Analisa dan Prediksi Cost Pada Food Mart Menggunakan Model Algoritma Random Forest Regression

Group 1
- Ray Maulida Muhammad
- Yosua Wenas
- Maira Mayasari
- Zulfatin Nafisah

Bussiness Understanding

Convenient Food Mart (CFM) adalah jaringan toko serba ada di Amerika Serikat, yang berdiri pada tahun 1958 di Chicago, Illinois. Kantor pusat perusahaan swasta berada di Mentor, Ohio, dan sekitar 325 toko berlokasi di AS untuk saat ini. CFM ini beroperasi pada sistem waralaba.

CFM adalah jaringan toko swalayan terbesar ketiga di negara itu tahun 1988. Namun, Bursa NASDAQ menjatuhkan CFM menjadi perusahaan gagal memenuhi persyaratan pelaporan keuangan.

Carden & Cherry mengiklankan CFM dengan karakter Ernest pada 1980-an. CFM menjual berbagai produk bahan makanan, minuman ringan, hingga makanan siap saji.

Tujuan

Memprediksi biaya akuisisi pelanggan atau yang dikenal dengan CAC(*Customer acquisition cost*). Hasil prediksi berupa *output* yaitu biaya akuisisi *customer* yang harus dikeluarkan oleh perusahaan berdasarkan nilai *input*nya.

Manfaat

- Memberikan rekomendasi algoritma yang baik dalam prediksi dari *dataset*.

- Mengetahui biaya akuisisi *customer* yang harus dikeluarkan oleh perusahaan.

Data Understanding

Menggunakan dataset media prediction and its cost. Data berisi customer yang melakukan pembelian di setiap produk dari CFM. Data tersebut mempunyai 40 kolom dan 60249 data.

NO		
1	food_category	Jenis Makanan
2	food_department	food_department termasuk jenis makanan
3	food_family	food_family family dari makanan
4	store_sales(in millions)	store_sales(dalam jutaan dolar)
5	store_cost(in millions)	biaya atau pengeluaran toko (dalam jutaan dolar)
6	unit_sales(in millions)	penjualan unit (dalam jutaan) di toko Kuantitas
7	promotion_name	Nama promosi yang dilakukan di media

Data Understanding

8	sales_country	Negara penjualan
9	marital_status	Status pernikahan pelanggan
10	gender	gender dari pelanggan
11	total_children	Total anak dirumah
12	education	Tingkat pendidikan pelanggan
13	member_card	Kartu anggota tersedia untuk pelanggan
14	occupation	Pekerjaan Pelanggan
15	houseowner	Pelanggan pemilik rumah atau bukan
16	avg_cars_at home(approx)	Rata – rata mobil dirumah (perkiraan)
17	avg. yearly_income	Rentang pendapatan tahunan pelanggan
18	num_children_at_home	Jumlah anak di rumah, diisi detail oleh pelanggan

Data Understanding

19	avg_cars_at home(approx)	Rata – rata mobil dirumah (perkiraan)
20	brand_name	Nama Merek Produk
21	SRP	Rekomendasi harga eceran
22	gross_weight	Berat kotor setiap item
23	net_weight	Berat bersih setiap item
24	recyclable_package	Makan kemasan daur ulang
25	low_fat	Makanan rendah lemak
26	units_per_case	Barang yang tersedia di rak toko
27	store_type	Tipe toko
28	store_city	Toko yang tersedia di kota
29	store_state	Toko yang hadir di negara

Data Understanding

30	store_sqft	Area Toko Tersedia dalam SQFT
31	grocery_sqft	Area Grocery Tersedia dalam SQFT
32	frozen_sqft	Area Frozen food tersedia dalam SQFT
33	meat_sqft	Area Daging Tersedia dalam SQFT
34	coffee_bar	Coffee Bar Tersedia di Toko
35	video_store	Toko Video/Toko Permainan tersedia
36	salad_bar	Salad Bar Tersedia di Toko
37	prepared_food	Makanan yang disiapkan Tersedia di Toko
38	florist	Rak Bunga Tersedia di Toko
39	media_type	Jenis Media Sumber Media Tersedia disini
40	cost	Biaya Untuk Memperoleh Pelanggan dalam Dollar

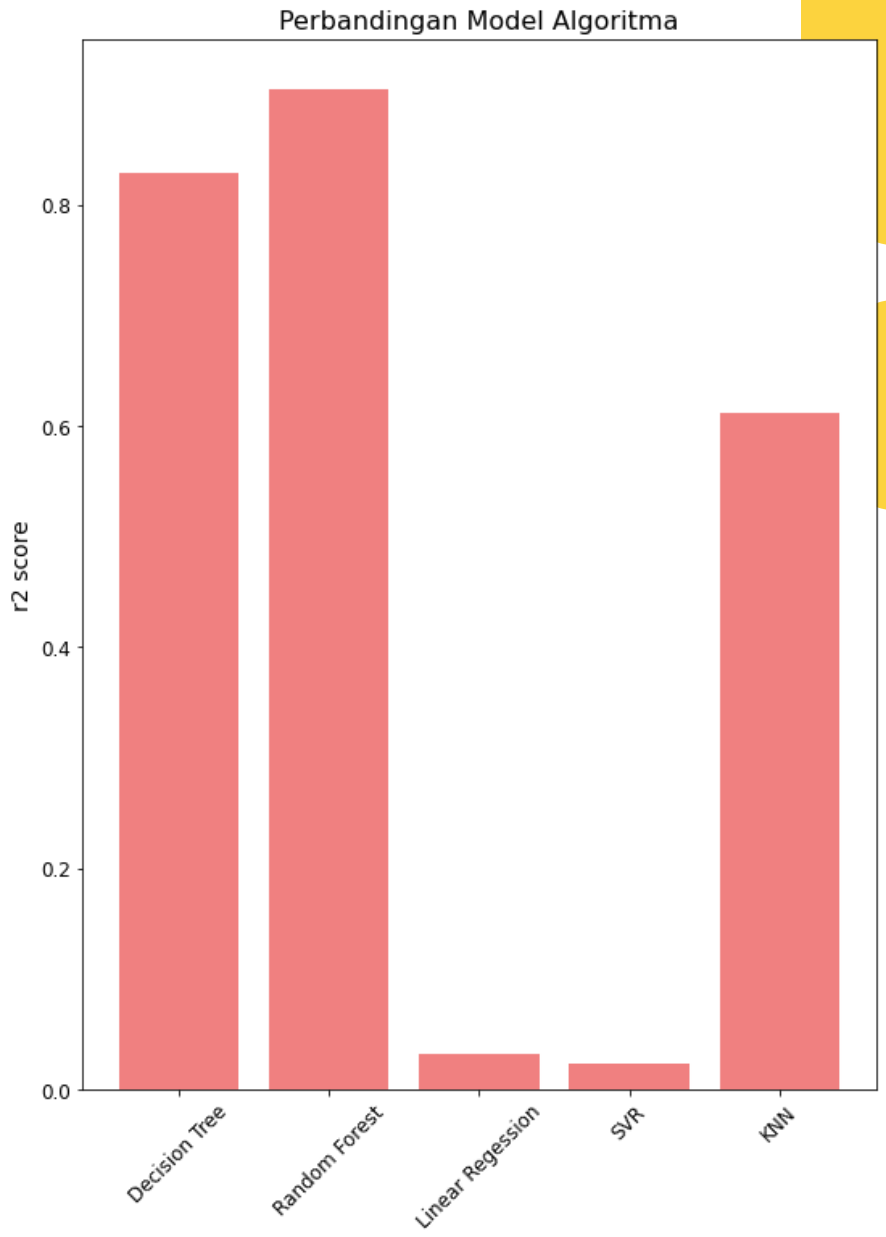
Data Preparation

Data yang digunakan tidak mencakup semua *columns* yang ada pada *dataset*. Beberapa *variable* yang tidak digunakan akan di dropping.
Sehingga data yang digunakan meliputi *variable*:

1. 'promotion_name'
2. 'sales_country'
3. 'marital_status',
4. 'education'
5. 'occupation'
6. 'avg_cars_at home(approx)'
7. 'avg_cars_at home(approx).1'
8. 'SRP'
9. 'gross_weight'
10. 'net_weight'
11. 'low_fat'
12. 'store_type'
13. 'store_city'
14. 'store_state'
15. 'cost'

Modeling

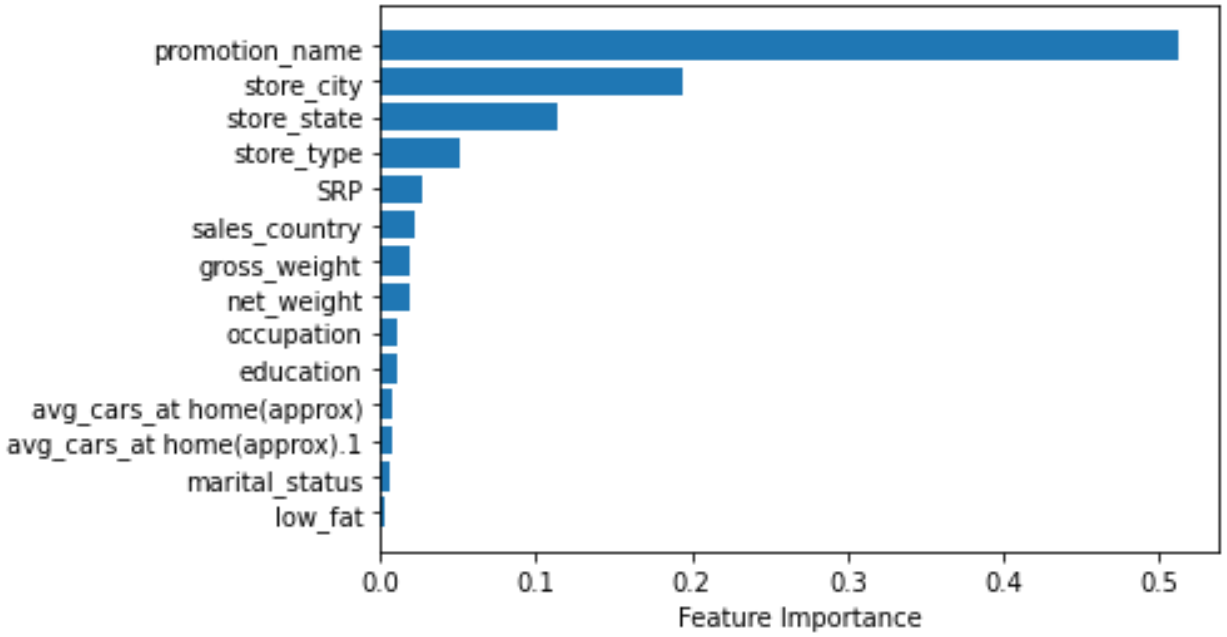
Algoritma yang digunakan untuk prediksi project ini adalah *Decission Tree Regression, Random Forest Regression, Linear Regression, Support Vector Regression (SVR), KNearestNeighbors Regression.* Kemudian didapatkan kesimpulan bahwa Algoritma *Random Forest Regression* memiliki performa yang baik dalam mengatasi prediksi dari dataset yang digunakan.



Modeling

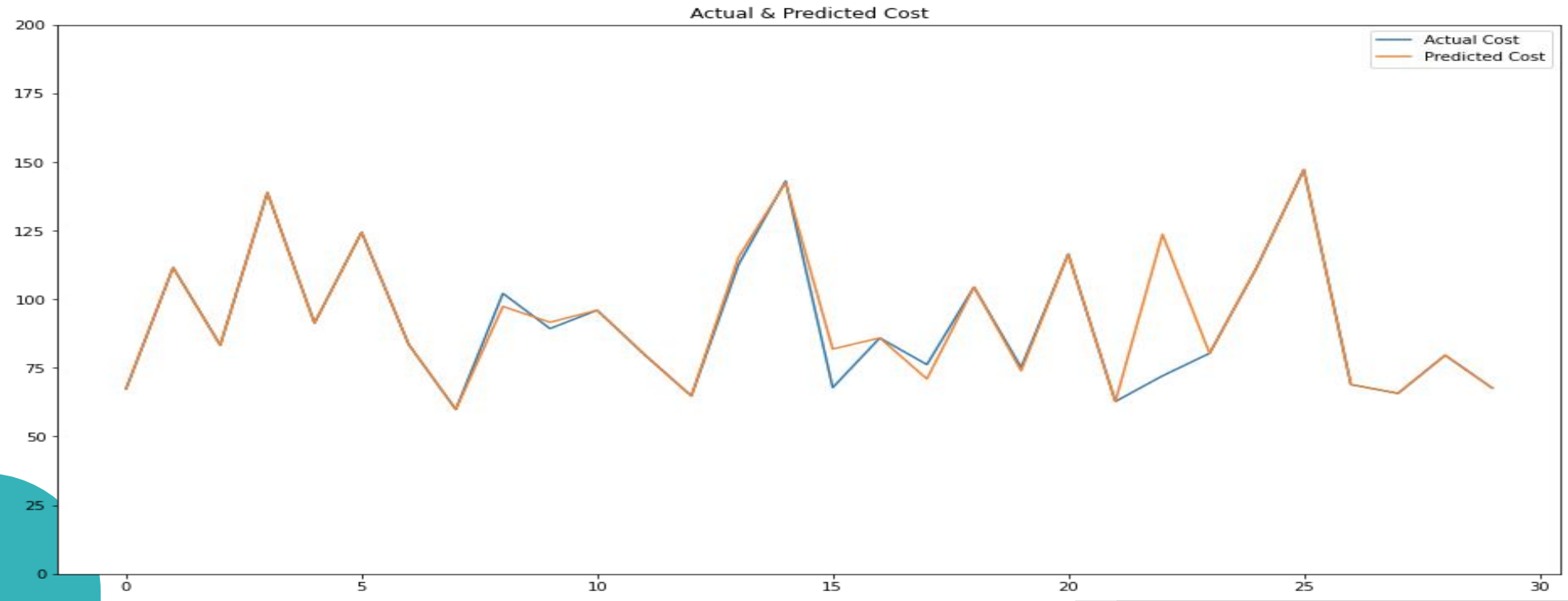
Menentukan *feature importance* untuk menentukan variabel mana yang memiliki nilai terpenting. Dari variabel yang diinputkan variabel promotion_name yang paling penting dengan skor 0,5 dan store_city dengan skor 0,2, serta variabel store_state dengan skor 0,11.

```
Feature: 0, Score: 0.51292
Feature: 1, Score: 0.02268
Feature: 2, Score: 0.00516
Feature: 3, Score: 0.01027
Feature: 4, Score: 0.01057
Feature: 5, Score: 0.00694
Feature: 6, Score: 0.00692
Feature: 7, Score: 0.02627
Feature: 8, Score: 0.01919
Feature: 9, Score: 0.01855
Feature: 10, Score: 0.00305
Feature: 11, Score: 0.05071
Feature: 12, Score: 0.19370
Feature: 13, Score: 0.11308
```



Modeling

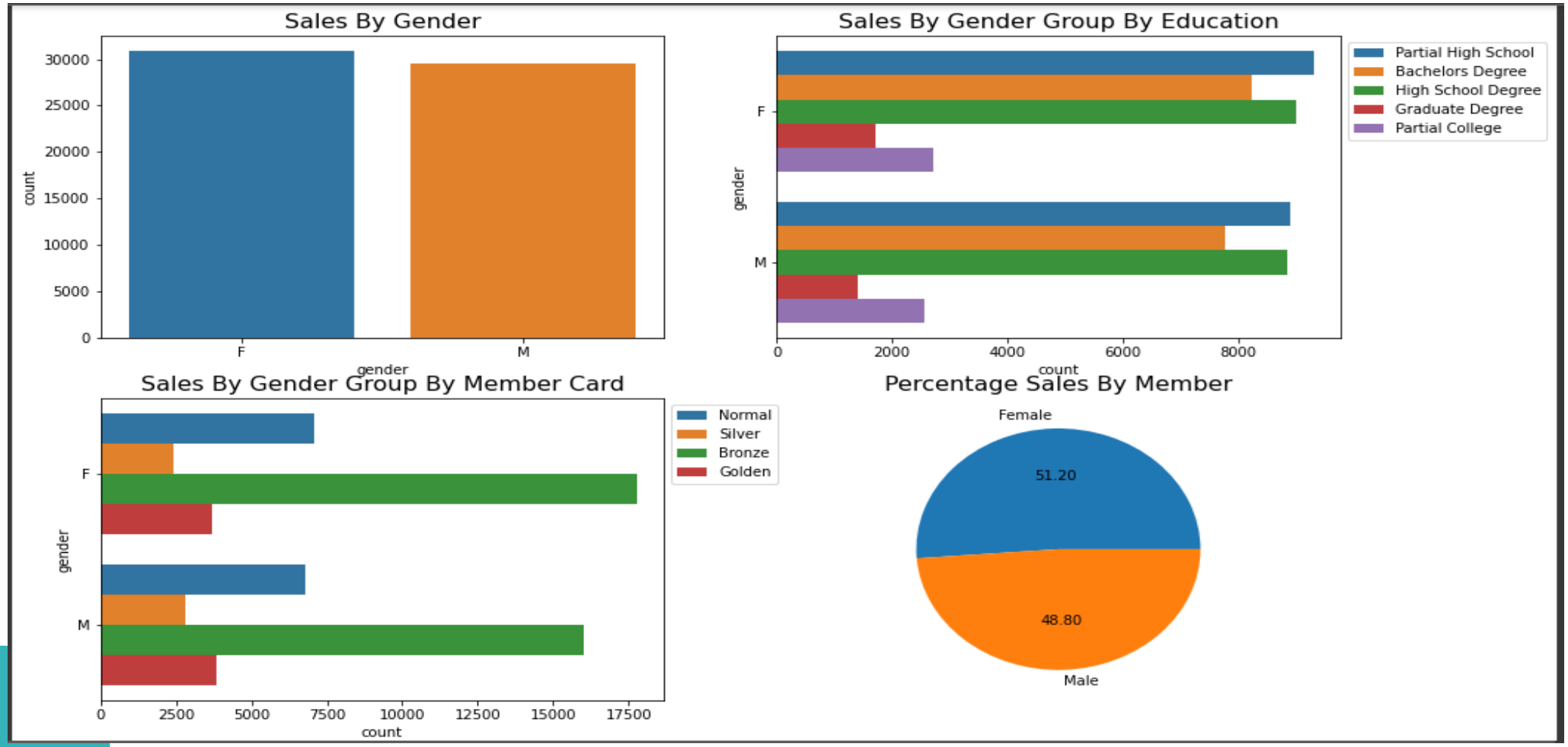
Pada hasil predisksi bisa kita lihat bahwa nilai aktual dan nilai prediksi mempunyai akurasi atau presisi yang kuat satu sama lainnya hal ini membuktikan bahwa model algoritma ini cocok pada dataset yang digunakan.



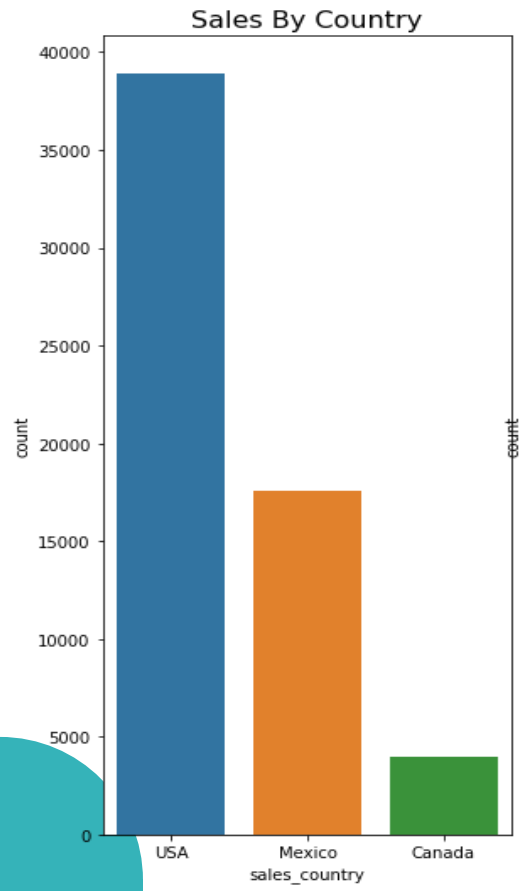
Visualisasi

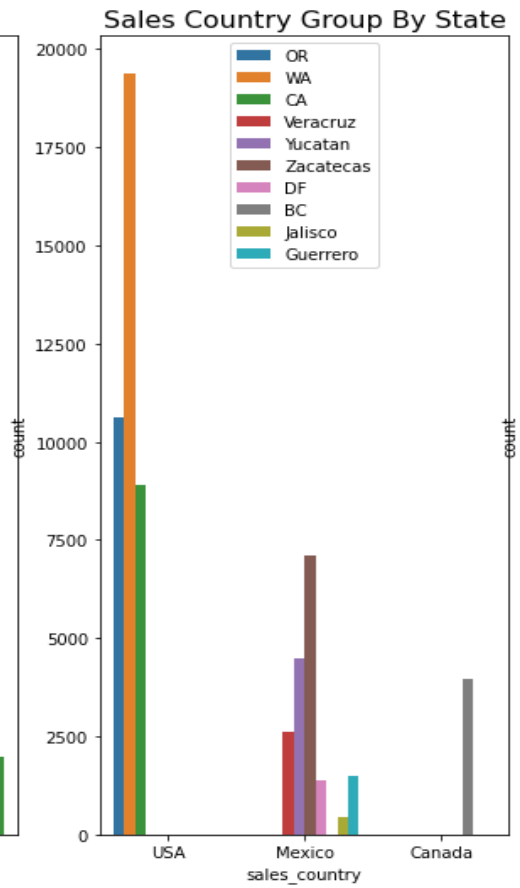
Visualisasi diambil berdasarkan gender, sales_country, avarage_sales_by_income, top_5_food_category. Kemudian, visualisasi juga diambil untuk melihat persebaran antara store_sales dan store_cost.

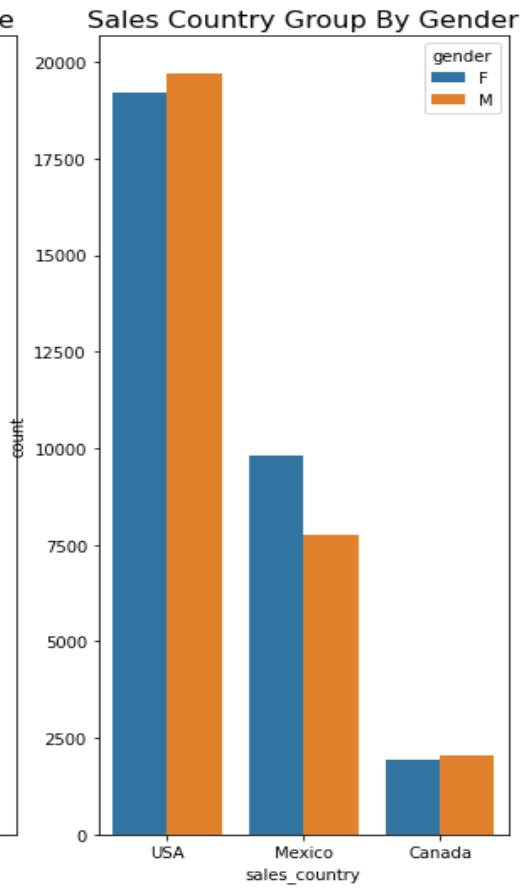
Visualisasi Berdasarkan Gender

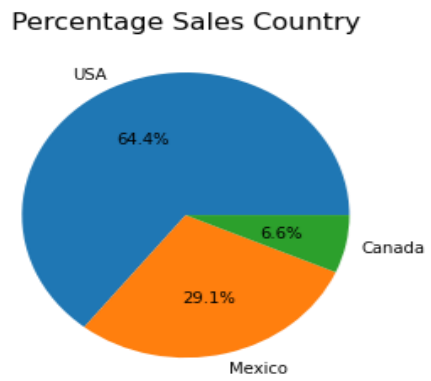


Visualisasi Berdasarkan Sales Country

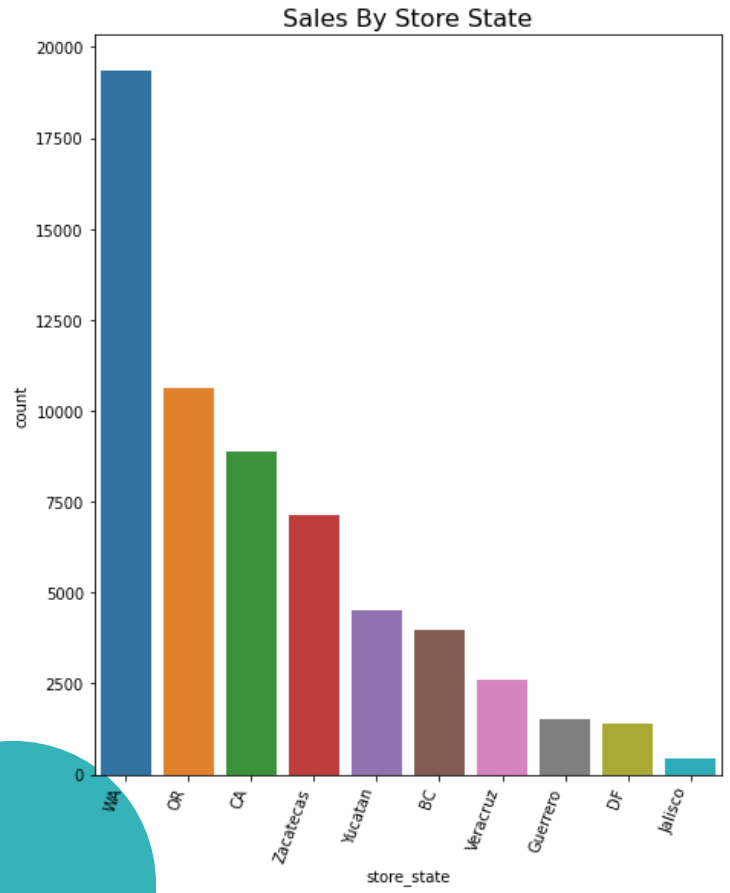


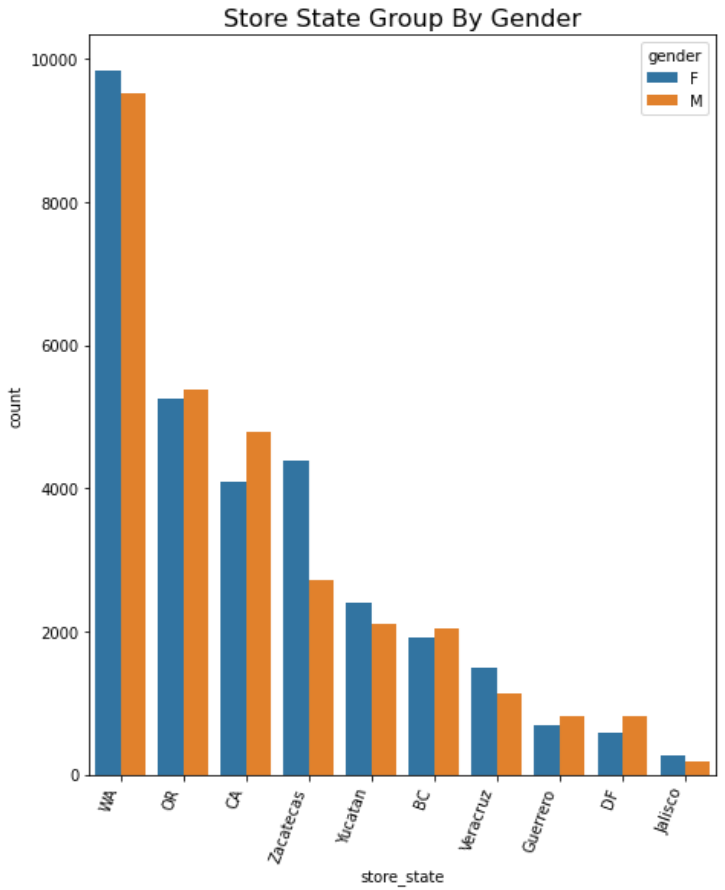


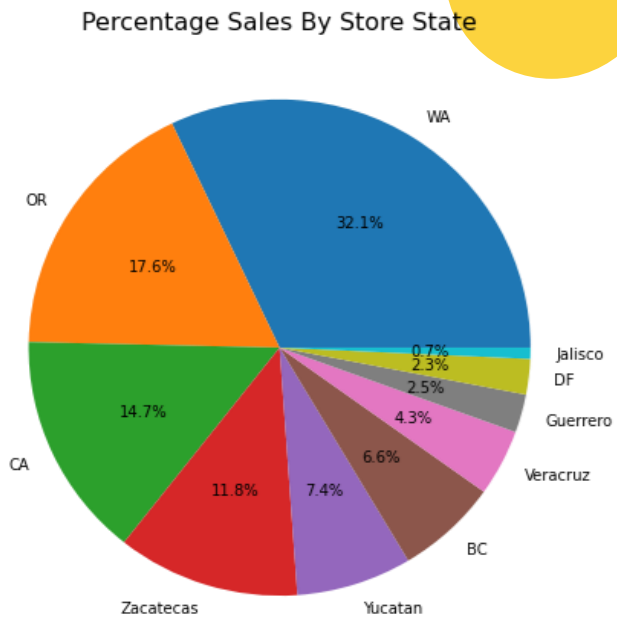




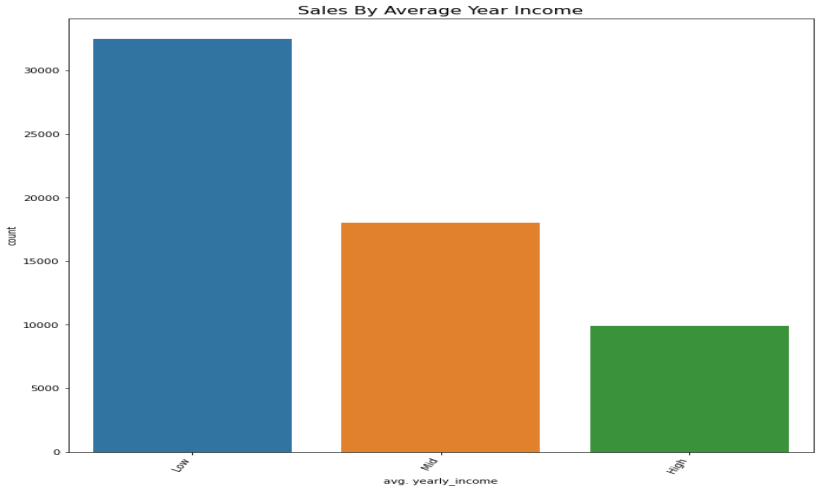
Visualisasi Berdasarkan Store State

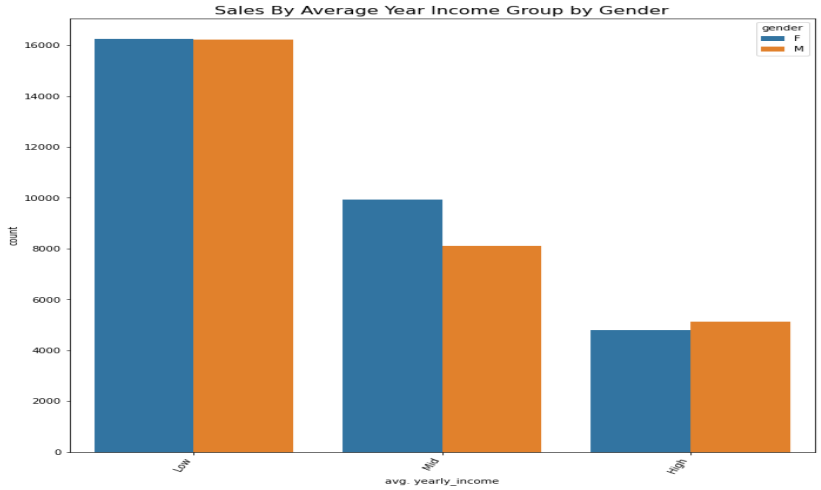


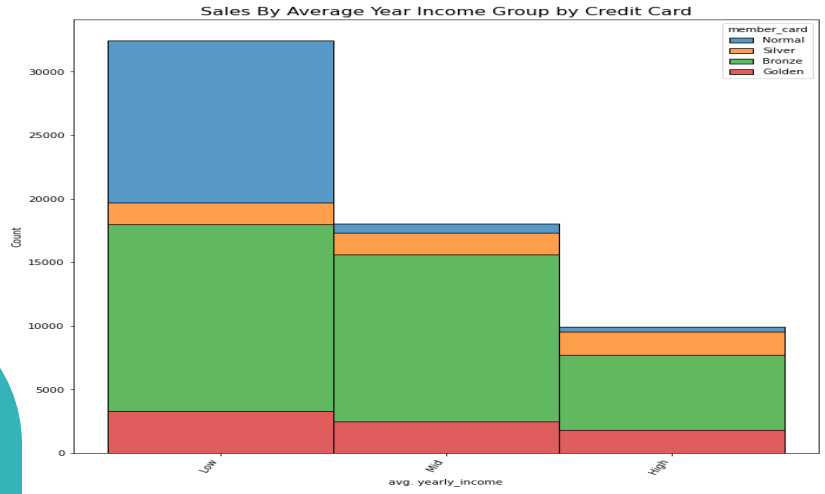


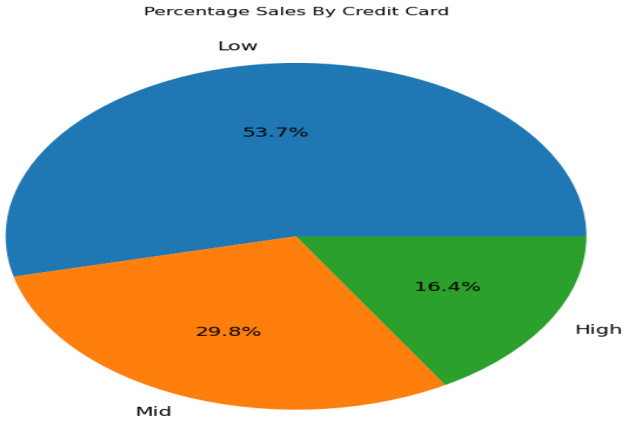


Visualisasi Berdasarkan Average Year Income



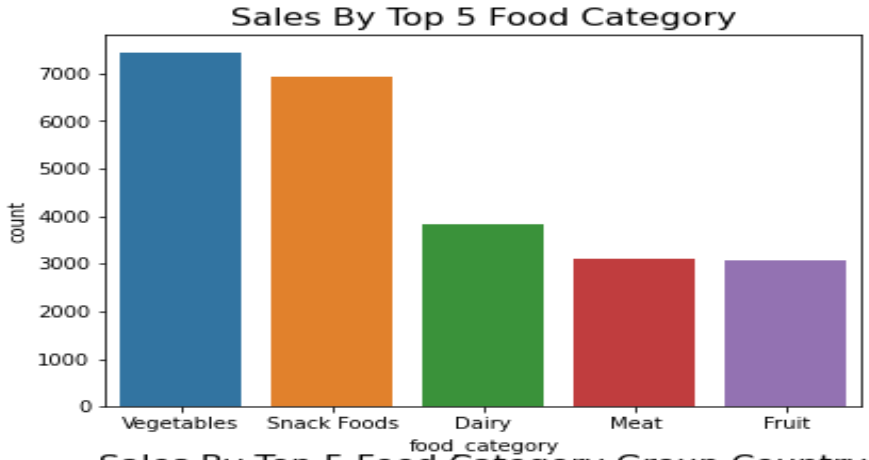


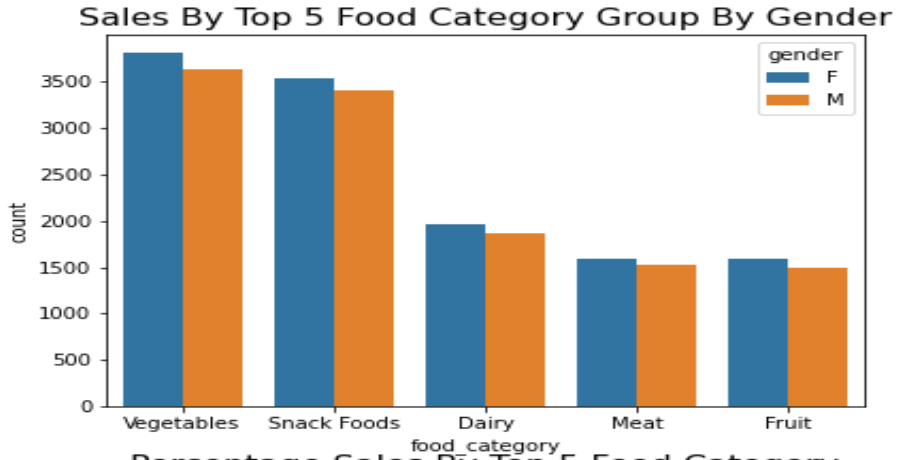


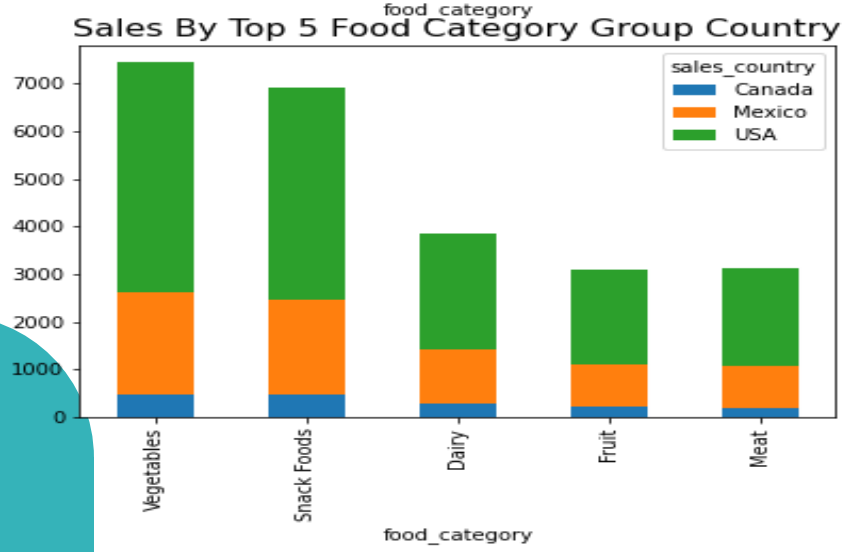


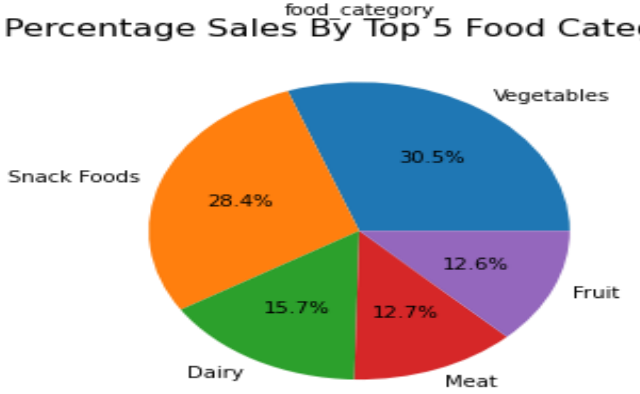
SIB Data Analytics

Visualisasi Berdasarkan Top 5 Food Category

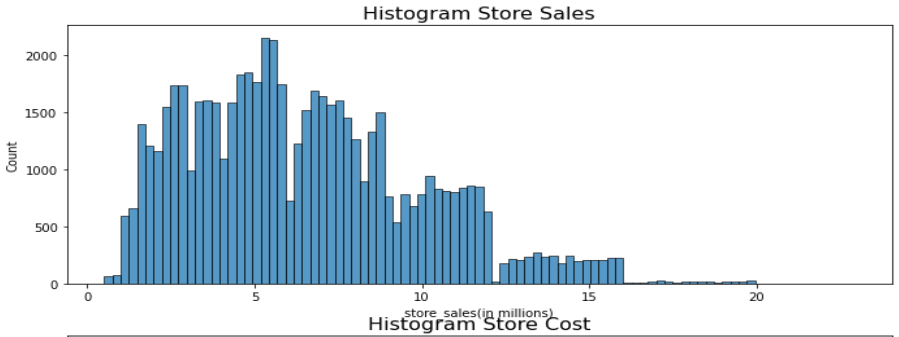


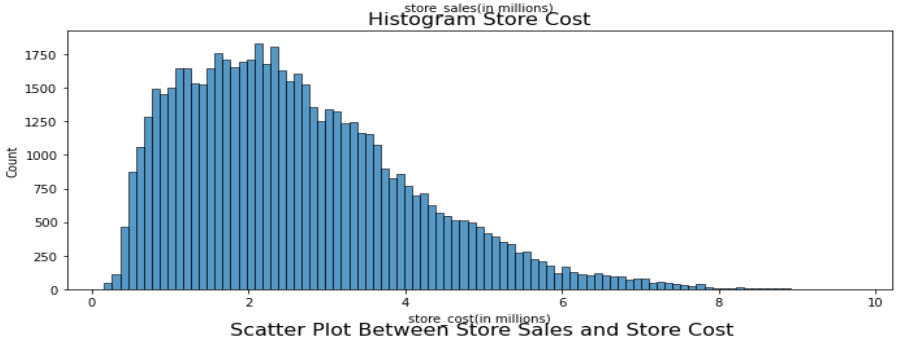


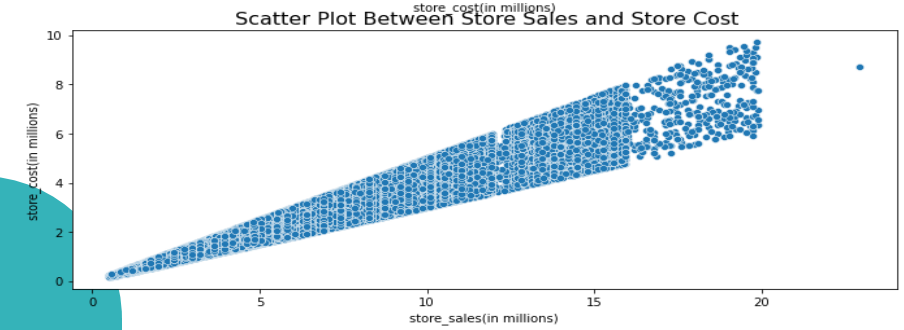


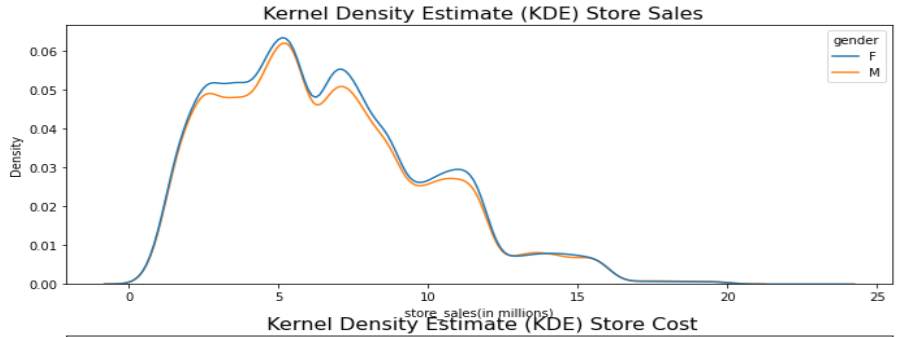


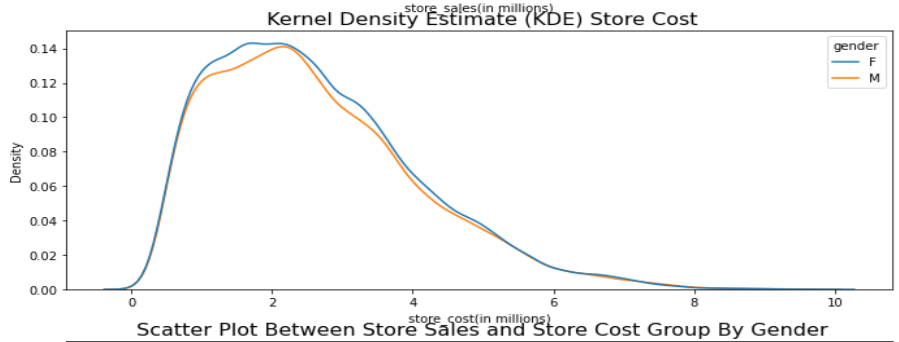
Visualisasi Berdasarkan Store Sales dan Store Cost

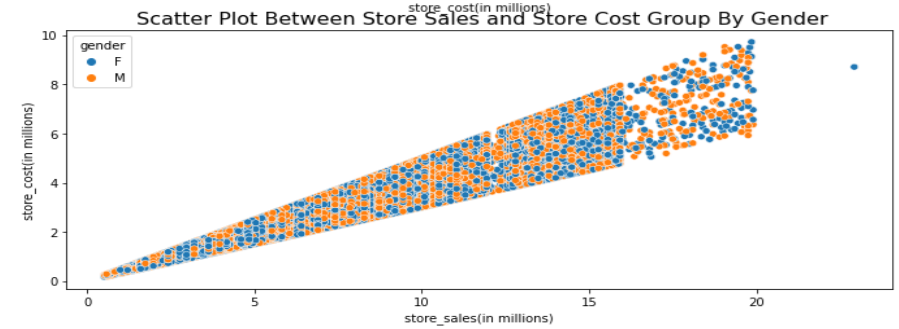




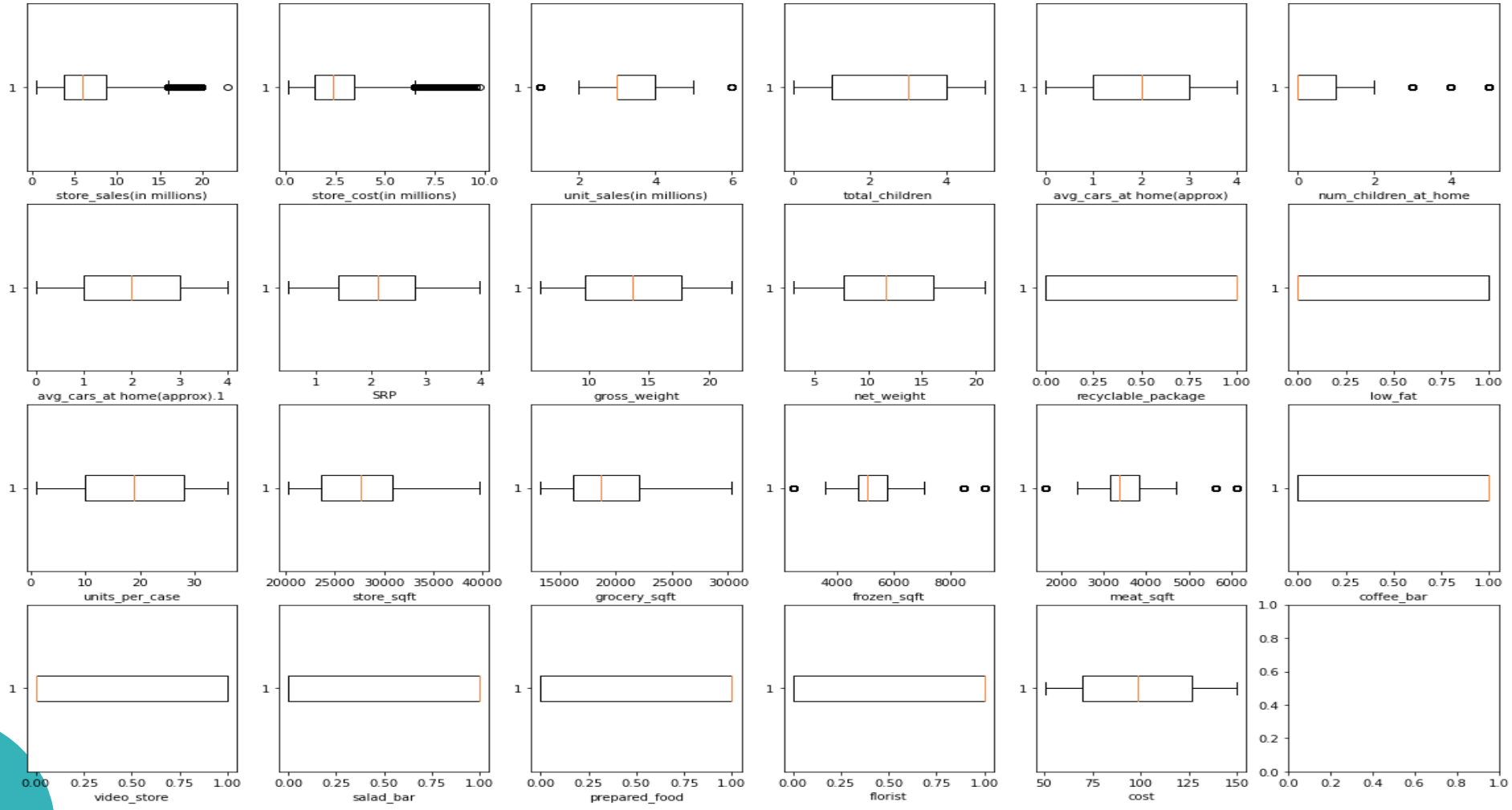








Visualisasi Boxplot



Kesimpulan

1. Dari variabel - variabel yang diambil, **variabel promotion_name** yang memiliki *feature importance* dengan **nilai 0,5** dan **variabel store_city** dengan **skor 0,2**. Dari kedua variabel tersebut promotion_name dapat dilakukan dengan berbagai media dan cara sehingga *cost* yang dikeluarkan pun juga besar. Begitu juga dengan store_city yang juga berpengaruh terhadap *cost*, dimana *store* yang berada pada kota yang maju dan lebih modern membuat *cost* juga lebih tinggi. Lalu, juga ada **variabel store_state** dengan **skor 0,11** dimana variabel ini juga memiliki hubungan dengan *cost* yang tinggi. Apabila *store* tersebut di negara yang maju pastinya membutuhkan biaya yang juga mahal.

Kesimpulan

2. Algoritma yang digunakan untuk prediksi project ini adalah ***Decission Tree Regression, Random Forest Regression, Linear Regression, Support Vector Regression (SVR),*** **dan** ***KNearestNeighbors Regression***. Dari kelima algoritma yang telah digunakan, **Algoritma *Random Forest Regression* dipilih** karena memiliki performa yang baik dalam mengatasi prediksi dari dataset yang digunakan. Untuk skor yang didapatkan dari algoritma tersebut adalah **0,905521.**

Thank you!

See u on the next event
☺