SUPPLY CHAIN DATASET ANALYSIS

Graduation Project for DEPi

Presented by : Sherif Mahmoud - Polus Fayez

Gehad Naser - Mohamed Ibrahim

Moamen Ahmed - Basem Ibrahim

Project Overview

OBJECTIE

Analyze a supply chain dataset to uncover insights and visualize key findings

PROCESS OVERVIEW

Data Cleaning and Preprocessing using Python

Analyzing data to answer key business questions

Visualizing insights with Tableau

TOOLS USED

Python (Pandas, Matplotlib), Tableau

Data Cleaning & Preprocessing

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt # for visualization
import plotly.express as px
import seaborn as sns
from scipy.stats import chi2_contingency
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor, GradientBoo
from statsmodels.tsa.arima.model import ARIMA
from sklearn.neural_network import MLPRegressor
from sklearn.metrics import mean_squared_error, mean_absolute_e
from sklearn.preprocessing import OneHotEncoder
pd.set_option('display.max_columns', None)
```

```
# Read CSV file into DataFrame
df = pd.read_csv("./supply_chain_data.csv")
```

```
df.head()
```

	Product type	SKU	Price	Availability	Number of products sold	Revenue generated	Customer demographics
0	haircare	SKU0	69.808006	55	802	8661.996792	Non-binary
1	skincare	SKU1	14.843523	95	736	7460.900065	Female
2	haircare	SKU2	11.319683	34	8	9577.749626	Unknown
3	skincare	SKU3	61.163343	68	83	7766.836426	Non-binary
4	skincare	SKU4	4.805496	26	871	2686.505152	Non-binary

Data Cleaning & Preprocessing

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 24 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Product type             100 non-null    object
 1   SKU                      100 non-null    object
 2   Price                    100 non-null    float64
 3   Availability             100 non-null    int64
 4   Number of products sold  100 non-null    int64
 5   Revenue generated        100 non-null    float64
 6   Customer demographics    100 non-null    object
 7   Stock levels             100 non-null    int64
 8   Lead times               100 non-null    int64
 9   Order quantities         100 non-null    int64
 10  Shipping times           100 non-null    int64
 11  Shipping carriers        100 non-null    object
 12  Shipping costs           100 non-null    float64
 13  Supplier name            100 non-null    object
 14  Location                 100 non-null    object
 15  Lead time                100 non-null    int64
 16  Production volumes       100 non-null    int64
 17  Manufacturing lead time  100 non-null    int64
 18  Manufacturing costs      100 non-null    float64
 19  Inspection results       100 non-null    object
 20  Defect rates             100 non-null    float64
 21  Transportation modes     100 non-null    object
 22  Routes                   100 non-null    object
 23  Costs                    100 non-null    float64
```

```
RangeIndex: 100 entries, 0 to 99
Data columns (total 24 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Product type             100 non-null    category
 1   SKU                      100 non-null    category
 2   Price                    100 non-null    float64
 3   Availability             100 non-null    int64
 4   Number of products sold  100 non-null    int64
 5   Revenue generated        100 non-null    float64
 6   Customer demographics    100 non-null    category
 7   Stock levels             100 non-null    int64
 8   Lead times               100 non-null    int64
 9   Order quantities         100 non-null    int64
 10  Shipping times           100 non-null    int64
 11  Shipping carriers        100 non-null    category
 12  Shipping costs           100 non-null    float64
 13  Supplier name            100 non-null    category
 14  Location                 100 non-null    category
 15  Lead time                100 non-null    int64
 16  Production volumes       100 non-null    int64
 17  Manufacturing lead time  100 non-null    int64
 18  Manufacturing costs      100 non-null    float64
 19  Inspection results       100 non-null    category
 20  Defect rates             100 non-null    float64
 21  Transportation modes     100 non-null    category
 22  Routes                   100 non-null    category
 23  Costs                    100 non-null    float64
dtypes: category(9), float64(6), int64(9)
```

Data Cleaning & Preprocessing

- Check for duplicates

Check for duplicate data

```
if df.duplicated().any():
    print(f"There are as many as {df.duplicated().sum()} duplicate data.")
else:
    print("There are no duplicate data.")
```

```
There are no duplicate data.
```

- Check for negative values

Check for negative values

```
df.describe()
# from min row no negative
```

	Price	Availability	Number of products sold	Revenue generated	Stock levels	Lead times	Order quantities	Shipping times
count	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000
mean	49.462461	48.400000	460.990000	5776.048187	47.770000	15.960000	49.220000	5.750000
std	31.168193	30.743317	303.780074	2732.841744	31.369372	8.785801	26.784429	2.724283
min	1.699976	1.000000	8.000000	1061.618523	0.000000	1.000000	1.000000	1.000000
25%	19.597823	22.750000	184.250000	2812.847151	16.750000	8.000000	26.000000	3.750000
50%	51.239831	43.500000	392.500000	6006.352023	47.500000	17.000000	52.000000	6.000000
75%	77.198228	75.000000	704.250000	8253.976921	73.000000	24.000000	71.250000	8.000000
max	99.171329	100.000000	996.000000	9866.465458	100.000000	30.000000	96.000000	10.000000

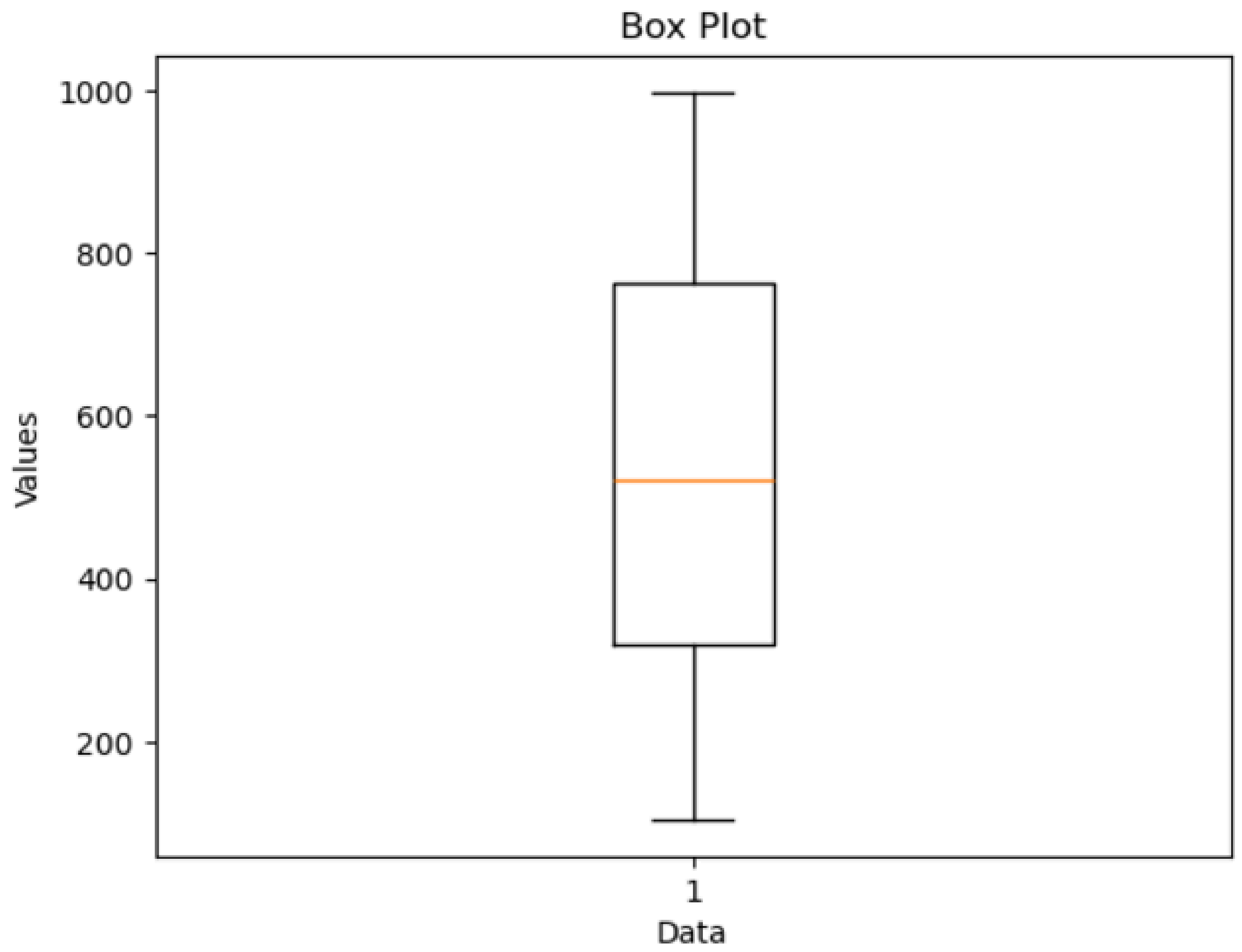
Data Cleaning & Preprocessing

- Detect Outlier

```
#Outlier detection
flag=0
for column in df.columns:
    if df[column].dtype=="int64" or df[column].dtype =="float64":
        max_value = df[column].max()
        Q1 = df[column].quantile(0.25)
        Q3 = df[column].quantile(0.75)
        IQR = Q3 - Q1
        outlier_threshold = Q3 + 1.5 * IQR
        if max_value > outlier_threshold :
            print(f"{column} has an outlier: {max_value}")
            flag=1
if flag==0:
     print("There is No Outlier")
```

```
There is No Outlier
```

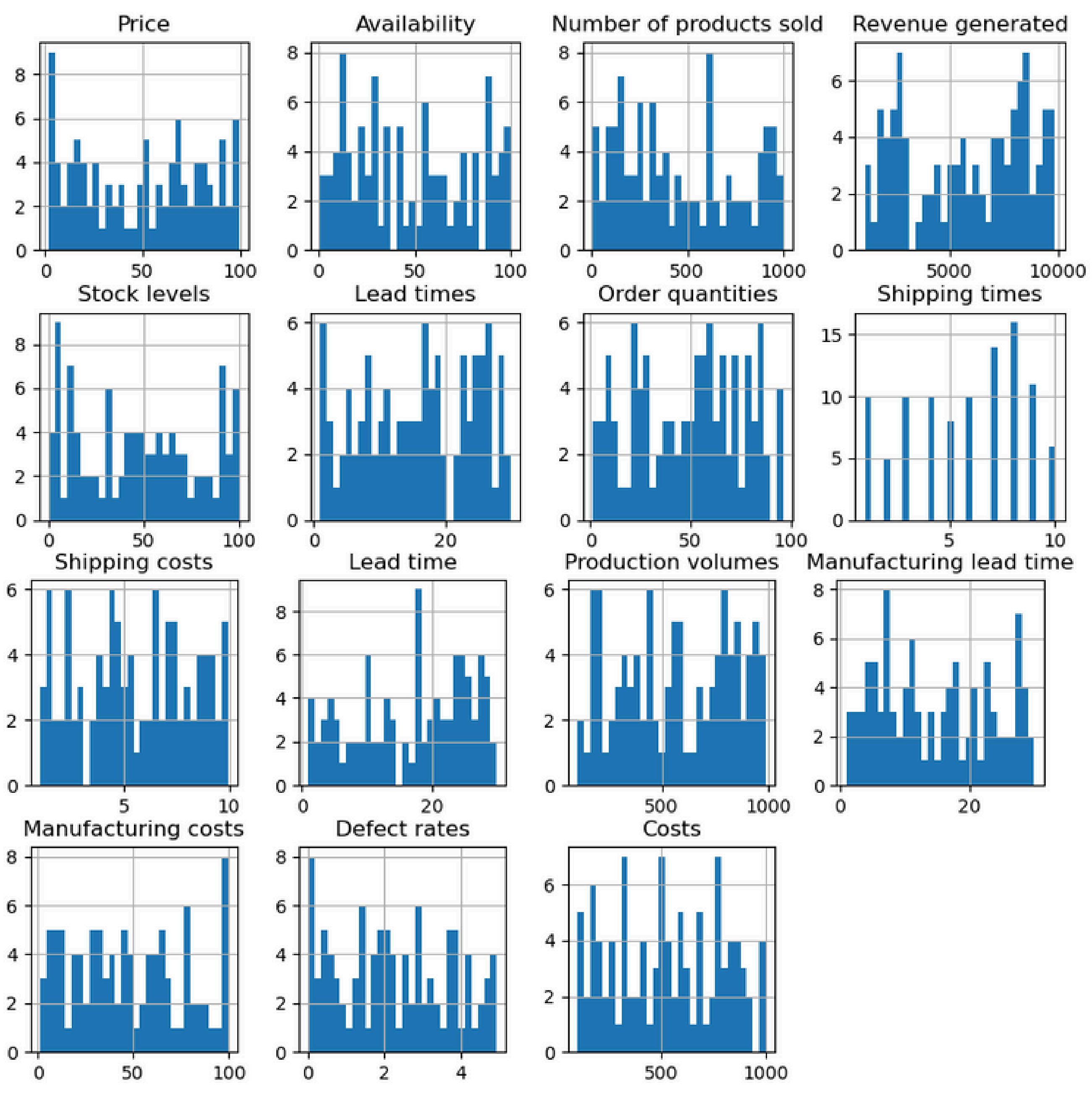
```
plt.boxplot(df["Costs"])
plt.xlabel('Data')
plt.ylabel('Values')
plt.title('Box Plot')
plt.show()
```



Data Cleaning & Preprocessing

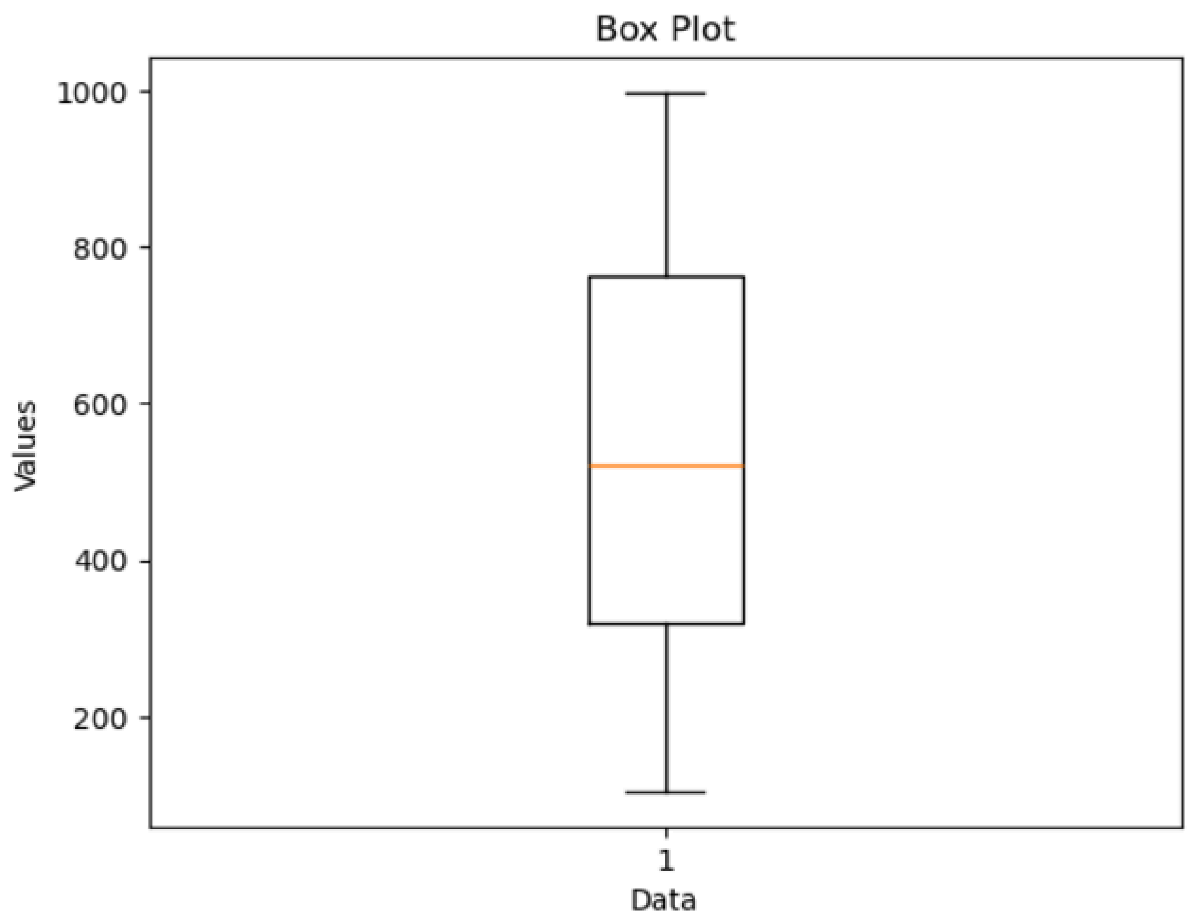
- Conduct Univariate Analysis (box plot & histograms)

```
df.hist(figsize=(10,10), bins=30);
```



From the code and the charts clearly that There is No Outlier

```
plt.boxplot(df["Costs"])
plt.xlabel('Data')
plt.ylabel('Values')
plt.title('Box Plot')
plt.show()
```

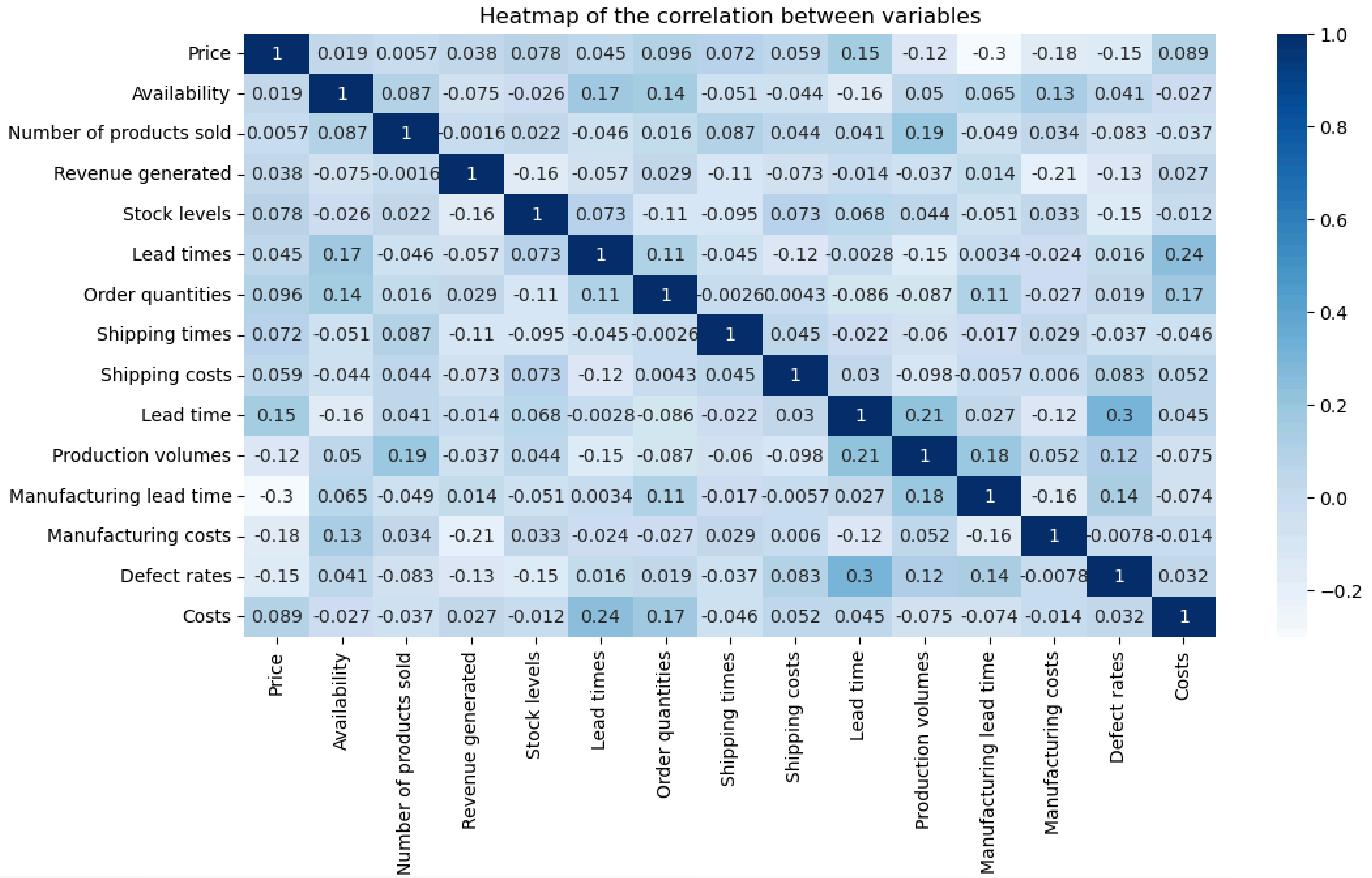


Data Cleaning & Preprocessing

- Check Correlation

There is no Correlation except
- a week one between Defect rates and Lead time
- a week negative one between Price and Manufacturing costs

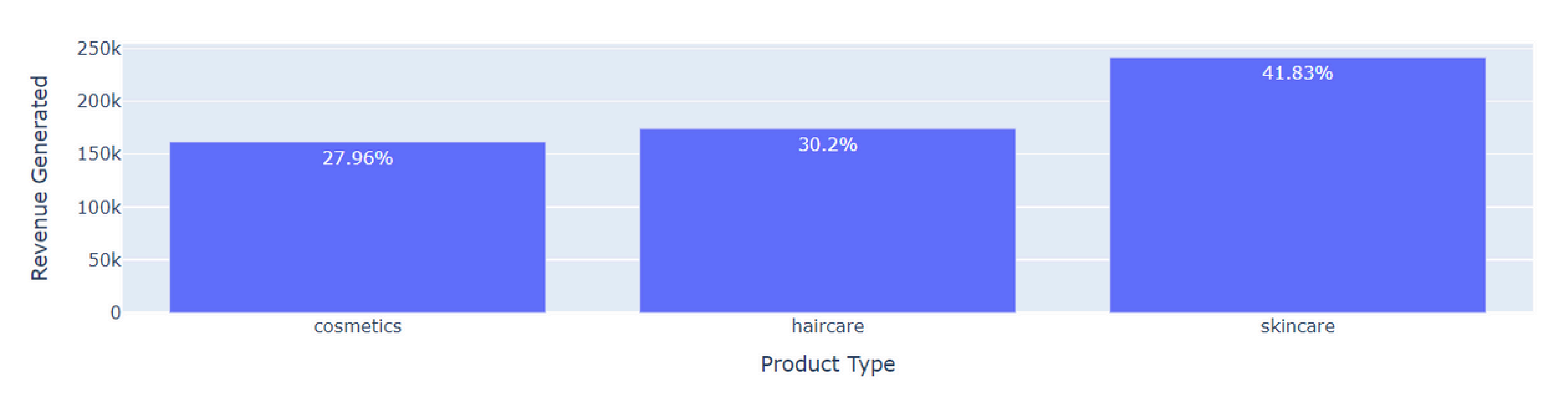
```
corr = df.corr(numeric_only=True)
plt.figure(figsize=(12,6))
sns.heatmap(corr, annot=True, cmap='Blues')
plt.title('Heatmap of the correlation between variables')
plt.show()
```

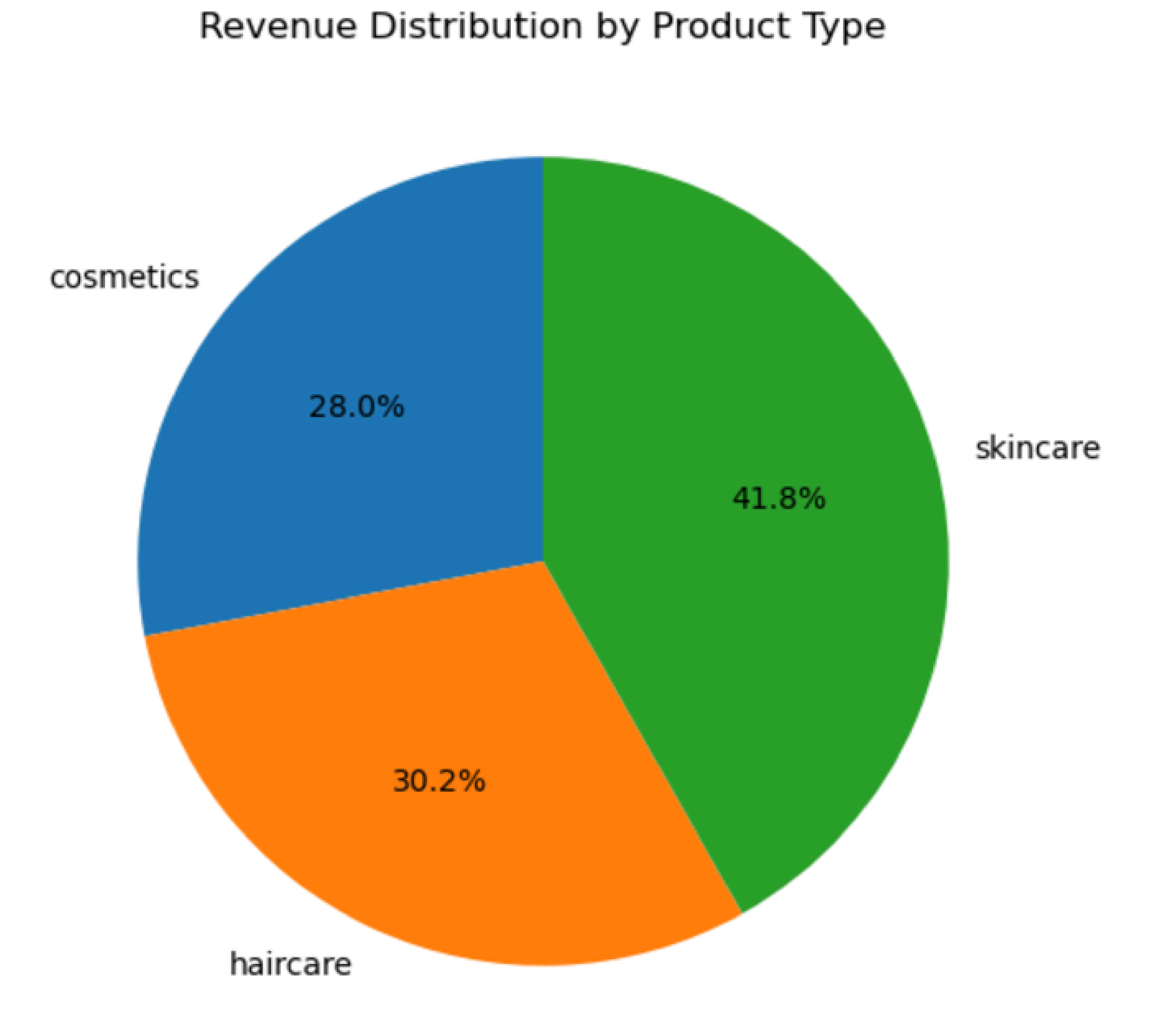


Data Analysis Questions

1- What is the impact of Product Category on Revenue?

```python
data = df.groupby('Product type', observed=True) ['Revenue generated'].sum()
#create the pie chart
plt.figure(figsize = (6, 6))
plt.pie(data, labels=data.index,autopct='%1.1f%%', startangle=90)
plt.title('Revenue Distribution by Product Type')
plt.show()
```

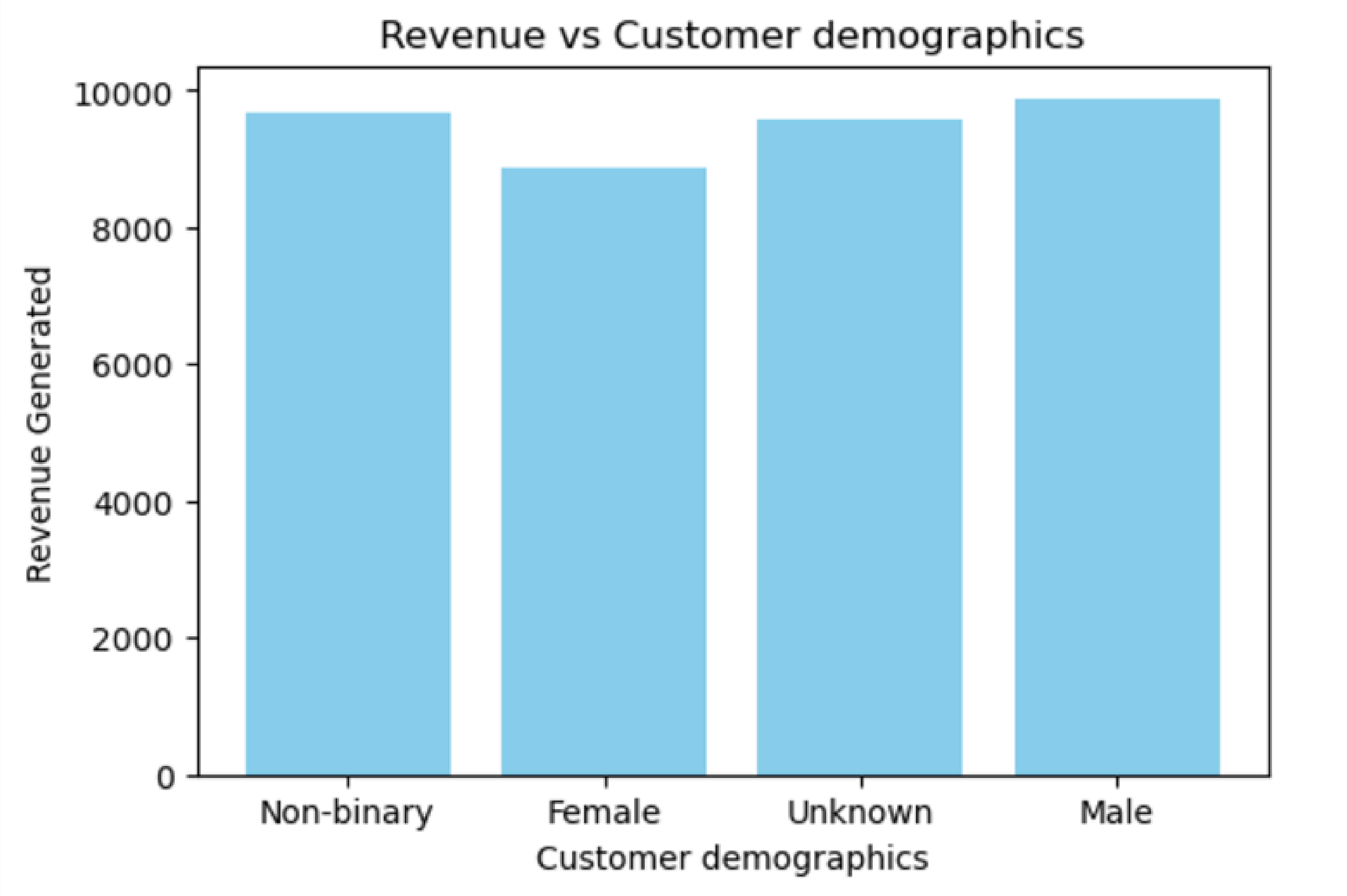




Data Analysis Questions

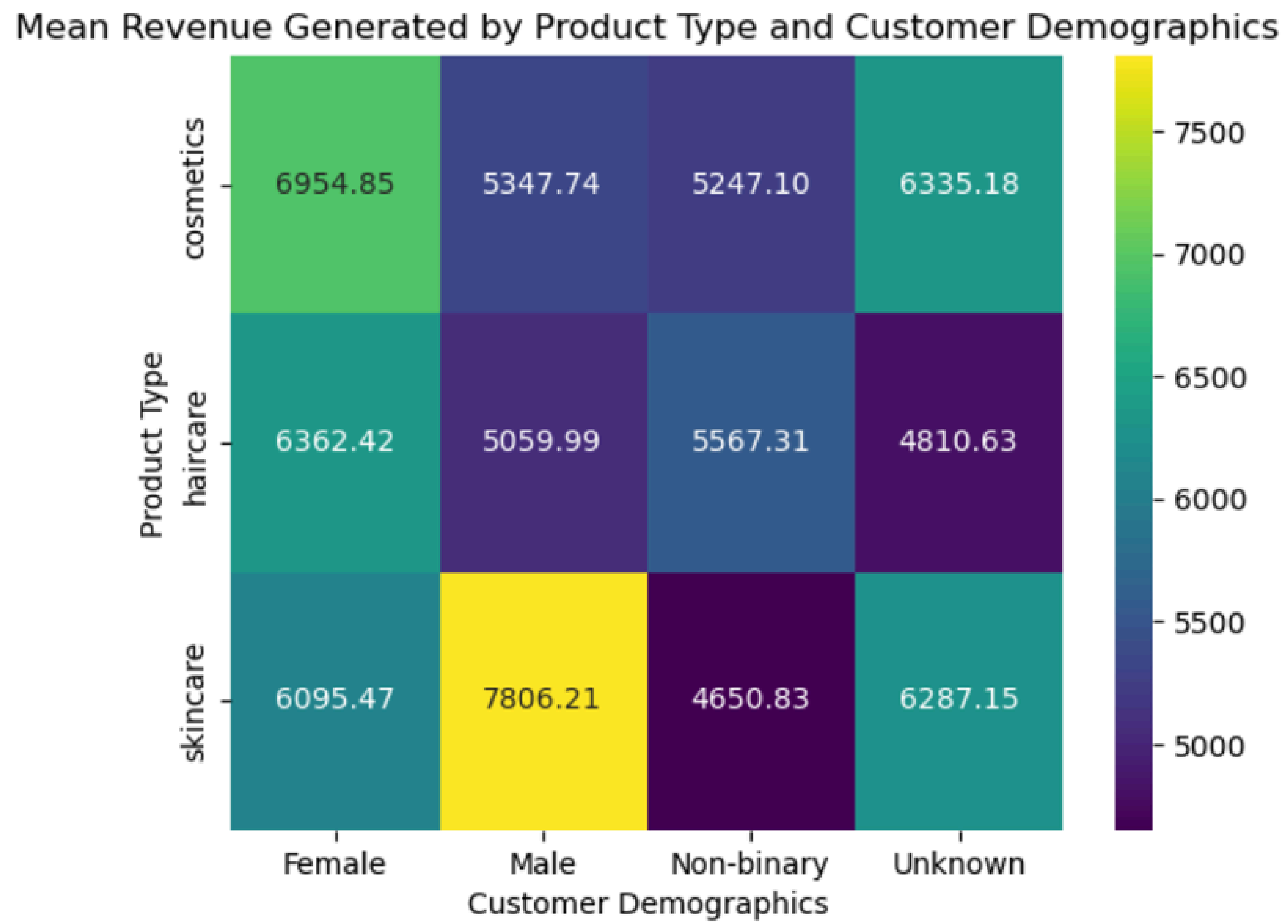
2- How do customers demographics influence purchasing behaviour?

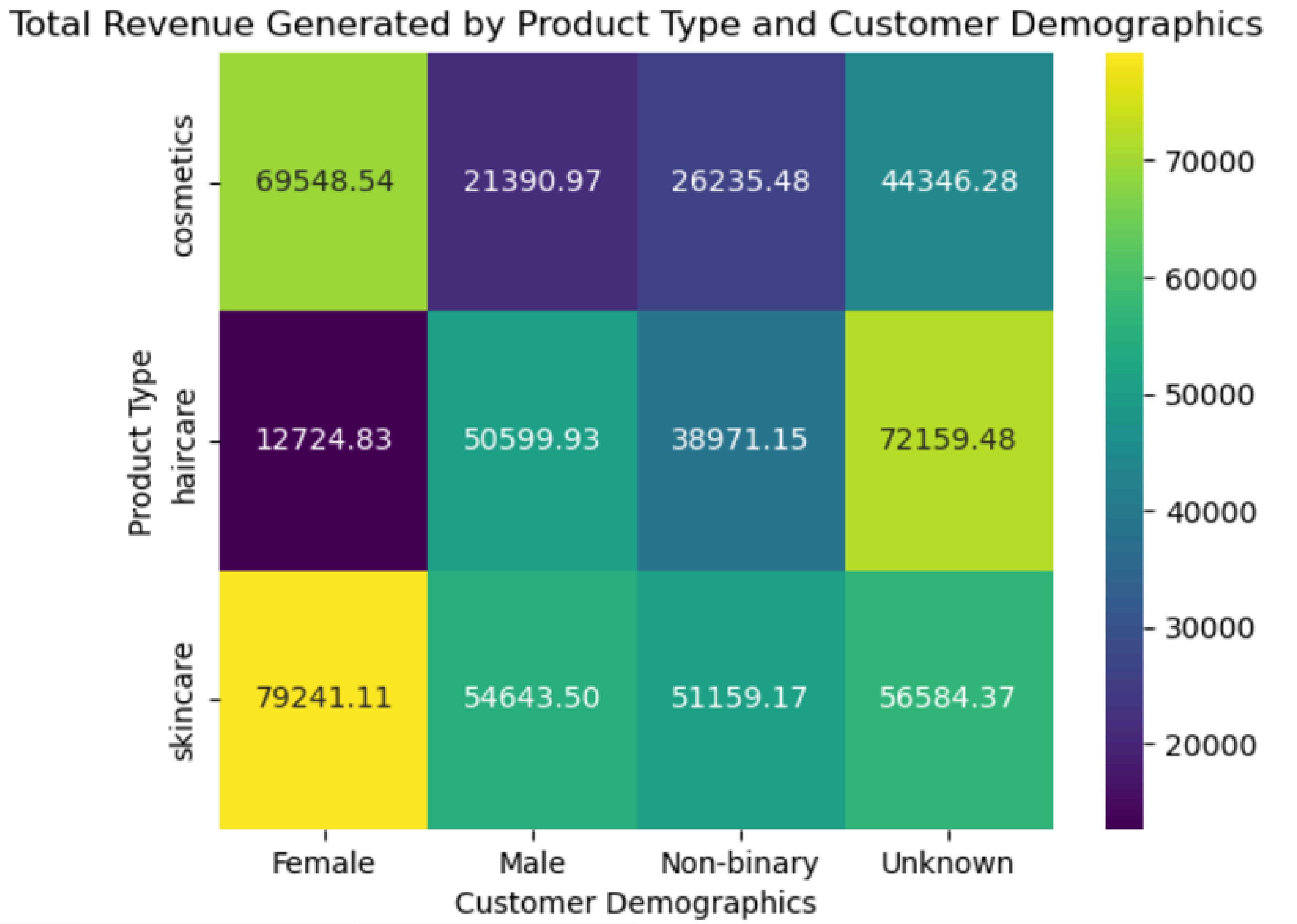
```
plt.figure(figsize=(6, 4))  # Set the size of the chart
plt.bar(df['Customer demographics'], df['Revenue generated'], color='skyblue')
# Add labels and title
plt.xlabel('Customer demographics')  # X-axis label
plt.ylabel('Revenue Generated')        # Y-axis label
plt.title('Revenue vs Customer demographics')  # Title
# Display the chart
plt.show()
```



Data Analysis Questions

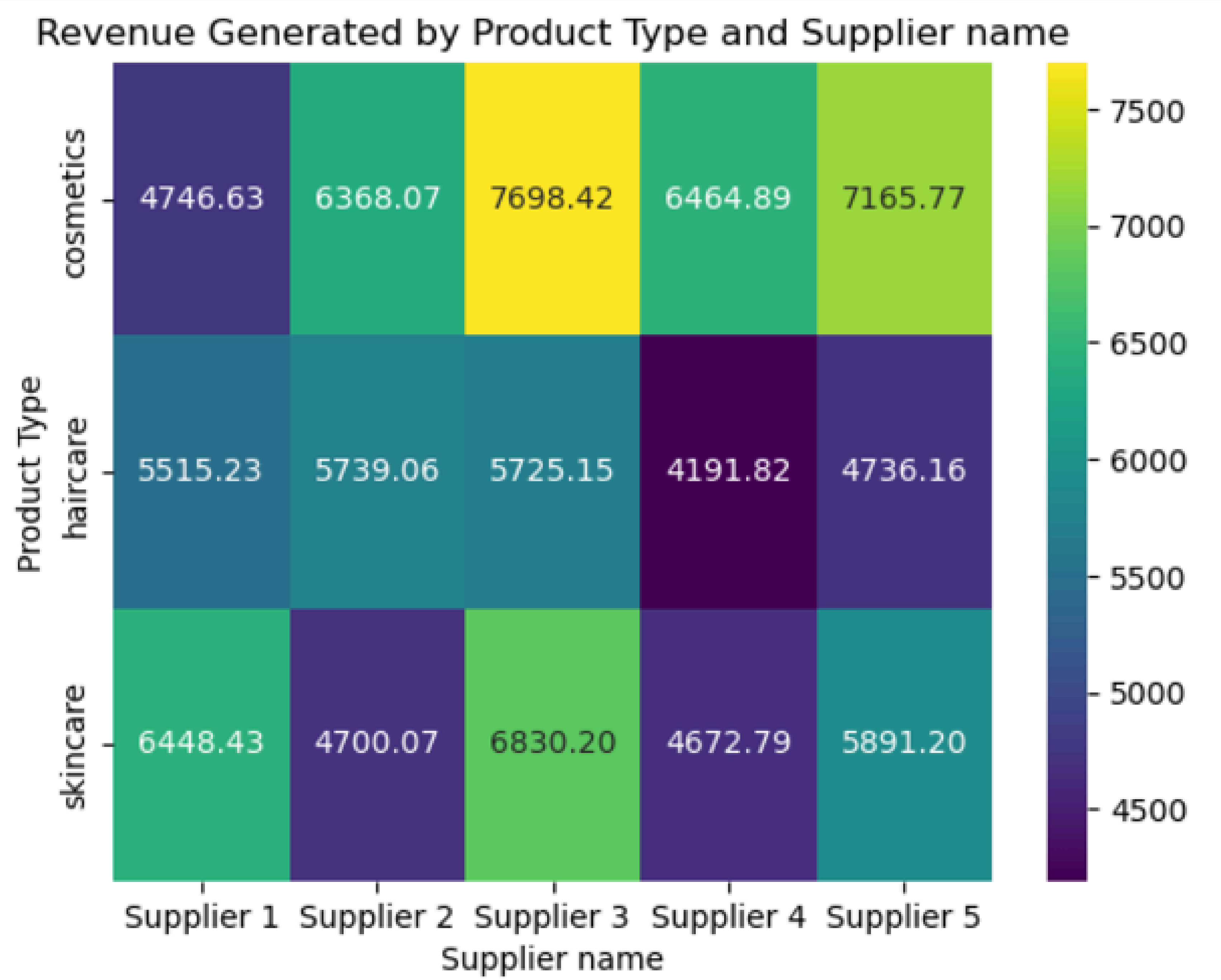
3- How do customers demographics influence purchasing behavior of different Product types?





Data Analysis Questions

4- How do Supplier influence purchasing behavior of different Product types?



Data Analysis Questions

5- What is the Impact of Transportation modes on Manufacturing lead time?

```
revenue_by_product = df.groupby('Transportation modes', observed=True)['Manufacturing lead time'].sum().reset_index()
# Calculate the total revenue
total_revenue = revenue_by_product['Manufacturing lead time'].sum()

# Calculate the percentage of total revenue for each product type
revenue_by_product['Percent of Total'] = ((revenue_by_product['Manufacturing lead time'] / total_revenue) * 100).round(2).astype(str) + '%'

fig = px.bar(revenue_by_product, x='Transportation modes', y='Manufacturing lead time', title='Impact of Transportation modes on Manufacturing lead time'
fig.update_xaxes(title_text='Transportation modes')  # Update x-axis label
fig.update_yaxes(title_text='Manufacturing lead time')  # Update y-axis label
fig.show()
```

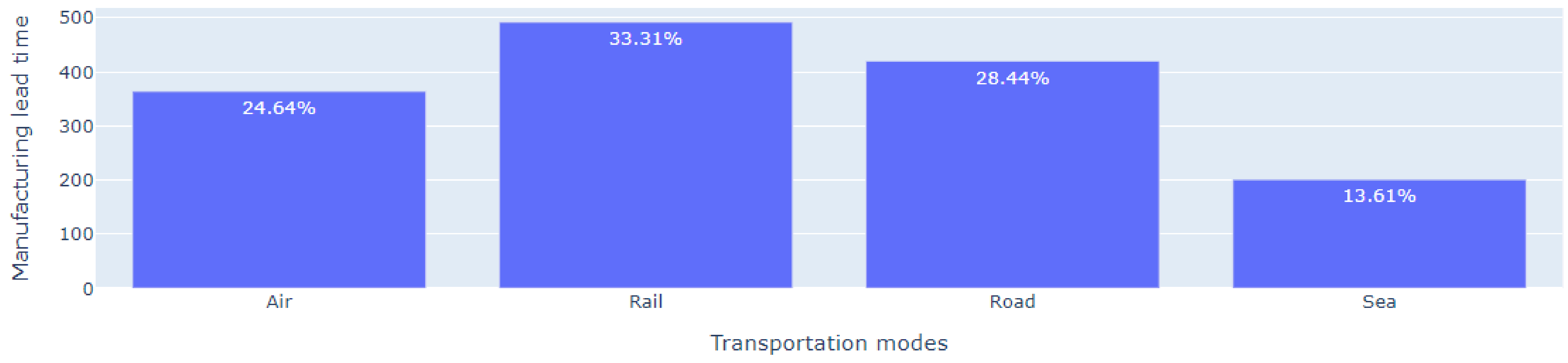


Tableau Dashboard

Question & Answer

THANK YOU