Final Project Presentation

Nomor Kelompok: 1
Nama Mentor: Rachmadio Noval L
Nama:

- Nadhira Ferita Kusuma
- Firda Zuhrotul Ma'wa

Machine Learning Class

Program Studi Independen Bersertifikat
Zenius Bersama Kampus Merdeka

Latar Belakang

Latar Belakang Project

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

Problem: **regression**

Tujuan:

- Mengetahui *insight* menarik dari penjualan di Walmart.
- Memprediksi pendapatan penjualan mingguan Walmart berdasarkan hari, cuaca, harga BBM, dan tingkat pengangguran.

Explorasi Data dan Visualisasi

Business Understanding

- Mengadopsi data penjualan dari salah satu toko ritel terkemuka di Amerika Serikat, yaitu **Walmart**, yang memiliki data penjualan sebanyak 45 toko.
- Tujuan utama dari proyek ini adalah untuk memprediksi perolehan penjualan mingguan *(weekly sales)* dengan menerapkan regresi menggunakan algoritma *machine learning*.
- Hasil prediksi dapat digunakan oleh marketer dalam menentukan strategi marketing yang relevan.

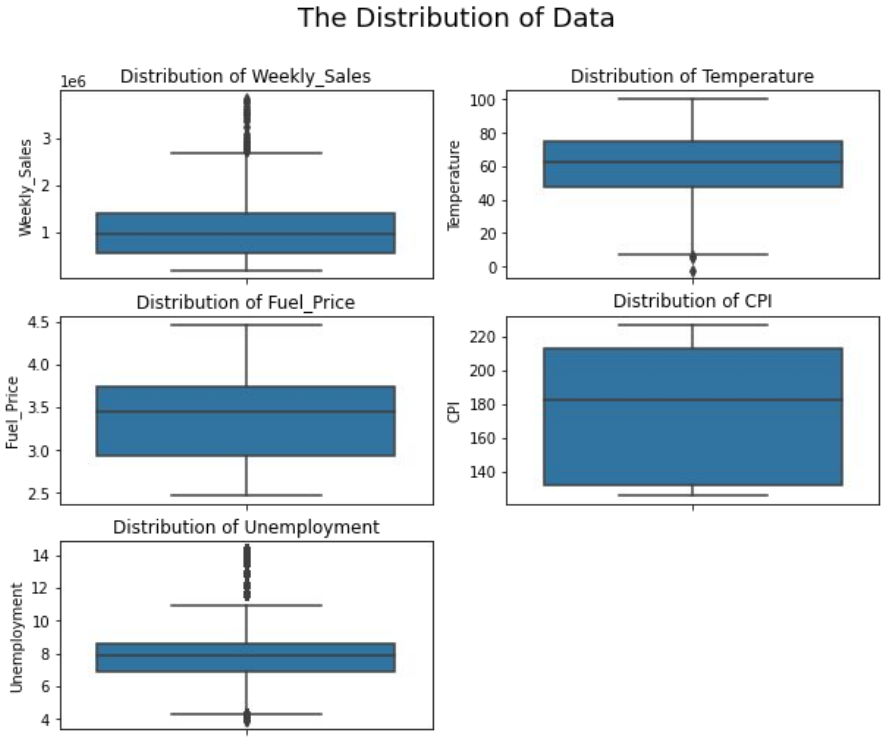
Weekly Sales

Weekly sales merupakan pendapatan yang diperoleh dari hasil penjualan tiap minggunya oleh masing-masing toko. Pendapatan yang diperoleh biasanya cenderung meningkat ketika adanya perayaan khusus atau hari libur nasional seperti Thanksgiving day, labour day, superbowl day, dan christmas.

Data Cleansing

- Jumlah Baris dan Kolom Data Walmart: **6435 baris dan 8 kolom (51480 data)**
- Kolom pada data: Store, Date, Weekly_Sales, Holiday_Flag, Temperature, Fuel_Price, CPI, Unemployment.
- **Tidak terdapat *missing data***
- Data pada Tahun 2012 tidak lengkap dimana akan berpengaruh dalam melihat insight data yang ada. Namun dalam pemodelan tidak berpengaruh.
- Melakukan Feature engineering:
    - Membuat 4 kolom baru berdasarkan kolom Date.
    - Pada kolom “Temperature” dibagi menjadi Hot, Warm, Cool, dan Cold.
    - Mengubah isi kolom “Holiday_Flag” dan kolom baru berdasarkan holiday event.

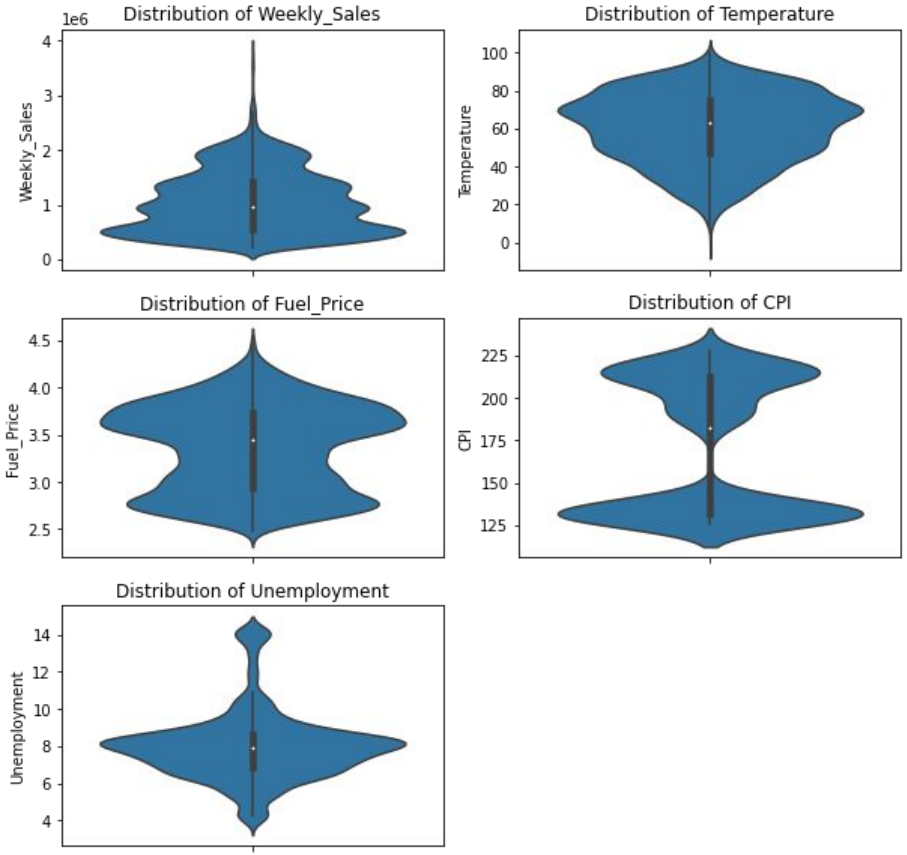
Data Cleansing



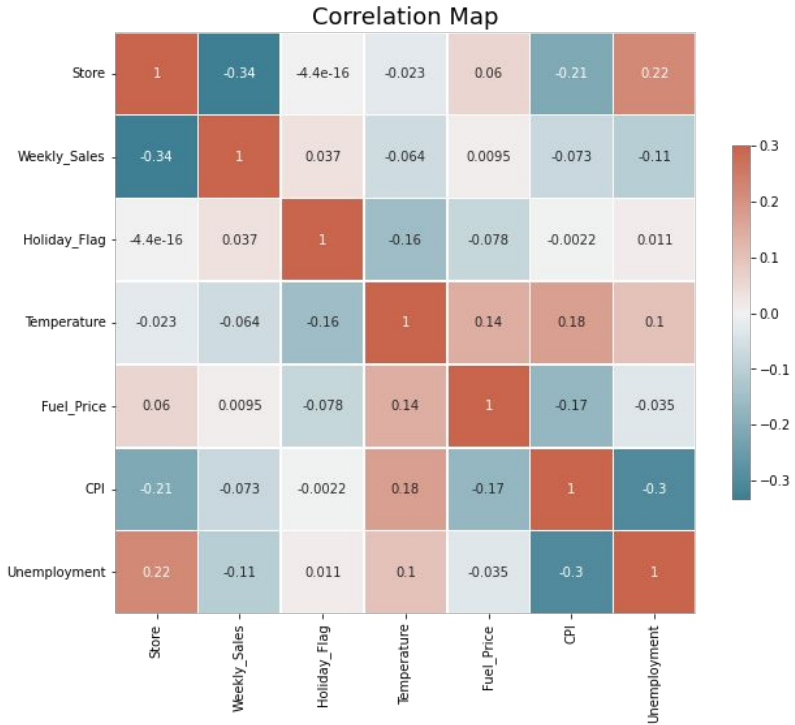
Berdasarkan boxplot, terdapat *outlier* di kolom Weekly Sales, Temperature, dan Unemployment. Namun tidak dilakukan penghapusan data karena outlier yang ada tidak janggal.

Data Cleansing

Pada kolom Fuel Price dan CPI distribusi yang terbentuk bimodal distribution. Maka dilakukan *Feature engineering* menentukan Fuel Price dan CPI terbagi menjadi “Low” dan “High” berdasarkan rata-rata pada data

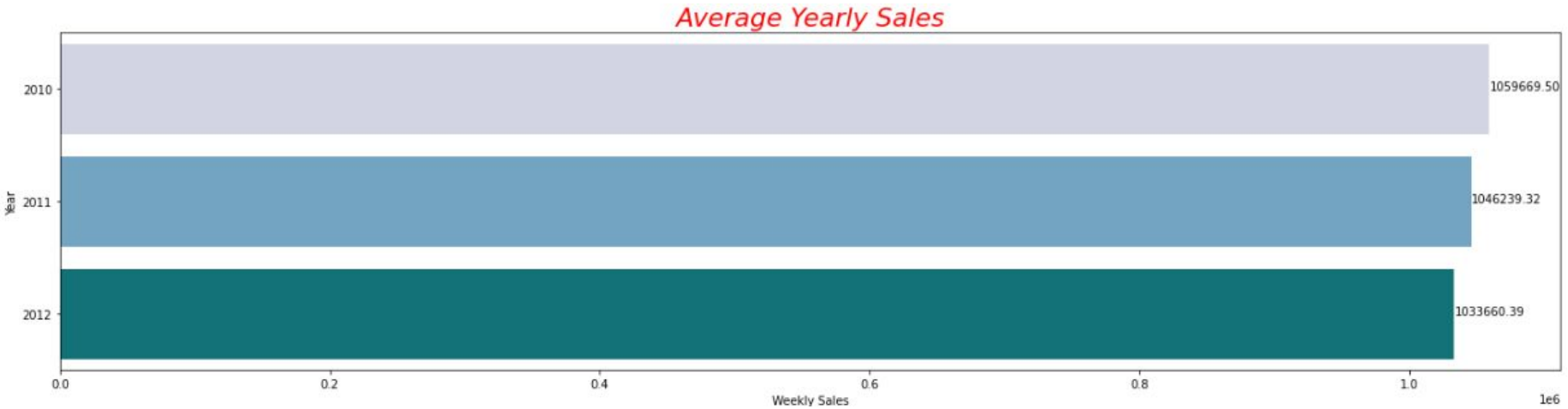


Exploratory Data Analysis



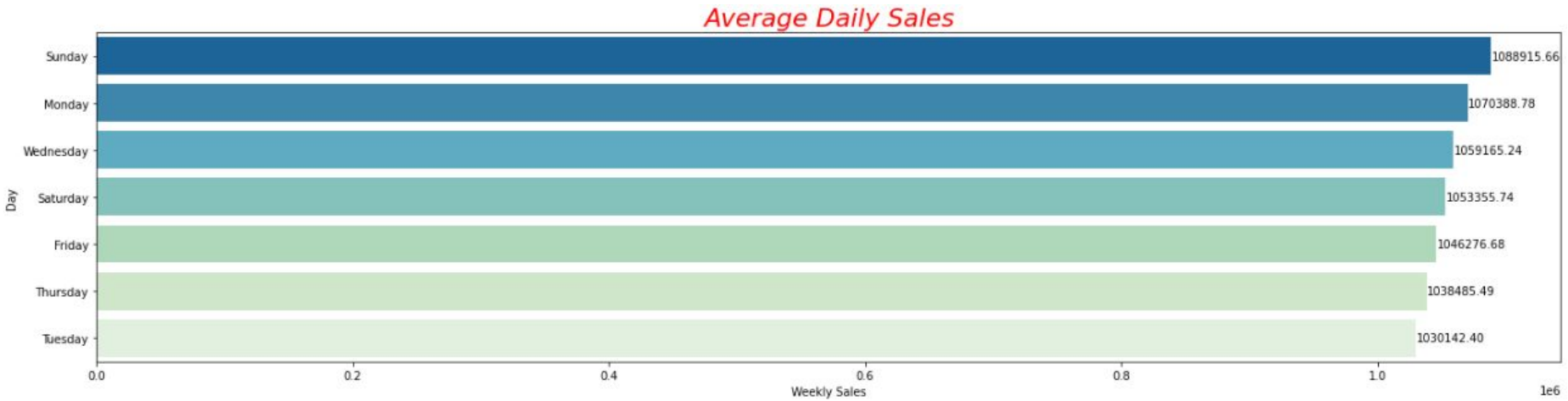
- Antara kolom target “Weekly Sales” dengan kolom lainnya memiliki korelasi tertinggi pada kolom “Store” dan “Unemployment”
- Antar variabel independen tidak terdapat multikolinearitas, sehingga semua variabel dapat digunakan dalam pemodelan.

Exploratory Data Analysis

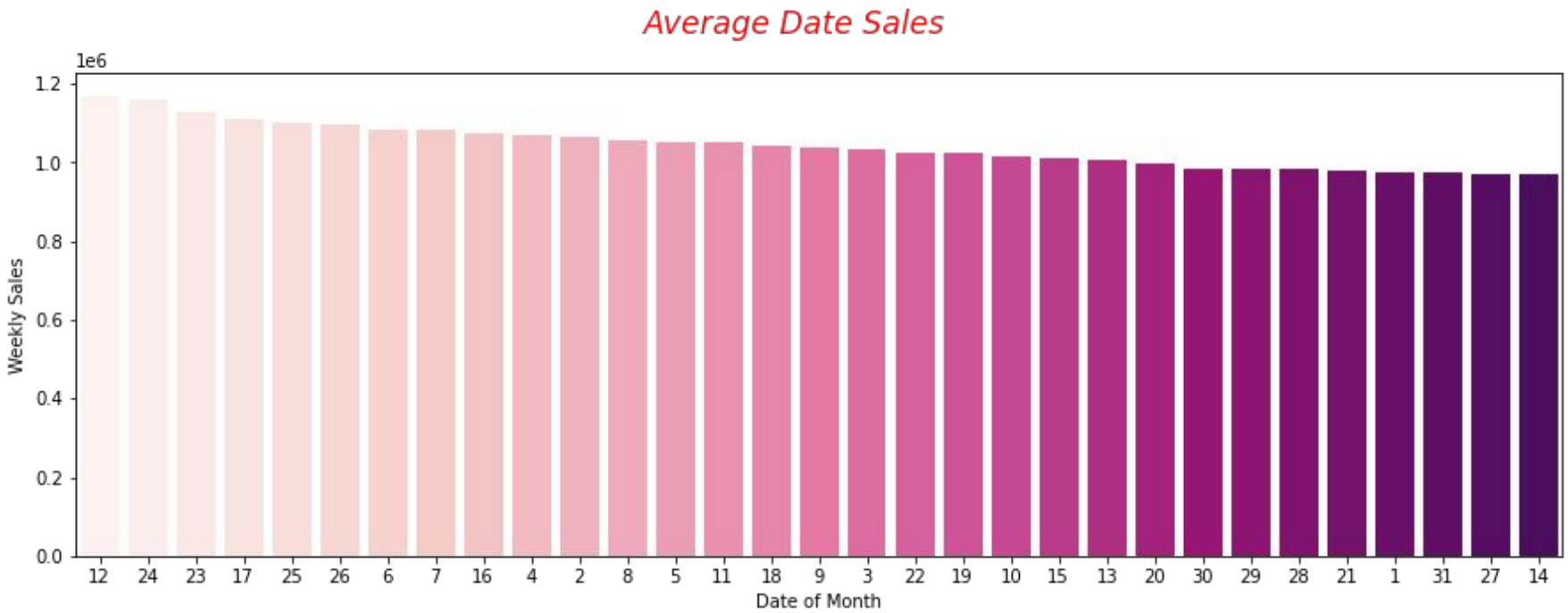


Exploratory Data Analysis

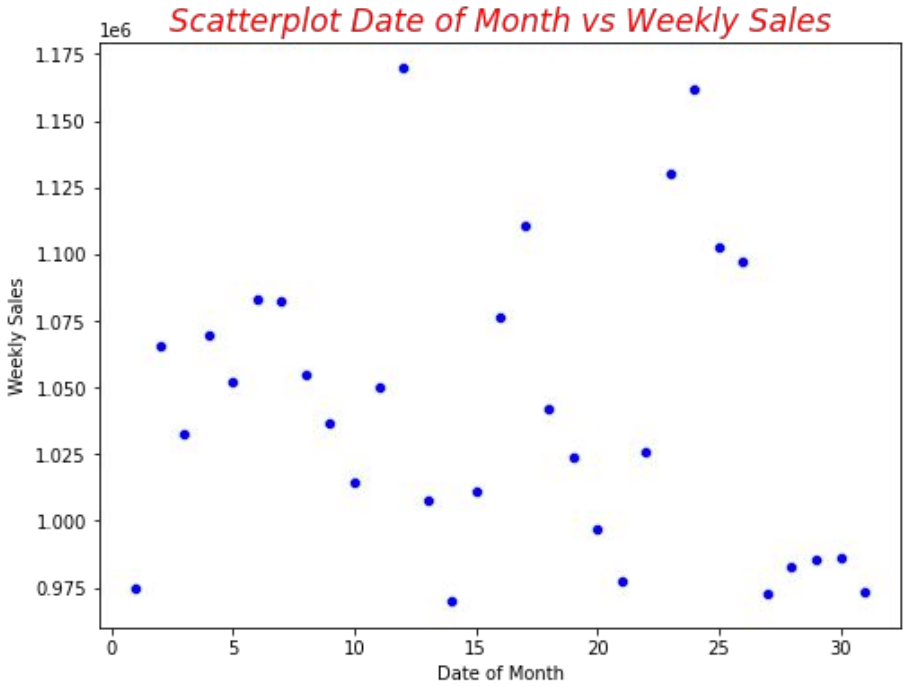
Exploratory Data Analysis

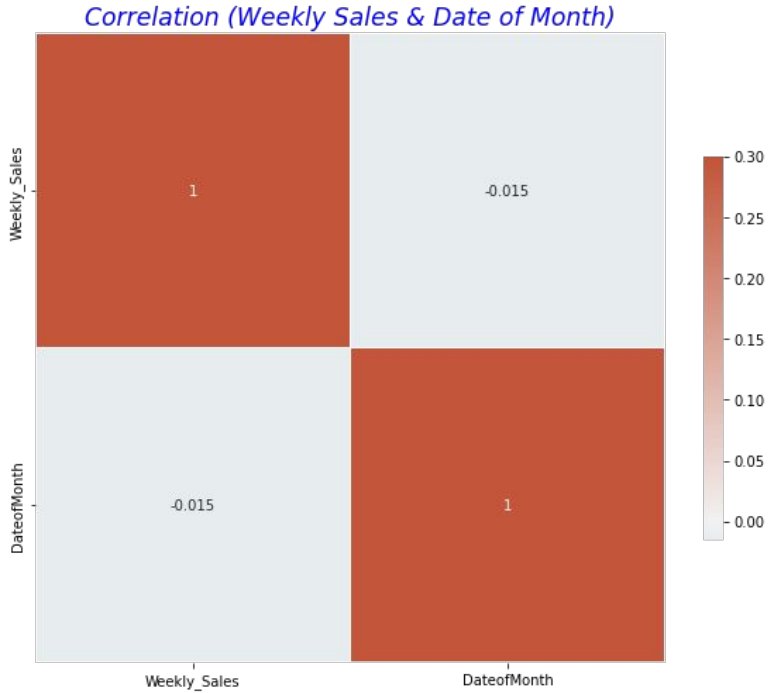


Exploratory Data Analysis



Exploratory Data Analysis

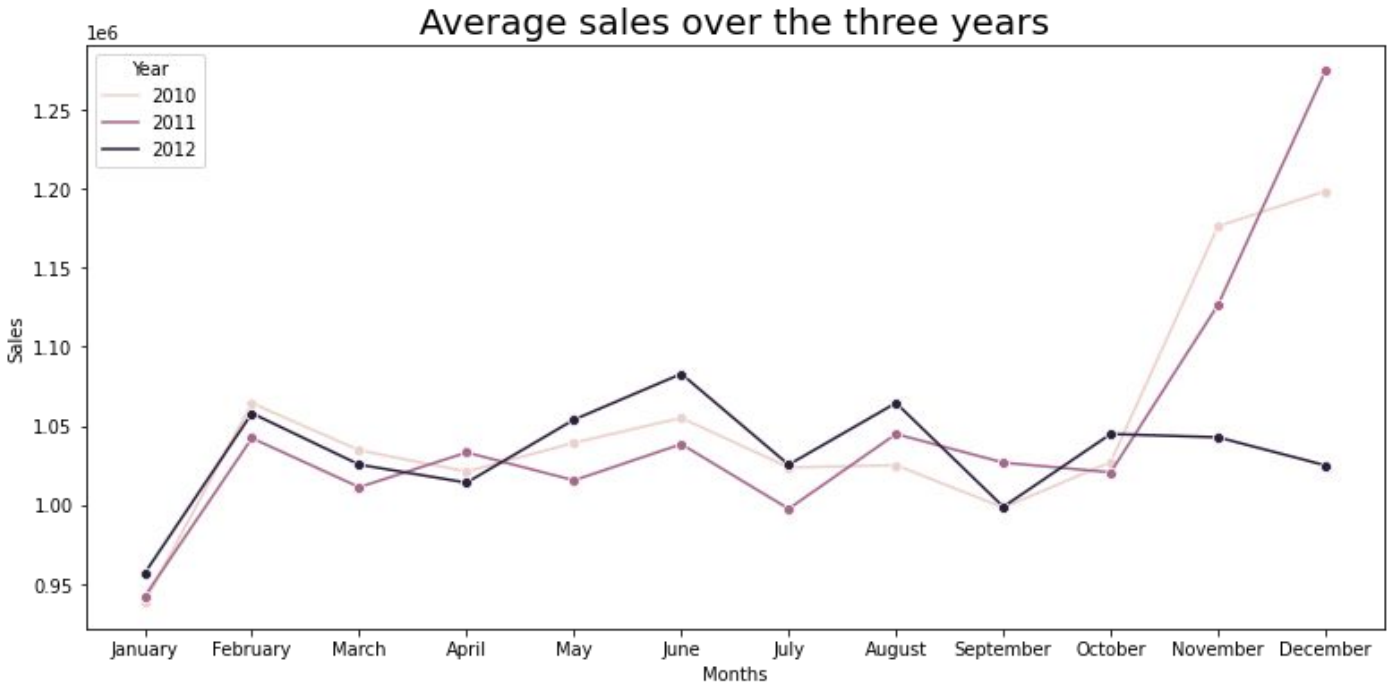




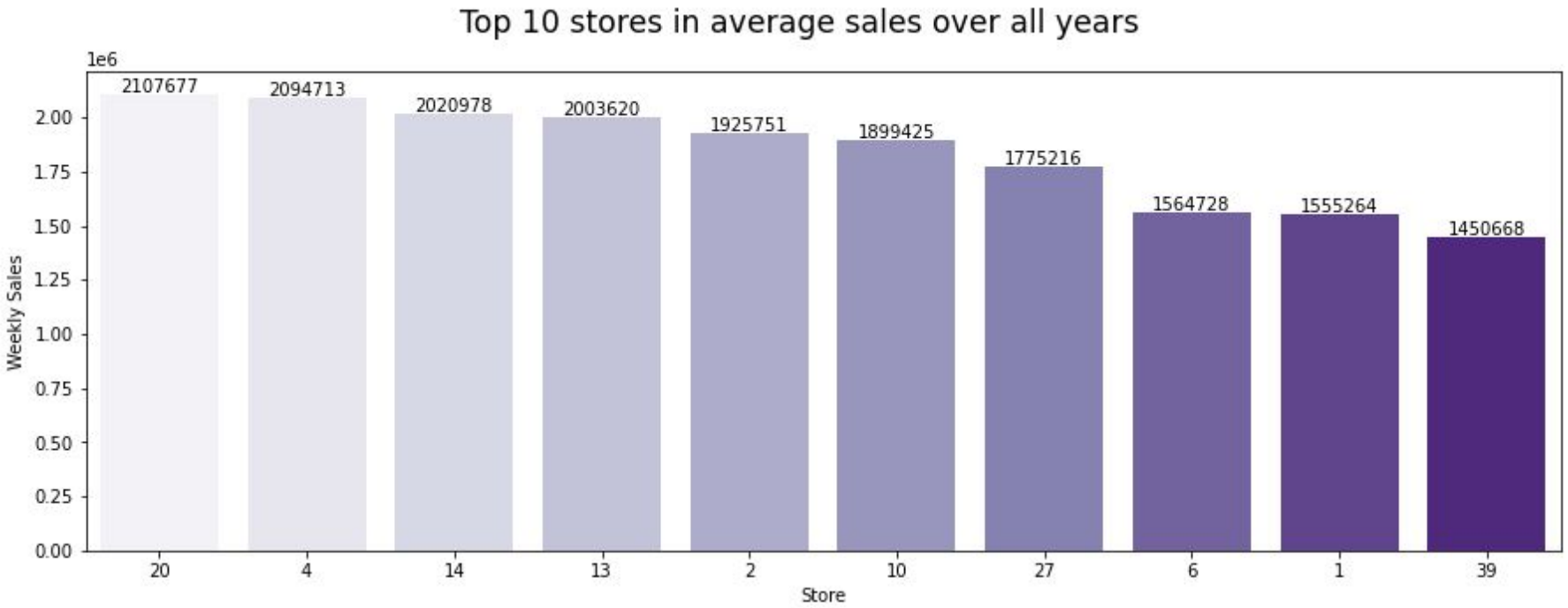
Exploratory Data Analysis

Exploratory Data Analysis

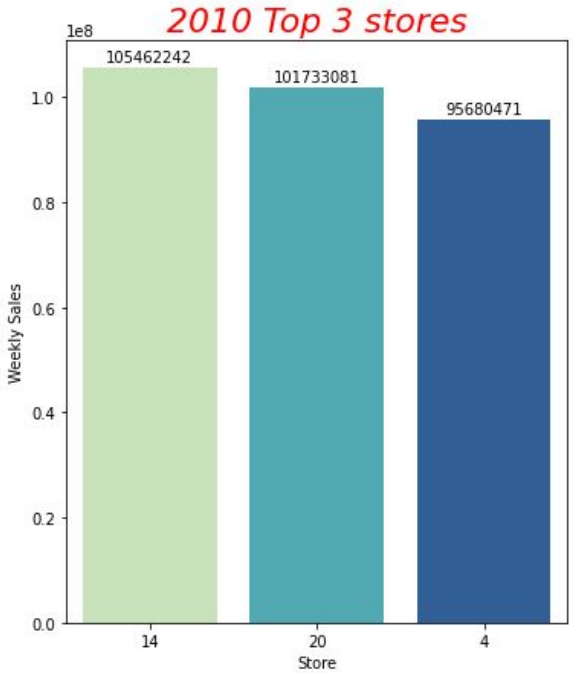
Exploratory Data Analysis

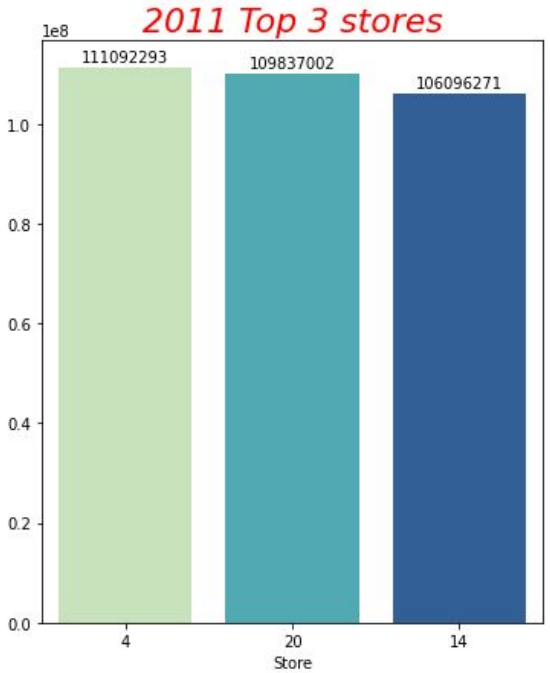


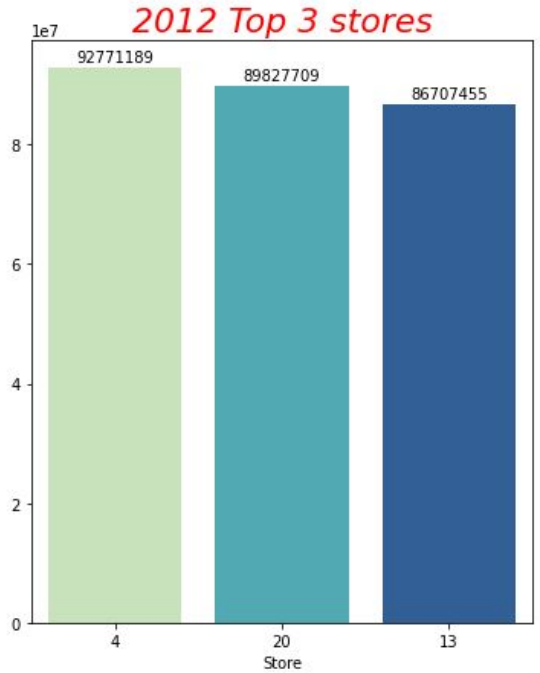
Exploratory Data Analysis



Exploratory Data Analysis







Exploratory Data Analysis

Average Weekly Sales vs Fuel Price

Fuel Price	Weekly Sales
High	1047613.232676
Low	1046208.936089

Average Weekly Sales vs CPI

CPI	Weekly Sales
Low	1082953.365245
High	1012541.106735

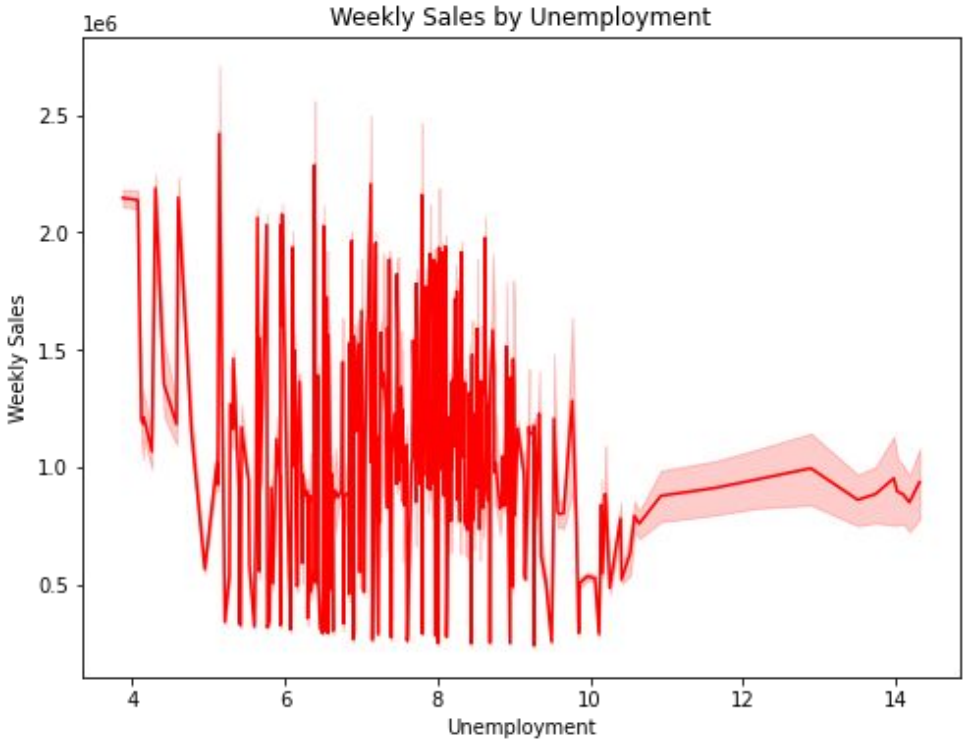
Average Weekly Sales vs Temperature

Temperature	Weekly Sales
Cool	1113007.774005
Warm	1061423.755676
Hot	1017418.437513
Cold	957897.996514

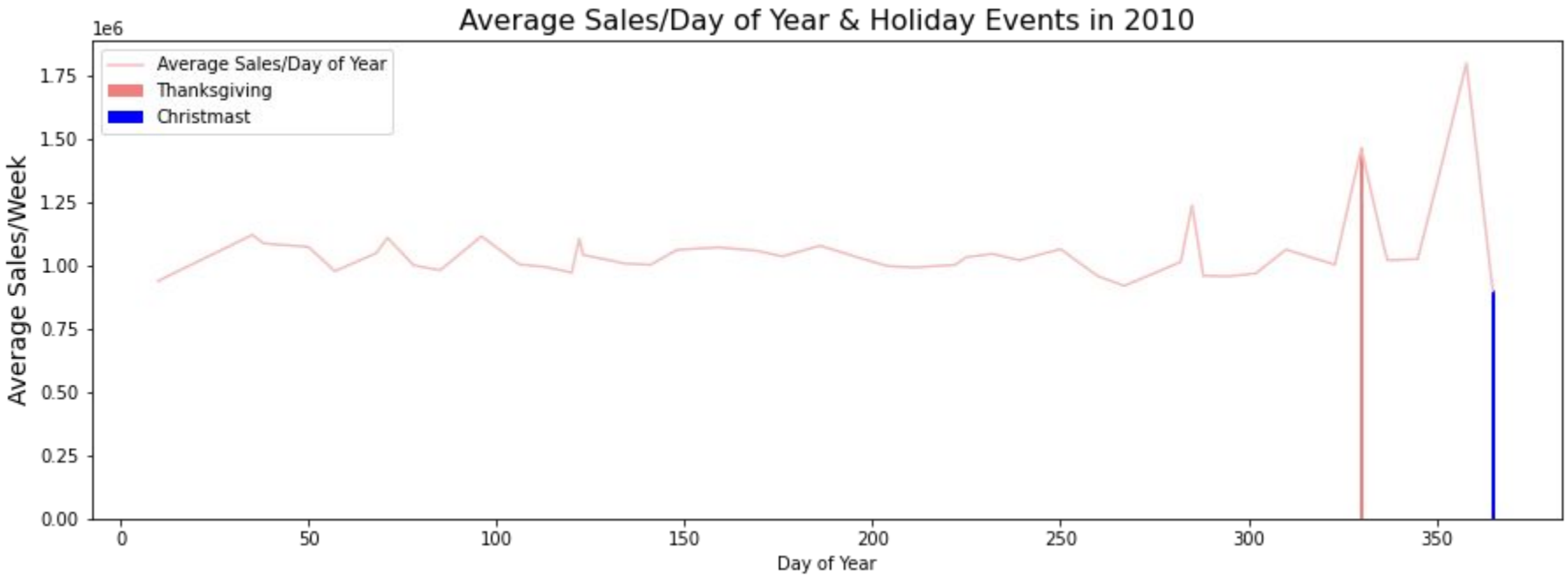
Average Weekly Sales vs Holiday

Holiday	Weekly Sales
Holiday	1122887.892356
Non-Holiday	1041256.380209

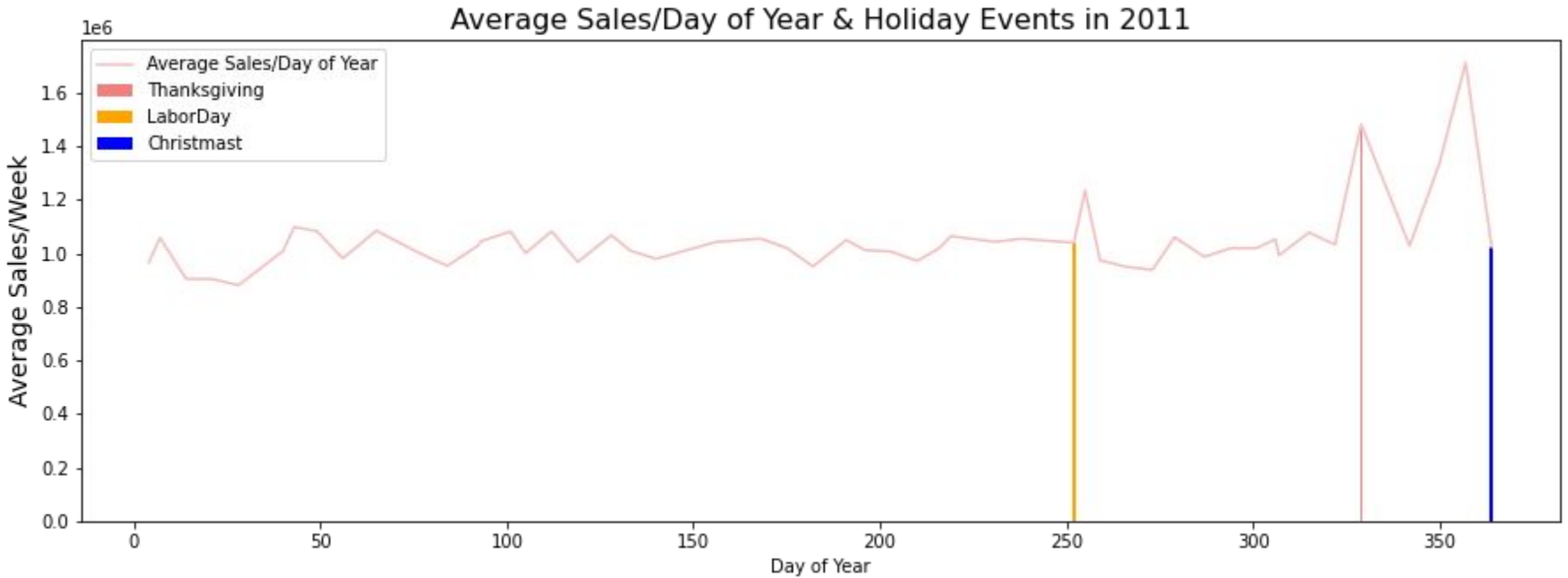
Exploratory Data Analysis



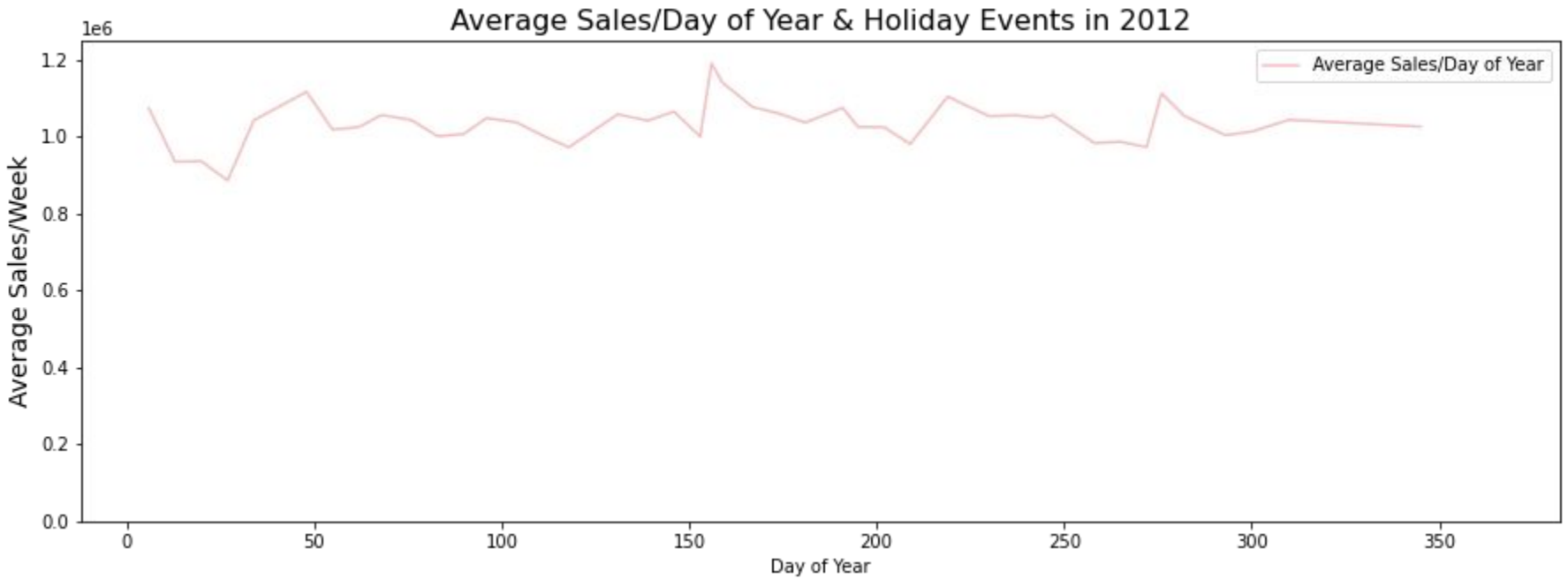
Exploratory Data Analysis



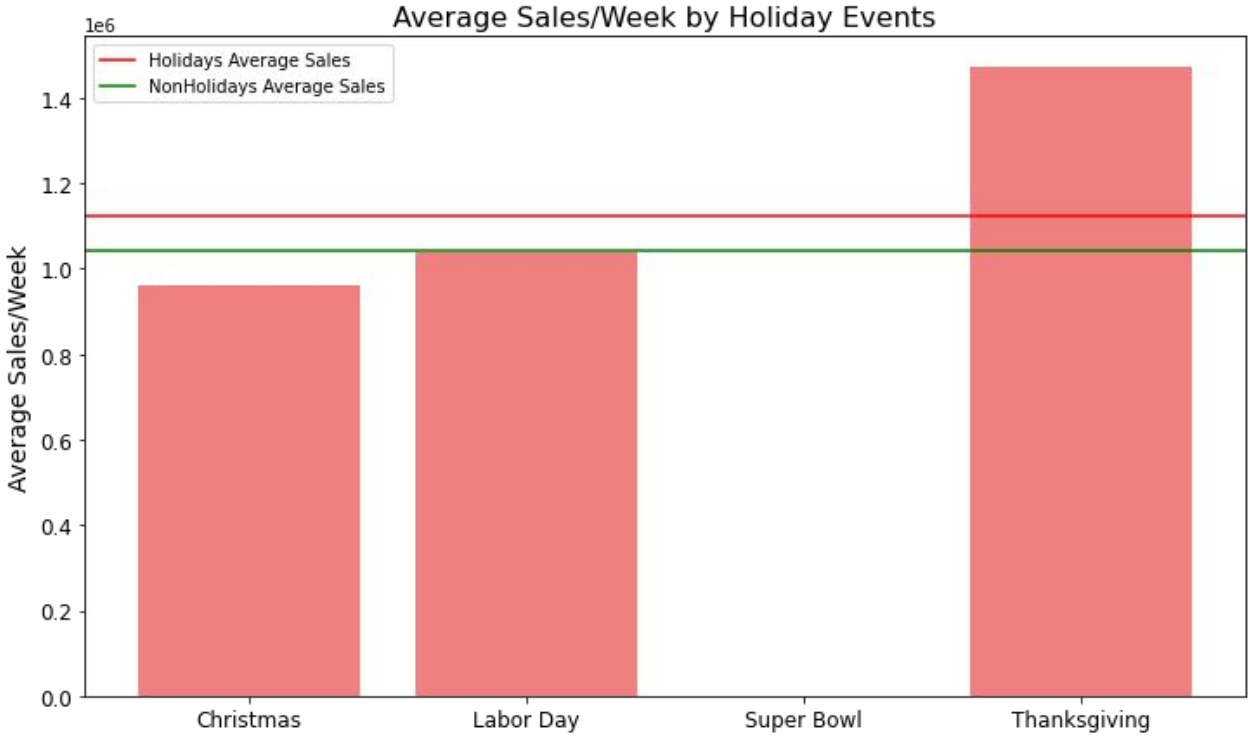
Exploratory Data Analysis



Exploratory Data Analysis



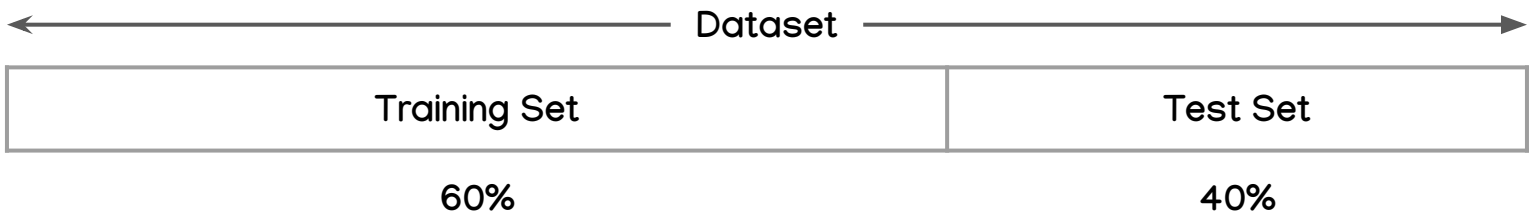
Exploratory Data Analysis



Modelling

Train Test Split

Dataset dibagi ke dalam data train dan data tes dengan test size sebesar 0.4



Matriks Evaluasi

- Root Mean Square Error (RMSE)
- Mean Square Error (MSE)
- Mean Absolute Evaluation (MAE)
- R Squared (R2)

Linear Regression

- **Model 1**

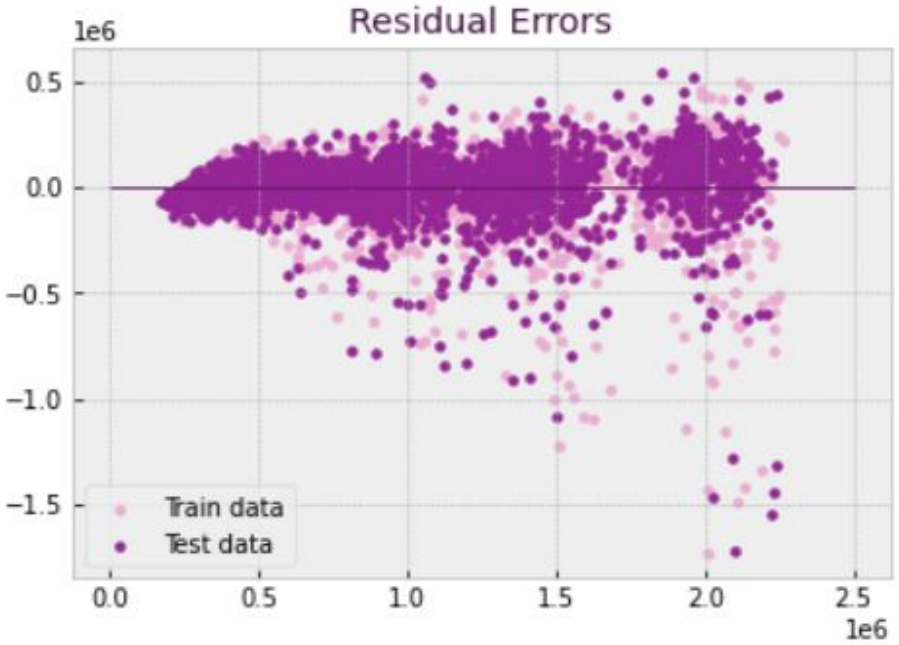
  Feature X : Store, Holiday_Flag, Temperature, Fuel_Price, Unemployment, CPI, DayOfYear, Day, Month, Year

  Feature Y : Weekly_Sales

- **Model 2**

  Feature X : Store, Holiday_Flag, Temperature, ~~Fuel_Price~~, Unemployment, CPI, DayOfYear, Day, Month, Year

  Feature Y : Weekly_Sales



Metode Regresi Lain yang Diujikan :

- Ridge Regression
- Decision Tree
- Random Forest Regressor

Hasil Pengujian Model 1 (Feature X : Store, Holiday_Flag, Temperature, Fuel_Price, Unemployment, CPI, DayOfYear, Day, Month, Year)

Jenis Regresi	MSE	MAE	RMSE	R2
Linear Regression	24514087285.731239	97375.474825	156569.752142	0.922840
Ridge Regression	24924341736.261620	97826.000887	157874.449283	0.921549
Decision Tree Regressor	28778944197.293335	88363.448710	169643.579888	0.909417
Random Forest Resgressor	20064018204.160213	71611.005292	141647.513936	0.936847

Hasil Pengujian Model 2 (Feature X : Store, Holiday_Flag, Temperature, ~~Fuel_Price~~, Unemployment, CPI, DayOfYear, Day, Month, Year)

Jenis Regresi	MSE	MAE	RMSE	R2
Linear Regression	24533119154.290142	97433.442109	156630.517953	0.922781
Ridge Regression	24969999125.138641	97883.186645	158018.983433	0.921405
Decision Tree Regressor	29978938907.623966	88938.009281	173144.271946	0.905639
Random Forest Resgressor	20543483549.635277	72621.730606	143329.981336	0.935338

Random Forest HyperParameter Tuning

- **Cross validation = 5**
- **n_itter = 50**

Setelah dilakukan Random Forest Hyperparameter Tuning, terjadi peningkatan akurasi terhadap model sebesar **6.8%**

Sehingga, final model memiliki nilai evaluasi R2 menjadi 0.943

Final Model

- Model terbaik yang menjadi final model pada projek ini adalah : **Random Forest Regressor Hyperparameter Tuning** dengan nilai R2 akhir sebesar 0.94
- Variabel independen adalah semua kolom, yaitu : **Store, Holiday_Flag, Temperature, Fuel_Price, Unemployment, DayOfYear, Day, Month, Year**
- Variabel dependen : **Weekly_Sales**

Conclusion

Insight

- Rata-rata penjualan di Walmart selama tiga tahun berturut-turut terus mengalami penurunan.
- Tanggal muda/tua tidak berpengaruh terhadap penjualan di Walmart.
- Penjualan terbanyak pada overall years terjadi pada bulan Desember yang terjadi karena merupakan akhir tahun dan perayaan natal, bulan November dimana ada thanksgiving dan black friday, bulan Juni yang biasanya merupakan bulan dimulainya libur musim panas.
- Penjualan pada bulan Januari selalu mengalami penurunan dan akan naik kembali pada bulan Februari.
- Harga BBM tidak mempengaruhi penjualan Walmart, sedangkan cuaca sangat mempengaruhi penjualan.
- Penjualan pada saat Labor Day meningkat di minggu setelahnya ketika buruh sudah mendapatkan tunjangan dan biasanya terdapat Labor Day Sales yaitu diskon besar-besaran.
- Penjualan meningkat di minggu Thanksgiving, hal ini karena harus membeli bahan makanan dan terdapat black friday sale.
- Penjualan Christmas Day meningkat di minggu sebelum event terjadi, hal ini bisa disebabkan karena pada pembeli membeli kebutuhan-kebutuhan seperti dekorasi, hadiah, dll.

Saran

Berdasarkan dari hasil insight penjualan yang didapatkan, pihak Walmart disarankan dapat memanfaatkan penjualan sebaik mungkin, seperti:

- Dapat memberikan promo new years sale agar pada bulan Januari penjualan meningkat.
- Memberikan promo-promo terbaik di waktu-waktu peak season.
- Tidak banyak membeli bahan/persediaan pada saat cuaca-cuaca ekstrim.

Terima kasih!

Ada pertanyaan?