Sports clothing and athleisure attire is a huge industry, worth approximately $193 billion in 2021 with a strong growth forecast over the next decade!
(
`Source: https://www.statista.com/statistics/254489/total-revenue-of-the-global-sport`
)

In this workbook, you will undertake the role of a product analyst for an online sports clothing company. The company is specifically interested in how it can improve revenue. You will dive into product data such as pricing, reviews, descriptions, and ratings, as well as revenue and website traffic, to produce recommendations for its marketing and sales teams.

The data:

You've been provided with four datasets to investigate:

`brands.csv`

Columns	Description
`product_id`	Unique product identifier
`brand`	Brand of the product

`finance.csv`

Columns	Description
`product_id`	Unique product identifier
`listing_price`	Original price of the product
`sale_price`	Discounted price of the product
`discount`	Discount off the listing price, as a decimal
`revenue`	Revenue generated by the product

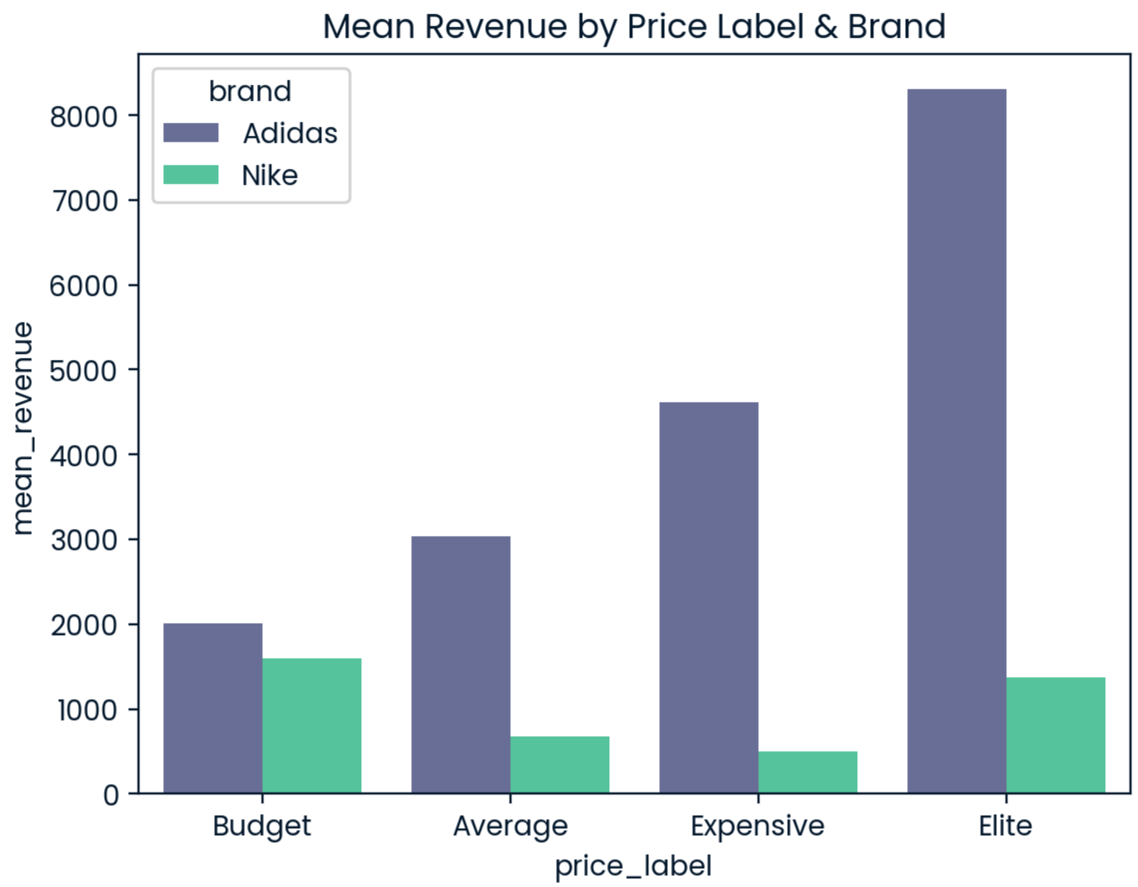
`info.csv`

Columns	Description
`product_name`	Name of the product
`product_id`	Unique product identifier
`description`	Description of the product

`reviews.csv`

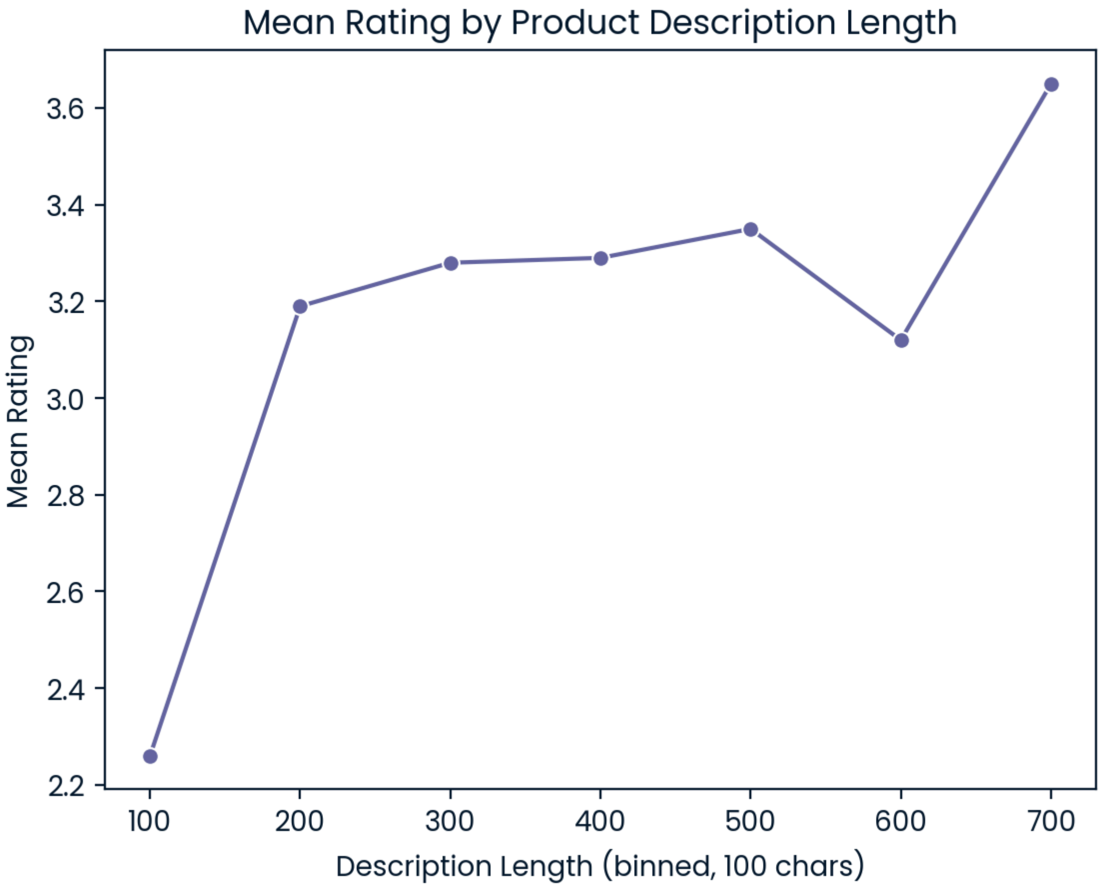
Columns	Description
`product_id`	Unique product identifier
`rating`	Average product rating
`reviews`	Number of reviews for the product

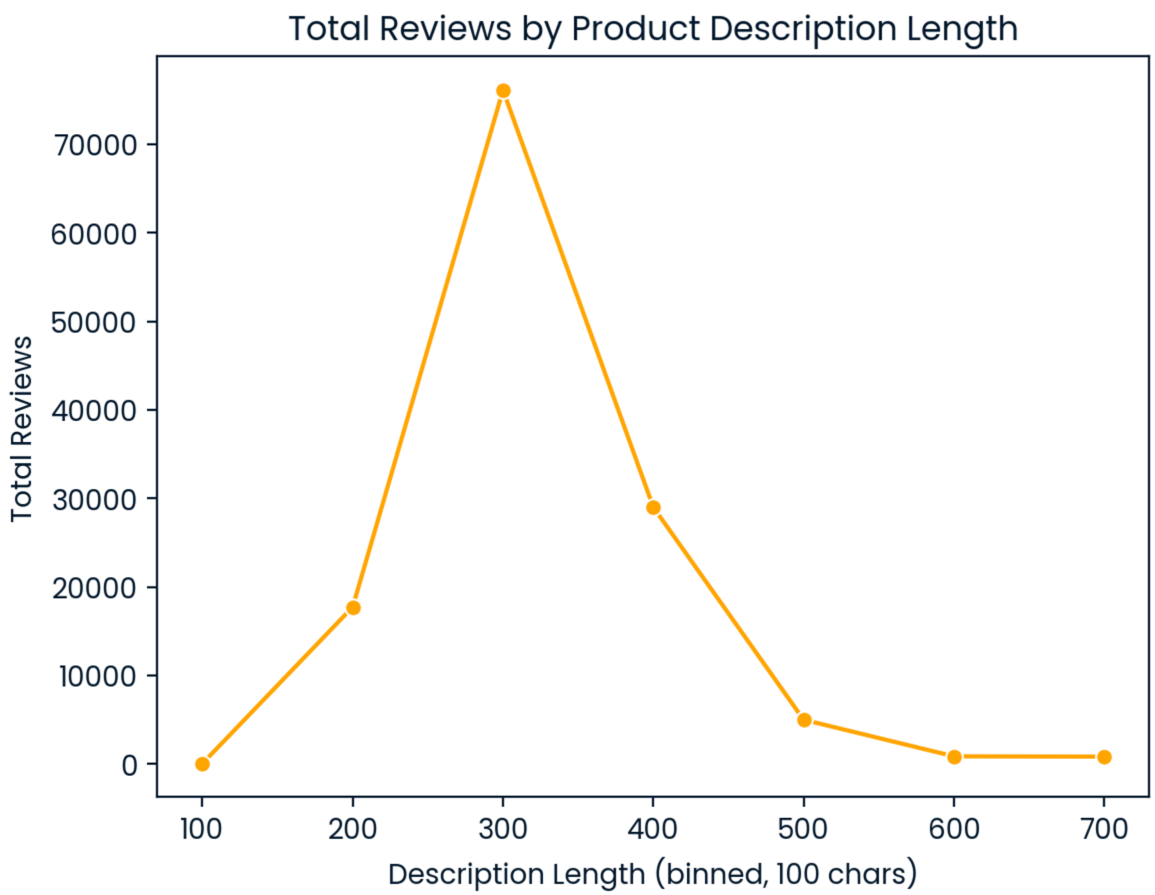
```python
# Importing libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Loading the data
brands = pd.read_csv("brands.csv")
finance = pd.read_csv("finance.csv")
info = pd.read_csv("info.csv")
reviews = pd.read_csv("reviews.csv")

# Merge datasets
df = pd.merge(brands, finance, on='product_id')
df = pd.merge(df, info, on='product_id')
df = pd.merge(df, reviews, on='product_id')

# Drop null values
df.dropna(inplace=True)

# volume of products and average revenue by pricing quartile
df["price_label"] = pd.qcut(df["listing_price"], q=4, labels=['Budget', 'Average', 
'Expensive', 'Elite'])

products = df.groupby(['brand', 'price_label'])['price_label'].count()
avg_rev = df.groupby(['brand', 'price_label'])['revenue'].mean().round(2)
adidas_vs_nike = pd.DataFrame()
adidas_vs_nike['num_products'] = products
adidas_vs_nike['mean_revenue'] = avg_rev
adidas_vs_nike = adidas_vs_nike.reset_index()

# differences between word count of a product description and its mean rating
# group product descriptions by word count
df["description_length"] = df["description"].str.len()
bins = [0, 100, 200, 300, 400, 500, 600, 700]
labels = ["100", "200", "300", "400", "500", "600", "700"]
df["description_length"] = pd.cut(df["description_length"], bins=bins, 
labels=labels)

# average rating and total review by description length
description_lengths = df.groupby("description_length", as_index=False).agg(
    mean_rating=("rating", "mean"),
    total_reviews=("reviews", "sum")
).round(2)

# mean revenue by price categories and brand
sns.barplot(data=adidas_vs_nike, x="price_label", y="mean_revenue", hue="brand")
plt.title("Mean Revenue by Price Label & Brand")
plt.show()
```



Adidas is generating higher revenue than Nike overall, and its revenue tends to increase with higher price ranges.

```python
sns.lineplot(data=description_lengths, x="description_length", y="mean_rating",
marker="o")
plt.title("Mean Rating by Product Description Length")
plt.ylabel("Mean Rating")
plt.xlabel("Description Length (binned, 100 chars)")
plt.show()

sns.lineplot(data=description_lengths, x="description_length", y="total_reviews",
marker="o", color="orange")
plt.title("Total Reviews by Product Description Length")
plt.ylabel("Total Reviews")
plt.xlabel("Description Length (binned, 100 chars)")
plt.show()
```





Ratings improve as descriptions get longer, stabilizing around 3.2–3.4 once descriptions are at least 200–500 chars. Reviews are heavily concentrated in the 200–300 character range, peaking around 300 chars (~75k reviews). This suggests that customers are most likely to engage with products that have medium-length descriptions, while very short or very long descriptions receive less attention.