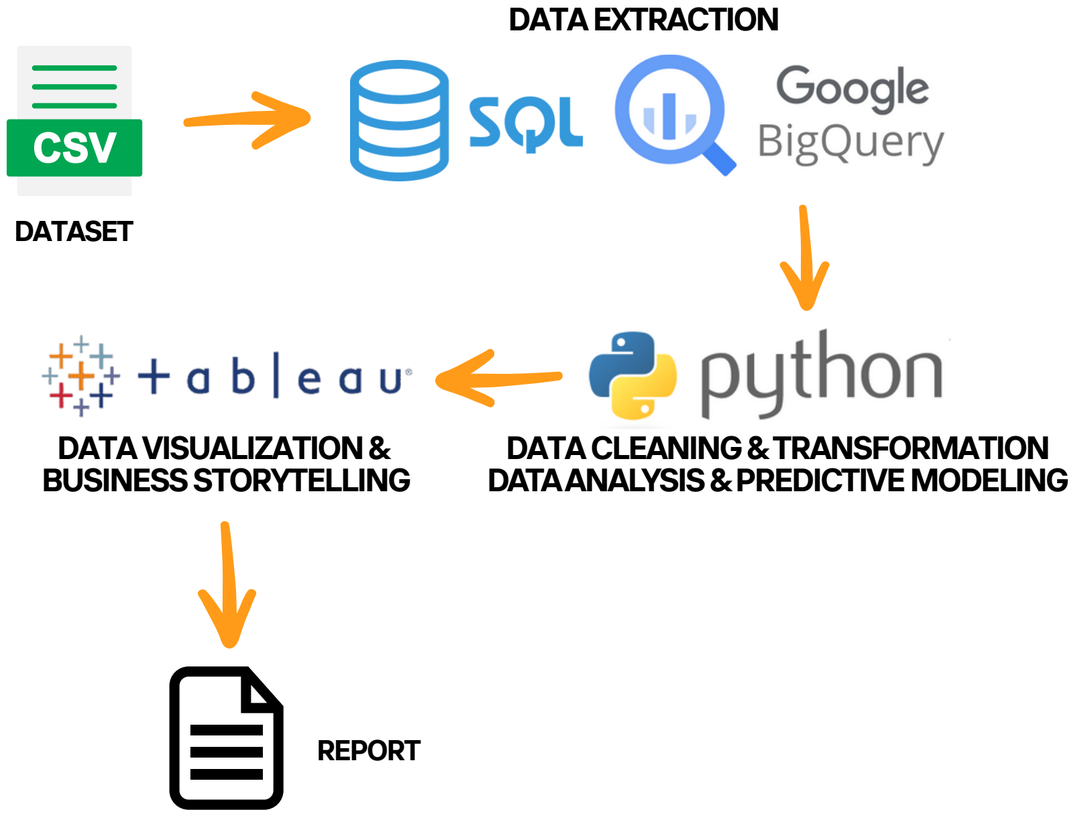
E-commerce Business Intelligence through Data Analytics

18 April, 2025

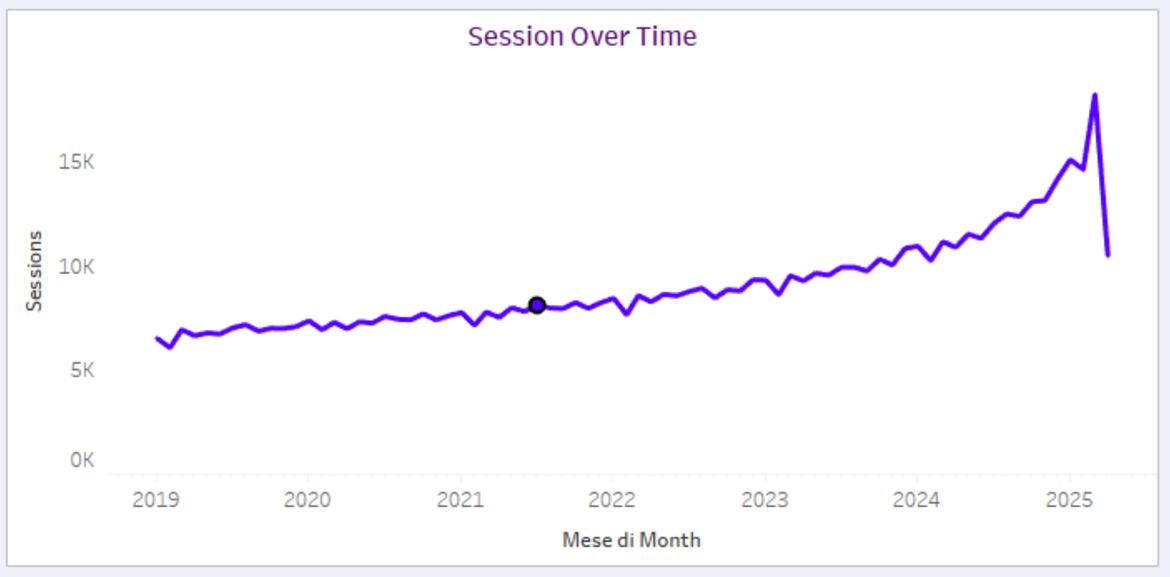
Introduction

In today's digital economy, e-commerce companies operate in highly competitive environments where understanding user behavior, product performance, and operational efficiency is vital to maintaining a competitive edge. Decision-makers need access to reliable data and insightful analytics to respond to market trends, optimize marketing strategies, streamline inventory, and enhance customer satisfaction.



- ***bigquery-public-data.thelook_ecommerce*** dataset
- **SQL** is used for data extraction from *Google BigQuery*
- **Python** is employed for analytical and predictive modeling,
- **Tableau** is used to design interactive dashboards for business storytelling.

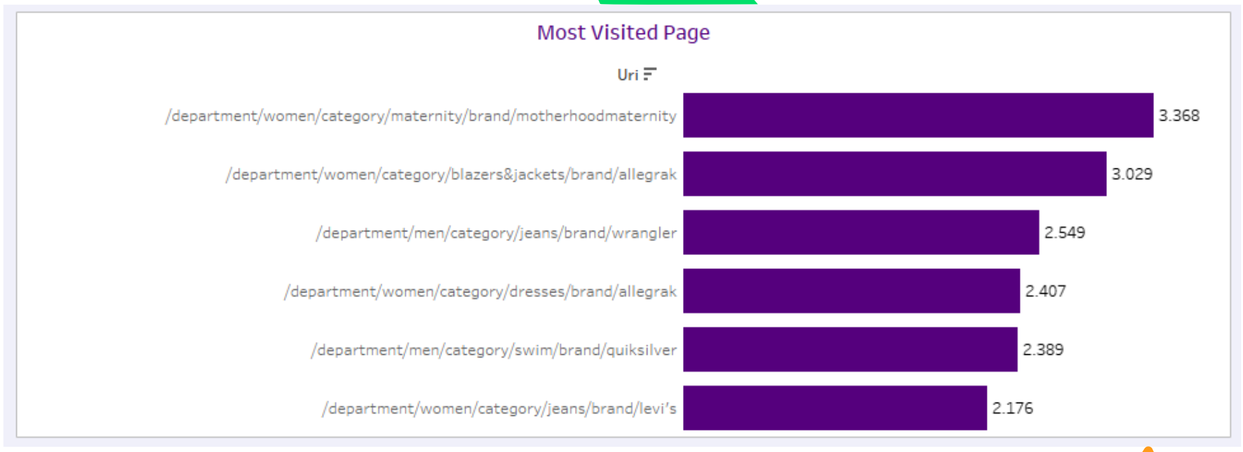
Website Activity




```
[ ] query = """
    SELECT FORMAT_TIMESTAMP('%Y-%m', created_at) AS month, COUNT(DISTINCT session_id) AS sessions
    FROM `bigquery-public-data.thelook_ecommerce.events`
    GROUP BY month
    ORDER BY month
    """
    df = client.query(query).to_dataframe()
    df.to_csv("/content/drive/MyDrive/ecommerce_analysis/website_sessions_over_time.csv", index=False)
    df.head()
```

	month	sessions
0	2019-01	6529
1	2019-02	6077
2	2019-03	6947
3	2019-04	6655
4	2019-05	6790

User engagement has increased over the years. The spike in 2025 suggests there was a significant event that temporarily boosted the activity which was followed by normalization or seasonal pattern.




```
[ ] query = """SELECT uri, COUNT(*) AS visits
            FROM `bigquery-public-data.thelook_ecommerce.events`
            GROUP BY uri
            ORDER BY visits DESC
            LIMIT 10"""
    df = client.query(query).to_dataframe()
    df.to_csv(f"/content/drive/MyDrive/ecommerce_analysis/top_visited_uris.csv", index=False)
    df.head()
```

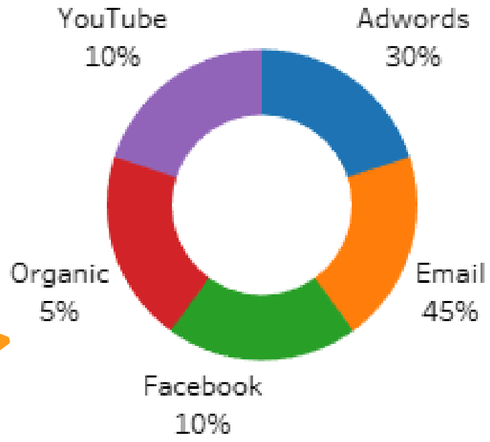
	uri	visits
0	/cart	591303
1	/purchase	180896
2	/cancel	124559
3	/home	87956
4	/department/women/category/maternity/brand/mot...	3368

There is a high level of users interest in fashion categories, with a notable emphasis on women's clothing and branded items.

Website Activity

```
[ ] query = """SELECT traffic_source, COUNT(*) AS sessions
               FROM `bigquery-public-data.thelook_ecommerce.events`
               GROUP BY traffic_source
               ORDER BY sessions DESC"""
    df = client.query(query).to_dataframe()
    df.to_csv(f"/content/drive/MyDrive/ecommerce_analysis/traffic_sources.csv", index=False)
    df.head()
```

	traffic_source	sessions
0	Email	1086920
1	Adwords	728717
2	YouTube	242310
3	Facebook	241415
4	Organic	119713

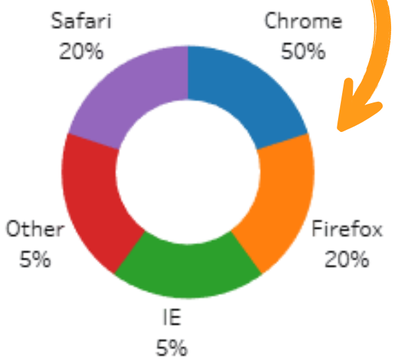


the reliance on email marketing highlights the effectiveness of targeted campaigns and suggests that maintaining a strong email strategy is important for continued engagement.

```
[ ] query = """
    SELECT browser, COUNT(*) AS visits
    FROM `bigquery-public-data.thelook_ecommerce.events`
    GROUP BY browser
    ORDER BY visits DESC
    """
    df = client.query(query).to_dataframe()
    df.to_csv("/content/drive/MyDrive/ecommerce_analysis/website_browser_usage.csv", index=False)
    df.head()
```

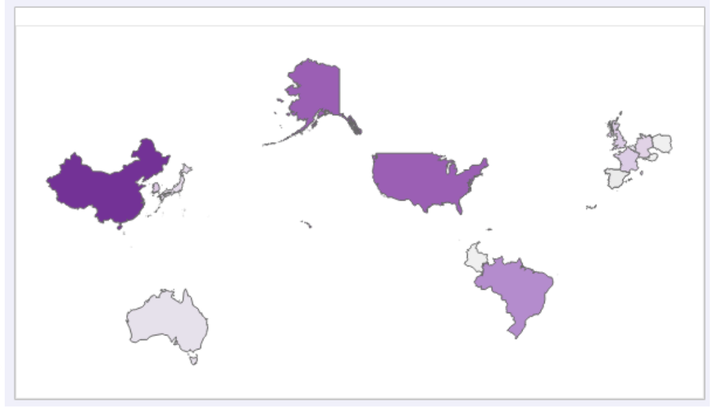
	browser	visits
0	Chrome	1210510
1	Firefox	483915
2	Safari	482977
3	IE	122017
4	Other	119656

website performance and design should be optimized primarily for Chrome, followed by Firefox and Safari to ensure a smooth user experience for the majority of visitors.



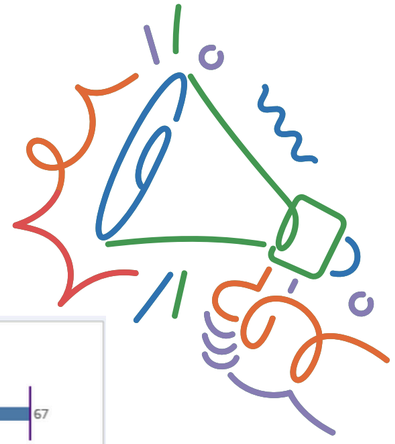

```
[ ] query = """
    SELECT u.country, COUNT(e.id) AS sessions
    FROM `bigquery-public-data.thelook_ecommerce.events` e
    JOIN `bigquery-public-data.thelook_ecommerce.users` u ON e.user_id = u.id
    GROUP BY u.country
    ORDER BY sessions DESC
    """
    df = client.query(query).to_dataframe()
    df.to_csv("/content/drive/MyDrive/ecommerce_analysis/website_sessions_by_country.csv", index=False)
    df.head()
```

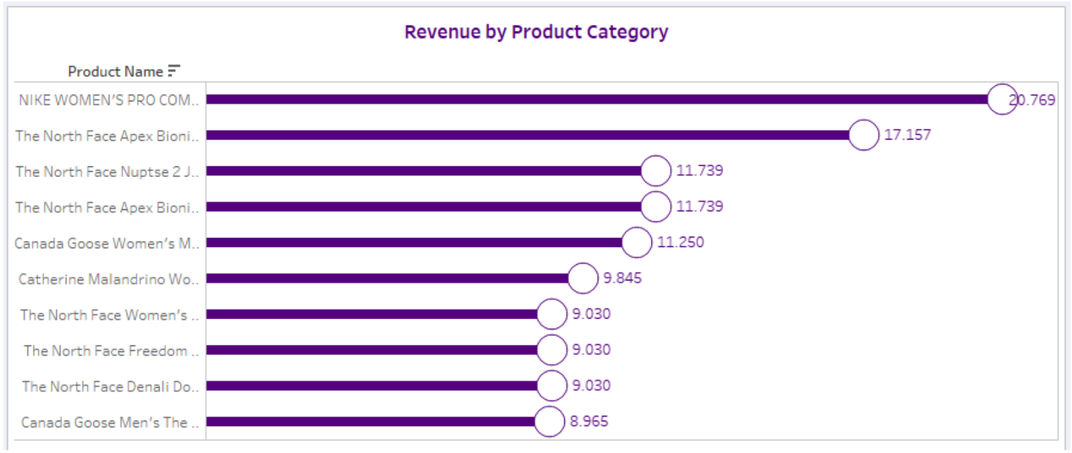
	country	sessions
0	China	440999
1	United States	290920
2	Brasil	186120
3	South Korea	70242
4	United Kingdom	60118



Targeted localization strategies and region-specific marketing could further enhance engagement in these key markets.

Product Composition

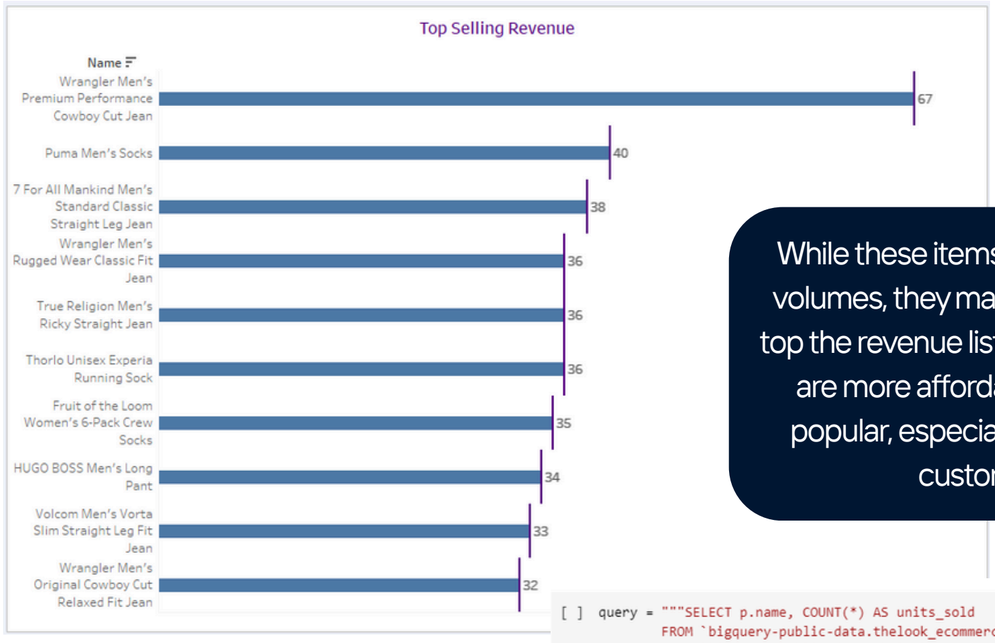





```
[ ] query = """
    SELECT p.name AS product_name, SUM(oi.sale_price) AS total_revenue
    FROM `bigquery-public-data.thelook_ecommerce.order_items` oi
    JOIN `bigquery-public-data.thelook_ecommerce.products` p ON oi.product_id = p.id
    GROUP BY product_name
    ORDER BY total_revenue DESC
    LIMIT 10
    """
    df = client.query(query).to_dataframe()
    df.to_csv("/content/drive/MyDrive/ecommerce_analysis/products_top_revenue.csv", index=False)
    df.head()
```

	product_name	total_revenue
0	NIKE WOMEN'S PRO COMPRESSION SPORTS BRA *Outst...	20769.0
1	The North Face Apex Bionic Soft Shell Jacket -...	17157.0
2	The North Face Nuptse 2 Jacket Deep Water Blue...	11739.0
3	The North Face Apex Bionic Jacket - Men's	11739.0
4	Canada Goose Women's Mystique	11250.0

the Revenue by Product Category chart indicates strong demand for activewear, especially among women. Several North Face and Canada Goose products also perform well, highlighting the popularity of premium outdoor and cold-weather apparel.

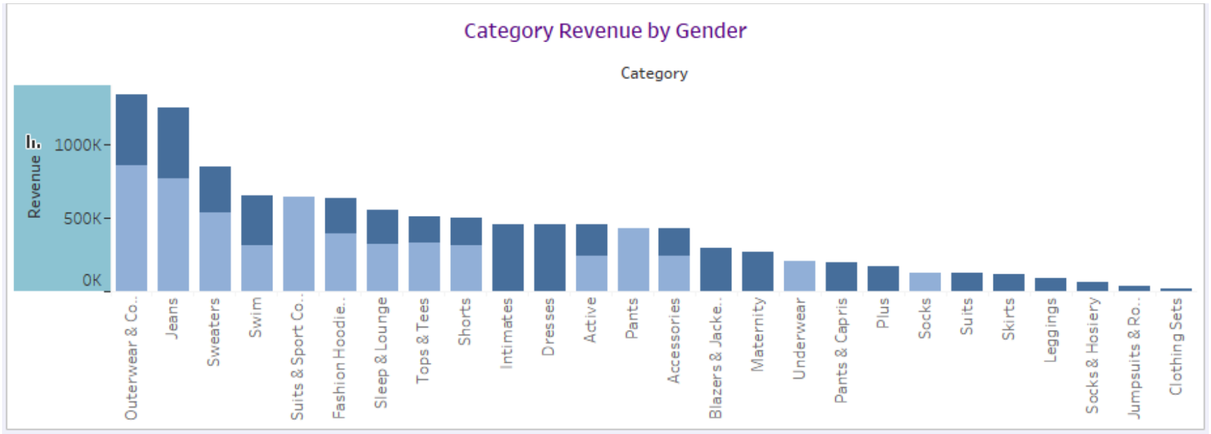


While these items have high sales volumes, they may not necessarily top the revenue list, suggesting they are more affordable but widely popular, especially among male customers.

```
[ ] query = """SELECT p.name, COUNT(*) AS units_sold
            FROM `bigquery-public-data.thelook_ecommerce.order_items` oi
            JOIN `bigquery-public-data.thelook_ecommerce.products` p ON oi.product_id = p.id
            GROUP BY p.name
            ORDER BY units_sold DESC
            LIMIT 10"""
    df = client.query(query).to_dataframe()
    df.to_csv(f"/content/drive/MyDrive/ecommerce_analysis/top_selling_products.csv", index=False)
    df.head()
```

	name	units_sold
0	Wrangler Men's Premium Performance Cowboy Cut ...	67
1	Puma Men's Socks	40
2	7 For All Mankind Men's Standard Classic Strai...	38
3	True Religion Men's Ricky Straight Jean	36
4	Thorlo Unisex Experia Running Sock	36

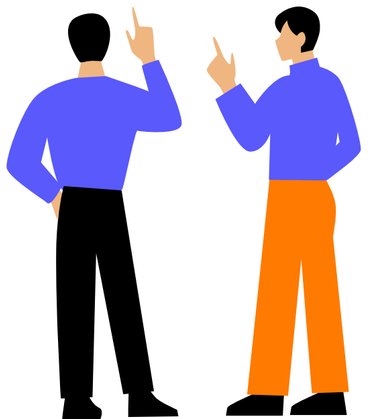
Product Composition

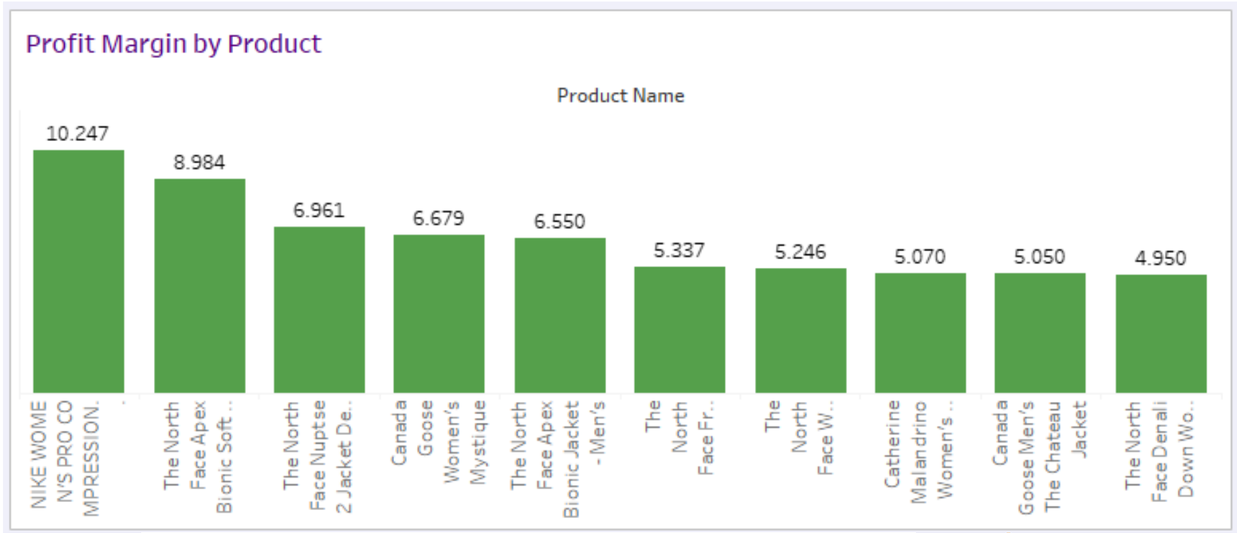



```
[ ] query = """
    SELECT p.category, o.gender, SUM(oi.sale_price) AS revenue
    FROM `bigquery-public-data.thelook_ecommerce.order_items` oi
    JOIN `bigquery-public-data.thelook_ecommerce.orders` o ON oi.order_id = o.order_id
    JOIN `bigquery-public-data.thelook_ecommerce.products` p ON oi.product_id = p.id
    GROUP BY p.category, o.gender
    ORDER BY revenue DESC
    """
    df = client.query(query).to_dataframe()
    df.to_csv("/content/drive/MyDrive/ecommerce_analysis/products_category_gender_revenue.csv", index=False)
    df.head()
```

	category	gender	revenue
0	Outerwear & Coats	M	854629.118807
1	Jeans	M	766627.420343
2	Suits & Sport Coats	M	640280.649470
3	Sweaters	M	535532.960245
4	Jeans	F	477295.711363

While both genders contribute to high-revenue categories, there is a clear difference in preference—Sleep & Lounge, Dresses, and Intimates perform better among female consumers, while Jeans and Activewear show a more balanced gender split.



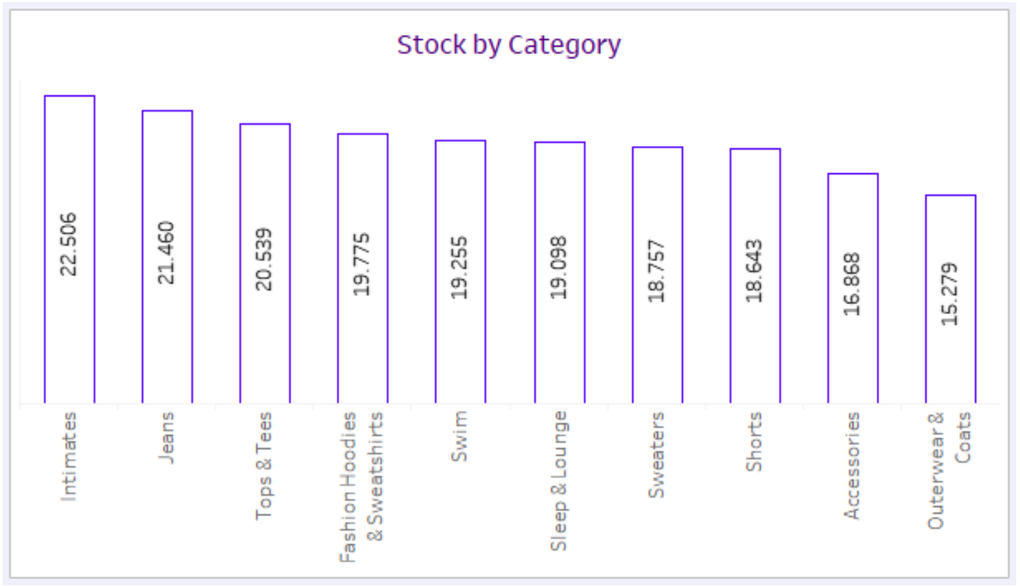


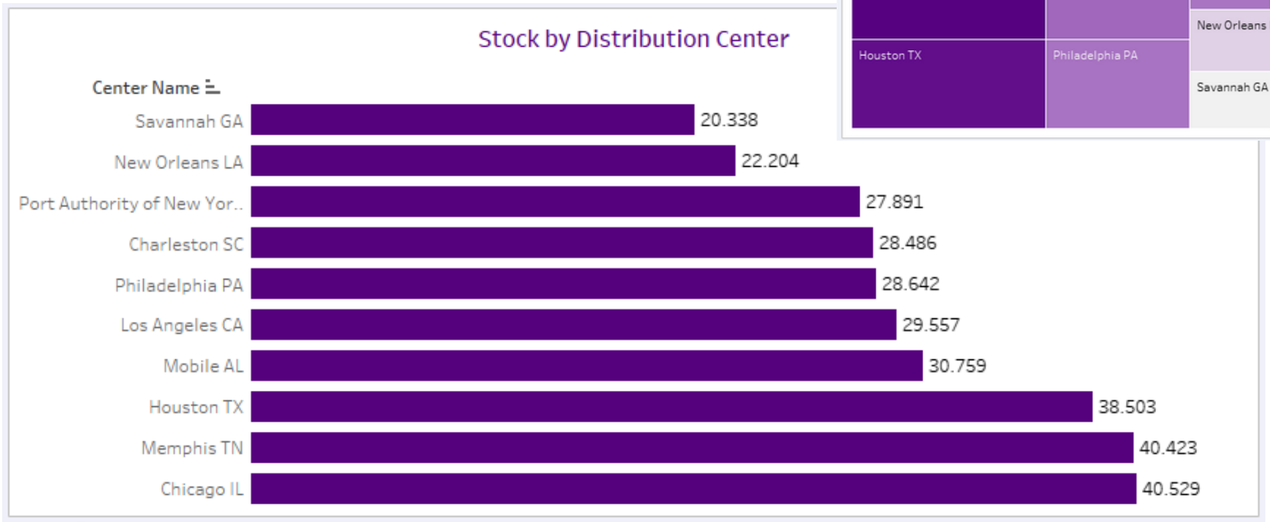

```
[ ] query = """
    SELECT p.name AS product_name, SUM(oi.sale_price - p.cost) AS total_profit
    FROM `bigquery-public-data.thelook_ecommerce.order_items` oi
    JOIN `bigquery-public-data.thelook_ecommerce.products` p ON oi.product_id = p.id
    GROUP BY product_name
    ORDER BY total_profit DESC
    LIMIT 10
    """
    df = client.query(query).to_dataframe()
    df.to_csv("/content/drive/MyDrive/ecommerce_analysis/products_top_profit.csv", index=False)
    df.head()
```

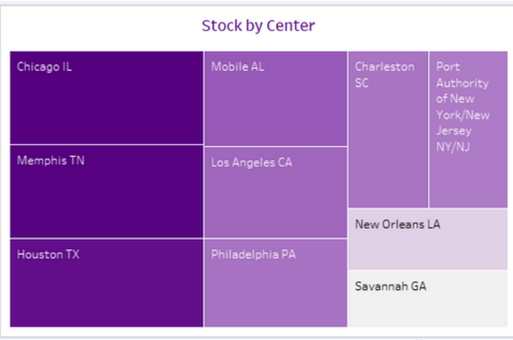
	product_name	total_profit
0	NIKE WOMEN'S PRO COMPRESSION SPORTS BRA *Outst...	10247.244008
1	The North Face Apex Bionic Soft Shell Jacket -...	8983.946972
2	The North Face Nuptse 2 Jacket Deep Water Blue...	6961.226982
3	Canada Goose Women's Mystique	6678.749991
4	The North Face Apex Bionic Jacket - Men's	6550.361983

Profit margins are generally higher for branded outerwear and performance gear, suggesting that while these items may have a higher cost, their markup allows for strong profitability.

Inventory Status








```
[ ] query = """
    SELECT product_category, COUNT(*) AS stock_available
    FROM `bigquery-public-data.thelook_ecommerce.inventory_items`
    WHERE sold_at IS NULL
    GROUP BY product_category
    ORDER BY stock_available DESC
    """
    df = client.query(query).to_dataframe()
    df.to_csv("/content/drive/MyDrive/ecommerce_analysis/inventory_stock_by_category.csv", index=False)
    df.head()
```

	product_category	stock_available
0	Intimates	22506
1	Jeans	21460
2	Tops & Tees	20539
3	Fashion Hoodies & Sweatshirts	19775
4	Swim	19255

items like Intimates, Jeans, and Tops & Tees are stocked in high numbers, with over 20,000 units each. On the other hand, Outerwear & Coats, which were top performers in revenue on the Product Performance Dashboard, are sitting at the bottom here with just over 15,000 units.

```
[ ] query = """
    SELECT d.name AS center_name, COUNT(i.id) AS available_stock
    FROM `bigquery-public-data.thelook_ecommerce.inventory_items` i
    JOIN `bigquery-public-data.thelook_ecommerce.distribution_centers` d
    ON i.product_distribution_center_id = d.id
    WHERE i.sold_at IS NULL
    GROUP BY center_name
    ORDER BY available_stock DESC
    """
    df = client.query(query).to_dataframe()
    df.to_csv("/content/drive/MyDrive/ecommerce_analysis/inventory_stock_by_center.csv", index=False)
    df.head()
```

	center_name	available_stock
0	Chicago IL	40529
1	Memphis TN	40423
2	Houston TX	38503
3	Mobile AL	30759
4	Los Angeles CA	29557

On the distribution centers, it's clear that Chicago, Memphis, and Houston are the heavy lifters, each holding upwards of 38,000 units. In contrast, Savannah GA and New Orleans LA are operating with much smaller stock levels, which might limit how quickly orders can be fulfilled in those regions. This is visualized in both the treemap and the bar chart.

Inventory Status

On the 10 Low Products in Stock table, several items, including women's outerwear and popular casual wear like Champion and Oakley products, are down to just one unit left. These could very easily go out of stock completely if no action is taken.

10 Low Products in Stock

Product Name	
BollÃ© Women's Essential Pleated Tennis Skirt	1
Carhartt Women's Cotton Web Belt	1
Champion 9 oz. 50/50 EcoSmart Open-Bottom P..	1
Enzyme Regular Solid Army Caps-Olive W35S45D (On..	1
Icebreaker Women's BF 200 Legging	1
Kipling Brownie Large Organizer Wallet	1
Knitted Beanie Crochet Winter Hat with Elegant K..	1
Oakley Men's Holbrook Iridium Sunglasses	1
Premium Super Soft Cashmere Feel Classic Plai..	1
Pro-CottonÂ® Fleece Full Zip Hood	1

```
[ ] query = """
    SELECT product_name, COUNT(*) AS stock
    FROM `bigquery-public-data.thelook_ecommerce.inventory_items`
    WHERE sold_at IS NULL
    GROUP BY product_name
    ORDER BY stock ASC
    LIMIT 10
    """
    df = client.query(query).to_dataframe()
    df.to_csv("/content/drive/MyDrive/ecommerce_analysis/inventory_low_stock.csv", index=False)
    df.head()
```

	product_name	stock
0	Champion 9 oz. 50/50 EcoSmart Open-Bottom Pant...	1
1	Kipling Brownie Large Organizer Wallet	1
2	Pro-CottonÂ® Fleece Full Zip Hood	1
3	Icebreaker Women's BF 200 Legging	1
4	Knitted Beanie Crochet Winter Hat with Elegant...	1

Stock Turnover Rate (Sold vs Available)

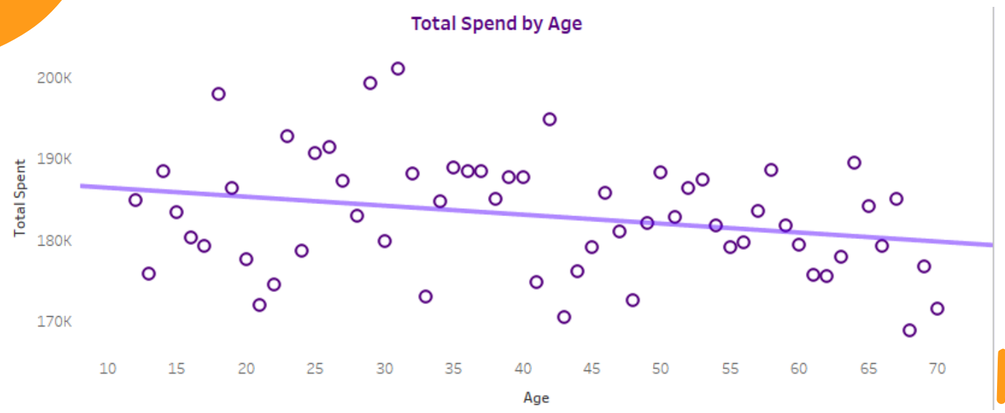
Product Name	Sold	Turnover Rate	Available
Bailey Curtis	2,000	0,500	2,000
Beige Knit Solid Color Circle Eternity Rin..	4,000	0,500	4,000
Black & Silver Double Grommet Holes Belt	2,000	0,500	2,000
Brooks Women's Infiniti Beanie	3,000	0,500	3,000
Calvin Klein Men's Smooth Leather Rever..	7,000	0,500	7,000
Carhartt Women's Cotton Web Belt	1,000	0,500	1,000
Carhartt Women's Dearborn Belt	3,000	0,500	3,000
DC Men's Clap Beanie	5,000	0,500	5,000
Dockers Men's Bridle Belt	3,000	0,500	3,000
Domo Men's Plush Wallet	2,000	0,500	2,000

```
[ ] query = """
    SELECT product_name,
      COUNTIF(sold_at IS NOT NULL) AS sold,
      COUNTIF(sold_at IS NULL) AS available,
      SAFE_DIVIDE(COUNTIF(sold_at IS NOT NULL), COUNT(*)) AS turnover_rate
    FROM `bigquery-public-data.thelook_ecommerce.inventory_items`
    GROUP BY product_name
    ORDER BY turnover_rate DESC
    LIMIT 10
    """
    df = client.query(query).to_dataframe()
    df.to_csv("/content/drive/MyDrive/ecommerce_analysis/inventory_turnover_rate.csv", index=False)
    df.head()
```

the Stock Turnover Rate section suggests the items that are moving slowly. Some may benefit from promotions or markdowns to clear out space and drive sales.

	product_name	sold	available	turnover_rate
0	Beige Knit Solid Color Circle Eternity Ring Scarf	4	4	0.5
1	Black & Silver Double Grommet Holes Belt	2	2	0.5
2	Domo Men's Plush Wallet	2	2	0.5
3	Dockers Men's Bridle Belt	3	3	0.5
4	Carhartt Women's Dearborn Belt	3	3	0.5

Demographic Composition



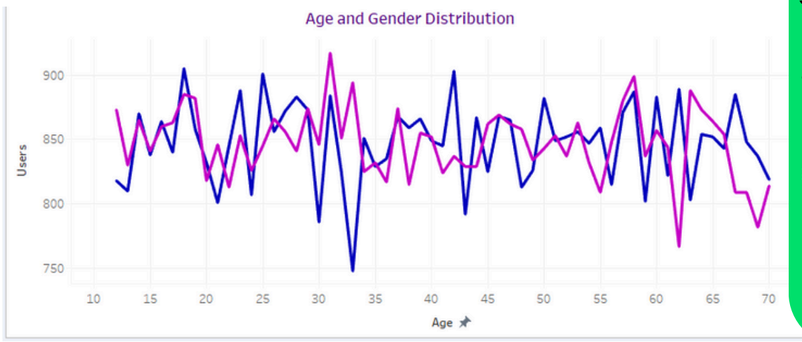

```
[ ] query = """
    SELECT u.age, SUM(oi.sale_price) AS total_spent
    FROM `bigquery-public-data.thelook_ecommerce.order_items` oi
    JOIN `bigquery-public-data.thelook_ecommerce.users` u ON oi.user_id = u.id
    GROUP BY u.age
    ORDER BY u.age
    """
    df = client.query(query).to_dataframe()
    df.to_csv("/content/drive/MyDrive/ecommerce_analysis/demographics_spend_by_age.csv", index=False)
    df.head()
```

	age	total_spent
0	12	184908.950229
1	13	175885.210222
2	14	188417.250204
3	15	183415.810149
4	16	180222.410136

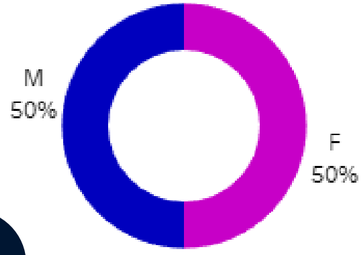
Although the R-squared value is low (0.07), indicating a weak model fit, the negative slope suggests younger users generally spend more than older ones, albeit marginally.

```
[ ] query = """SELECT age, gender, COUNT(*) AS users
            FROM `bigquery-public-data.thelook_ecommerce.users`
            GROUP BY age, gender
            ORDER BY age"""
    df = client.query(query).to_dataframe()
    df.to_csv(f"/content/drive/MyDrive/ecommerce_analysis/age_gender_distribution.csv", index=False)
    df.head()
```

	age	gender	users
0	12	F	873
1	12	M	818
2	13	F	830
3	13	M	810
4	14	M	870



The Age and Gender Distribution line chart shows a fairly even age distribution across genders, with noticeable spikes in both segments between the ages of 20 to 35. Some fluctuations exist at different ages, but no single gender appears to dominate overall. This reinforces a balanced gender presence on the platform.




```
[ ] query = """
    SELECT gender, COUNT(*) AS total_orders
    FROM `bigquery-public-data.thelook_ecommerce.orders`
    GROUP BY gender
    """
    df = client.query(query).to_dataframe()
    df.to_csv("/content/drive/MyDrive/ecommerce_analysis/demographics_orders_by_gender.csv", index=False)
    df.head()
```

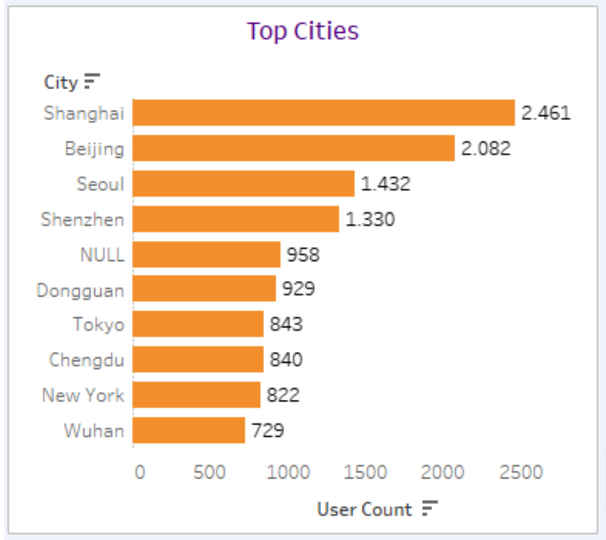
	gender	total_orders
0	F	62698
1	M	62464

The Order by Gender Distributions donut chart suggests that marketing efforts can be uniformly targeted toward both genders without needing significant customization based on gender identity.

Demographic Composition

Top 20 Loyal Customers

First Name	Last Name	User Id	
Alicia	Prince	77	4
Brandon	Nielsen	161	4
Cheryl	Mason	195	4
Cynthia	Holt	221	4
Deborah	Hudson	227	4
Emily	Stewart	235	4
Erica	Brown	142	4
Fernando	Snyder	74	4
Heather	Gibson	149	4
Jason	Miller	116	4
Jeffrey	Stevenson	340	4
Jennifer	Mcknight	86	4
Jessica	Gonzalez	352	4
Julia	Mitchell	436	4
Maria	Miranda	426	4
Nathan	Bailey	323	4
Sean	Yates	225	4
Sherry	Rivera	204	4
Stephanie	Mosley	8	4
Tina	Johnson	224	4



the Top Cities bar chart identifies Shanghai (2,461 users) and Beijing (2,082 users) as the most populated user bases, followed by Seoul, Shenzhen, and Dongguan. Major cities in China and East Asia dominate the top list, affirming the platform's strong presence in these regions. Additionally, New York appears among the top cities, indicating a reach that extends into the western market.

```
[ ] query = """
    SELECT u.first_name, u.last_name, o.user_id, COUNT(DISTINCT o.order_id) AS num_orders
    FROM `bigquery-public-data.thelook_ecommerce.orders` o
    JOIN `bigquery-public-data.thelook_ecommerce.users` u ON o.user_id = u.id
    GROUP BY u.first_name, u.last_name, o.user_id
    HAVING num_orders > 1
    """
    df = client.query(query).to_dataframe()
    df.to_csv("/content/drive/MyDrive/ecommerce_analysis/demographics_loyal_customers.csv", index=False)
    df.head()
```

	first_name	last_name	user_id	num_orders
0	Jennifer	Taylor	166	2
1	Miranda	Alvarado	213	2
2	Erin	Mccoy	258	2
3	Heather	Kelley	317	2
4	Tammy	Gallagher	327	2

The Top 20 Loyal Customers table highlights individuals with the highest repeat engagement, each making 4 orders. This includes a mix of male and female customers, indicating that loyalty spans across both demographics.

```
[ ] query = """
    SELECT city, COUNT(*) AS user_count
    FROM `bigquery-public-data.thelook_ecommerce.users`
    WHERE city IS NOT null
    GROUP BY city
    ORDER BY user_count DESC
    LIMIT 10
    """
    df = client.query(query).to_dataframe()
    df.to_csv("/content/drive/MyDrive/ecommerce_analysis/demographics_top_cities.csv", index=False)
    df.head()
```

	city	user_count
0	Shanghai	2461
1	Beijing	2082
2	Seoul	1432
3	Shenzhen	1330
4	null	958

Predictive Modelling
Linear Regression...

```
[ ] from sklearn.model_selection import train_test_split
    from sklearn.metrics import mean_absolute_error, mean_squared_error, accuracy_score, classification_report, ConfusionMatrix
    import numpy as np
    from sklearn.metrics import r2_score


    query = """
    SELECT u.age,
           u.gender,
           u.country,
           u.traffic_source,
           COUNT(DISTINCT o.order_id) AS total_orders,
           COUNT(DISTINCT oi.product_id) AS unique_products,
           COUNT(oi.id) AS total_items,
           SUM(oi.sale_price) AS total_spent,
           AVG(oi.sale_price) AS avg_sale_price,
           COUNT(DISTINCT p.department) AS num_departments,
           COUNT(DISTINCT p.category) AS num_categories,
           EXTRACT(HOUR FROM o.created_at) AS order_hour,
           EXTRACT(DAYOFWEEK FROM o.created_at) AS order_weekday
    FROM `bigquery-public-data.thelook_ecommerce.order_items` oi
    JOIN `bigquery-public-data.thelook_ecommerce.users` u ON oi.user_id = u.id
    JOIN `bigquery-public-data.thelook_ecommerce.orders` o ON oi.order_id = o.order_id
    JOIN `bigquery-public-data.thelook_ecommerce.products` p ON oi.product_id = p.id
    GROUP BY u.age, u.gender, u.country, u.traffic_source, order_hour, order_weekday
    """
    df = client.query(query).to_dataframe()

    # One-hot encode categorical variables
    categorical = ['gender', 'country', 'traffic_source']
    df_encoded = pd.get_dummies(df, columns=categorical)

    # Normalize numerical features
    from sklearn.preprocessing import StandardScaler
    features_to_scale = ['total_orders', 'unique_products', 'total_items', 'total_spent', 'avg_sale_price', 'num_departments',
    scaler = StandardScaler()
    df_encoded[features_to_scale] = scaler.fit_transform(df_encoded[features_to_scale])

    # Train linear regression model
    from sklearn.linear_model import LinearRegression
    X = df_encoded.drop(columns=['age'])
    y = df_encoded['age']

    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
    model = LinearRegression()
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)

    # Evaluate
    mae = mean_absolute_error(y_test, y_pred)
    rmse = np.sqrt(mean_squared_error(y_test, y_pred))
    r2 = r2_score(y_test, y_pred)

    print(f"Linear Regression MAE: {mae:.2f}, RMSE: {rmse:.2f}, R²: {r2 * 100:.2f}%")
```

```
Linear Regression MAE: 14.69, RMSE: 16.97, R²: -0.00%
```

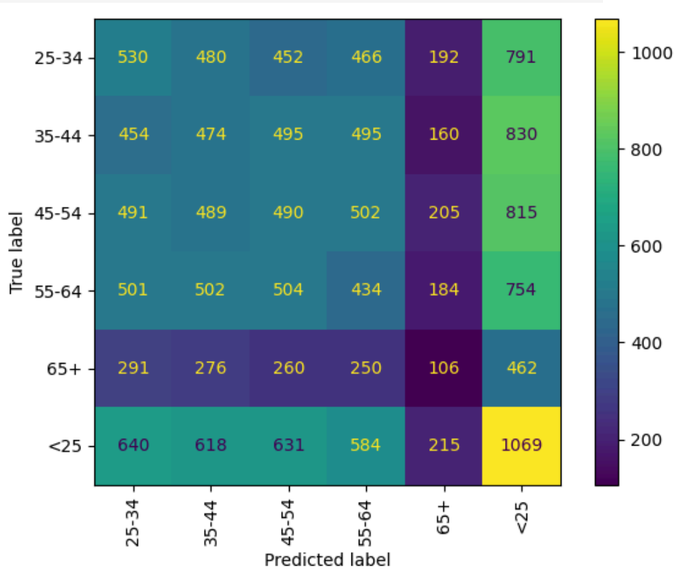
It seems that the R^2 value is negative, which indicates that the linear regression model is not performing well. Therefore, i will try a classification model using Random forest as shown belo

A linear regression was used to predict users' exact age based on their shopping behavior—things like total orders, number of unique products purchased, how much they spent, their average order value, the variety of product categories and departments they bought from, along with their gender, country, traffic source, and even the day and time they placed orders. However, the model didn't perform well at all. With an R-squared value of -0.00%, it essentially meant the model couldn't explain any variation in age from the available data, and the predictions were almost as good as guessing.

Predictive Modelling
Linear Regression...

```python
from sklearn.metrics import accuracy_score, classification_report, ConfusionMatrixDisplay
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.preprocessing import StandardScaler
import pandas as pd

# Assuming the query to get the data is already executed and `df` is the DataFrame
# Binning ages into groups (classification)
bins = [0, 24, 34, 44, 54, 64, 100]
labels = ['<25', '25-34', '35-44', '45-54', '55-64', '65+']
df['age_group'] = pd.cut(df['age'], bins=bins, labels=labels)

# Encode categorical variables
categorical = ['gender', 'country', 'traffic_source']
df_encoded = pd.get_dummies(df.drop(columns=['age']), columns=categorical)

# Standardize numerical features
features_to_scale = ['total_orders', 'unique_products', 'total_items', 'total_spent', 'avg_sale_price', 'num_departments', 'num_categories', 'order_hour', 'order_weekday']
scaler = StandardScaler()
df_encoded[features_to_scale] = scaler.fit_transform(df_encoded[features_to_scale])

# Classification
X = df_encoded.drop(columns=['age_group'])
y = df_encoded['age_group']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
clf = RandomForestClassifier(n_estimators=100, random_state=42)
clf.fit(X_train, y_train)
y_pred = clf.predict(X_test)

# Calculate Accuracy
accuracy = accuracy_score(y_test, y_pred) * 100  # Convert to percentage
print(f"Classification Accuracy: {accuracy:.2f}%")

# Evaluation
print(classification_report(y_test, y_pred))

# Confusion Matrix
ConfusionMatrixDisplay.from_estimator(clf, X_test, y_test, xticks_rotation='vertical')
plt.tight_layout()
plt.show()
```

```
Classification Accuracy: 18.15%
              precision    recall  f1-score   support

       25-34       0.18      0.18      0.18      2911
       35-44       0.17      0.16      0.16      2908
       45-54       0.17      0.16      0.17      2992
       55-64       0.16      0.15      0.15      2879
         65+       0.10      0.06      0.08      1645
         <25       0.23      0.28      0.25      3757

    accuracy                           0.18     17092
   macro avg       0.17      0.17      0.17     17092
weighted avg       0.18      0.18      0.18     17092
```



Given these poor results, we took a different approach: instead of trying to predict a specific age, we grouped ages into broader categories (like "<25", "25–34", "35–44", and so on) and used a classification model—specifically, a Random Forest classifier. Using the same features as before, we trained the model to classify users into age groups rather than predict an exact number. This approach gave us slightly better results, with an overall accuracy of 18.15%. While this is still relatively low, it suggests that age is not strongly reflected in transaction-level behavior alone.

Thank You

Group 6

- Mohammad Diaby
- Gabriele Garattoni
- Valentina Grisolia
- Massimiliano Piccolo