Final Project Presentation

Nomor Kelompok: 2
Nama Mentor: Ramdhan Hidayat
Nama:

- Dwi Wulandari
- Riska Budiyanti

Machine Learning Class

Program Studi Independen Bersertifikat
Zenius Bersama Kampus Merdeka

Petunjuk

- Waktu presentasi adalah 5 menit (tentatif, tergantung dari banyaknya kelompok yang mendaftarkan diri)
- Waktu tanya jawab adalah 5 menit
- Silakan menambahkan gambar/visualisasi pada slide presentasi
- Upayakan agar tetap dalam format poin-poin (ingat, ini presentasi, bukan esai)
- Jangan masukkan *code* ke dalam slide presentasi (tidak usah memasukan screenshot jupyter notebook)

Latar Belakang

Latar Belakang Project

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

Problem: **regression**

Tujuan:

- Mengetahui pola (model) yang baik dalam memprediksi penjualan mingguan Walmart
- Mengetahui faktor-faktor apa saja yang memengaruhi penjualan mingguan Walmart

Explorasi Data dan Visualisasi

Business Understanding

Walmart merupakan perusahaan retail multinasional yang berasal dari Amerika Serikat. Perusahaan ini sudah berdiri cukup lama, yaitu sejak tahun 1962. Singkat cerita, Walmart saat ini sudah berkembang begitu besar hingga jumlah konsumennya mencapai sekitar **245 juta** orang. Mereka juga memiliki sekitar **10 ribu** gerai yang tersebar di seluruh dunia.

Business Understanding

Melalui bisnisnya, Walmart fokus menjual barang-barang yang murah dengan volume penjualan yang tinggi. Hingga saat ini Walmart masih setia dengan strategi **EDLP** (*Every Day Low Price*).

Dengan peningkatan skala bisnis yang begitu pesat, tentu Walmart harus memutar otak agar kegiatan operasional bisnisnya menjadi lebih efisien. Dalam hal ini, solusi yang mereka pilih adalah memanfaatkan **data**.

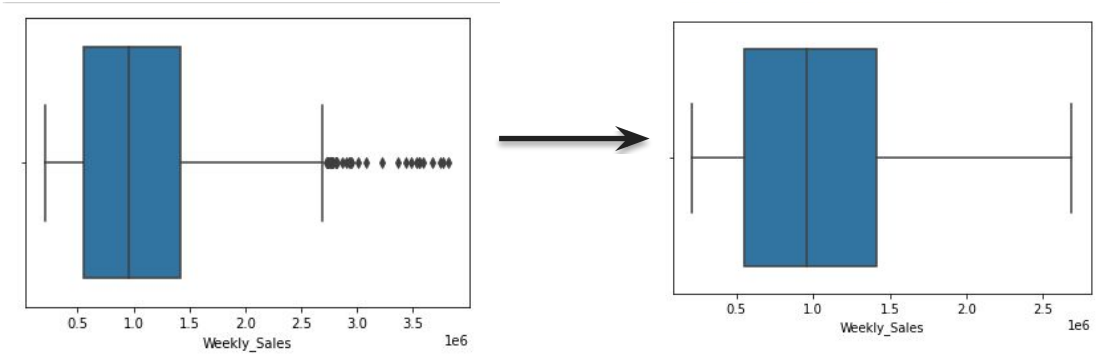
Data Cleansing

Data Walmart tidak perlu dibersihkan. Rincian dimensi datanya sendiri yaitu terdiri dari 6435 baris dengan 8 kolom dan tidak terdapat data yang missing. Berikut ini penjelasan masing-masing kolom:

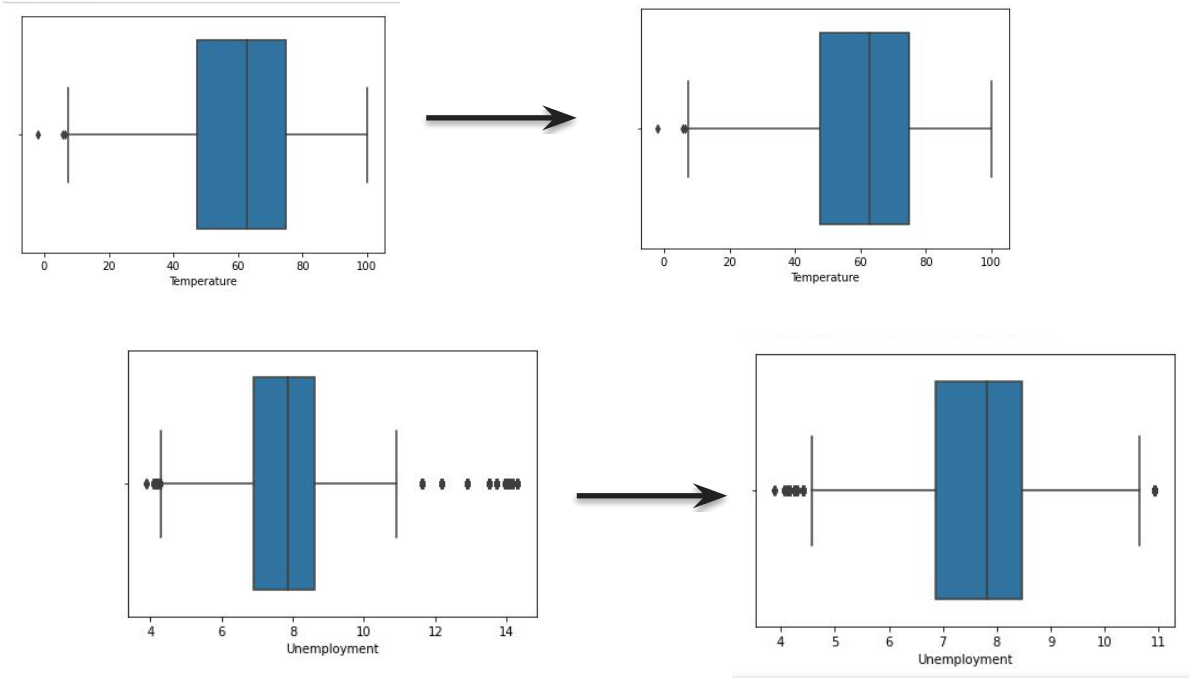
- Store: nomor toko
- Date: tanggal penjualan
- Weekly_Sales: penjualan mingguan toko
- Holiday_Flag: banyaknya hari libur pada minggu tersebut
- Temperature: suhu pada minggu penjualan
- Fuel_Price: biaya bahan bakar di daerah tersebut
- CPI: (Consumer Price Index) indeks harga konsumen yang berlaku
- Unemployment: tingkat pengangguran yang berlaku

Data Cleansing

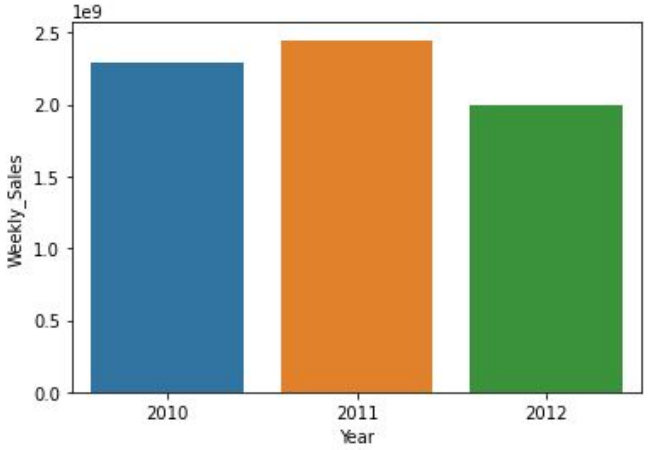
Namun, pada data Walmart mempunyai outliers yang terdapat pada kolom ‘Weekly_Sales’, ‘Temperature’, dan ‘Unemployment’. Solusi dari kami untuk mengatasi outliers tersebut yaitu dengan melakukan removal ketiga kolom tersebut. Dan didapatkan visualisasi boxplot sebelum dan sesudah removal sebagai berikut:

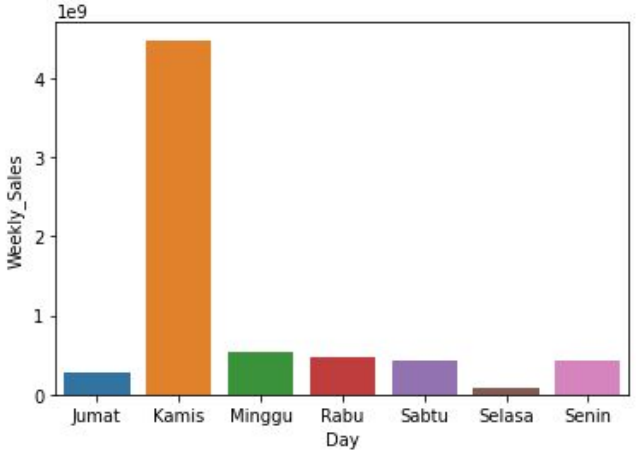


Data Cleansing



Exploratory Data Analysis

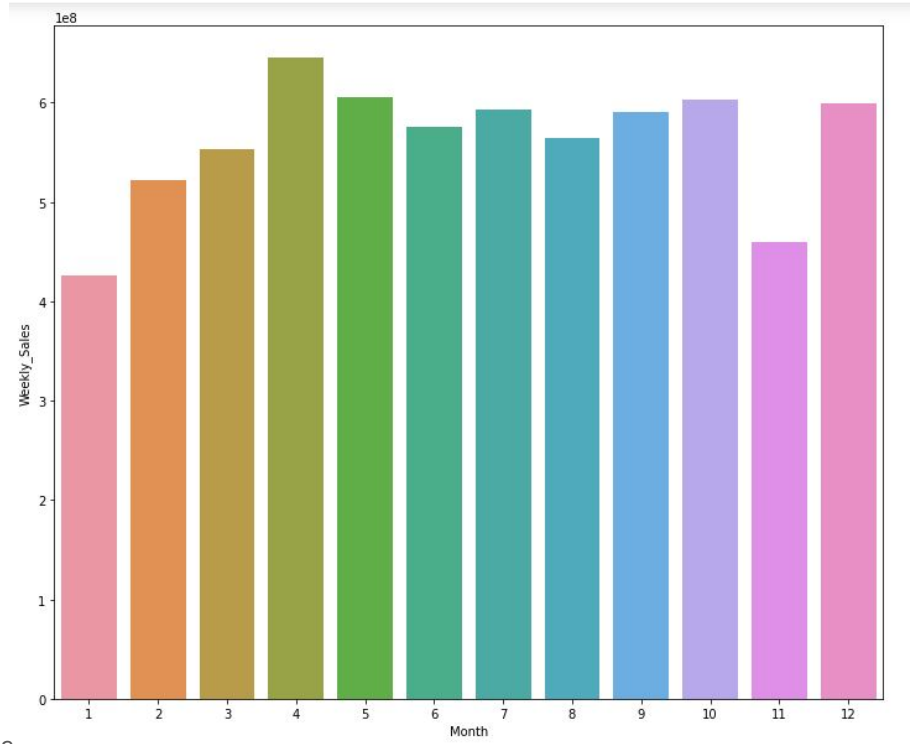




- Tahun 2011 mencatat jumlah penjualan tertinggi kemudian diikuti oleh tahun 2010 dan tahun 2012.
- Penjualan tertinggi terjadi pada Hari Kamis yaitu lebih dari 50%.

Exploratory Data Analysis

- Bulan April(4) tercatat sebagai penjualan tertinggi dan Bulan Januari(1) sebagai penjualan terendah.



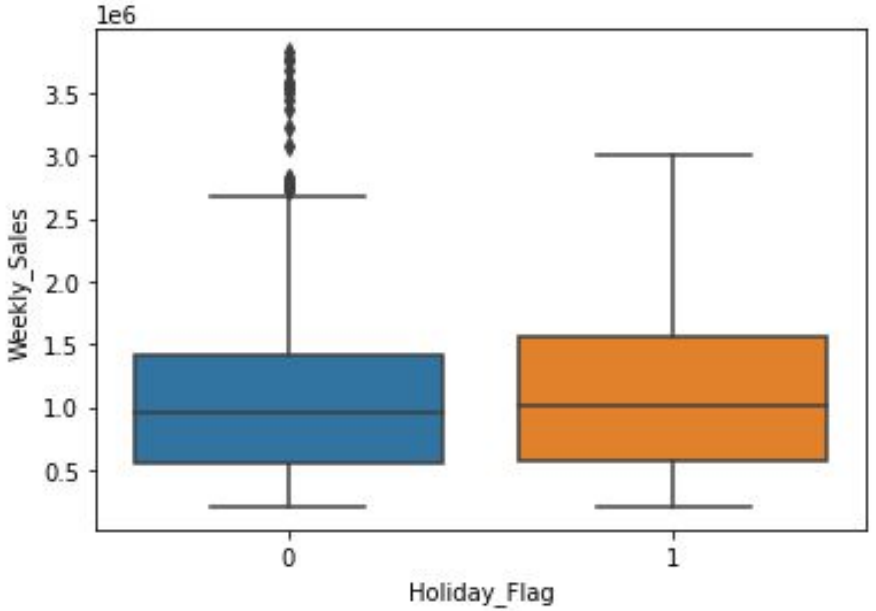
Exploratory Data Analysis

- Berdasarkan *Spearman correlation*, kolom 'Year' dan 'Fuel_Price' merupakan dua variabel yang memiliki korelasi tinggi. Hal ini make sense karena, biaya bahan bakar di daerah tertentu pasti berbeda setiap tahunnya.

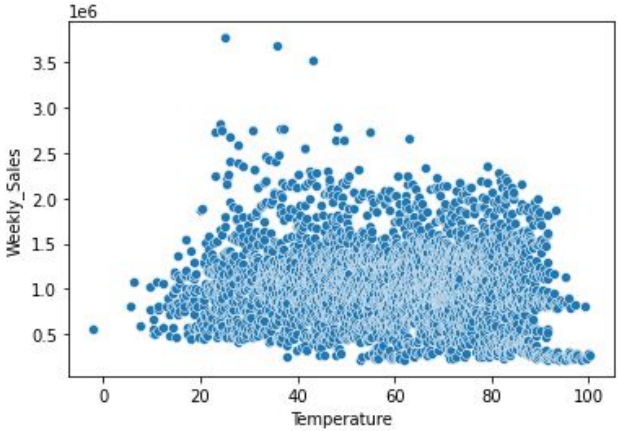
	Store	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment	Month	Year	Day_Number
Store	1.000000	-0.309227	0.000000	-0.026392	0.064878	-0.238852	0.304139	0.000000	0.000000	0.000000
Weekly_Sales	-0.309227	1.000000	0.027774	-0.070962	0.025471	-0.055040	-0.062354	0.045589	-0.006395	-0.018558
Holiday_Flag	0.000000	0.027774	1.000000	-0.143588	-0.080111	-0.004752	0.011177	0.329995	-0.056426	-0.091559
Temperature	-0.026392	-0.070962	-0.143588	1.000000	0.128624	0.165957	0.038833	0.070467	0.064448	0.004193
Fuel_Price	0.064878	0.025471	-0.080111	0.128624	1.000000	-0.045867	-0.064725	-0.056058	0.762505	-0.060913
CPI	-0.238852	-0.055040	-0.004752	0.165957	-0.045867	1.000000	-0.388563	0.005928	0.220834	-0.031950
Unemployment	0.304139	-0.062354	0.011177	0.038833	-0.064725	-0.388563	1.000000	-0.001387	-0.279020	0.043187
Month	0.000000	0.045589	0.329995	0.070467	-0.056058	0.005928	-0.001387	1.000000	-0.135752	-0.128210
Year	0.000000	-0.006395	-0.056426	0.064448	0.762505	0.220834	-0.279020	-0.135752	1.000000	-0.128449
Day_Number	0.000000	-0.018558	-0.091559	0.004193	-0.060913	-0.031950	0.043187	-0.128210	-0.128449	1.000000

Exploratory Data Analysis

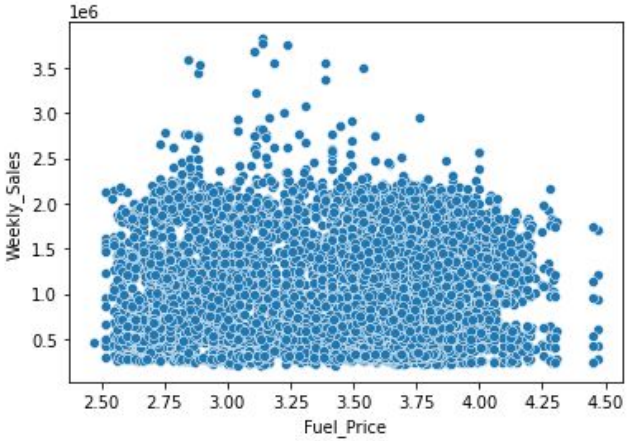
- Korelasi antara ‘Holiday_Flag’ dan ‘Weekly_Sales’ dikatakan rendah karena dari boxplot terlihat perbedaan penjualan di hari libur dan hari tidak libur tidak jauh beda.



Exploratory Data Analysis



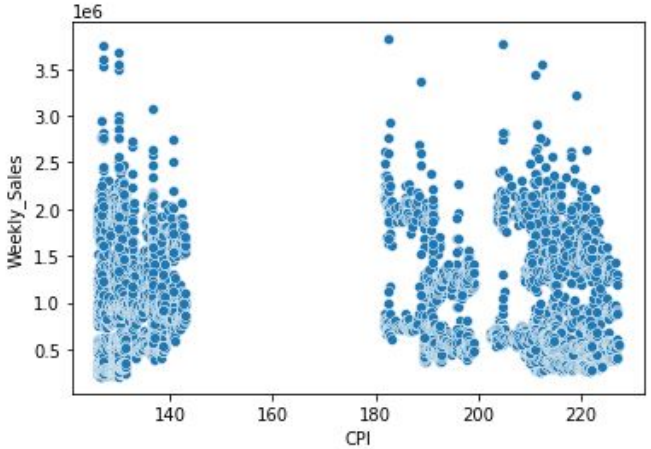
- Temperature pada selang 60-70 **mempengaruhi** penjualan, dimana terlihat dari tingginya histogram pada selang tersebut.



- **Tidak ada korelasi** antara Fuel_Price dan Weekly_Sales.

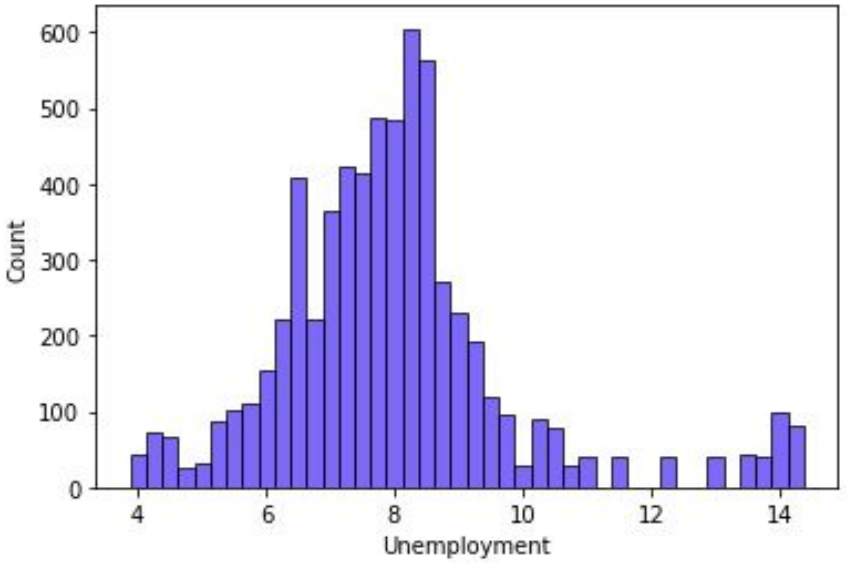
Exploratory Data Analysis

- Rentang CPI yang berbeda memiliki distribusi penjualan yang sama. Sehingga disimpulkan bahwa perbedaan Indeks Harga Konsumen (CPI) **tidak akan mempengaruhi** penjualan.



Exploratory Data Analysis

- **Terdapat korelasi** antara Unemployment dan Weekly_Sales bahwa tingkat pengangguran yang rendah akan **berpengaruh** pada penjualan (terindikasikan dari tingginya 'bar' diantara rentang 6-10).



Modelling

Modelling Walmart

- Metode train test split / cross validation yang digunakan

Data training yang kelompok kami gunakan sebesar 80% dan data testing yang sebesar 20%.

- Metrik dalam melakukan evaluasi yang kelompok kami gunakan adalah MAE, MSE, RMSE.
- Jenis model awal yang dicoba

Model awal yang kelompok kami coba adalah linear regression.

- Model lain yang juga dicoba untuk dataset ini yaitu ridge regression dan random forest. Serta untuk mencoba random forest tuning untuk menambah akurasi model.
- Model final yang kelompok kami gunakan adalah random forest tuning.
- Kolom yang menjadi prediktor dalam model kelompok kami adalah semua kolom, kecuali kolom target yaitu kolom ‘Weekly_Sales’.

Conclusion

Berdasarkan evaluasi masing-masing model diantaranya:

- Linear Regression
  - MAE: 430365.87206897634
  - MSE: 268123763332.19513
  - RMSE: 517806.6852911375
- Ridge Regression
  - MAE: 430364.895750012
  - MSE: 268123443576.494
  - RMSE: 517806.3765313189
- Random Forest
  - MAE: 223185.77432712252
  - MSE: 100089287268.63945
  - RMSE: 316368.91008542455
- Random Forest Tuning
  - MAE: 68056.89373772778
  - MSE: 13972246369.597878
  - RMSE: 118204.25698593886

dikarenakan MAE pada **model random forest tuning** dengan 'n_estimators': 100, 'min_samples_split': 10, 'min_samples_leaf': 4, 'max_features': 'auto', 'max_depth': 50, 'bootstrap': True paling kecil, maka disimpulkan bahwa **model random forest tuning** merupakan model yang bagus/baik untuk dataset Walmart.

Terima kasih!

Ada pertanyaan?