

Project Data Analytics
“Optimizing Sales Strategies : Leveraging Monthly Shopping Patterns and Voucher Preferences”



Oleh:

2602104806 - Diandra Nathania Auwlia

2602152515 - Ratu Faradiba Adiazahra

2602108760 - Viola

Kelas : LB01

Kelompok : 7

Jurusan : Computer Science

Dosen : D6659 - EKO SETYO PURWANTO, S.Pd., M.Kom.

Binus University

Tahun Ajaran 2023/2024

DAFTAR ISI

DAFTAR ISI.....	1
BAB I.....	2
PENDAHULUAN.....	2
1.1 Latar Belakang.....	2
1.2 Tujuan dan Manfaat.....	2
1.3 Metode yang digunakan.....	3
BAB II.....	4
STUDI LITERATUR.....	4
2.1 Referensi Dataset.....	4
2.2 Hasil Analisis Referensi.....	4
BAB III.....	6
METODE PROYEK.....	6
3.1 Metode yang diusulkan dan digunakan.....	6
BAB IV.....	8
HASIL IMPLEMENTASI.....	8
4.1 Filter data untuk hanya mencakup 10 hari pertama setiap bulan.....	8
4.2 Menentukan jumlah data training, testing, dan validation.....	9
4.3 Menghitung jumlah pembelian yang dilakukan oleh pria dan wanita selama periode tersebut.....	9
4.3 Menghitung jumlah penggunaan kupon selama periode tersebut.....	10
4.4 Menganalisis preferensi penggunaan kupon berdasarkan jenis kelamin.....	13
4.5 Menganalisis lokasi pembelian selama periode tersebut.....	14
4.6 Hasil Cross Validation & Linear Regression.....	15
BAB V.....	17
KESIMPULAN PROYEK.....	17

5.1 Kesimpulan Hasil Implementasi Analisis Dataset.....	17
5.1.1 Hasil Implementasi.....	17
5.1.2 Peningkatan Hasil Analisis dari Analisis Sebelumnya.....	17
REFERENSI.....	18
LAMPIRAN.....	19

BAB I

PENDAHULUAN

1.1 Latar Belakang

Dalam era digital dan persaingan bisnis yang semakin ketat, pemahaman yang mendalam tentang perilaku pelanggan (*customer behavior*) dan strategi penjualan yang efektif menjadi kunci keberhasilan. Dengan adanya teknologi dan data yang melimpah, kesempatan terbuka lebar bagi kita untuk memanfaatkan informasi yang ada guna meningkatkan strategi penjualan dan meraih keunggulan kompetitif. Dalam konteks ini, kami memilih proyek yang berjudul “Optimizing Sales Strategies: Leveraging Monthly Shopping Patterns and Voucher Preferences.”

Pada proses online shopping, pengguna atau pembeli cenderung memutuskan untuk membeli suatu barang jika terdapat voucher yang sesuai dengan budget dan preferensi mereka. Sehingga, fokus utama dalam proyek analisis ini adalah untuk mengetahui preferensi penggunaan voucher berdasarkan *customer behavior* yang ditentukan dan didukung oleh faktor-faktor tertentu salah satunya jenis kelamin dan lokasi tempat tinggal mereka. Hasil analisis ini kemudian dapat digunakan untuk suatu perusahaan *e-commerce* atau *online shopping* dalam menentukan penyediaan voucher bagi pengguna tertentu berdasarkan *shopping patterns* yang sudah dianalisis.

10 hari pertama di setiap bulan sering dikatakan sebagai *peak days*, dimana tingkat pembelian sedang melambung tinggi. Oleh karena itu, analisis berdasarkan faktor *peak days* ini diperlukan untuk dapat membantu perusahaan menentukan stok, pemberian voucher kepada user tertentu berdasarkan data yang sudah dianalisis.

1.2 Tujuan dan Manfaat

Kami memilih proyek tersebut dengan tujuan untuk mengumpulkan data tentang perilaku belanja pelanggan selama 10 hari pertama setelah penggajian bulanan. Tujuan kami juga meliputi pemahaman yang lebih baik tentang preferensi penggunaan voucher oleh pelanggan, baik voucher itu digunakan, diklik, atau tidak digunakan sama sekali. Dengan data-data yang dikumpulkan, kami akan mengidentifikasi faktor-faktor yang mempengaruhi keputusan pembelian pelanggan, seperti jenis produk, harga, promosi, dan lain sebagainya.

Manfaat dari proyek ini adalah untuk meningkatkan efisiensi penjualan, mengurangi biaya promosi yang tidak efektif, mengambil keputusan berbasis data, meningkatkan daya saing, dan memperkuat hubungan dengan pelanggan.

1.3 Metode yang digunakan

Metode yang digunakan dalam proyek ini adalah Descriptive Analytics. Metode ini tergolong sederhana dan bertujuan untuk memahami apa yang telah terjadi dalam data dengan cara yang dapat dipahami dengan mudah. Tujuan utama dari descriptive analytics adalah memberikan pemahaman tentang tren, pola, dan hubungan dalam data yang telah ada. Dalam metode ini, beberapa metode yang dilakukan antara lain deskripsi data, filtering data, grouping and aggregation, visualisasi data, dan interpretasi. Oleh karena itu, tidak ada penggunaan teknik machine learning seperti regresi linier atau K-Means clustering. Analisis lebih berfokus pada pemahaman pola dan tren dalam data dengan menggunakan metode deskriptif dan visualisasi.

BAB II

STUDI LITERATUR

2.1 Referensi Dataset

Referensi dataset yang diambil untuk proyek ini menggunakan kaggle.com dengan topik Online Shopping Dataset. Secara spesifik, referensi dataset yang kita gunakan berjudul “*EDA and Visualizing Shopping Dataset*” oleh Jackson Divakar R. Link Referensi (Kaggle) :

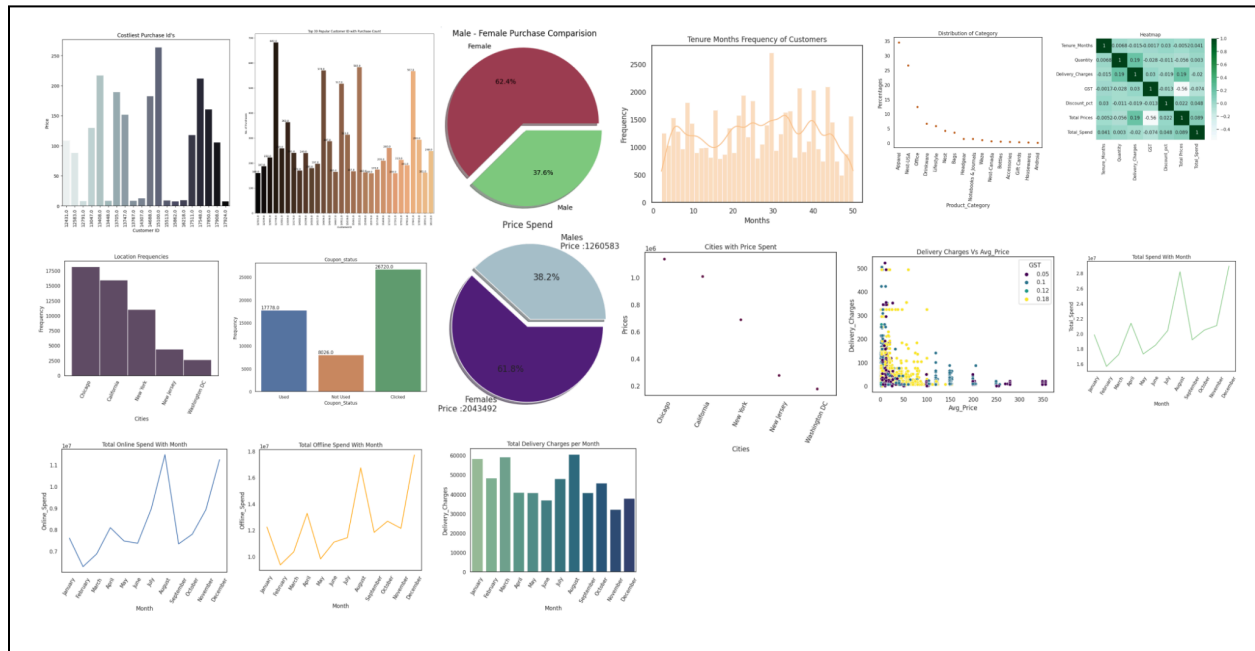
- Dataset (file.csv)
kaggle.com/code/jacksondivakarr/eda-and-visualizing-shopping-dataset/input
- Code (Notebook)
kaggle.com/code/jacksondivakarr/eda-and-visualizing-shopping-dataset/notebook

2.2 Hasil Analisis Referensi

Algoritma yang digunakan oleh referensi tersebut bertujuan untuk mengambil beberapa analisis antara lain:

- TOP 20 Costliest Purchase ID's,
- TOP 30 Popular Customer ID With Purchase Count
- Total Male-Female Purchase Comparison
- Tenure Months Frequency Of Customers
- Product Category Frequency Analysis
- Heatmap For Correlation
- Location Frequencies
- Coupon Status
- Male-Female Price Spent
- Cities With Price Spent
- Delivery Charges Vs Average Price
- Total Spend
- Online Spend
- Offline Spend
- Total Delivery Charges Per Month

Metode yang digunakan dalam referensi tersebut adalah Descriptive Analytics karena analisis pada referensi tersebut lebih berfokus pada pemahaman pola dan tren dalam data dengan menggunakan metode deskriptif dan visualisasi.



Gambar 2.1 Hasil Analisis Referensi Dataset

Hasil analisis referensi dataset tersebut dapat dilihat pada link ipynb berikut:

- (OneDrive)

https://binusianorg-my.sharepoint.com/personal/eko_purwanto_binus_ac_id/_layouts/15/guestaccess.aspx?share=EWm2IzkeTJdOr1-ZfX3VJ94ByAS_TgE4MuD887hiJha_-Q&e=Afaogx

- (Google Colab)

<https://colab.research.google.com/drive/1lhGAgP0IT5SyuLniLexo0H5InOrbwF6H?usp=sharing>

BAB III

METODE PROYEK

3.1 Metode yang diusulkan dan digunakan

Dari dataset yang sudah ada, analisis yang ingin kita lakukan dan modifikasi dari analisis sebelumnya antara lain:

- Filter data untuk hanya mencakup 10 hari pertama setiap bulan.
- Menghitung jumlah pembelian yang dilakukan oleh pria dan wanita selama periode tersebut.
- Menghitung jumlah penggunaan kupon (menggunakan, mengklik, tidak menggunakan) selama periode tersebut.
- Menganalisis preferensi penggunaan kupon berdasarkan jenis kelamin.
- Menganalisis lokasi pembelian selama periode tersebut.

Metode yang akan digunakan untuk menganalisis dataset tersebut adalah *descriptive analytics* yang sederhana dan bertujuan untuk memahami apa yang telah terjadi dalam data dengan cara yang mudah dipahami serta memberikan pemahaman tentang tren, pola, dan hubungan dalam data yang telah ada. Metode yang digunakan dalam descriptive analytics meliputi:

- Deskripsi Data: Pada awalnya, data dijelaskan untuk memahami struktur dan sifatnya, termasuk informasi tentang tipe data, jumlah entri, dan statistik deskriptif seperti mean, median, dan jumlah.
- Filtering Data: Data di filter untuk hanya mencakup 10 hari pertama setiap bulan menggunakan fungsi pemfilteran pada Pandas DataFrame.
- Grouping and Aggregation: Data dikelompokkan berdasarkan kategori tertentu seperti gender atau status kupon, dan kemudian dilakukan agregasi (misalnya, menghitung jumlah, rata-rata, atau total) untuk mendapatkan informasi yang lebih terperinci.
- Visualisasi Data: Visualisasi data digunakan untuk mewakili informasi dalam bentuk grafik, plot, atau diagram, seperti bar plot, pie chart, histogram, atau scatter plot. Visualisasi membantu dalam memahami pola dan tren dalam data dengan lebih jelas.

- Interpretasi: Hasil analisis kemudian diinterpretasikan untuk mendapatkan pemahaman yang lebih dalam tentang perilaku pembelian, preferensi penggunaan kupon, dan pola lainnya dalam data.

Dengan demikian, analisis ini lebih berfokus pada pemahaman pola dan tren dalam data dengan menggunakan metode deskriptif dan visualisasi, tanpa menggunakan teknik machine learning seperti regresi linier atau K-Means clustering.

BAB IV

HASIL IMPLEMENTASI

4.1 Filter data untuk hanya mencakup 10 hari pertama setiap bulan

```
[ ] data['Transaction_Date'] = pd.to_datetime(data['Transaction_Date'])
data_filtered = data[data['Transaction_Date'].dt.day <= 10]
```

Gambar 4.1 Kode Filter Data

[] data

	CustomerID	Gender	Location	Tenure_Months	Transaction_ID	Transaction_Date	Product_SKU	Product_Description	Product_Category	Quantity	...	Coupon_Status	GST	Date	Offline_Spend	Online_Spend	Mon
0	17850.0	M	Chicago	12.0	16679.0	2019-01-01	GGOENEBJ079499	Nest Learning Thermostat 3rd Gen-USA - Stainless Steel	Nest-USA	1.0	...	Used	0.10	1/1/2019	4500.0	2424.50	
1	17850.0	M	Chicago	12.0	16680.0	2019-01-01	GGOENEBJ079499	Nest Learning Thermostat 3rd Gen-USA - Stainless Steel	Nest-USA	1.0	...	Used	0.10	1/1/2019	4500.0	2424.50	
2	17850.0	M	Chicago	12.0	16696.0	2019-01-01	GGOENEBQ078999	Nest Cam Outdoor Security Camera - USA	Nest-USA	2.0	...	Not Used	0.10	1/1/2019	4500.0	2424.50	
3	17850.0	M	Chicago	12.0	16699.0	2019-01-01	GGOENEBQ079099	Nest Protect Smoke + CO White Battery Alarm-USA	Nest-USA	1.0	...	Clicked	0.10	1/1/2019	4500.0	2424.50	
4	17850.0	M	Chicago	12.0	16700.0	2019-01-01	GGOENEBJ079499	Nest Learning Thermostat 3rd Gen-USA - Stainless Steel	Nest-USA	1.0	...	Clicked	0.10	1/1/2019	4500.0	2424.50	
...
52919	13155.0	F	California	8.0	22504.0	2019-03-10	GGOEGGCX056399	Gift Card - \$250.00	Gift Cards	1.0	...	Clicked	0.05	3/10/2019	2500.0	1294.22	
52920	18077.0	M	Chicago	34.0	24250.0	2019-03-28	GGOEGGCX056299	Gift Card - \$25.00	Gift Cards	1.0	...	Used	0.05	3/28/2019	2000.0	1066.12	
52921	16085.0	M	California	15.0	39991.0	2019-10-06	GGOEGGCD078399	Google Leather Perforated Journal	Notebooks & Journals	1.0	...	Clicked	0.05	10/6/2019	3000.0	2230.76	
52922	16085.0	M	California	15.0	39991.0	2019-10-06	GGOEGGCR078499	Google Spiral Leather Journal	Notebooks & Journals	1.0	...	Used	0.05	10/6/2019	3000.0	2230.76	
52923	13659.0	F	Chicago	8.0	39998.0	2019-10-06	GGOEGGCC077999	Google Spiral Journal with Pen	Notebooks & Journals	1.0	...	Not Used	0.05	10/6/2019	3000.0	2230.76	

52524 rows x 22 columns

Gambar 4.2 Data sebelum di filter menjadi 10 hari pertama di setiap bulan

[] data_filtered

	CustomerID	Gender	Location	Tenure_Months	Transaction_ID	Transaction_Date	Product_SKU	Product_Description	Product_Category	Quantity	...	Coupon_Status	GST	Date	Offline_Spend	Online_Spend	Mon
0	17850.0	M	Chicago	12.0	16679.0	2019-01-01	GGOENEBJ079499	Nest Learning Thermostat 3rd Gen-USA - Stainless Steel	Nest-USA	1.0	...	Used	0.10	1/1/2019	4500.0	2424.50	
1	17850.0	M	Chicago	12.0	16680.0	2019-01-01	GGOENEBJ079499	Nest Learning Thermostat 3rd Gen-USA - Stainless Steel	Nest-USA	1.0	...	Used	0.10	1/1/2019	4500.0	2424.50	
2	17850.0	M	Chicago	12.0	16696.0	2019-01-01	GGOENEBQ078999	Nest Cam Outdoor Security Camera - USA	Nest-USA	2.0	...	Not Used	0.10	1/1/2019	4500.0	2424.50	
3	17850.0	M	Chicago	12.0	16699.0	2019-01-01	GGOENEBQ079099	Nest Protect Smoke + CO White Battery Alarm-USA	Nest-USA	1.0	...	Clicked	0.10	1/1/2019	4500.0	2424.50	
4	17850.0	M	Chicago	12.0	16700.0	2019-01-01	GGOENEBJ079499	Nest Learning Thermostat 3rd Gen-USA - Stainless Steel	Nest-USA	1.0	...	Clicked	0.10	1/1/2019	4500.0	2424.50	
...
52903	16367.0	M	California	20.0	27406.0	2019-05-06	GGOEWCKQ085457	Waze Pack of 9 Decal Set	Accessories	1.0	...	Used	0.10	5/6/2019	3000.0	2100.89	
52919	13155.0	F	California	8.0	22504.0	2019-03-10	GGOEGGCX056399	Gift Card - \$250.00	Gift Cards	1.0	...	Clicked	0.05	3/10/2019	2500.0	1294.22	
52921	16085.0	M	California	15.0	39991.0	2019-10-06	GGOEGGCD078399	Google Leather Perforated Journal	Notebooks & Journals	1.0	...	Clicked	0.05	10/6/2019	3000.0	2230.76	
52922	16085.0	M	California	15.0	39991.0	2019-10-06	GGOEGGCR078499	Google Spiral Leather Journal	Notebooks & Journals	1.0	...	Used	0.05	10/6/2019	3000.0	2230.76	
52923	13659.0	F	Chicago	8.0	39998.0	2019-10-06	GGOEGGCC077999	Google Spiral Journal with Pen	Notebooks & Journals	1.0	...	Not Used	0.05	10/6/2019	3000.0	2230.76	

16865 rows x 22 columns

Gambar 4.3 Data setelah di filter menjadi 10 hari pertama di setiap bulan

Dari filter data diatas, dapat disimpulkan bahwa data awal yang berjumlah 52524 rows di filterisasi dan berkurang menjadi 16865 rows saja. Tujuan dari adanya filter data ini adalah agar kita menganalisis data selama 10 hari saja di setiap bulannya. Alasan hanya menggunakan data pada periode tersebut karena merupakan periode dimana pekerja cenderung mendapatkan gaji mereka, sehingga bisa dikatakan sebagai *peak days*, dimana tingkat pembelian sedang melambung tinggi.

4.2 Menentukan jumlah data training, testing, dan validation


```
[39] from sklearn.model_selection import train_test_split

# Split the data into training and testing sets
X_train, X_temp, y_train, y_temp = train_test_split(X, y, test_size=0.4, random_state=42)

# Split the temporary set into testing and validation sets
X_test, X_validation, y_test, y_validation = train_test_split(X_temp, y_temp, test_size=0.5, random_state=42)

num_train_samples = len(X_train)
num_test_samples = len(X_test)
num_validation_samples = len(X_validation)

print(f"Jumlah data training: {num_train_samples}")
print(f"Jumlah data testing: {num_test_samples}")
print(f"Jumlah data validation: {num_validation_samples}")
```



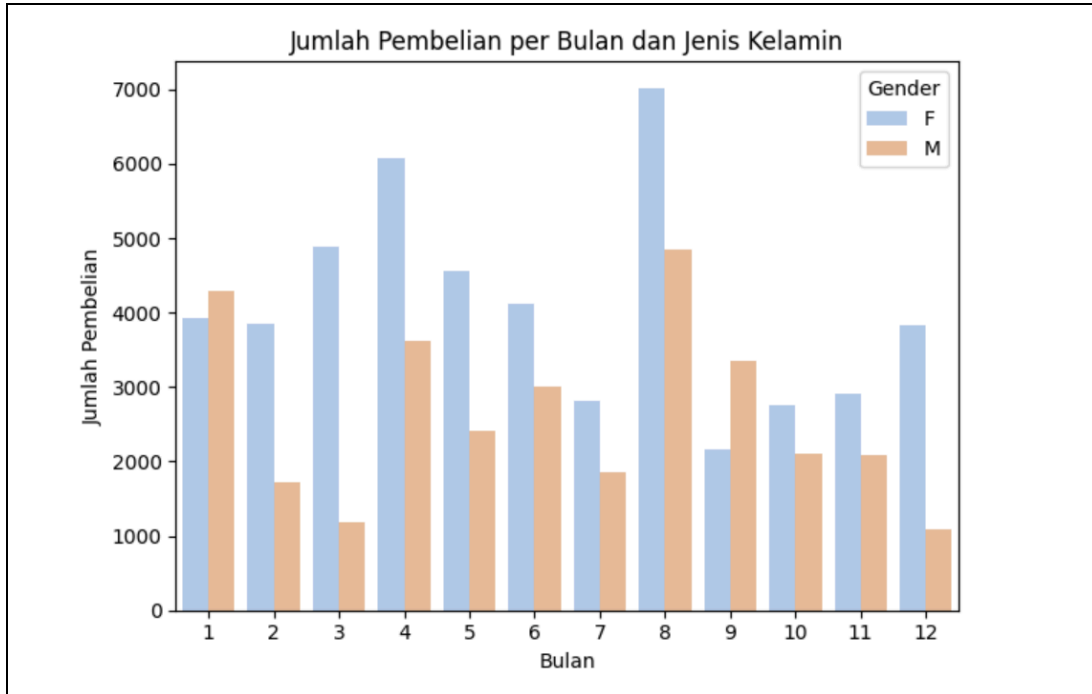
```
Jumlah data training: 60
Jumlah data testing: 20
Jumlah data validation: 20
```

4.3 Menghitung jumlah pembelian yang dilakukan oleh pria dan wanita selama periode tersebut

```
[ ] jppg = data_filtered.groupby(['Gender', 'Month'])['Quantity'].sum().reset_index()

sb.barplot(x='Month', y='Quantity', hue='Gender', data=jppg, palette='pastel')
plt.title('Jumlah Pembelian per Bulan dan Jenis Kelamin')
plt.xlabel('Bulan')
plt.ylabel('Jumlah Pembelian')
plt.legend(title='Gender')
plt.show()
```

Gambar 4.4 Kode jumlah pembelian per bulan dan jenis kelamin



Gambar 4.5 Visualisasi data jumlah pembelian per bulan dan jenis kelamin

Dari data tersebut, dapat kita analisis bahwa Female atau perempuan cenderung melakukan lebih banyak pembelian per bulannya. Hal ini ditunjukkan pada lebih tingginya angka jumlah pembelian oleh kelamin Female atau perempuan pada bulan 2, bulan 3, bulan 4, bulan 5, bulan 6, bulan 7, bulan 8, bulan 10, bulan 11, dan bulan 12. Sedangkan sisanya yaitu pada bulan 1 dan bulan 9, angka jumlah pembelian didominasi oleh Male atau pria. Namun perbandingannya adalah 10 : 2 dimana Female atau perempuan lebih banyak melakukan pembelian. Hal tersebut dapat memberi kesimpulan bahwa perempuan cenderung lebih banyak melakukan pembelian.

4.3 Menghitung jumlah penggunaan kupon selama periode tersebut

Penggunaan kupon meliputi parameter seperti clicked, used, dan not used.

a). Kode Tipe 1

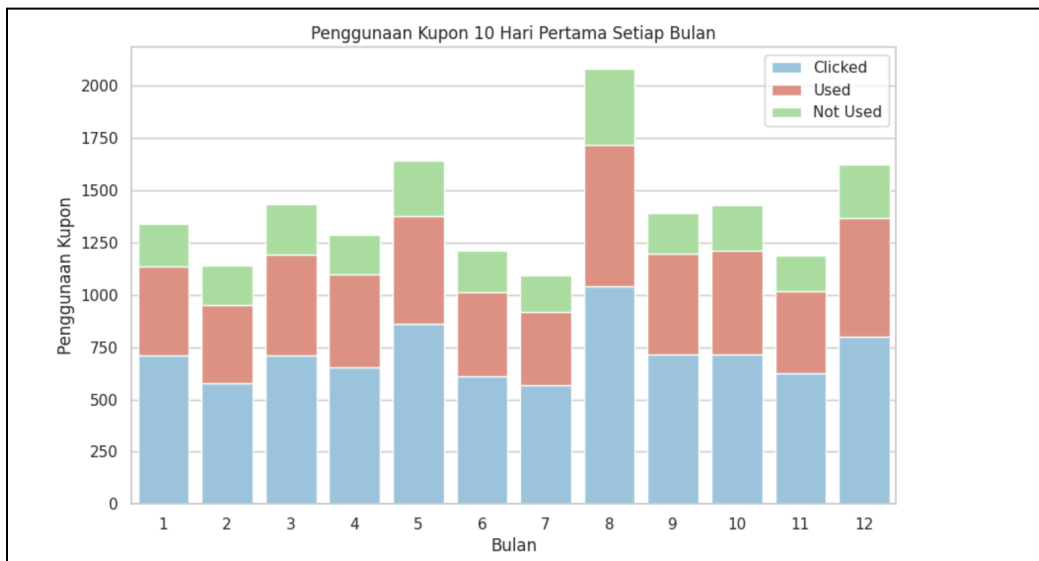
```
[ ] # Menghitung jumlah penggunaan kupon berdasarkan status
coupon_usage = data_filtered.groupby(['Month', 'Coupon_Status']).size().unstack(fill_value=0)

# Membuat grafik
plt.figure(figsize=(10, 6))
sb.barplot(x=coupon_usage.index, y=coupon_usage['Clicked'], color='skyblue', label='Clicked')
sb.barplot(x=coupon_usage.index, y=coupon_usage['Used'], color='salmon', label='Used', bottom=coupon_usage['Clicked'])
sb.barplot(x=coupon_usage.index, y=coupon_usage['Not Used'], color='lightgreen', label='Not Used',
           bottom=coupon_usage['Used'] + coupon_usage['Clicked'])

plt.xlabel('Bulan')
plt.ylabel('Penggunaan Kupon')
plt.title('Penggunaan Kupon 10 Hari Pertama Setiap Bulan')

plt.show()
```

Gambar 4.6 Kode penggunaan kupon 10 hari pertama setiap bulan



Gambar 4.7 Visualisasi data penggunaan kupon 10 hari pertama setiap bulan

b). Kode Tipe II

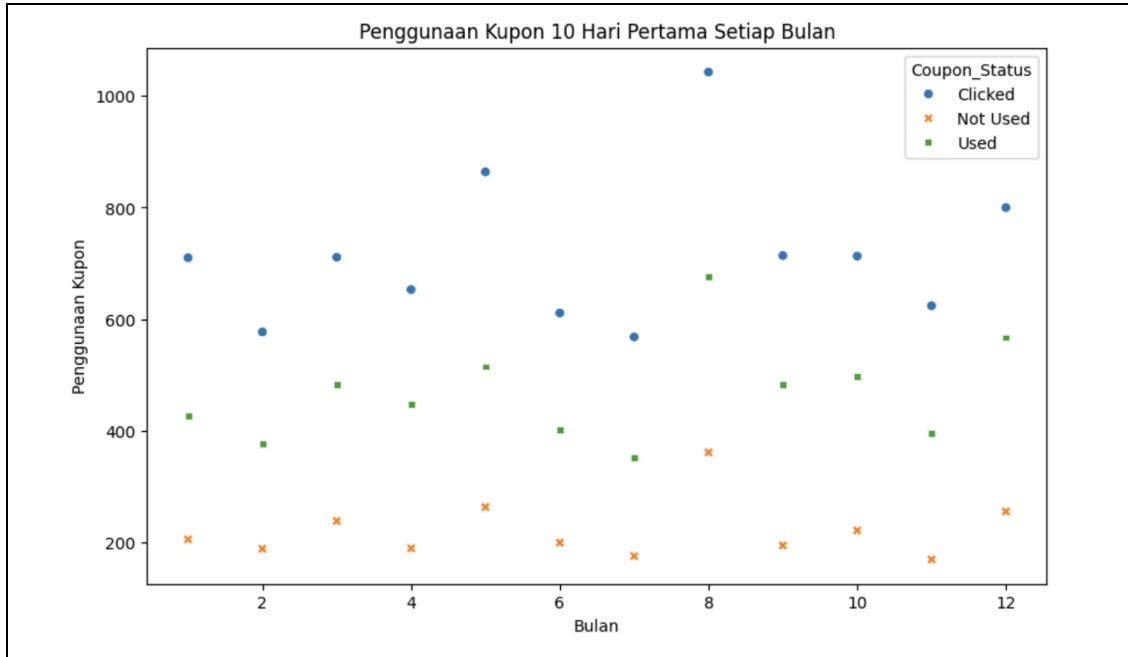
```
[ ] coupon_usage = data_filtered.groupby(['Month', 'Coupon_Status']).size().unstack()

plt.figure(figsize=(10, 6))
sb.scatterplot(data=coupon_usage, marker='o')

plt.xlabel('Bulan')
plt.ylabel('Penggunaan Kupon')
plt.title('Penggunaan Kupon 10 Hari Pertama Setiap Bulan')

plt.show()
```

Gambar 4.8 Kode penggunaan kupon 10 hari pertama setiap bulan



Gambar 4.9 Visualisasi data penggunaan kupon 10 hari pertama setiap bulan

Dari hasil data tersebut, dapat kita analisis bahwa penggunaan kupon atau coupon status cenderung pada 10 hari pertama di setiap bulan lebih banyak diklik atau *clicked* dibandingkan digunakan atau *used* dan tidak digunakan atau *not used*. Hal ini dapat disebabkan oleh beberapa faktor seperti kurang relevannya penggunaan kupon dengan kebutuhan pengguna. Pada visualisasi data tersebut, yaitu dalam bentuk *bar chart* dan *dot*, keduanya menyimpulkan bahwa urutan penggunaan atau status kupon dari yang terbanyak *adalah clicked, used, kemudian not used*.

4.4 Menganalisis preferensi penggunaan kupon berdasarkan jenis kelamin

```
[36] data['Transaction_Date'] = pd.to_datetime(data['Transaction_Date'])

# Filter the data for the first 10 days and 'Used' coupon status
data_coupon = data[(data['Transaction_Date'].dt.day <= 10) & (data['Coupon_Status'] == 'Used')]

# Get the gender value counts
val3 = data_coupon.Gender.value_counts()

# Define colors
c1 = plt.get_cmap('Blues') # Example of using a colormap
c2 = plt.get_cmap('Oranges') # Example of using a colormap

plt.pie(val3, labels=['Perempuan', 'Laki - Laki'], autopct="%1.1f%", shadow=True,
        explode=(0.1, 0), colors=[c2(0.7), c1(0.1)])
plt.axis('equal')
plt.title('Perbandingan Penggunaan Kupon 10 Hari Pertama Setiap Bulan')

sb.set(style='white')
plt.show()
```

Gambar 4.10 Kode analisis preferensi penggunaan kupon berdasarkan jenis kelamin



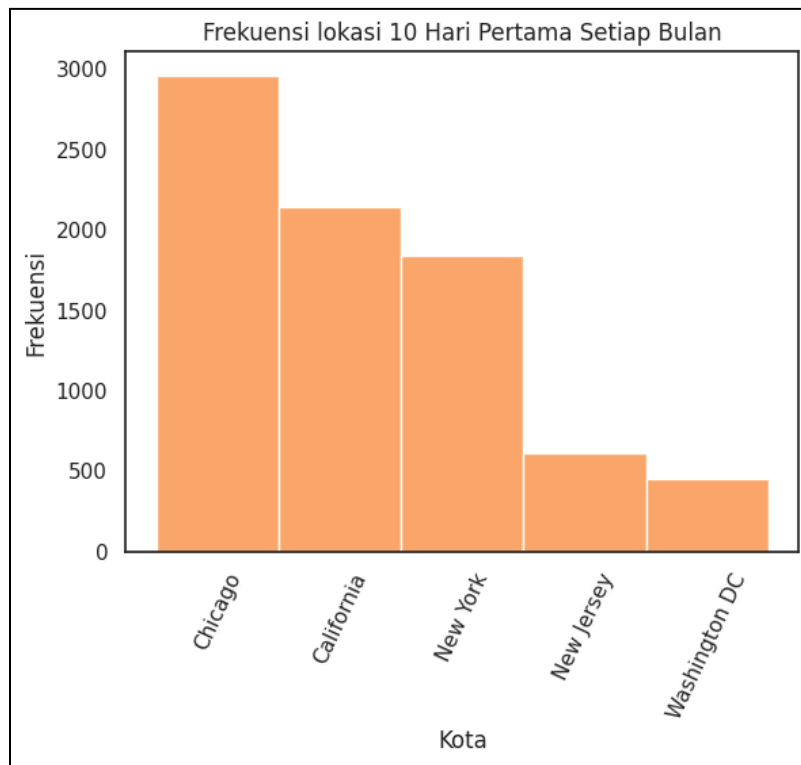
Gambar 4.11 Visualisasi data preferensi penggunaan kupon berdasarkan jenis kelamin

Dari analisis tersebut, dapat disimpulkan bahwa visualisasi data dalam bentuk *pie chart* tersebut menunjukkan perbandingan penggunaan kupon 10 hari pertama setiap bulan yang didominasi oleh jenis kelamin perempuan sebanyak 58.3% atau sekitar 9832 data dari dataset yang ada.

4.5 Menganalisis lokasi pembelian selama periode tersebut

```
[ ] sb.histplot(data_filtered.Location,color=c2(0.5))
plt.ylabel('Frekuensi')
plt.xlabel('Kota')
plt.xticks(rotation=65)
plt.title('Frekuensi lokasi 10 Hari Pertama Setiap Bulan')
```

Gambar 4.12 Kode analisis lokasi pembelian selama periode tersebut



Gambar 4.13 Visualisasi data analisis lokasi pembelian selama periode tersebut

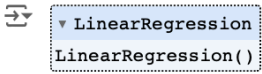
Dari hasil visualisasi data tersebut, dapat kita analisis dan simpulkan bahwa di 10 hari pertama di setiap bulan, persebaran lokasi pembeli paling banyak berasal dari Chicago, dan kemudian disusul oleh California, New York, New Jersey, dan terakhir Washington DC. Analisis frekuensi ini dapat memberikan asumsi bahwa Chicago merupakan kota dengan tingkat impulsif yang tinggi karena memiliki jumlah pembelian tertinggi. Hal ini dapat berguna bagi perusahaan untuk lebih banyak memberikan manfaat atau fasilitas seperti voucher atau kupon kepada user dengan domisili kota tersebut.

4.6 Hasil Cross Validation & Linear Regression

```
[ ] # Create some example data
    np.random.seed(0)
    X = 2 * np.random.rand(100, 1)
    y = 4 + 3 * X + np.random.randn(100, 1)

[ ] X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

[ ] model = LinearRegression()
    model.fit(X_train, y_train)
```



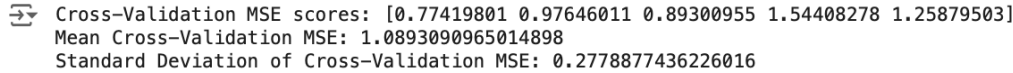
The image shows a Jupyter Notebook cell with a dropdown menu for 'LinearRegression' and the class name 'LinearRegression()' displayed below it.

Gambar 4.14 Kode Linear Regression

```
[ ] # Melakukan K-Fold Cross-Validation
    kf = KFold(n_splits=5, shuffle=True, random_state=42)
    mse_scorer = make_scorer(mean_squared_error)

[ ] # Mendapatkan hasil cross-validation
    cv_scores = cross_val_score(model, X_train, y_train, cv=kf, scoring=mse_scorer)

[ ] # Mencetak hasil cross-validation
    print(f"Cross-Validation MSE scores: {cv_scores}")
    print(f"Mean Cross-Validation MSE: {cv_scores.mean()}")
    print(f"Standard Deviation of Cross-Validation MSE: {cv_scores.std()}")
```



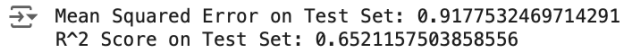
The image shows the output of the cross-validation step, displaying the MSE scores, mean, and standard deviation.

```
Cross-Validation MSE scores: [0.77419801 0.97646011 0.89300955 1.54408278 1.25879503]
Mean Cross-Validation MSE: 1.0893090965014898
Standard Deviation of Cross-Validation MSE: 0.2778877436226016
```

```
[ ] # Melatih model pada seluruh training set
    model.fit(X_train, y_train)

# Melakukan prediksi pada test set
y_pred = model.predict(X_test)

# Evaluasi model pada test set
mse_test = mean_squared_error(y_test, y_pred)
r2_test = model.score(X_test, y_test)
print(f"Mean Squared Error on Test Set: {mse_test}")
print(f"R^2 Score on Test Set: {r2_test}")
```



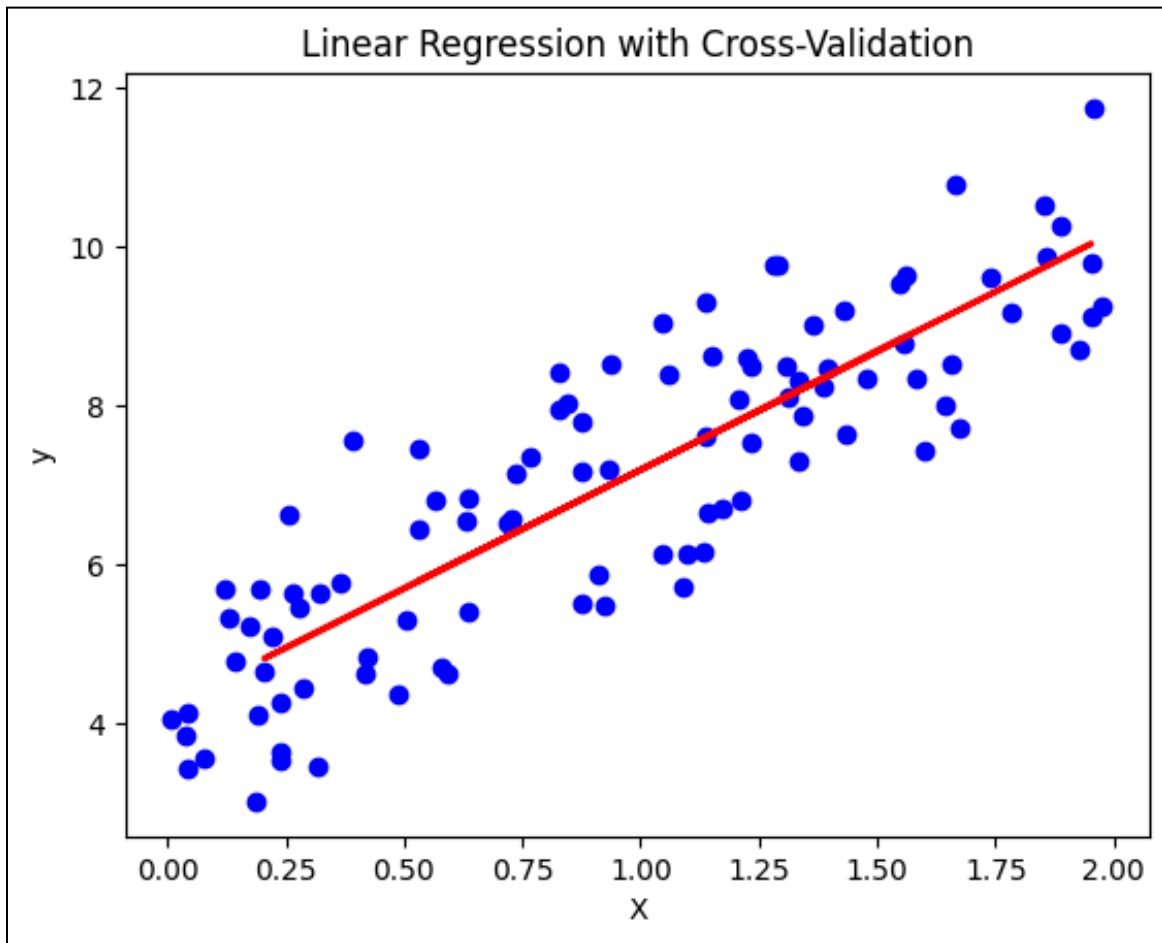
The image shows the output of the model evaluation on the test set, displaying the Mean Squared Error and the R-squared score.

```
Mean Squared Error on Test Set: 0.9177532469714291
R^2 Score on Test Set: 0.6521157503858556
```

Gambar 4.15 Kode Cross Validation


```
[ ] # Visualisasi hasil
plt.scatter(X, y, color='blue')
plt.plot(X_test, y_pred, color='red', linewidth=2)
plt.xlabel('X')
plt.ylabel('y')
plt.title('Linear Regression with Cross-Validation')
plt.show()
```

Gambar 4.16 Kode Plotting Linear Regression



Gambar 4.17 Plotting Linear Regression

1. **Evaluasi Model dengan Cross-Validation:** Kode pada Gambar 4.15 menunjukkan penggunaan teknik cross-validation untuk mengevaluasi model regresi linier. Hasil evaluasi menunjukkan nilai Mean Squared Error (MSE) rata-rata dari cross-validation

sebesar 1.0893099690514898 dan standar deviasi sebesar 0.2778877436220616. Hal ini menunjukkan variasi performa model di berbagai fold data

2. **Pelatihan Model pada Seluruh Training Set:** Model regresi linier dilatih pada seluruh data training (Gambar 4.15). Prediksi dilakukan pada data test, dan hasil evaluasi menunjukkan nilai MSE pada data test sebesar 0.91775532469174291 serta nilai R^2 sebesar 0.6521157503858556. Nilai R^2 yang lebih dari 0.6 menunjukkan bahwa model cukup baik dalam menjelaskan variabilitas data
3. **Visualisasi Hasil:** Gambar 4.16 menunjukkan kode untuk membuat visualisasi hasil regresi linier dengan cross-validation. Pada plot di Gambar 4.17, titik-titik data asli ditampilkan dalam warna biru, sementara garis regresi hasil prediksi ditampilkan dalam warna merah. Visualisasi ini menunjukkan hubungan linear antara variabel X dan Y, dengan garis regresi yang sesuai dengan pola data

Kesimpulannya adalah model regresi linier yang dibangun memiliki performa yang cukup baik dalam memprediksi variabel target pada data test, dengan nilai R^2 sebesar 0.652. Teknik cross-validation yang digunakan membantu dalam mengevaluasi kestabilan dan keandalan model. Visualisasi hasil menunjukkan bahwa model mampu menangkap pola linear dalam data dengan baik.

BAB V

KESIMPULAN PROYEK

5.1 Kesimpulan Hasil Implementasi Analisis Dataset

5.1.1 Hasil Implementasi

Hasil implementasi analisis pada dataset, ditemukan beberapa temuan penting:

- Filter data dilakukan untuk hanya mencakup 10 hari pertama setiap bulan, yang merupakan periode dengan tingkat pembelian yang tinggi.
- Perempuan cenderung melakukan lebih banyak pembelian daripada pria selama periode yang diteliti.
- Penggunaan kupon cenderung lebih banyak dalam status "clicked" dibandingkan "used" atau "not used".
- Preferensi penggunaan kupon didominasi oleh jenis kelamin perempuan.
- Lokasi pembelian terbanyak berasal dari Chicago, menunjukkan tingkat impulsif yang tinggi dalam pembelian.

5.1.2 Peningkatan Hasil Analisis dari Analisis Sebelumnya

Analisis ini lebih terperinci dan menyeluruh, termasuk deskripsi data, pemfilteran data, pengelompokan dan agregasi, visualisasi data, dan interpretasi, karena pemilihan data yang spesifik dari periode peak days. Berfokus pada 10 hari pertama setiap bulan, yang sering dianggap sebagai peak days, memberikan kesempatan untuk mengamati tren dan pola pembelian yang mungkin lebih signifikan dan terukur. Dengan membatasi data pada periode peak days, penelitian ini dapat memberikan pemahaman yang lebih mendalam tentang perilaku pembelian pelanggan selama periode yang kritis tersebut. Interpretasi hasil analisis yang lebih rinci dan jelas, terutama dalam menafsirkan preferensi penggunaan kupon berdasarkan jenis kelamin dan lokasi pembelian, dapat diberikan karena fokus pada periode waktu yang relevan dan signifikan dalam konteks strategi penjualan. Dengan demikian, penggunaan data dari peak days saja memberikan dasar yang lebih kokoh untuk analisis yang lebih terperinci dan interpretasi yang lebih mendalam, yang pada gilirannya dapat memberikan pemahaman yang lebih baik tentang implikasi hasil analisis dalam konteks strategi penjualan.

REFERENSI

- kaggle.com/code/jacksondivakarr/eda-and-visualizing-shopping-dataset/input
- <https://colab.research.google.com/drive/1lhGAgP0IT5SyuLniLexo0H5InOrbwF6H?usp=sharing>
- kaggle.com/code/jacksondivakarr/eda-and-visualizing-shopping-dataset/notebook
- kaggle.com/code/ibrahimelgmmal/online-shopping-analysis/notebook
- kaggle.com/code/ahmedismaail/online-shopping-sales
- <https://www.jaspersoft.com/articles/what-is-descriptive-analytics>

LAMPIRAN

Link Google Colab (Code): [\[ANALISIS KAMI\] Data Analytics-Kelompok 7.ipynb](#)

Link Dataset: [click here](#)

