Walmart

Eksplorasi Data dan Visualisasi

Business Understanding

Walmart merupakan salah satu perusahaan *retail* multinasional terbesar di dunia. Walmart memiliki banyak pesaing yang bergerak di bidang *retail* sehingga diperlukan keputusan yang strategis agar bisa mempertahankan posisinya.

Business Understanding

Resource/Dataset:

- Gabungan data dari 45 toko termasuk informasi toko dan penjualan mingguan.
- Data disediakan setiap minggu
- Terdapat 4 minggu liburan (Natal, Thanksgiving, Super bowl, Hari Buruh)

Business Understanding

Business Objectives:

- Apakah terdapat insights pada data? Sehingga kita bisa...

Increase Profit

Reduce Cost

Reduce Future Loss

Side Questions: Bagaimana faktor waktu dan perekonomian negara bisa mempengaruhi penjualan mingguan?

Data Cleansing

Dimensi data: **6435 baris dan 8 kolom**

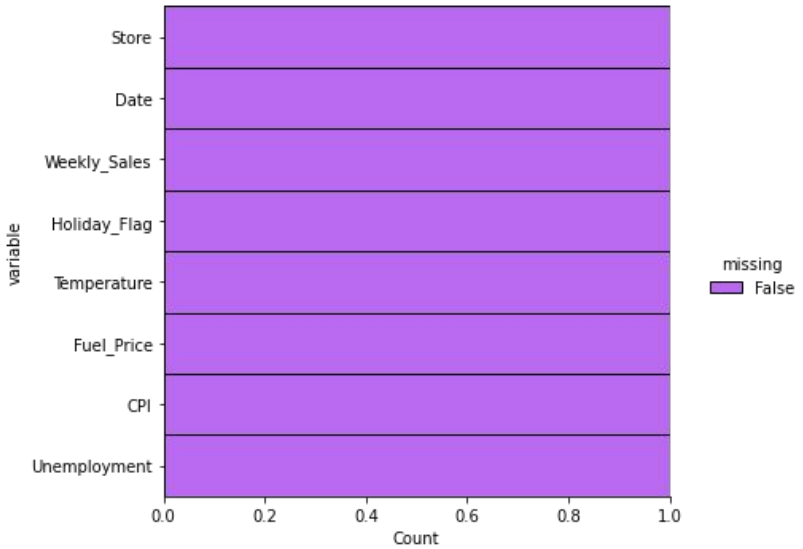
Kolom target: **Weekly_Sales**

Missing values: **0**

Duplicated values: **0**

Jumlah toko: **45**

Hari unik: **Jum'at***



*Pencatatan data dilakukan setiap hari Jum'at

Data Cleansing

Info Hari Libur Besar dari tahun 2010-2012:

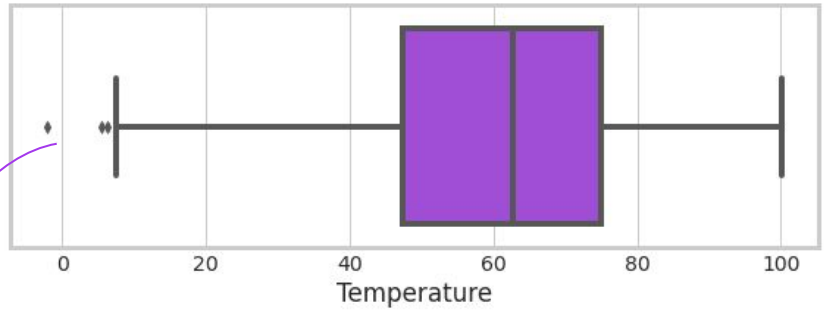
- Super Bowl: Seluruhnya terdapat dalam dataset
- Labor Day: Seluruhnya terdapat dalam dataset
- Thanksgiving: **Tidak ada data tahun 2012*** — Harusnya ada di bulan November 2012
- Christmas: **Tidak ada data tahun 2012*** — Harusnya ada di bulan Desember 2012

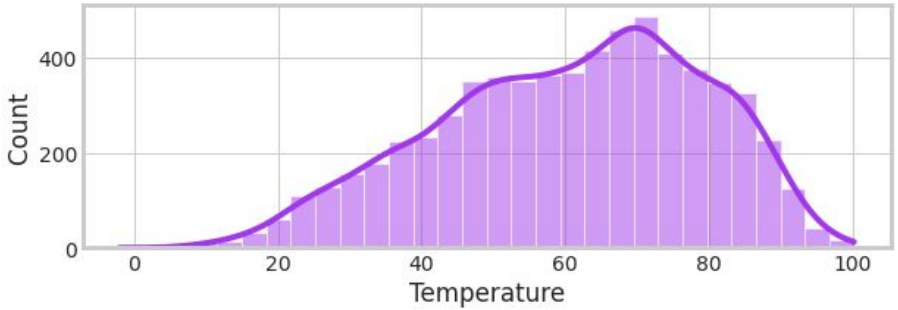
Range data: **5 Februari 2010 - 6 Oktober 2012**

*akibat dari range data yang kurang lengkap. Data yang tidak ada berada di luar range data

Data Cleansing

Outliers: Kolom Temperature = 3 data





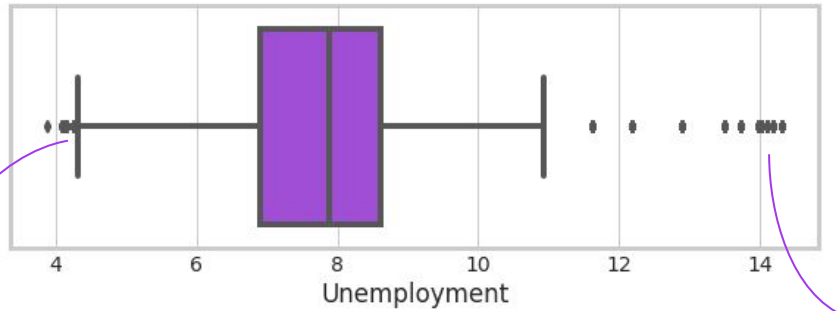
Wajar. Pencilan ada di suhu terendah dan berada di Bulan Januari dan Februari.

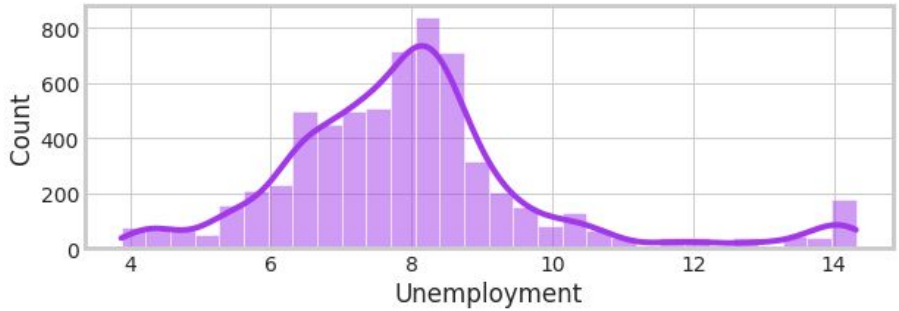
Asumsi: Data temperatur pada hari itu diambil ketika musim dingin

Data Pendukung: Pada tahun 2011, musim dingin dimulai dari awal Desember 2010 dan berakhir di akhir Februari 2011

Data Cleansing

Outliers: Kolom Unemployment = 481 data



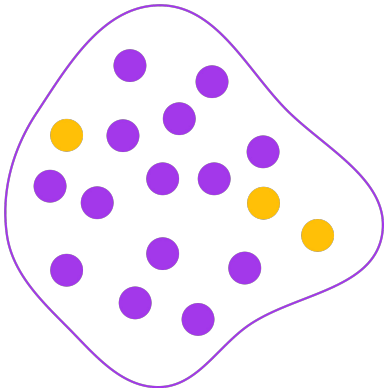


Unemployment rendah terdapat pada tahun 2012. Sementara itu, Unemployment tinggi terdapat pada semua tahun. **Perlu dilakukan handling outlier lebih jauh**

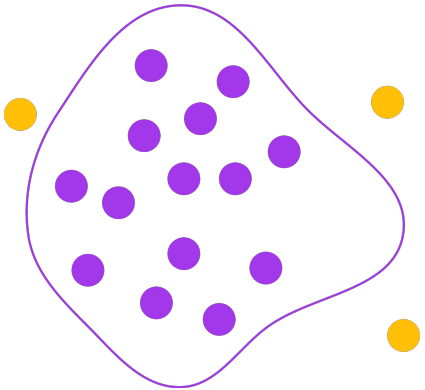
Handling Outliers

3 Skenario handling outliers

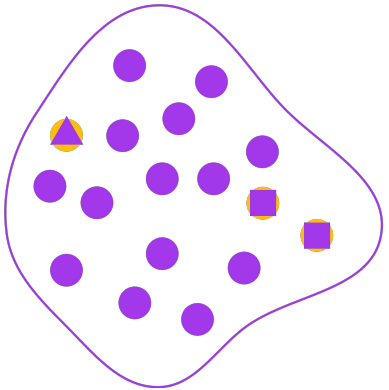
Outlier dibiarkan



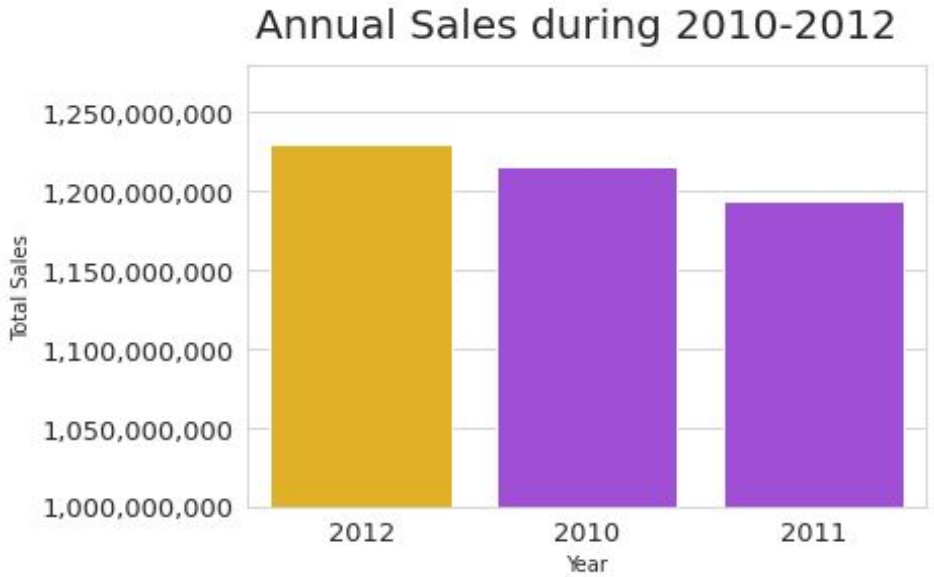
Outlier dihapus



Transformasi outlier



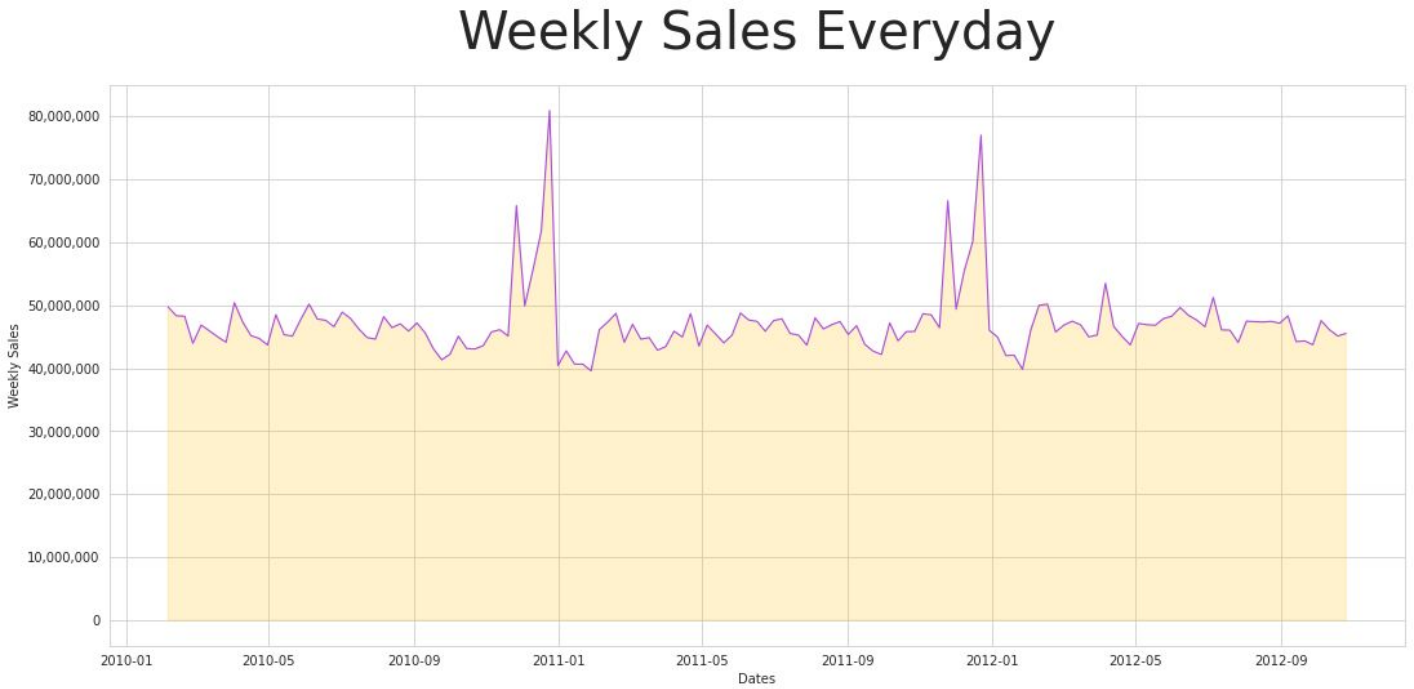
Exploratory Data Analysis



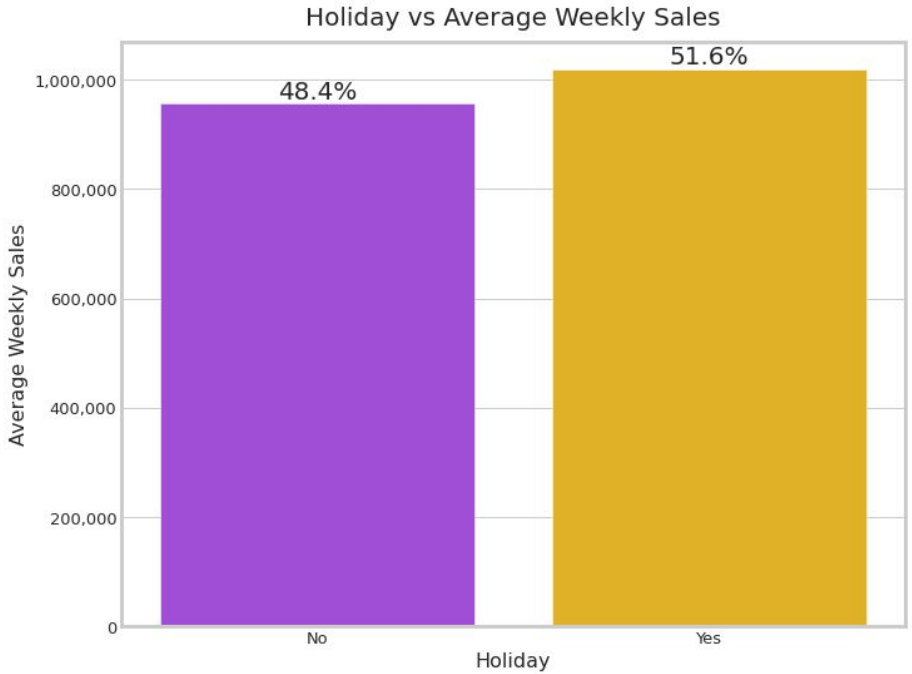
Karena range data setiap tahun tidak seimbang, Total Weekly Sales dihitung dari bulan Februari hingga Juli saja untuk melihat perbandingannya.

Exploratory Data Analysis

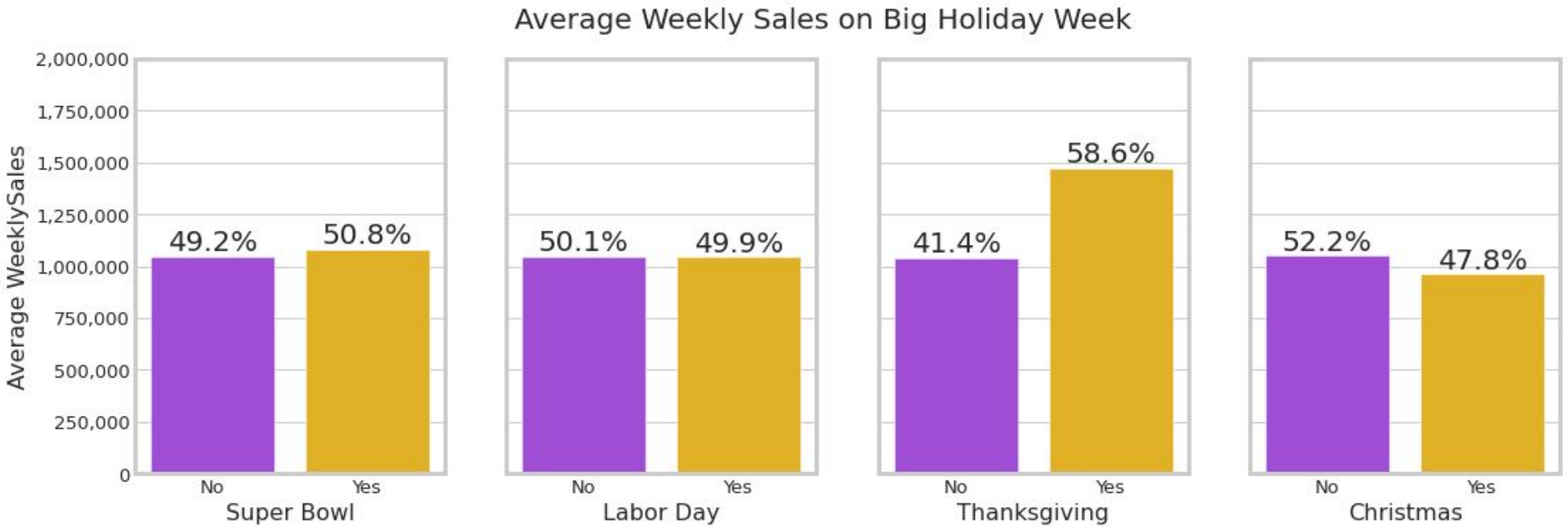
Exploratory Data Analysis



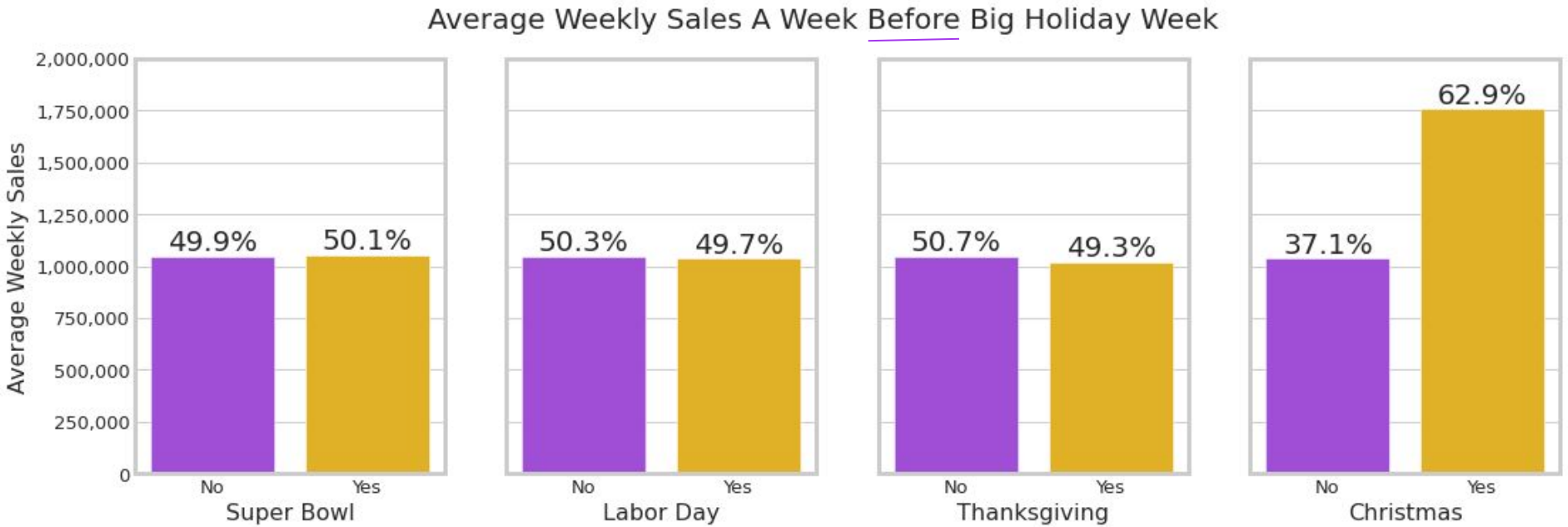
Exploratory Data Analysis



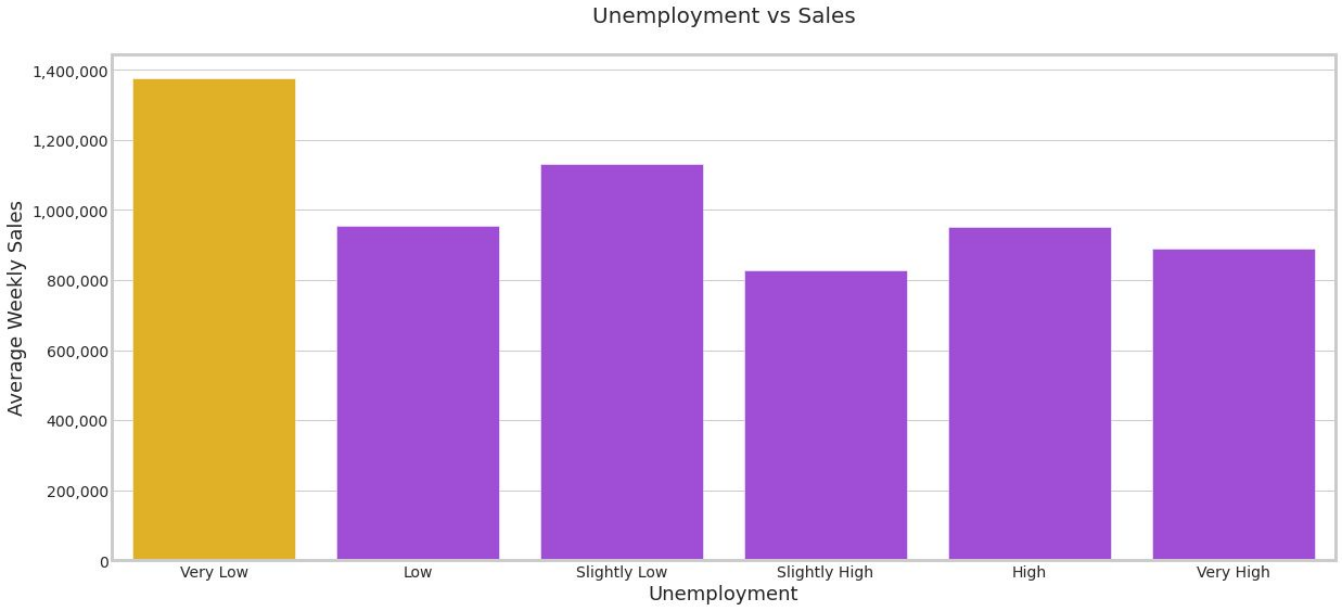
Exploratory Data Analysis



Exploratory Data Analysis



Exploratory Data Analysis

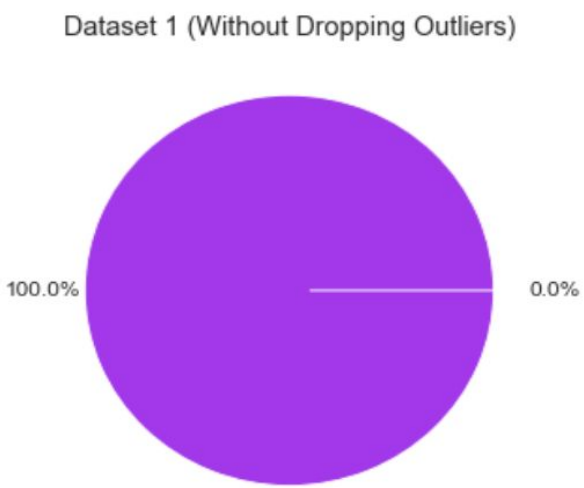


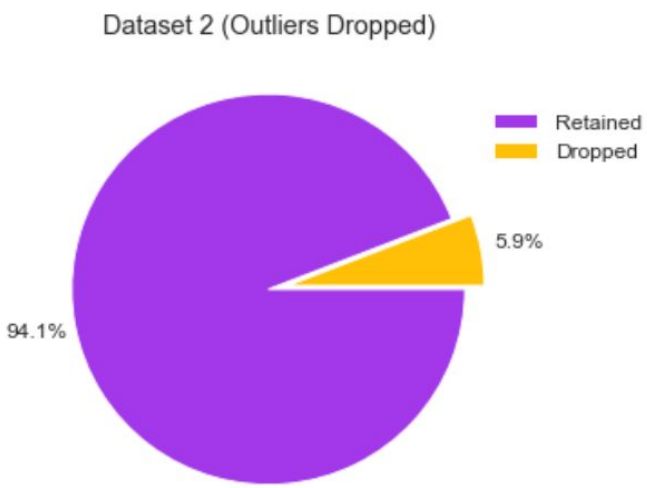
Kategori Unemployment Rate:

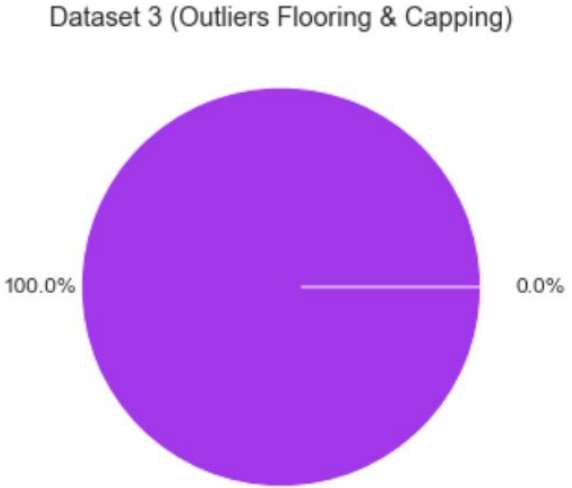
8.0 - 9.9	: 1 (Very Low)
10.0 - 10.9	: 2 (Low)
11.0 - 11.9	: 3 (Slightly Low)
12.0 - 12.9	: 4 (Slightly High)
13.0 - 13.9	: 5 (High)
14.0 - 14.9	: 6 (Very High)

Modelling

Final Dataset





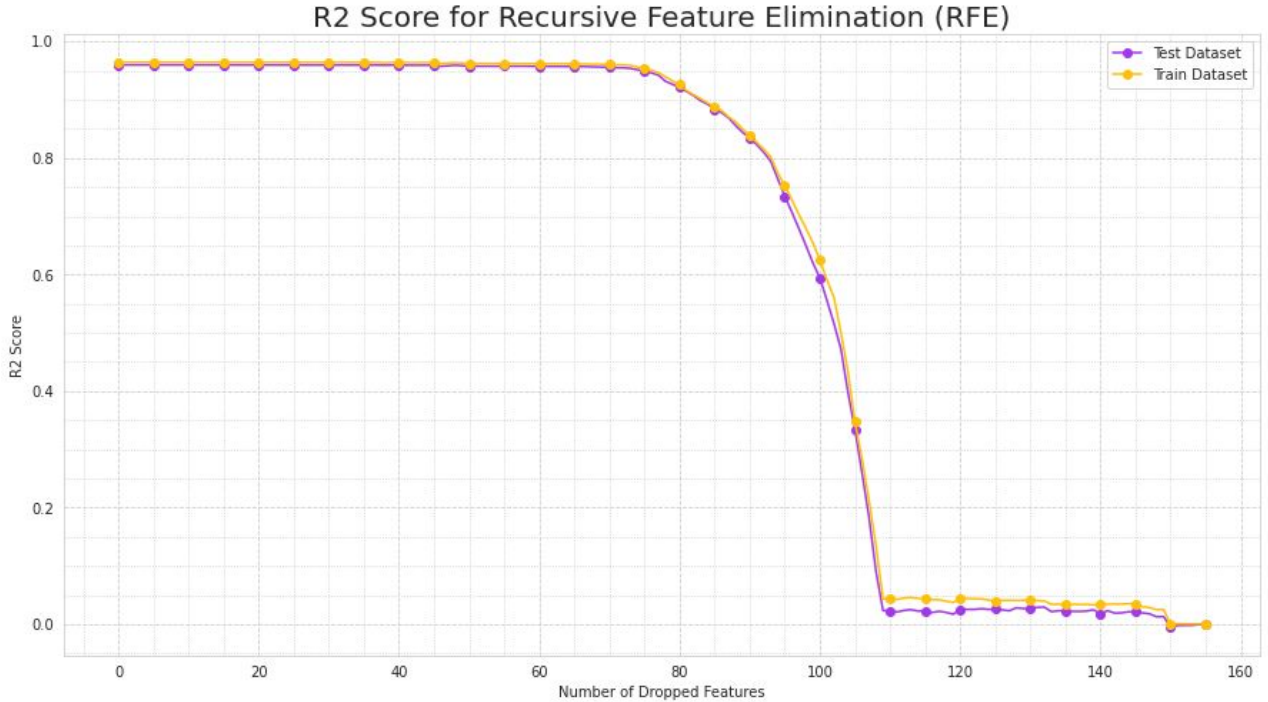


Feature Selection & Decomposition

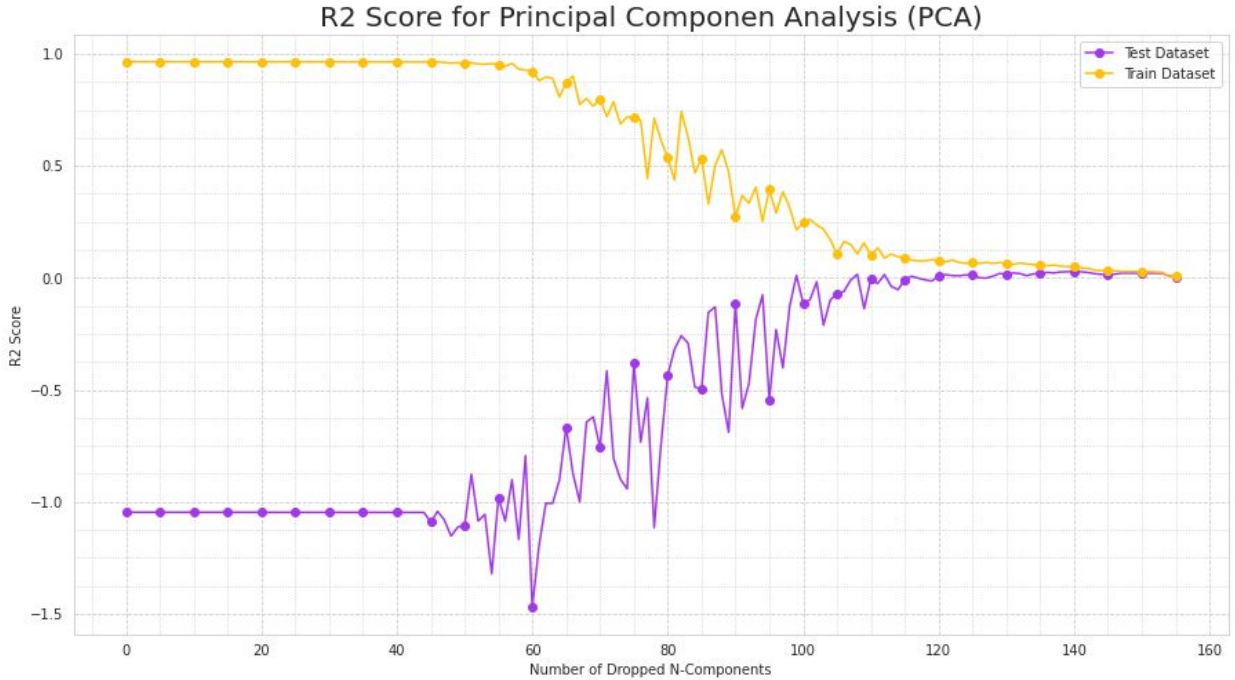
- **Recursive Feature Elimination (RFE)**
  menyeleksi fitur berdasarkan estimator secara rekursif

- **Principal Component Analysis (PCA)**
  mereduksi fitur dengan melakukan kombinasi berdasarkan nilai varians

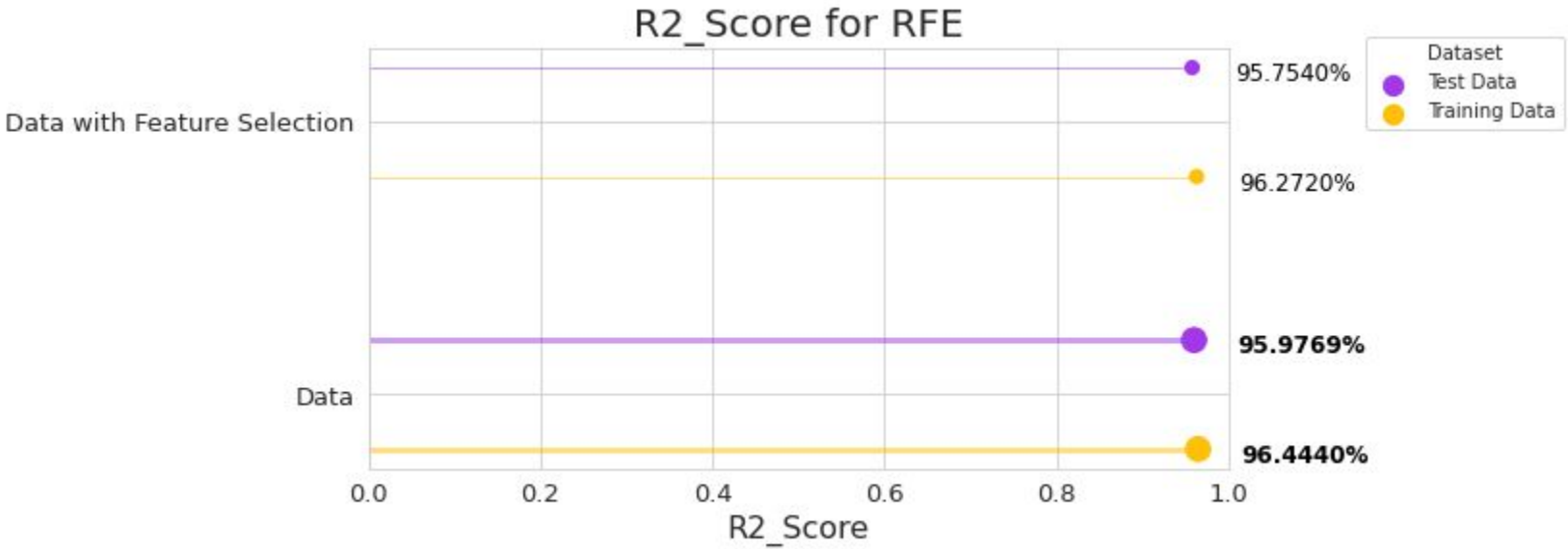
Feature Selection & Decomposition



Feature Selection & Decomposition



Feature Selection & Decomposition



Model Selection

Dataset Splitting = Train (60%), Test (40%)

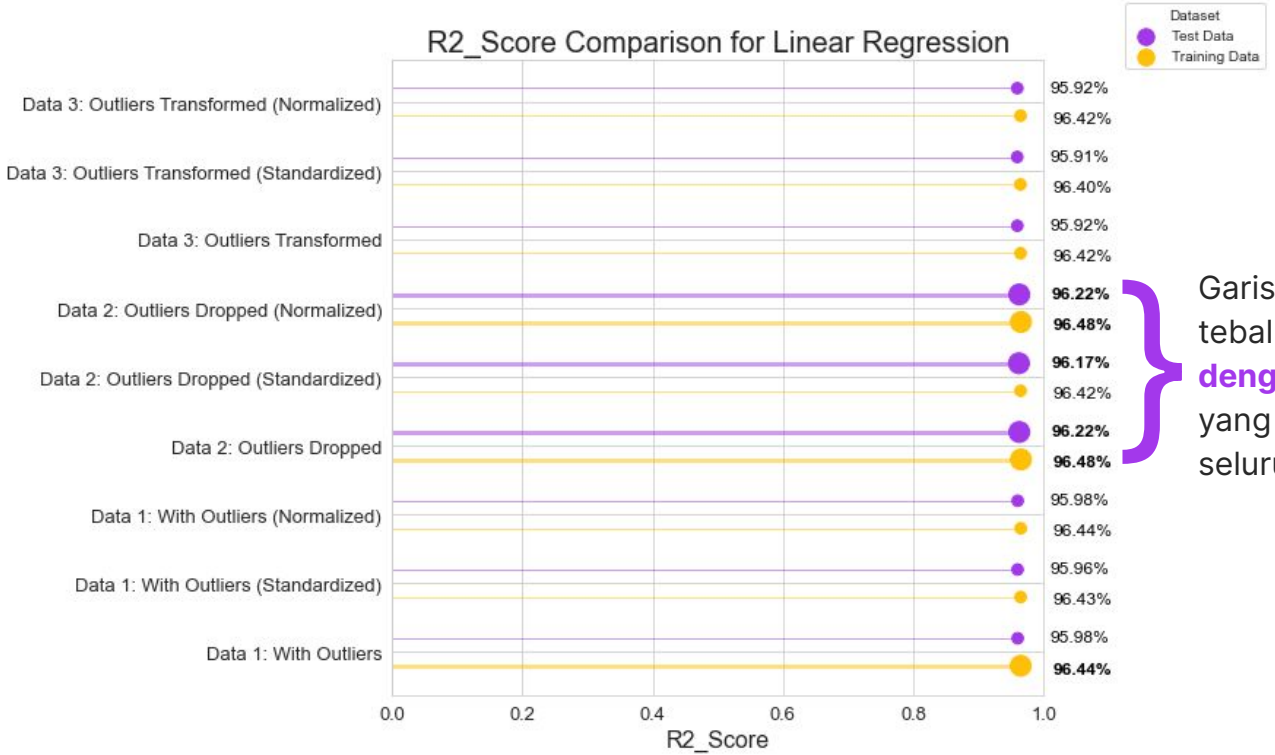
Model:

- Linear Regression
- Ridge Regression
- Lasso Regression

Evaluation Metric

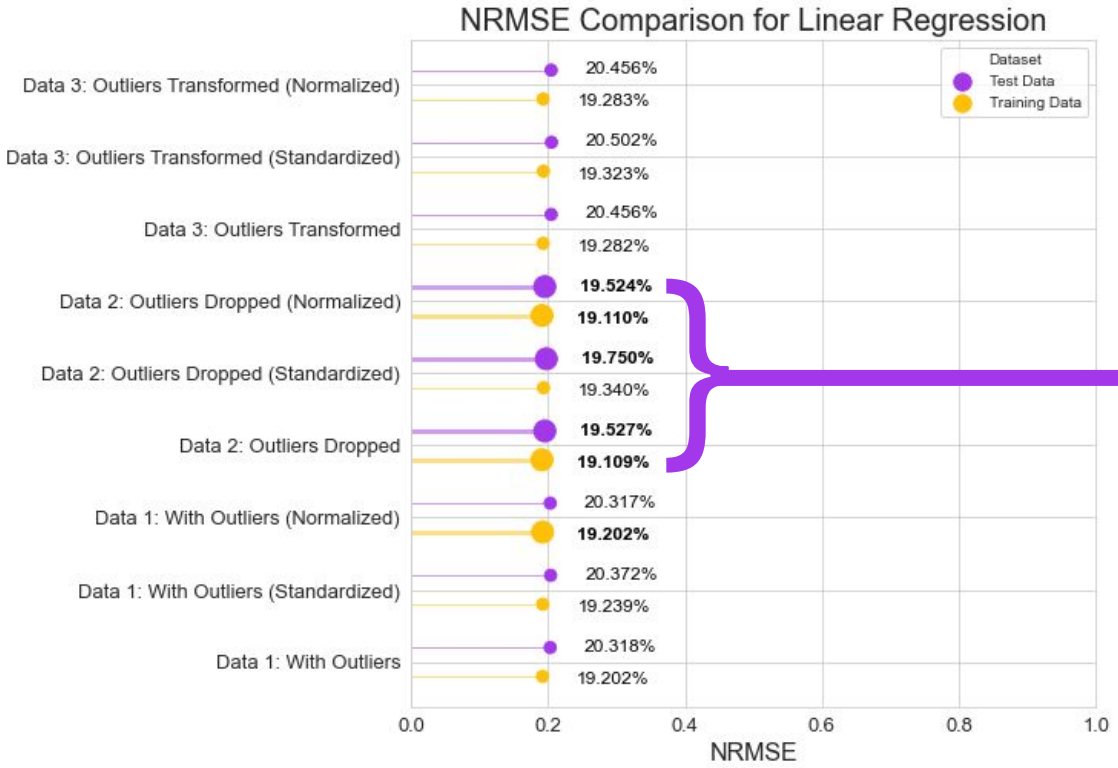
- MAE (Mean Absolute Error)
- MSE (Mean Squared Error)
- RMSE (Root Mean Squared Error)
- MAPE (Mean Absolute Percentage Error)
- NRMSE (Normalized Root Mean Squared Error)
- R2-Score

Model: Linear Regression



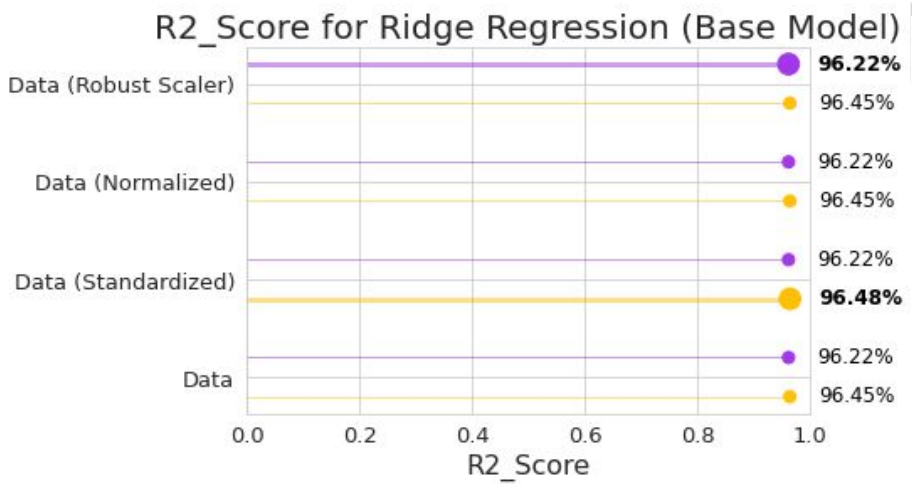
Garis dan ukuran lingkaran yang tebal merepresentasikan **3 data dengan R2 Score terbesar**, yang mana hampir dimiliki seluruh **dataset kedua**

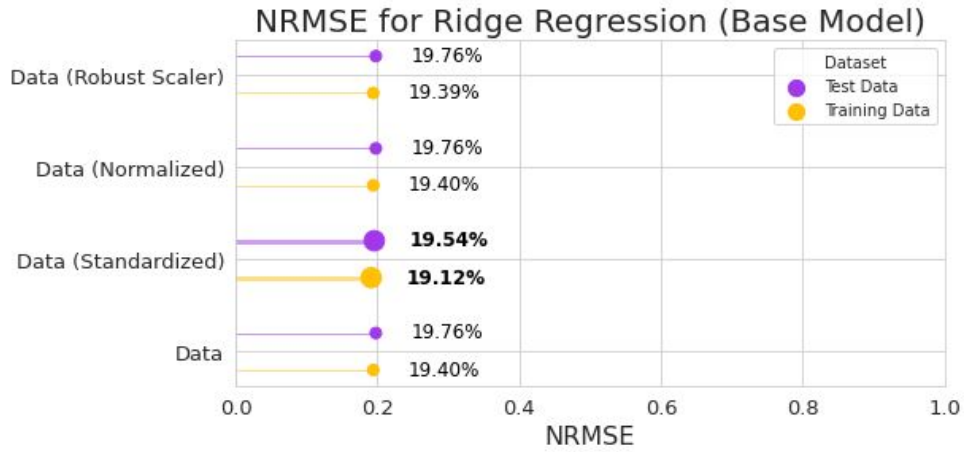
Model: Linear Regression



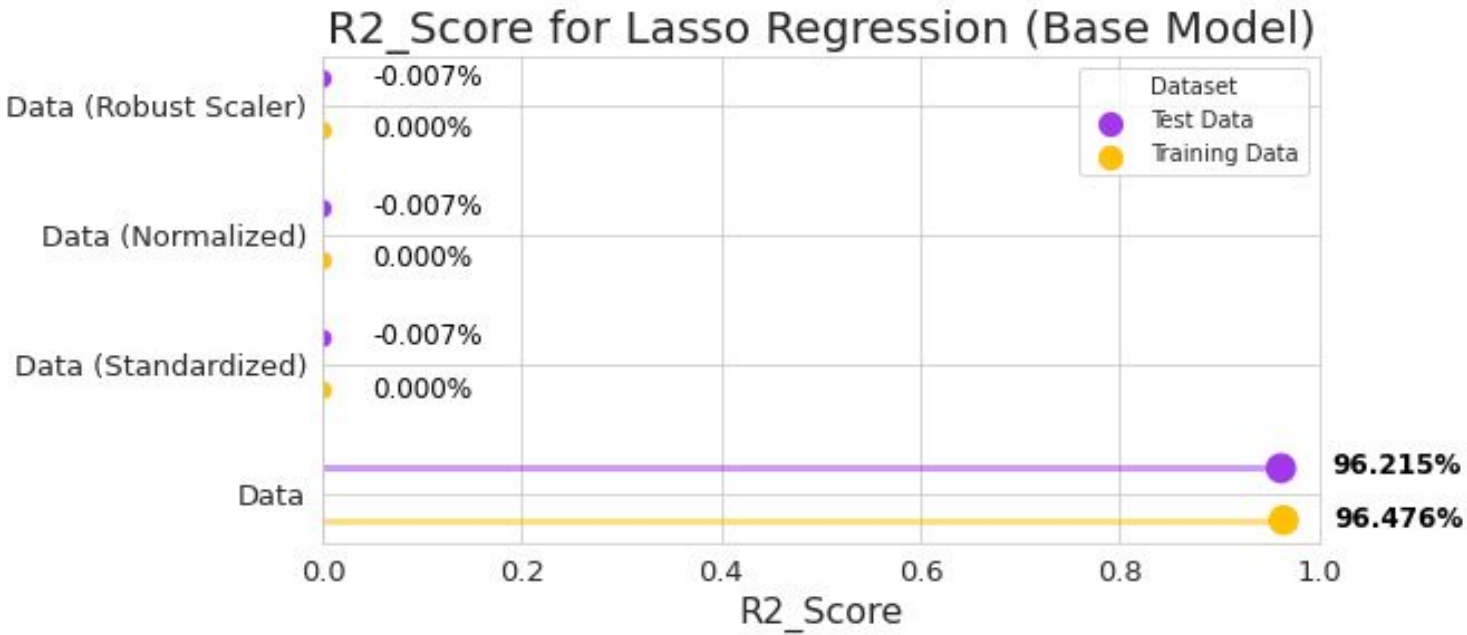
Garis dan ukuran dot yang tebal merepresentasikan **3 data dengan NMRSE terendah**, yang mana hampir dimiliki seluruh **dataset kedua**

Model: Ridge Regression

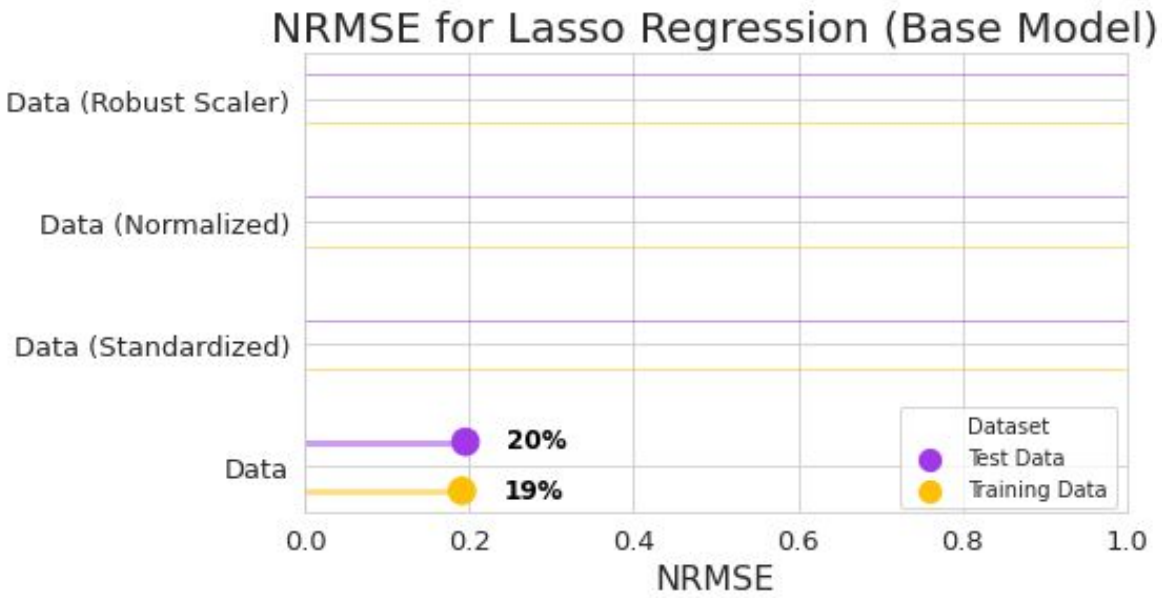




Model: Lasso Regression



Model: Lasso Regression



4713697309056279552%

2349629902499330560%

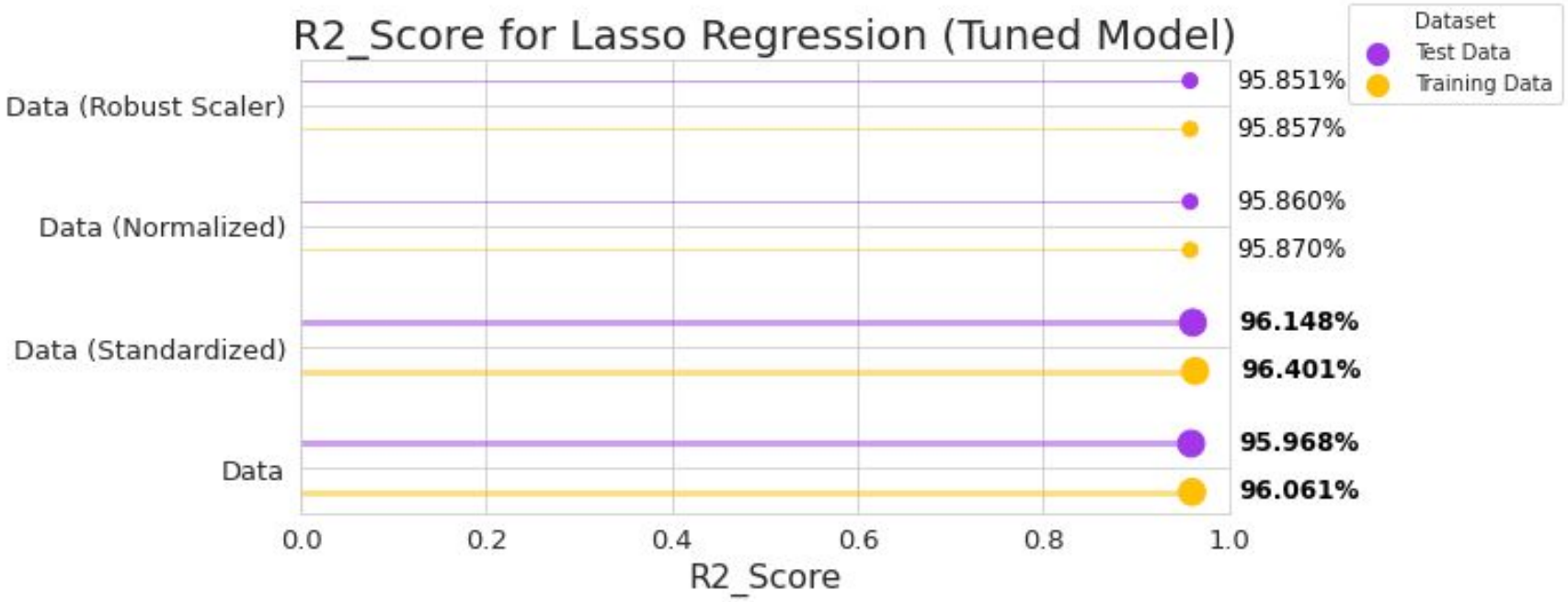
289751527351716672%

288864052281180352%

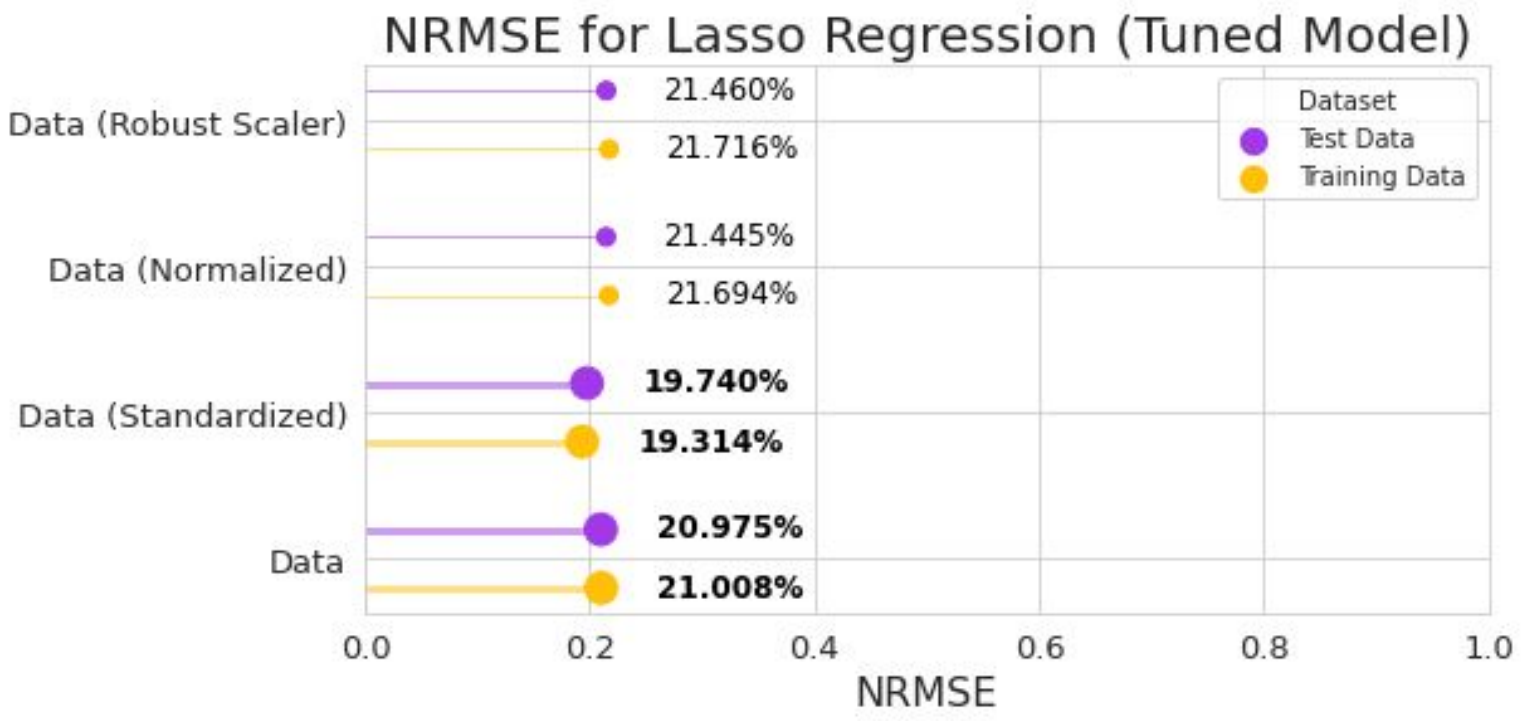
406894462264945485091032247828480 0%

202824096036516704239472512860160 0%

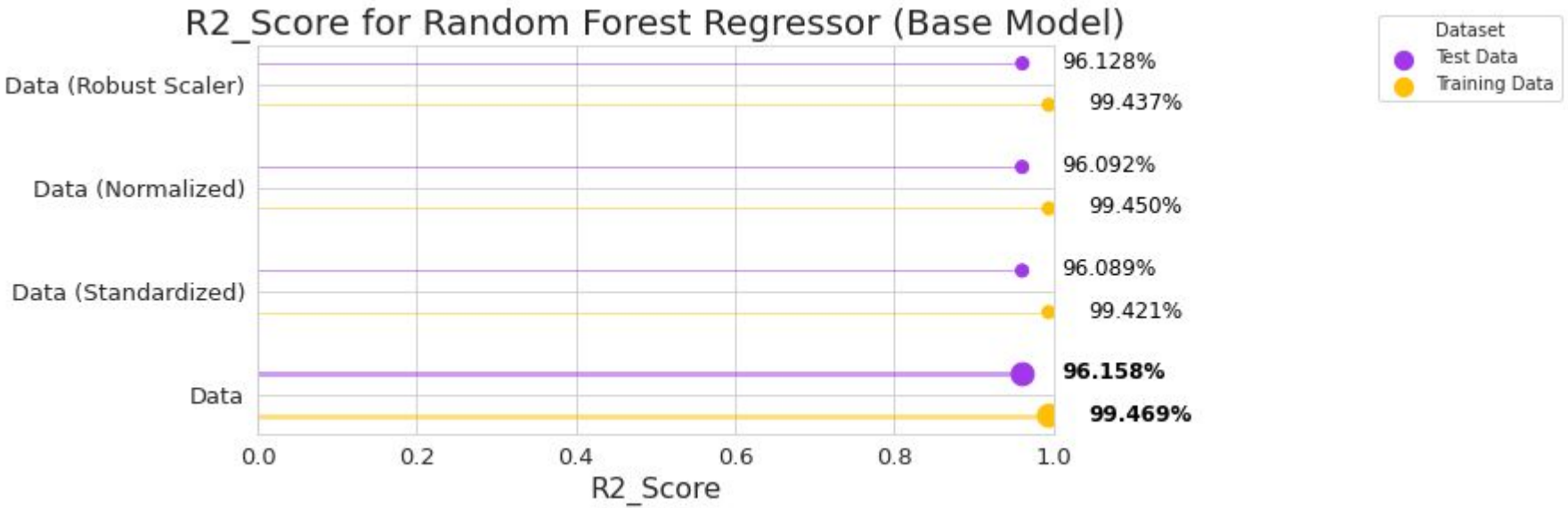
Hyperparameter Tuning: Lasso Regression



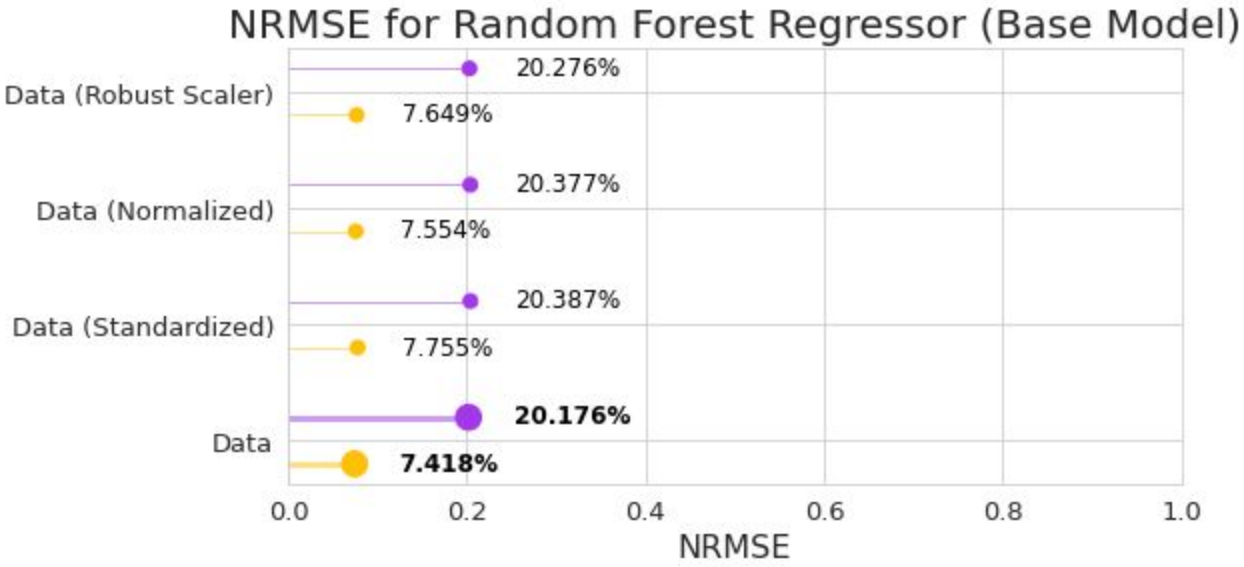
Hyperparameter Tuning: Lasso Regression



Model: Random Forest Regressor



Model: Random Forest Regressor



Conclusion

Conclusion

- Model terbaik untuk memprediksi penjualan mingguan adalah Ridge Regression. Model ini dapat memberikan informasi yang membantu manajer bisnis mengidentifikasi dan memahami kelemahan dalam perencanaan bisnis.

Rekomendasi:

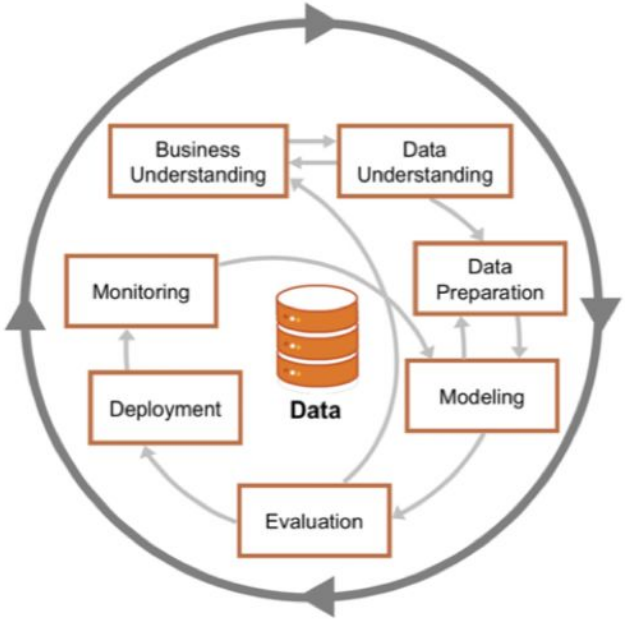
- Divisi Marketing sebaiknya meningkatkan advertising ketika Minggu-minggu sebelum Christmas dan saat Thanksgiving.
- Perlu penambahan orang dari divisi Logistik di bulan November-Desember karena penjualannya meningkat signifikan dibandingkan bulan-bulan lainnya.
- Melakukan re-stock barang secukupnya di hari biasa untuk meminimalkan production cost.

Terima kasih!

Ada pertanyaan?

LAMPIRAN

CRISP-DM Methodology



Feature Selection & Decomposition

- **Recursive Feature Elimination (RFE)**
  menyeleksi fitur berdasarkan estimator secara rekursif

  (+) Implementasinya mudah
  (+) Mudah diinterpretasikan'

  (-) Sangat bergantung pada estimator
  (-) Fitur yang berpotensi bisa saja tidak lolos seleksi

- **Principal Component Analysis (PCA)**
  mereduksi fitur dengan melakukan kombinasi berdasarkan nilai varians

  (+) Implementasinya mudah
  (+) Fitur akan dikombinasikan, bukan dieliminasi

  (-) Hasil sulit diinterpretasikan