

## **Final Project in Data Science - Full Data Exploration (EDA)**

### **introduction:**

This document presents the data characterization and analysis process of Northwind, a global wholesale distributor specializing in importing and marketing food and beverage products to businesses worldwide. The analysis, conducted by us, data analysts, is designed to provide a strategic answer to the key business question: "How can Northwind increase its profits through data-driven decisions?"

To generate insights, a process was also performed to calculate gross profit: the UnitPrice field in the Products table was defined as the buying cost for the company, while the UnitPrice field in the Order Details table represents the actual selling price to the customer.

Analysis of the company's key metrics shows that Northwind is on a trajectory of accelerated growth, with an impressive jump in revenue from 200K in 1996 to ~650K in 1997. The company's main economic powerhouse is the US market (\$1.2M in sales), with the highest-yielding flagship product being Côte de Blaye wine. This characterization will detail the three research channels we will present to you to identify operational inefficiencies and leverage the company's growth engines to improve the bottom line.

### **Carrier Study 1: Product Profitability Analysis and Inventory Management**

**The main question:** How can Northwind increase its profits by optimizing its various products and improving inventory operational efficiency?

#### **Sub-questions guiding the analysis:**

Which products and categories generate the highest profit margins, and not just the highest revenue? What are the products whose marketing has been discontinued but which still take up space in inventory?

### **Findings and analysis**

From analyzing the SQL queries we performed on the company's database, the following findings emerged:

#### **1Revenue vs. Profitability:**

The product "Cote de Blaye" leads the revenue chart with an amount of approximately \$149,984.2. It is important to note that for the purpose of this calculation, we used the critical definition from the project guidelines: the column UnitPrice in the product table represents the cost of purchasing from the supplier, while the UnitPrice in the Order Details table represents the sales price to the customer. The Beverages category is the strongest category with total sales of approximately \$500,000.

#### **Identifying capital stuck in inventory:**

We identified several discontinued products that still have physical inventory, such as Mishi Kobe Niku from the meat category, with 29 units in stock, with a monetary value of \$1,406.

Another example is Rössle Sauerkraut with stock worth \$592.80.

In total, there are discontinued products that do not generate active cash flow worth \$2,200, even though the company has already paid suppliers for them.

### **Business conclusions and recommendations - courses of action**

Realizing inventory that does not promote profits: A liquidation operation should be carried out on products from the Discontinued category. The goal is to quickly turn the inventory into liquid capital to prevent a complete loss of the investment in these products.

Reinvestment in profitable products and optimization of liquid capital for profit: We recommend investing the capital freed up from inventory liquidation in purchasing additional inventory and marketing promotion of products with proven profitability, such as Côte de Blaye and products from the Beverages category.

Another important note we came to is that it is recommended to ensure that products that are no longer in use do not remain in inventory for long, thus maintaining higher liquidity of the company's capital and maximizing profits. It is important to implement a process to avoid such examples.

The code:

```
SELECT
    p.ProductName,(p.UnitsOnOrder * o.UnitPrice)(1-o.Discount) AS
    Gross_profit_after_discount
FROM [Northwind].[Order Details] o
JOIN [Northwind].[Orders] w ON o.OrderID = w.OrderID
JOIN [Northwind].[Products] p ON p.ProductID = o.ProductID
JOIN [Northwind].[Categories] c ON c.CategoryID = p.CategoryID
where p.UnitsOnOrder != 0
```

```
SELECT
    p.ProductName,
    c.CategoryName,
    SUM(od.UnitPrice * od.Quantity) AS TotalRevenue,
    SUM((od.UnitPrice - p.UnitPrice) * od.Quantity) AS TotalGrossProfit,
    ROUND(
        SUM((from.UnitPrice - from.UnitPrice) * from.Quantity)
        / NULLIF(SUM(od.UnitPrice * od.Quantity), 0) * 100,
        2
    ) AS ProfitMarginPercentage
FROM Northwind.Products p
JOIN Northwind.[Order Details] od
    ON p.ProductID = od.ProductID
JOIN Northwind.Categories c
    ON p.CategoryID = c.CategoryID
GROUP BY p.ProductName, c.CategoryName
ORDER BY TotalGrossProfit DESC;
```

```
SELECT
    p.ProductName,
    c.CategoryName,
    p.UnitsInStock,
    p.UnitPrice AS BuyingCost,
    (p.UnitsInStock * p.UnitPrice) AS ValueLockedInStock
FROM Northwind.Products p
JOIN Northwind.Categories c ON p.CategoryID = c.CategoryID
WHERE p.Discontinued = 1 AND p.UnitsInStock > 0
ORDER BY ValueLockedInStock DESC;
```

Output - highest values:

	ProductName	Gross_profit_after_discount			
1	Chang	486.4			
2	Mascarpone Fabioli	819.2			
3	Outback Lager	90			
4	Outback Lager	120			
5	Ipoh Coffee	368			
6	Scottish Longbreads	90			
7	Maxilaku	960			
8	Queso Cabrales	504			
9	Ipoh Coffee	368			
	ProductName	CategoryName	TotalRevenue	TotalGrossProfit	ProfitMarginPercentage
1	C9te de Blaye	Beverages	149984.20	67903.95	45.27
2	Th•ringer Rostbratwurst	Meat/Poultry	87736.40	41562.73	47.37
3	Raclette Courdavault	Dairy Products	76296.00	35156.00	46.07
4	Camembert Pierrot	Dairy Products	50286.00	23477.00	46.68
5	Tarte au sucre	Confections	49827.90	23131.95	46.42
6	Manjimup Dried Apples	Produce	44742.60	21263.60	47.52
7	Gnocchi di nonna Alice	Grains/Cereals	45121.20	21124.20	46.81
8	Alice Mutton	Meat/Poultry	35482.20	16411.20	46.25
9	Camaron Tigres	Seafood	31987.50	15143.75	47.34
10	R5ssle Sauerkraut	Produce	26865.60	12273.60	45.68
11	Ipoh Coffee	Beverages	25079.20	11739.20	46.80
12	Mozzarella di Giovanni	Dairy Products	25738.80	11714.40	45.51
13	Gudbrandsdalsost	Dairy Products	24307.20	11455.20	47.12
14	Uncle Bob's Organic ...	Produce	22464.00	11019.00	49.05
15	Sir Rodney's Marmalade	Confections	23635.80	10959.30	46.36
16	Wimmers gute Semme...	Grains/Cereals	23009.00	10706.50	46.53
17	Ikura	Seafood	22140.20	10639.20	48.05
18	Gumb1r Gummib1rchen	Confections	21534.90	9776.805	45.39
19	Perth Pasties	Meat/Poultry	21510.20	9669.40	44.95
20	Flotemysost	Dairy Products	20876.50	9513.75	45.57
21	Boston Crab Meat	Seafood	19048.30	8900.70	46.72
22	P1t1 chinois	Meat/Poultry	19512.00	8676.00	44.46
23	Pavlova	Confections	18748.05	8644.50	46.10
24	Chang	Beverages	18559.20	8517.70	45.89
25	Lakkalik55ri	Beverages	16794.00	7965.00	47.42
26	Vegie-spread	Condiments	17696.30	7928.55	44.80
27	Sirup d'5rable	Condiments	16438.80	7846.05	47.72
	ProductName	CategoryName	UnitsInStock	BuyingCost	ValueLockedInStock
1	Mishi Kobe Niku	Meat/Poultry	29	48.50	1406.50
2	R5ssle Sauerkraut	Produce	26	22.80	592.80
3	Singaporean Ho...	Grains/Cereals	26	7.00	182.00
4	Guaran1 Fant1st...	Beverages	20	2.25	45.00

## Output lowest values:

	ProductName	Gross_profit_after_discount			
441	Wimmers gute Semmelkn�del	2660			
442	Wimmers gute Semmelkn�del	2660			
443	Sir Rodney's Scones	400			
444	Maxilaku	1200			
445	Sir Rodney's Scones	400			
446	Scottish Longbreads	125			
447	Outback Lager	150			
448	Sir Rodney's Scones	400			

	ProductName	CategoryName	TotalRevenue	TotalGrossProfit	ProfitMarginPercentage
51	Tofu	Produce	8630.40	3933.90	45.58
52	Grandma's Boysenberr...	Condiments	7345.00	3582.50	48.77
53	Gustaf's Kn�ttebr�d	Grains/Cereals	7232.40	3578.40	49.47
54	Ravioli Angelo	Grains/Cereals	7807.80	3576.30	45.80
55	Sasquatch Ale	Beverages	6678.00	3136.00	46.96
56	Escargots de Bourgog...	Seafood	6664.75	3127.00	46.91
57	Spegesild	Seafood	6144.00	2856.00	46.48
58	Teatime Chocolate Bis...	Confections	6159.50	2833.70	46.00
59	Chef Anton's Gumbo ...	Condiments	5801.15	2620.00	45.16
60	Konbu	Seafood	5234.40	2561.40	48.93
61	Rogede sild	Seafood	4740.50	2327.50	49.09
62	Tourti�re	Meat/Poultry	5121.00	2308.625	45.08
63	Guaran� Fant�stica	Beverages	4782.60	2251.35	47.07
64	T�nnbr�d	Grains/Cereals	4840.20	2230.20	46.07
65	Zaanse koeken	Confections	4358.60	2054.85	47.14
66	R�d Kaviar	Seafood	4200.00	2002.50	47.67
67	NuNuCa Nu�-Nougat-...	Confections	4051.60	1825.60	45.05
68	Filo Mix	Grains/Cereals	3383.80	1633.80	48.28
69	Valkoinen suklaa	Confections	3510.00	1600.625	45.60
70	Louisiana Hot Spiced ...	Condiments	3519.00	1487.50	42.27
71	Aniseed Syrup	Condiments	3080.00	1440.00	46.75
72	Gravad lax	Seafood	3047.20	1422.20	46.67
73	Laughing Lumberjack ...	Beverages	2562.00	1274.00	49.72
74	Longlife Tofu	Produce	2566.00	1081.00	42.12
75	Genen Shouyu	Condiments	1813.50	868.00	47.86
76	Geitost	Dairy Products	1713.50	769.75	44.92
77	Chocolade	Confections	1542.75	663.00	42.97

	ProductName	CategoryName	UnitsInStock	BuyingCost	ValueLockedInStock
1	Mishi Kobe Niku	Meat/Poultry	29	48.50	1406.50
2	R�ssle Sauerkraut	Produce	26	22.80	592.80
3	Singaporean Ho...	Grains/Cereals	26	7.00	182.00
4	Guaran� Fant�st...	Beverages	20	2.25	45.00

## **Research Question 2 Topic: Segmenting top customers and analyzing purchasing behavior.**

### **1. The main question:**

How can profitability be maximized by identifying top customers and analyzing the relationship between order frequency and gross profit?

### **2. Sub-questions:**

Who are the top 5 customers in terms of total revenue and gross profit?

Is there a statistical correlation between the number of orders placed by a customer and the profit it generated for the company?

What is the geographic distribution of top customers, and where is Northwind's global economic power concentrated?

### **3. Work plan:**

Tables used: Customers, Orders, Order Details, and Products.

**Tools:** Python (Pandas) for data processing, correlation calculation, (Seaborn, matplotlib) for visualization, and Plotly for producing a world map by gross profitability for countries.

Calculation logic: Merging the tables to cross-reference the sales price with the purchase cost (which appears in the product table), calculating the Gross Profit at the row level, and aggregating by customer.

### **4. Findings and conclusions**

1. Three clients: QUICK-Stop (Germany), Save-a-lot Markets (USA), and Ernst Handel (Austria), are the main engines of the company, with each of them generating a gross profit of over \$53,000, an amount more than twice that of the other clients on the list.

2. A very strong correlation of 0.89 was found between order frequency and gross profit. This means that each additional order from an existing customer contributes to the profit line, indicating operational efficiency in repeat sales.

Geographically, according to the map, the world map clearly shows that the bulk of the economic power is in the United States (\$115K) and the German region (\$110K).

The South American market (Brazil) shows significant potential but lower profitability.

### **5. Actions and recommendations**

Establishing a VIP customer club: Given the large share of top customers, they should be assigned dedicated account managers (from the Employees table) and ensure that they receive personal attention to prevent them from leaving the company.

Frequency Incentive Program: Since we have proven a high correlation to order volume, we recommend introducing benefits such as "free shipping" or tiered discounts from the 10th order per year, to push customers to become regular customers.

Penetration into potential markets: The map shows Brazil (50K) and France (36K) as growth markets. It is recommended to direct targeted marketing budgets to these areas to increase order volume there.

The following code is in Python and executed in Jupyter Notebook

The code is separated into blocks in the order of writing in the notebook.

After each block of code, if there is output, then the output was pasted directly from the notebook.

The code was created in the VS CODE workspace:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

customers = pd.read_csv("Northwind- Customers.csv")
orders = pd.read_csv("Northwind - Orders.csv")
order_details = pd.read_csv("Northwind - Order Details.csv")
products = pd.read_csv("Northwind - Products.csv")
```

```
order_details["total_revenue"] = (order_details["UnitPrice"] *
order_details["Quantity"])
```

```
df_merged = pd.merge(customers, orders, on='CustomerID')
df_merged = pd.merge(df_merged, order_details, on='OrderID')
df_merged = pd.merge(df_merged, products, on='ProductID',
suffixes=('_sold', '_bought'))
```

```
df_merged['TotalRevenue'] = df_merged['UnitPrice_sold'] *
df_merged['Quantity']
```

```
df_merged['TotalGrossProfit'] = ((df_merged['UnitPrice_sold'] -
df_merged['UnitPrice_bought']) * df_merged['Quantity']) * (1 -
df_merged['Discount'])
```

```
customer_analysis = df_merged.groupby(['CustomerID', 'CompanyName']).agg(
    TotalRevenue=('TotalRevenue', 'sum'),
    TotalGrossProfit=('TotalGrossProfit', 'sum'),
    OrderCount=('OrderID', 'nunique')
).reset_index()
```

```
top_5_customers = customer_analysis.sort_values(by='TotalRevenue',
ascending=False).head(5)
top_5_customers
```

output:

	CustomerID	CompanyName	TotalRevenue	TotalGrossProfit	OrderCount
60	QUICK	QUICK-Stop	117483.39	52930.37250	28
68	SAVEA	Save-a-lot Markets	115673.39	50010.30450	31
19	ERNSH	Ernst Handel	113236.68	49193.76175	30
35	HUNGO	Hungry Owl All-Night Grocers	57317.39	23668.08350	19
62	RATTC	Rattlesnake Canyon Grocery	52245.90	22439.05025	18

```
correlation =
customer_analysis['OrderCount'].corr(customer_analysis['TotalGrossProfit'])

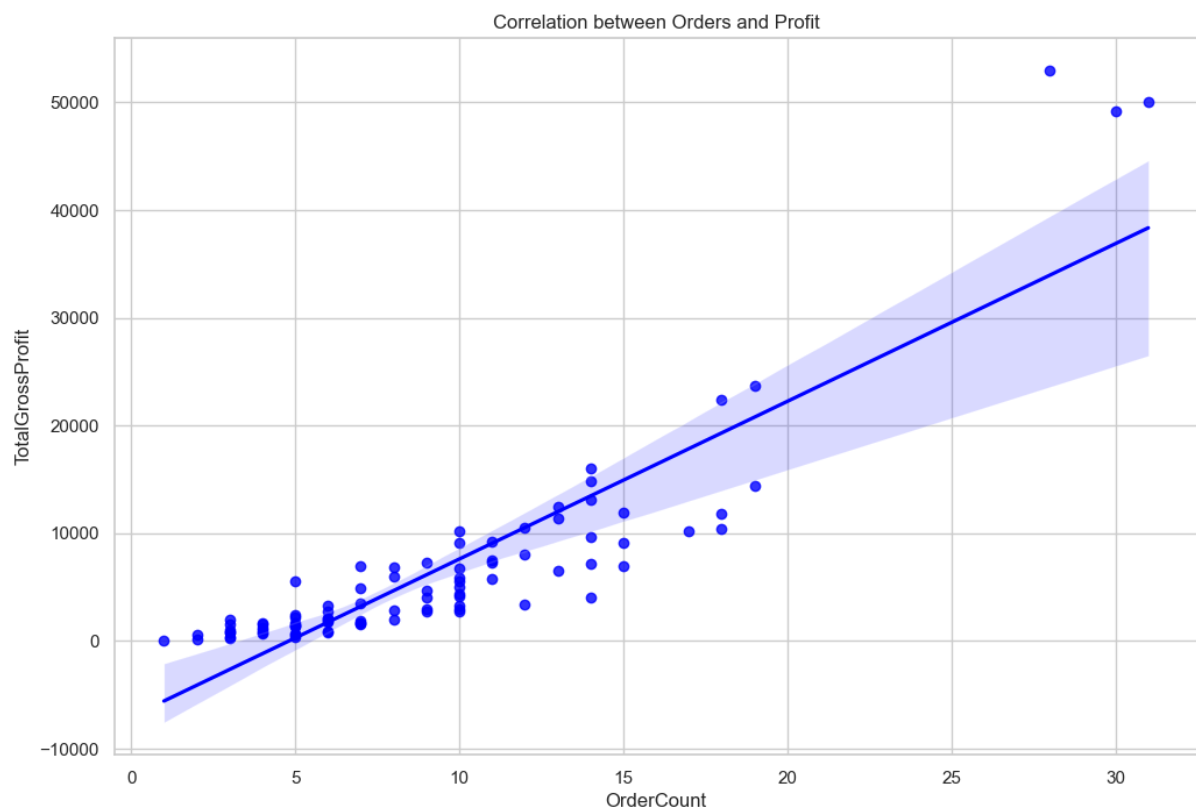
print(f"\nCorrelation between Order Count and Profit: {correlation:.2f}")
```

output:

```
Correlation between Order Count and Profit: 0.89
```

```
sns.set_theme(style="whitegrid")
plt.figure(figsize=(12, 8))
sns.regplot(data=customer_analysis, x = 'OrderCount', and =
'TotalGrossProfit', color='blue')
plt.title('Correlation between Orders and Profit')
```

output:



```
import plotly.express as px

country_profit =
df_merged.groupby('Country')['TotalGrossProfit'].sum().reset_index()

fig = px.choropleth(
    data_frame=country_profit,
    locations="Country",
    locationmode="country names",
    color="TotalGrossProfit",
    hover_name="Country",
    color_continuous_scale=px.colors.sequential.Plasma,
    title='Total Gross Profit by Country')
```

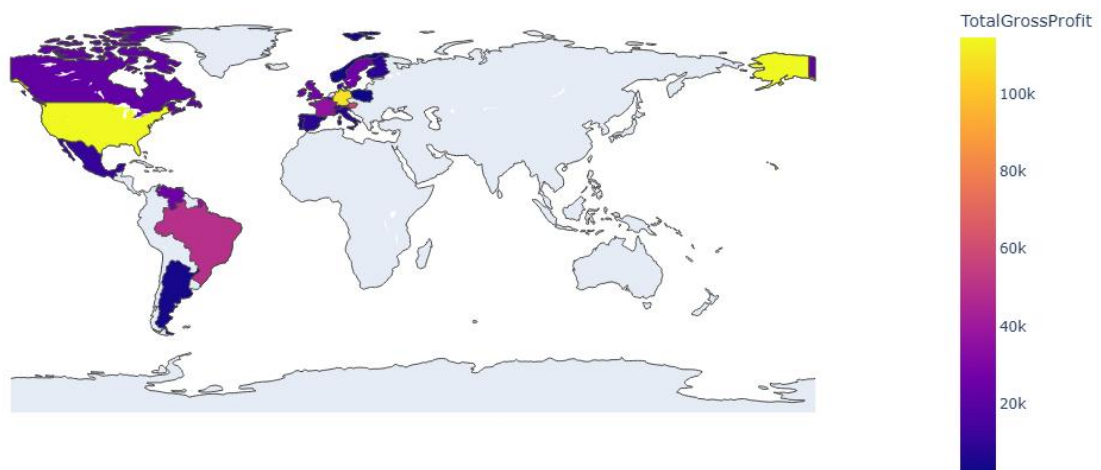
```

fig.update_layout(
    geo=dict(showframe=False, showcoastlines=True,
            projection_type='equiangular'),
            margin={"r":0,"t":40,"l":0,"b":0},
            paper_bgcolor='white',
            plot_bgcolor='white'
    )

```

output:

Total Gross Profit by Country



output:

country_profit		
	Country	TotalGrossProfit
0	Argentina	3963.66000
1	Austria	59367.08175
2	Belgium	15355.14250
3	Brazil	49607.09700
4	Canada	22289.74000
5	Denmark	14246.53725
6	Finland	8496.45625
7	France	36737.16375
8	Germany	109037.08925
9	Ireland	23668.08350
10	Italy	7436.83500
11	Mexico	11046.90875
12	Norway	2734.97500
13	Poland	1708.60000
14	Portugal	5036.05000
15	Spain	8575.45500
16	Sweden	26165.03875
17	Switzerland	15062.22700
18	UK	27416.59250
19	USA	114827.89150
20	Venezuela	26162.93200

**Research Question 3 Topic:** Analyzing human growth drivers and characterizing employee sales excellence

### 1The central question

How can sales excellence be characterized, and opportunities for strengthening employees in weak areas be identified to maximize the company's profitability?

### 2Guiding sub-questions



Who are the top 5 employees in terms of gross profit (Total Profit) they generated for the company?  
Is there a correlation between employees who excel in sales and using more discounts to close a deal? (Identifying employees who maintain high profit margins).

What is the level of specialization of employees or marketing in different product categories, and where are there gaps in professional knowledge?

### 3Work plan and analysis

Tables used: Employees, Orders, Order Details, and Categories.

Visualization tool: Tableau.

How to do it:

Creating a Bar Chart to rank employees by gross profit.

Constructing a Scatter Plot to Examine the Relationship Between Avg. Discount to Sum of Profit.

Heatmap generation that crosses between Employee ID To Category Name, by sales volume/profit.

### 4. Findings and conclusions

Employee number 4 (Margaret Peacock) is the standout, with a total gross profit of about \$232,000, well above the rest of the team.

Smart Discount Strategy: The scatter plot shows that the top employees (4, 3, 1) manage to generate high profits while maintaining a low average discount. In contrast, employee number 7 provides very high discounts but generates significantly low profit, indicating inefficiency in sales and negotiations.

Professional gaps (from the Heatmap analysis):

Works 4 Extra Strong -Beverages and in-Dairy Products.

There are points for improvement, for example, Employee 9 is rarely mentioned in the Produce category.

This gap presents an opportunity for growth through targeted training.

### 5. Actions and recommendations

Margaret Peacock and Janet Loverling are the main growth engine and should be given profit-based bonuses (not just volume) to retain them

It is recommended that Employee 4 (high profit, low discount) provide oversight to employees like Employee 7, to teach them how to close deals without reducing profit margins with high discounts, thereby optimizing and maximizing profits.

Training should be provided for employees in weak categories. For example, employee 4 can train employee 9 in selling beverages and dairy products, and here too, try to get the most out of each category in terms of profits by overlapping, training, and workshops for employees.

The Heatmap is used as a quick tool to identify gaps in the category. Dark colors show strength and light colors indicate opportunities for improvement, so that each employee knows how to target even the weaker areas.



[Link to the full interactive tableau](#)

## Conclusion:

[Link to our presentation](#)- Google Slides

Our comprehensive data analysis of Northwind demonstrates that the company is on an accelerated growth path, but realizing its full potential requires adopting a profitability-focused strategy. The company's long-term success depends on three key pillars: first, operational optimization by clearing "dead" inventory and freeing up untapped capital to invest in growth products; second, retaining top customers through a frequency-based loyalty program, while targeting potential markets such as Brazil and Canada; and finally, upgrading the skills of its human capital and transitioning its sales team to a compensation model that prioritizes gross profit maximization over sales volume alone. Implementing these recommendations will allow Northwind to transform data into actionable insights to streamline, increase profits, and improve the pillars on which the company relies.

tableau<sup>++</sup>public



pandas

matplotlib

plotly

seaborn

