UNIFIED MENTOR

YOUR SKILL, SUCCESS & JOURNEY

Supply Chain Data Analysis: Insights into Efficiency, Costs, and Risks

A **Data Analyst** project by

AMIT BHADE

*Tools Used – **Google Colab, Excel***

*Techniques – **Python, SQL, Machine Learning***

OBJECTIVES

- **Data Exploration:** Gain insights into the supply chain data by exploring production volumes, stock levels, order quantities, revenue, costs, lead times, shipping costs, transportation routes, risks, and sustainability factors.
- **Visualization:** Create informative visualizations to better understand the relationships and distributions within the data.
- **Insights & Recommendations:** Summarize findings into actionable insights to guide supply chain decision-making.

DATASET

Click here to download the dataset

ABOUT DATASET

The dataset contains supply chain records covering products, suppliers, logistics, costs, and quality metrics. It can be used to analyze performance, efficiency, and risks in supply chain operations.

- **Product type**: Category of product (e.g., Skincare, Haircare, Cosmetics).
- **SKU**: Unique product identifier.
- **Price**: Selling price per unit.
- **Availability**: Availability status or quantity of product in stock.
- **Number of products sold**: Total units sold (demand).
- **Revenue generated**: Sales revenue = Price × Units sold.
- **Customer demographics**: Information about customer segment (age, gender, etc.).
- **Stock levels**: Current quantity available in inventory.
- **Lead times**: Time to replenish stock (days).
- **Order quantities**: Number of units ordered by customers.
- **Shipping times**: Delivery duration (days).
- **Shipping carriers**: Logistics providers (e.g., Carrier A, Carrier B).
- **Shipping costs**: Cost incurred in delivering products.
- **Supplier name**: Supplier providing raw materials or products.
- **Location**: Market or regional location (e.g., Mumbai, Kolkata).
- **Lead time**: Procurement lead time (days).
- **Production volumes**: Quantity of goods produced.
- **Manufacturing lead time**: Time taken for manufacturing cycle (days).
- **Manufacturing costs**: Cost of production.
- **Inspection results**: Quality control outcome (Pass/Fail/Pending).
- **Defect rates**: Percentage of defective items.
- **Transportation modes**: Mode of transport (Air, Road, Rail).
- **Routes**: Shipping routes (Route A, Route B, etc.).
- **Costs**: Overall costs (including manufacturing + logistics).

PROJECT SUMMERY

This project work aims on, how to perform an in-depth analysis of supply chain data. The main focus is on key metrics that define the performance and efficiency of supply chain operations, with special attention to products, suppliers, logistics, costs, and risks.

PROJECT STEPS:

1. **Data Overview**

- Dataset structure, missing values, descriptive statistics.
- Basic KPIs such as total revenue, units sold, average price, average costs, lead times, and defect rates.

2. **Exploratory Data Analysis (EDA)**

- Distributions of key measures (revenue, costs, shipping times, defect rates).
- Relationships between variables such as costs vs revenue, price vs manufacturing costs, and stock levels vs lead times.

3. **Business Visualizations**

- Revenue by product type and by location.
- Costs by supplier and inspection results.
- Orders by transportation mode.
- Shipping costs by carrier.
- Production volumes by location.
- Average lead time by product type.
- Transportation route frequency.
- Shipping time distribution by product.
- Supplier performance metrics (lead time, defect rate, costs, efficiency).

4. **Risk and Sustainability**

- Supply chain risk distribution by defect rates and inspection results.
- A proxy sustainability view using transportation modes (Air, Road, Rail) to approximate emissions impact.

5. **Insights & Recommendations**

- Clear takeaways from the analysis.
- Practical recommendations for improving supplier quality, reducing costs, optimizing routes, and improving sustainability.

STEP BY STEP PROJECT IMPLEMENTATION

Step 1: Import Libraries

```
import pandas as pd
import numpy as np
```

Step 2: Load Dataset

```
from google.colab import files
uploaded = files.upload()
df = pd.read_csv('supply_chain_data.csv')
print("Shape:", df.shape)
```

Step 3: Overview

```
print(df.info())
print("\nMissing values per column:\n", df.isna().sum())
df.describe(include='all').T
```

Step 4: KPIs

```
def kpi(x, digits=2):
    return f"{x:,.{digits}f}" if isinstance(x, float) else f"{x:,}"

kpis = {
    "Total Revenue": df['Revenue generated'].sum(),
    "Total Units Sold": df['Number of products sold'].sum(),
    "Avg Price": df['Price'].mean(),
    "Avg Shipping Cost": df['Shipping costs'].mean(),
    "Avg Lead Time": df['Lead time'].mean(),
    "Avg Defect Rate": df['Defect rates'].mean()
}
for k, v in kpis.items():
    print(f"{k:>20}: {kpi(v)}")
```

Step 5: Distributions

```
# Products sold
df['Number of products sold'].hist(bins=20)
plt.xlabel('Number of products sold'); plt.ylabel('Frequency'); plt.title('Distribution:
Products Sold')
plt.show()

# Revenue
df['Revenue generated'].hist(bins=20)
plt.xlabel('Revenue'); plt.ylabel('Frequency'); plt.title('Distribution: Revenue')
plt.show()

# Shipping costs
df['Shipping costs'].hist(bins=20)
plt.xlabel('Shipping costs'); plt.ylabel('Frequency'); plt.title('Distribution: Shipping Costs')
plt.show()

# Defect rates
df['Defect rates'].hist(bins=20)
plt.xlabel('Defect rate'); plt.ylabel('Frequency'); plt.title('Distribution: Defect Rates')
plt.show()
```

Step 6: Relationships

```
# Price vs Manufacturing costs
plt.scatter(df['Price'], df['Manufacturing costs'])
plt.xlabel('Price'); plt.ylabel('Manufacturing costs'); plt.title('Price vs Manufacturing Costs')
plt.show()

# Costs vs Revenue
corr = df[['Costs','Revenue generated']].corr().iloc[0,1]
print(f"Correlation between Costs & Revenue: {corr:.3f}")
plt.scatter(df['Costs'], df['Revenue generated'])
plt.xlabel('Costs'); plt.ylabel('Revenue'); plt.title('Costs vs Revenue')
plt.show()

# Stock vs Lead Times
plt.scatter(df['Stock levels'], df['Lead times'])
plt.xlabel('Stock levels'); plt.ylabel('Lead times'); plt.title('Stock Levels vs Lead Times')
plt.show()
```

Step 7: Business Grouped Charts

7.1: Revenue by product type

```
rev_by_pt = df.groupby('Product type')['Revenue
generated'].sum().sort_values(ascending=False)
print(rev_by_pt)
rev_by_pt.plot(kind='bar', title='Revenue by Product Type')
plt.ylabel('Total revenue'); plt.show()
```

7.2: Cost by inspection result

```
cost_by_insp = df.groupby('Inspection
results')['Costs'].mean().sort_values(ascending=False)
print(cost_by_insp)
cost_by_insp.plot(kind='bar', title='Avg Cost by Inspection Result')
plt.ylabel('Average cost'); plt.show()
```

7.3: Cost by supplier

```
cost_by_supplier = df.groupby('Supplier
name')['Costs'].mean().sort_values(ascending=False)
print(cost_by_supplier)
cost_by_supplier.plot(kind='bar', title='Avg Cost by Supplier')
plt.ylabel('Average cost'); plt.show()
```

7.4: Orders by transport mode

```
orders_by_mode = df.groupby('Transportation modes')['Order
quantities'].sum().sort_values(ascending=False)
print(orders_by_mode)
orders_by_mode.plot(kind='bar', title='Orders by Transport Mode')
plt.ylabel('Total orders'); plt.show()
```

7.5: Average defect rate by product

```
defect_by_prod = df.groupby('Product type')['Defect
rates'].mean().sort_values(ascending=False)
print(defect_by_prod)
defect_by_prod.plot(kind='bar', title='Avg Defect Rate by Product')
plt.ylabel('Avg defect rate'); plt.show()
```

7.6: Cost efficiency by supplier (Revenue / Cost)

```
agg = df.groupby('Supplier name').agg(total_rev=('Revenue generated','sum'),
                                      total_cost=('Costs','sum'))
agg['efficiency'] = agg['total_rev'] / agg['total_cost']
print(agg.sort_values('efficiency', ascending=False))
agg['efficiency'].sort_values(ascending=False).plot(kind='bar', title='Cost Efficiency by
Supplier')
plt.ylabel('Revenue / Cost'); plt.show()
```

7.7: Average lead time by product

```
lead_by_prod = df.groupby('Product type')['Lead
time'].mean().sort_values(ascending=False)
print(lead_by_prod)
lead_by_prod.plot(kind='bar', title='Avg Lead Time by Product')
plt.ylabel('Lead time'); plt.show()
```

5.8 Transportation route frequency

```
route_freq = df['Routes'].value_counts()
print(route_freq)
route_freq.plot(kind='bar', title='Transportation Route Frequency')
plt.ylabel('Count'); plt.show()
```

5.9 Shipping cost by carrier

```
ship_cost_carrier = df.groupby('Shipping carriers')['Shipping
costs'].mean().sort_values(ascending=False)
print(ship_cost_carrier)
ship_cost_carrier.plot(kind='bar', title='Avg Shipping Cost by Carrier')
plt.ylabel('Avg shipping cost'); plt.show()
```

5.10 Production by location

```
prod_by_loc = df.groupby('Location')['Production
volumes'].sum().sort_values(ascending=False)
print(prod_by_loc)
prod_by_loc.plot(kind='bar', title='Production by Location')
plt.ylabel('Production volume'); plt.show()
```

5.11 Shipping time distribution by product

```
groups = [g['Shipping times'].dropna().values for _, g in df.groupby('Product type')]
labels = df['Product type'].unique()
plt.boxplot(groups, labels=labels)
plt.xlabel('Product type'); plt.ylabel('Shipping time'); plt.title('Shipping Time by Product')
plt.show()
```

Step 8: Risk & Supplier Performance

```
# Defect rate by supplier
defect_by_supplier = df.groupby('Supplier name')['Defect
rates'].mean().sort_values(ascending=False)
print(defect_by_supplier)
defect_by_supplier.plot(kind='bar', title='Avg Defect Rate by Supplier')
plt.ylabel('Avg defect rate'); plt.show()

# Inspection result distribution
insp_dist = df['Inspection results'].value_counts(normalize=True) * 100
print(insp_dist)
insp_dist.plot(kind='bar', title='Inspection Results %')
plt.ylabel('%'); plt.show()

# Supplier performance matrix
perf = df.groupby('Supplier name').agg(
    avg_lead_time=('Lead time','mean'),
    avg_defect_rate=('Defect rates','mean'),
    avg_cost=('Costs','mean'),
    avg_shipping_time=('Shipping times','mean')
)
print(perf)
```

Step 9: Sustainability Proxy

```
emission_factor = {'Air': 3.0, 'Road': 1.5, 'Rail': 1.0}
df['EmissionsIndex'] = df['Transportation modes'].map(emission_factor).fillna(1.5)
df['EmissionsImpact'] = df['EmissionsIndex'] * df['Order quantities']

emis_by_mode = df.groupby('Transportation
modes')['EmissionsImpact'].mean().sort_values(ascending=False)
print(emis_by_mode)
emis_by_mode.plot(kind='bar', title='Sustainability Proxy by Transport Mode')
plt.ylabel('Emissions impact (proxy)'); plt.show()
```

KEY INSIGHTS

- Skincare is top revenue driver.
- Kolkata and Mumbai lead in production and orders.
- Carrier B is fastest, Carrier A costlier.
- Supplier 5 has highest defects and lead time; Supplier 1 best overall.
- Route B is costliest; Routes A/C cheaper.
- Air transport has highest emissions proxy, Rail lowest.

CONCLUSION

What we did

- Explored the Supply Chain dataset (100 records, 24 features).
- Conducted data cleaning checks (dtypes, missing values, duplicates).
- Computed KPIs: total revenue, units sold, avg price, avg shipping cost, avg lead time, avg defect rate.
- Performed Exploratory Data Analysis (EDA): distributions, relationships (cost vs revenue, price vs manufacturing cost, stock vs lead time).
- Created business visualizations:
  • Revenue by product type
  • Costs by inspection result & supplier
  • Orders by transport mode
  • Shipping costs by carrier
  • Production by location
  • Transportation route frequency
  • Avg defect rate & lead time by product
  • Supplier performance metrics
- Added risk analysis (inspection results, defect rates).
- Included a sustainability proxy using transport modes (Air, Road, Rail).
- Summarized findings into actionable insights & recommendations.

Key Insights

- **Products:** Skincare is the top revenue driver; cosmetics show higher defect rates.
- **Locations:** Mumbai & Kolkata dominate production and orders.
- **Suppliers:** Supplier 5 has highest defect rate and longest lead times; Supplier 1 is most efficient.
- **Logistics:** Carrier B delivers faster and cheaper; Route B is costlier than Routes A/C.
- **Costs & Revenue:** Positively correlated ($r \approx 0.7$).
- **Sustainability:** Air transport is fastest but least sustainable (proxy index).

Limitations

- **Small dataset (100 rows)** → not representative of real operations.
- **Cross-sectional snapshot** → no time series (no seasonality or trend analysis).
- **Sustainability proxy** → based only on assumed emission factors (not real CO_2 data).

Future Work

- Collect time-stamped data to analyze seasonality and forecast demand.
- Add supplier financials, distance, and emissions data for more accurate cost and sustainability analysis.
- Deploy an interactive dashboard (Streamlit / Tableau / Power BI) for real-time monitoring.
- Incorporate optimization models (e.g., route or supplier selection).

Business Impact

- Helps identify top revenue drivers (products, locations).
- Highlights supplier risks (defects, delays, costs).
- Supports logistics optimization (choosing best carriers, transport modes, routes).
- Frames sustainability discussions (shift volume away from air freight).
- Provides a foundation for cost reduction and service level improvement.

REFERENCE