Machine Learning Final Project

PREDIKSI VOLUME PENJUALAN PRODUK

The Team

Rakdim
Data preparation

Daffa Alif
Data Cleaning

Aliya
Data Cleaning

Bisma
EDA

Irfan
Feature Engineering

Rizcy
Create Model

Sakanti
Model Evaluation

Our Case

Seorang manajer pemasaran di perusahaan ritel FMCG ingin meningkatkan penjualan produk, namun kesulitan menentukan faktor-faktor yang paling berpengaruh. Walaupun perusahaan telah memiliki data historis terkait harga, promosi, posisi produk, jumlah pengunjung, dan faktor lainnya, informasi tersebut belum dimanfaatkan secara optimal. Oleh karena itu, dibutuhkan pendekatan machine learning guna memprediksi volume penjualan dan mendukung pengambilan keputusan yang lebih tepat.

Key Components

Strategy to Reach our Goals

Describing Data

Penjelasan cara mendapatkan karakteristik dataset

Exploratory Data Analysis

Penjelasan cara mendapatkan hubungan antar data da;a, dataset

Feature Engineering

Penjelasan strategi mendapatkan fitur beserta cara menghindari masalahnya

Modelling + Evaluation

Penjelasan karakter model + cara mengevaluasinya

Describing Data

- **Noise Data** : Tidak terdapat Noise di data tersebut
- **Tipe Data** :

  -Tipe: int64 float64

  Digunakan untuk data angka yang bisa dihitung dan dianalisis statistiknya.

  Tipe: category

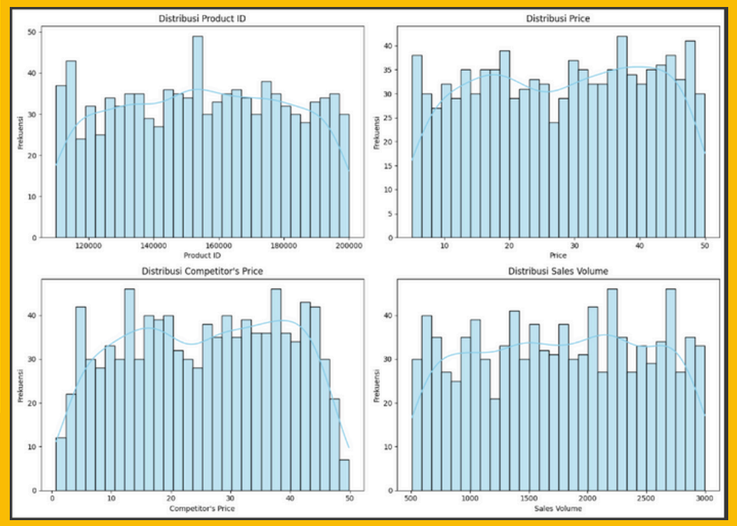
  Data non-numerik yang merepresentasikan label atau kategori, cocok untuk klasifikasi.
- **Volume Data** : 10 Columns, 1000 Entries
- **Variabel Target** : memprediksi volume penjualan dan mendukung pengambilan keputusan yang lebih tepat.

```
Product ID                int64
Product Position       category
Price                   float64
Competitor's Price      float64
Promotion              category
Foot Traffic           category
Consumer Demographics  category
Product Category       category
Seasonal               category
Sales Volume              int64
dtype: object

Missing values per column:
Product ID                0
Product Position          0
Price                     0
Competitor's Price        0
Promotion                 0
Foot Traffic              0
Consumer Demographics     0
Product Category          0
Seasonal                  0
Sales Volume              0
dtype: int64
```

```
df.describe()
```

	Product ID	Price	Competitor's Price	Sales Volume
count	1000.000000	1000.000000	1000.000000	1000.000000
mean	154899.862000	28.020010	25.550110	1769.311000
std	25795.563607	13.067876	13.156466	718.386603
min	110033.000000	5.060000	0.720000	507.000000
25%	133164.500000	16.917500	14.277500	1136.500000
50%	154694.500000	28.680000	26.145000	1791.500000
75%	176954.250000	39.332500	37.125000	2363.750000
max	199976.000000	49.980000	49.850000	2999.000000



Data Cleaning

- Remove Duplicates

  untuk menghapus data yang muncul lebih dari sekali
  agar analisis tidak salah.

- Detect and remove outliers

  untuk menghilangkan nilai yang jauh berbeda dari data lainnya,
  karena outlier bisa menyebabkan model machine learning tidak akurat

- Remove Irrelevant data

  untuk menghapus data yang tidak dibutuhkan atau
  tidak berpengaruh pada tujuan analisis

- Remove Duplicates

```
print("jumlah data sebelum remove duplicates:", df.shape[0])
df = df.drop_duplicates()
print("jumlah data setelah remove duplicates:", df.shape[0])
df
```

```
jumlah data sebelum remove duplicates: 1000
jumlah data setelah remove duplicates: 1000
```

- Detect and remove outliers

```
from scipy import stats
numeric_cols = ["Price", "Competitor\'s Price", "Sales Volume"]
z_scores = np.abs(stats.zscore(df[numeric_cols]))
df = df[(z_scores<3).all(axis=1)]
print("data setelah remove outliers:", df.shape)
df
```

```
data setelah remove outliers: (1000, 10)
```

- Remove Irrelevant data

```
irrelevant_cols = ["Product ID", "Notes"]
df = df.drop(columns=[col for col in irrelevant_cols if col in df.columns])
df
```

Data Cleaning

- Standardize capitalization

  untuk menghindari error saat memproses string sehingga data lebih konsisten dan mudah diolah saat pemodelan.

- Convert Data Type

  untuk mengubah tipe data agar sesuai dengan fungsi atau analisis yang akan dilakukan sehingga memastikan data bisa diolah dengan benar.

- Clear Formatting

  menghapus format yang tidak perlu dari data untuk memastikan struktur data bersih, bebas dari karakter atau spasi yang bisa menyebabkan kesalahan saat diproses, dan agar analisis berjalan lancar.

- Standardize capitalization

```
for col in df.select_dtypes(include="object").columns:
  df[col] = df[col].str.strip().str.title()
df
```

- Convert Data Type

```
df['Promotion'] = df['Promotion'].replace({'Yes': True, 'No': False})
df['Seasonal'] = df['Seasonal'].replace({'Yes': True, 'No': False})

df
```

```
<ipython-input-19-5b1194356982>:1: FutureWarning: Downcasting behavior in `replace` is deprecated and will be removed in a future version. To retain the old
  df['Promotion'] = df['Promotion'].replace({'Yes': True, 'No': False})
<ipython-input-19-5b1194356982>:1: FutureWarning: The behavior of Series.replace (and DataFrame.replace) with CategoricalDtype is deprecated. In a future ver
  df['Promotion'] = df['Promotion'].replace({'Yes': True, 'No': False})
<ipython-input-19-5b1194356982>:2: FutureWarning: Downcasting behavior in `replace` is deprecated and will be removed in a future version. To retain the old
  df['Seasonal'] = df['Seasonal'].replace({'Yes': True, 'No': False})
<ipython-input-19-5b1194356982>:2: FutureWarning: The behavior of Series.replace (and DataFrame.replace) with CategoricalDtype is deprecated. In a future ver
  df['Seasonal'] = df['Seasonal'].replace({'Yes': True, 'No': False})
```

- Clear Formatting

```
[ ]  df.reset_index(drop=True, inplace=True)
```

- return type: Mengembalikan DataFrame baru dengan indeks yang baru.
- drop: Jika diatur ke True, indeks lama dibuang daripada ditambahkan sebagai kolom.
- inplace: Jika diatur ke True, operasi mengubah DataFrame yang ada tanpa mengembalikan yang baru.

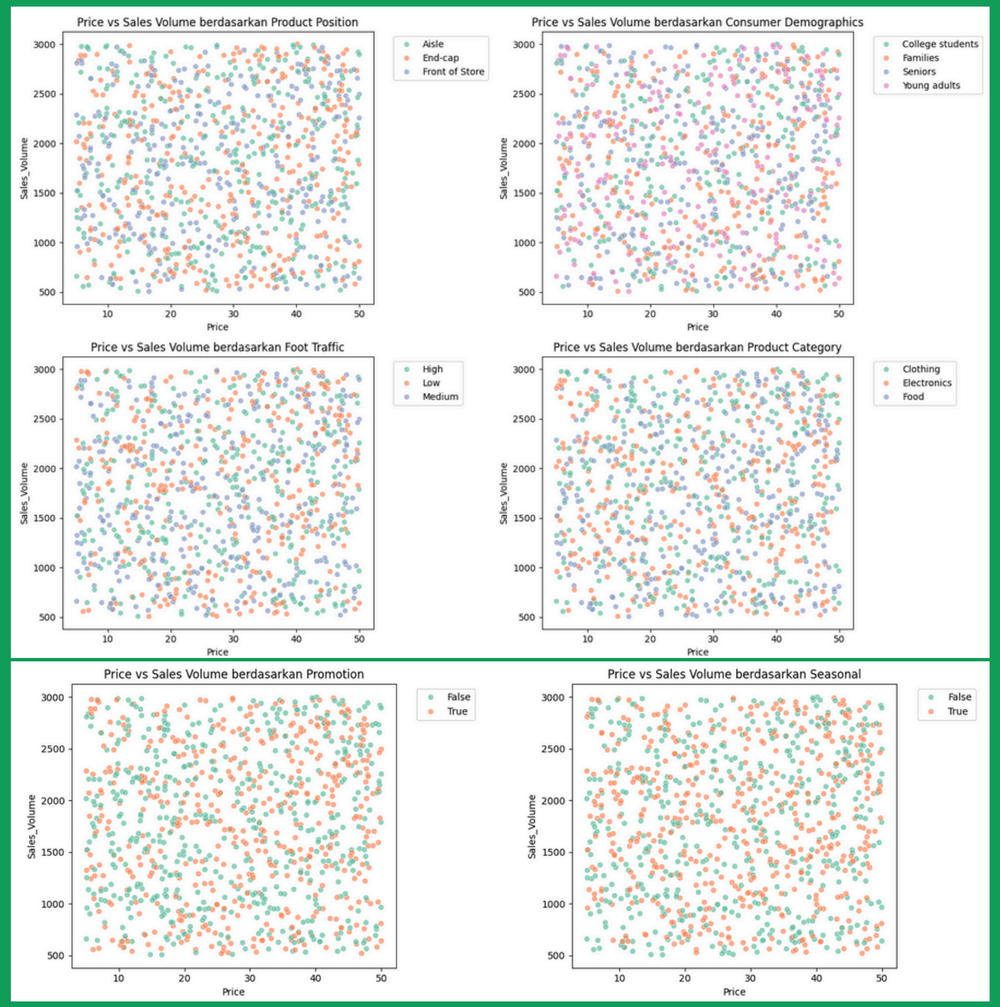
```
[ ]  df.columns = df.columns.str.strip()  # Menghapus spasi di awal/akhir
     df.columns = df.columns.str.replace(" ", "_")  # Mengganti spasi " " dengan garis bawah "_"
```

Clean Column name

EDA

- Pada scatter plot di samping, terlihat bahwa hubungan antara Price dengan Sales Volume sangat acak, sehingga belum ditemukannya korelasi yang kuat untuk menggambarkan data

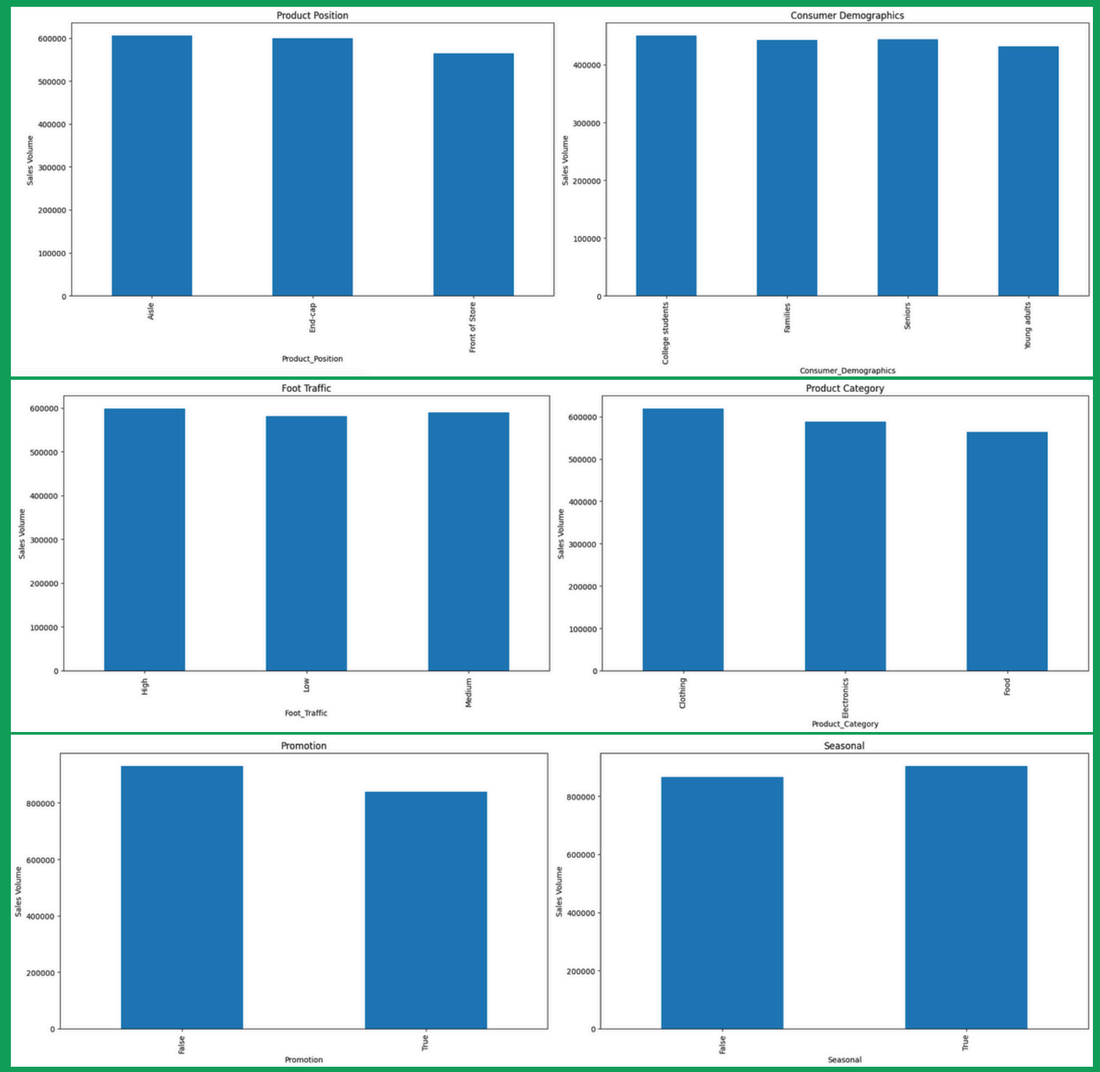
Scatter Plot Price vs Sales Volume dilihat dari berbagai category



EDA

- Dan pada box plot di samping, terlihat bahwa tiap category memiliki total penjualan yang hampir sama, sehingga masih belum diketahuinya category mana yang memengaruhi penjualan produk

Bar Plot Category vs Total Sales Volume

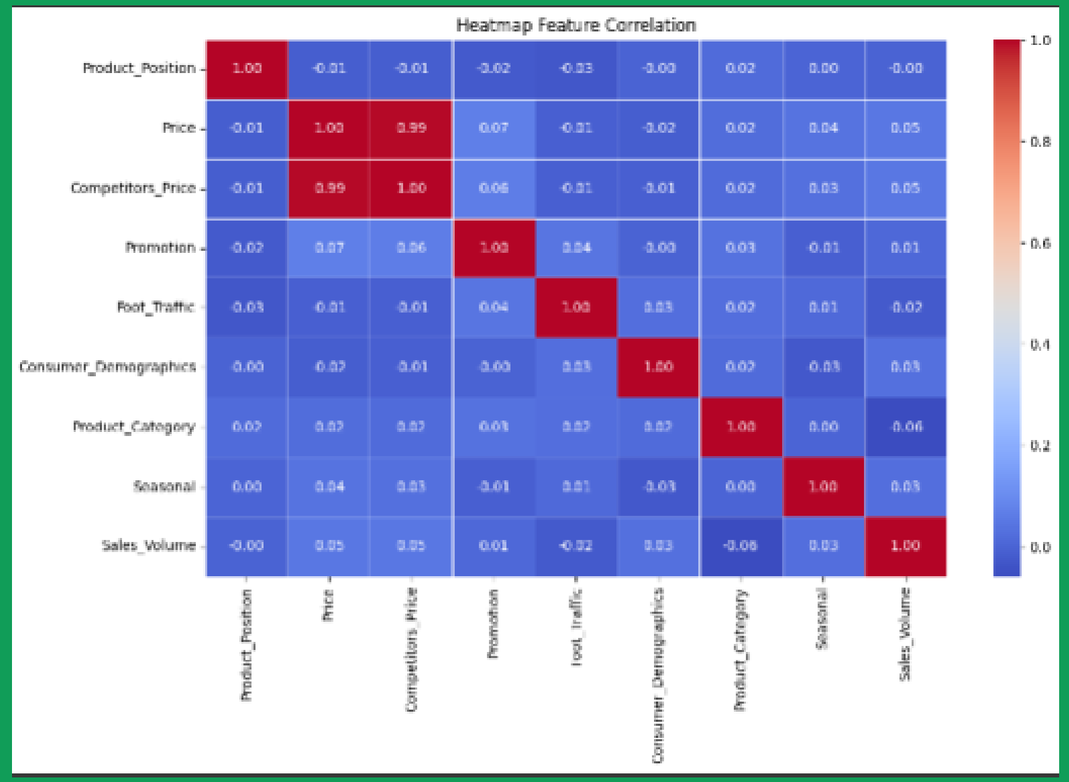


EDA

Berikut urutan korelasi dari terkuat sampe terlemah:

- Harga dengan Promosi
- Harga dengan Volume Penjualan
- Foot Traffic dengan Promosi
- Harga dengan Barang Musiman

Heatmap Korelasi antar fitur



Feature Engineering

Feature Engineering adalah proses menciptakan fitur baru atau mengubah fitur yang sudah ada untuk meningkatkan kinerja model pembelajaran mesin. Proses ini melibatkan pemilihan informasi yang relevan dari data mentah dan mengubahnya menjadi format yang mudah dipahami oleh model. Tujuannya adalah untuk meningkatkan akurasi model dengan menyediakan informasi yang lebih bermakna dan relevan. Berikut beberapa feature engineering yang digunakan:

📊Encoding

Converting categories (like text) into structured numerical formats (one-hot, label encoding).

⚙️Extraction

Creating new features from existing ones to provide more relevant information to the machine learning model.

Feature Engineering

📊Encoding (Label Encoding)

Feature Encoding ada berbagai macam jenis, pada dataset kali ini menggunakan feature encoding dengan jenis label encoding disebabkan oleh beberapa kondisi pada dataset sehingga penggunaan label encoding lebih sesuai.

Product_Position	Promotion	Foot_Traffic	Consumer_Demographics	Product_Category	Seasonal
0	0	1	1	0	0
0	0	0	2	0	0
1	1	1	3	1	1
0	1	0	1	0	1
1	0	1	0	0	1
1	0	1	2	0	0
2	1	2	0	0	1
0	0	0	0	0	0
0	1	2	1	1	1
0	0	2	1	1	1
0	1	2	0	2	1

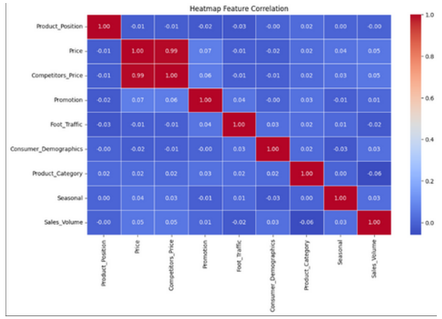
Feature Engineering

⚙Extraction

Feature Extraction dapat menggabungkan beberapa fitur menjadi satu atau membuat fitur baru lainnya. Dapat dilihat pada gambar bahwa terdapat penambahan beberapa fitur baru dengan menggabungkan fitur sebelumnya.

Is_Promo_Seasonal	Relative_Price	Is_Overprice	Relative_Price_By_Category	Is_Overprice_By_Category	Percentage_Diff
0	-10.95001	0	-10.680000	0	5.631188
0	-10.61001	0	-10.340000	0	32.597106
1	15.13999	1	15.308899	1	12.483711
1	14.23999	1	14.510000	1	8.414572
0	19.91999	1	20.190000	1	5.154639
...	...	...	...	...	...
0	-16.46001	0	-16.291101	0	34.262485
0	19.69999	1	19.245951	1	2.009406
0	-6.72001	0	-6.551101	0	13.842865
0	-5.20001	0	-5.031101	0	19.289075
0	-21.64001	0	-21.471101	0	21.523810

Model Building



Challenges - x%

Saat membangun model, tantangan utama yang dihadapi adalah nilai R^2 score yang negatif akibat rendahnya korelasi antar fitur dalam data.

```
## Training model
model_lr = LinearRegression()
model_lr.fit(X_train_scaled, y_train)
```

```
model_rf = RandomForestRegressor(random_state=42)
model_rf.fit(X_train, y_train)
```

```
model_dt = DecisionTreeRegressor()
model_dt.fit(X_train, y_train)
```

Strategy - x%

Untuk mengatasi masalah tersebut, kami menambahkan fitur baru dan mencoba beberapa model machine learning untuk menemukan pendekatan yang lebih sesuai.

```
R² Score: -0.02
Mean Absolute Error (MAE): 601.46
Mean Squared Error (MSE): 496686.70
Root Mean Squared Error (RMS): 704.76
```

Linear Regression

```
R² Score: -1.19
Mean Absolute Error (MAE): 859.09
Mean Squared Error (MSE): 1060206.97
Root Mean Squared Error (RMS): 1029.66
```

Decision Tree

```
R² Score: -0.18
Mean Absolute Error (MAE): 640.95
Mean Squared Error (MSE): 572251.31
Root Mean Squared Error (RMS): 756.47
```

Random Forest

Solution - x%

Kami melakukan improvisasi dalam pemilihan fitur serta evaluasi model untuk meningkatkan performa. Meskipun R^2 score masih negatif, proses ini memberikan pemahaman yang lebih baik terhadap data dan tantangan pemodelan.

Model Evaluation

Model Evaluation Linier Regression

```
R² Score: -0.0237
Mean Absolute Error (MAE): 601.46
Mean Squared Error (MSE): 496686.70
Root Mean Squared Error (RMS): 704.76
```

Model Evaluation Random Forest Regression

```
R² Score: -0.1795
Mean Absolute Error (MAE): 640.95
Mean Squared Error (MSE): 572251.31
Root Mean Squared Error (RMS): 756.47
```

Model Evaluation Decision Tree Regression

```
R² Score: -1.0045
Mean Absolute Error (MAE): 808.95
Mean Squared Error (MSE): 972546.56
Root Mean Squared Error (RMS): 986.18
```

Linear Regression menghasilkan performa terbaik di antara ketiganya, dengan nilai error lebih kecil dibandingkan dengan model lainnya dan memiliki R2 score tertinggi.

Model Conclusion

Linear Regression menjadi **model terbaik** dari tiga model yang diuji, meskipun performanya tetap kurang baik secara keseluruhan dengan nilai R^2 score bernilai negatif.

Kemungkinan besar, **fitur yang digunakan belum cukup kuat menjelaskan target** (**penjualan**), Hasil eksplorasi dan heatmap sebelumnya menunjukkan bahwa **mayoritas fitur memiliki korelasi rendah** terhadap Sales_Volume.

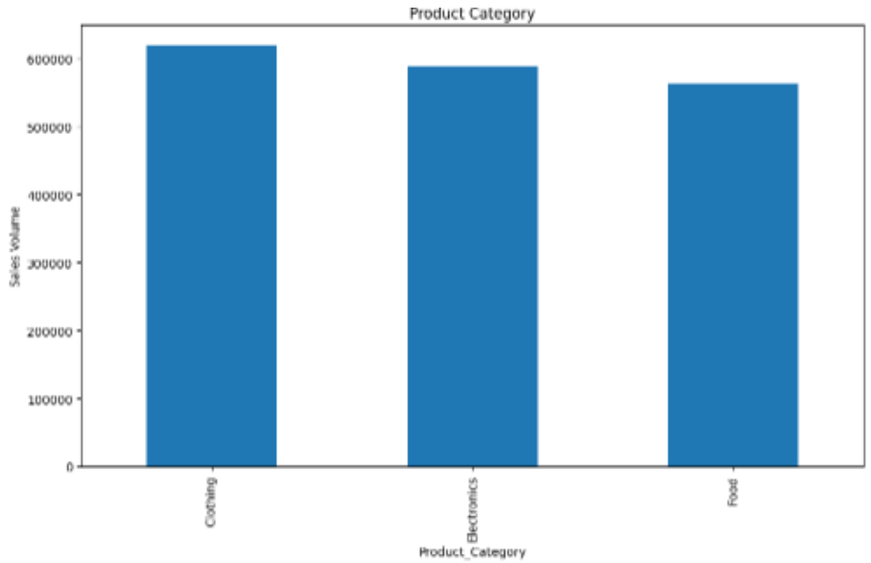
Sehingga Model tidak mampu menjelaskan variansi Sales_Volume secara akurat dari fitur yang diberikan. Model kesulitan menemukan pola yang relevan dan hanya menebak secara acak di sekitar rata-rata.

Model Usability

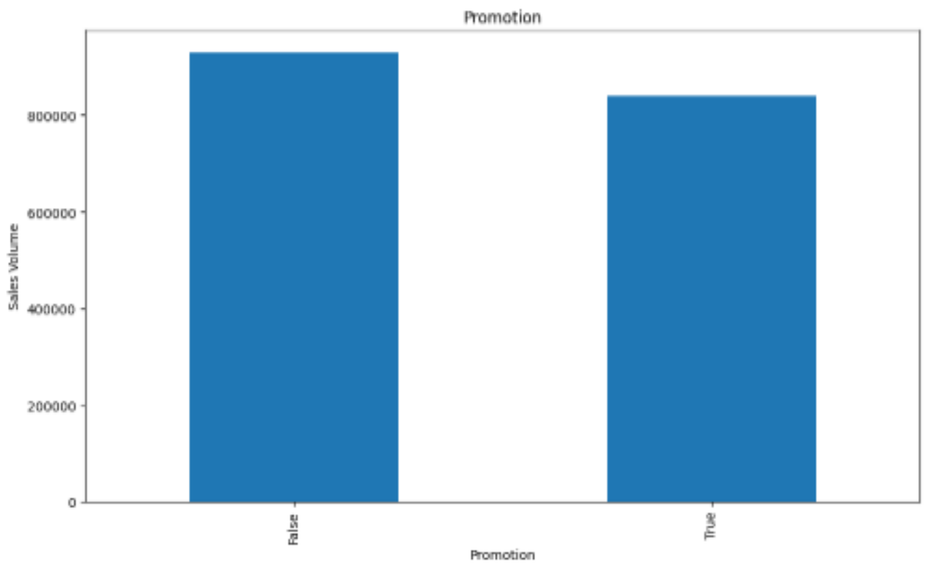
Meskipun Model Linier Regression performanya paling baik, namun R^2 score masih negatif sehingga model belum layak dipakai untuk prediksi akurat.

Tidak ada model yang cukup baik untuk langsung digunakan dalam prediksi nyata atau pengambilan keputusan pada dataset Prediksi Volume Penjualan Produk ini.

Data Analysis

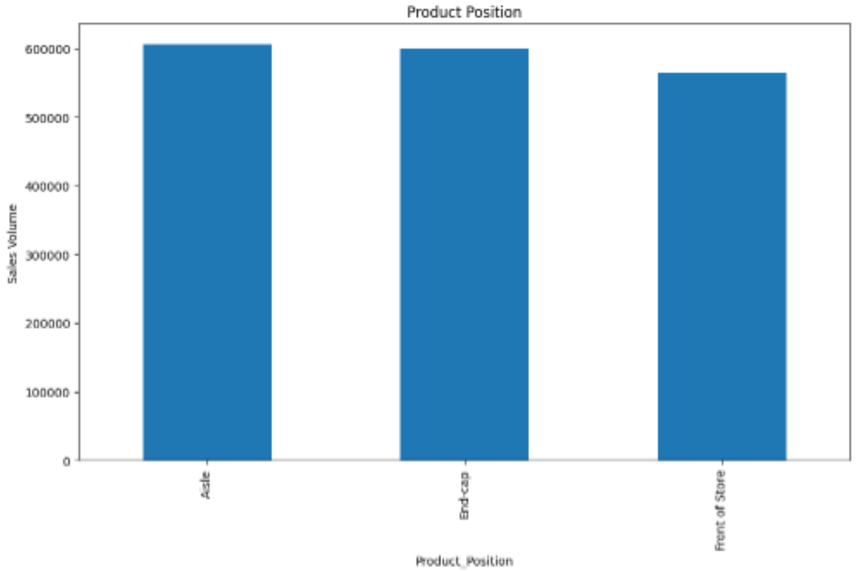


Product Category terjual paling banyak ialah baju (Clothing)

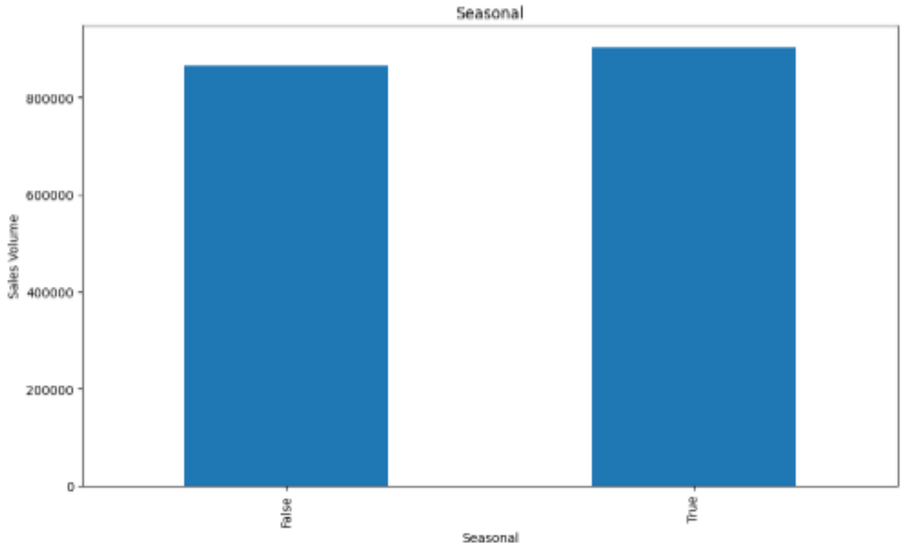


Terdapat lebih banyak produk yang tidak dipromosi

Data Analysis

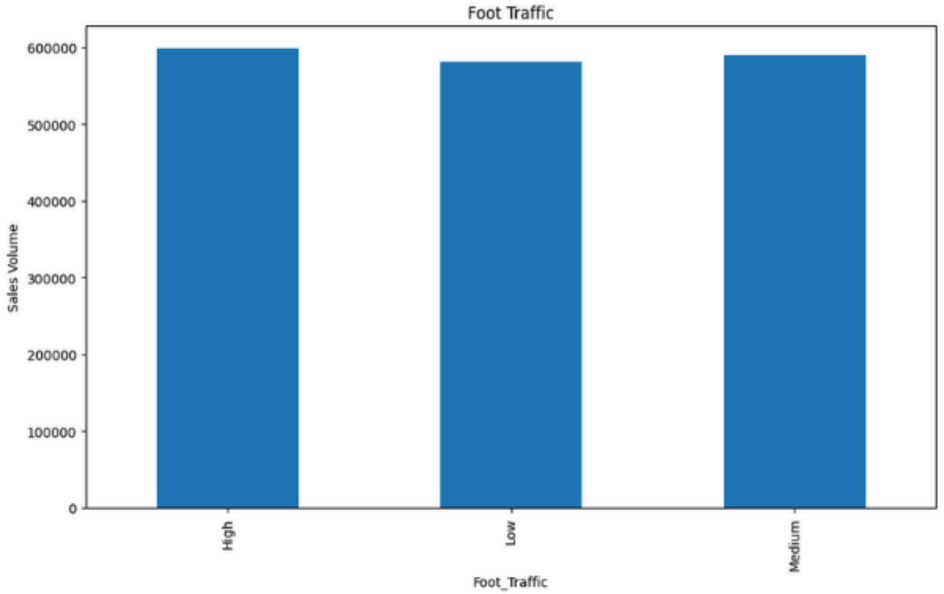


Posisi barang tidak terlalu memengaruhi penjualan, namun produk yang ditaruh di Aisle memiliki penjualan lebih banyak

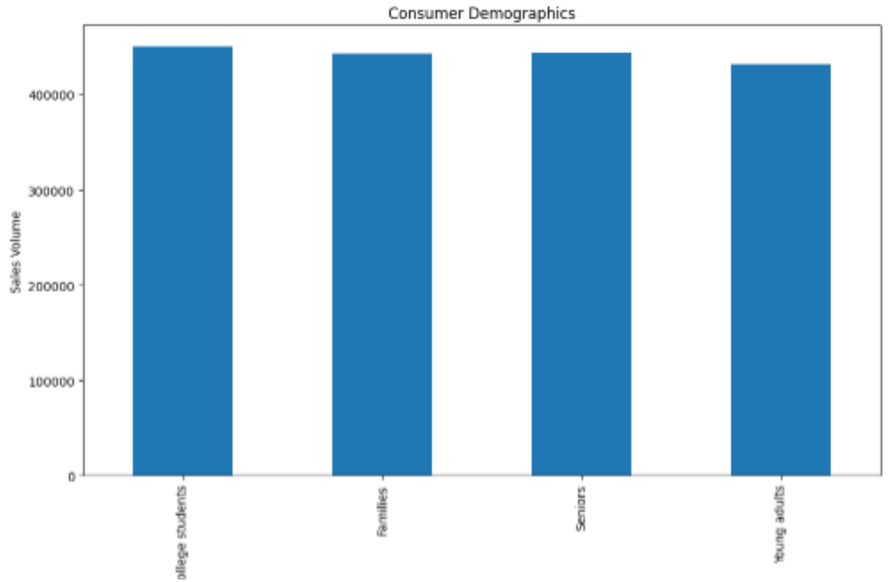


Terdapat lebih banyak penjualan pada produk musiman

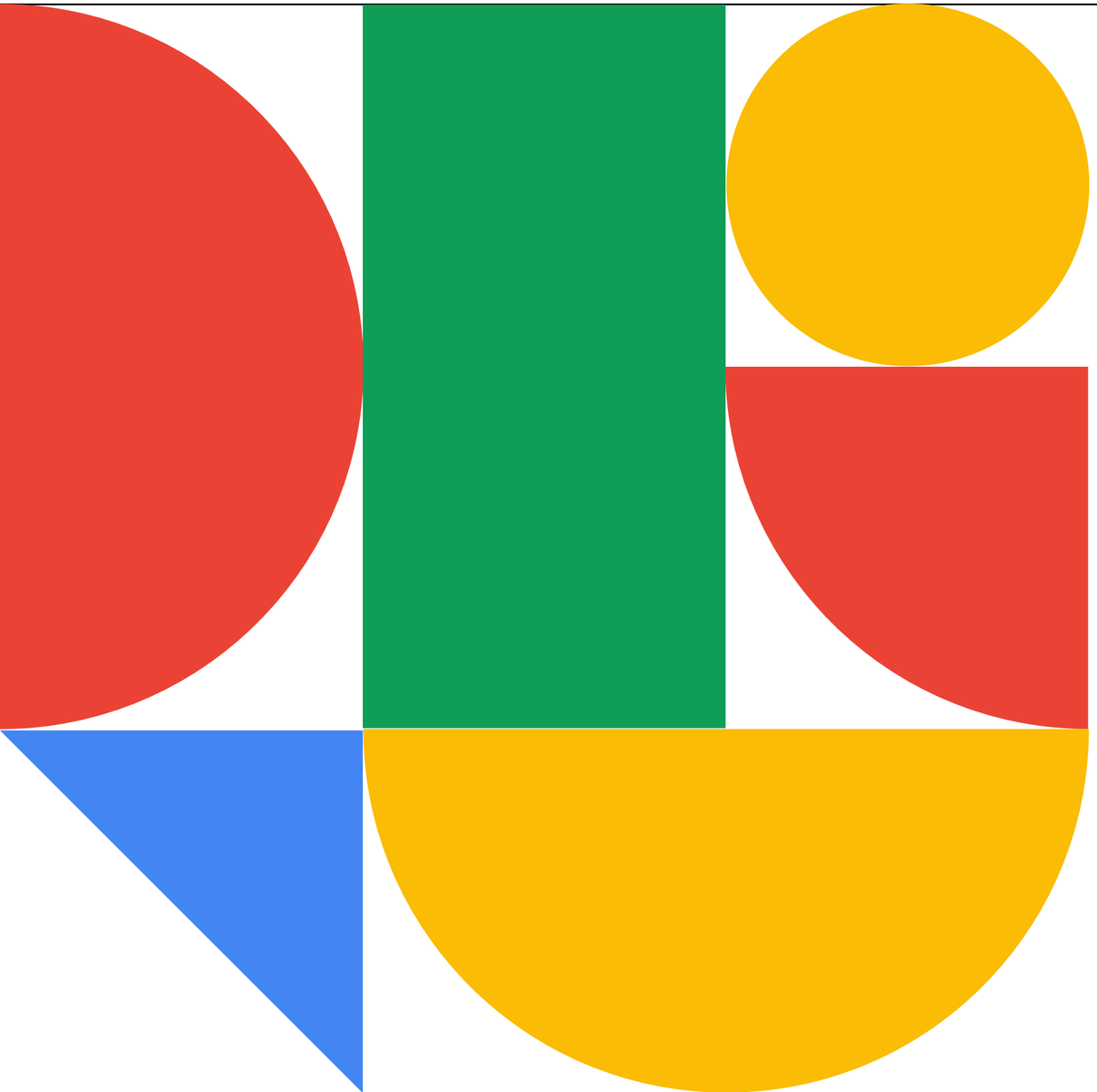
Data Analysis



toko atau lokasi yang sering dilalui pejalan kaki (High Foot Traffic) cenderung memiliki volume penjualan lebih tinggi



Pembeli terbanyak dari kalangan college student



Questions?
Reactions?

Feel free to get in touch with us.