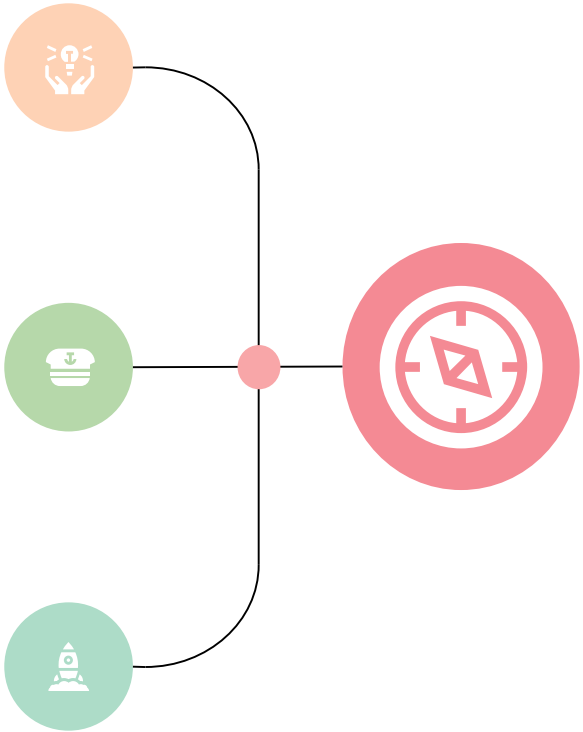
EME Graduation Project

Tourism Dataset Analysis



Agenda

1. Introduction

Project Objective & Project Goals.

2. Python

Data Preparation (Cleaning, preprocessing) &Questions (KPIs) &Code Implementation Details.

3. Excel

Data Cleaning & Exploration Using Excel & Create 10 business-oriented question

4. Dashboard

Exploring the Data Analytics Dashboard & The Role of the Dashboard in Data Visualization

1 Introduction

Project Objective :

The main goal of this project is to analyze a large, dirty tourism dataset to uncover insights about tourist behavior, demographics, and preferences. The project also aims to help the country in encouraging tourism by identifying key patterns and trends that can inform strategic decisions in tourism promotion, targeted marketing, and service improvements. This will involve data cleaning, visualization, and interactive dashboard creation using various tools

Project Goals :

1- Explore Travel Patterns

2- Evaluate Visitor Feedback

3- Identify Popular Purposes of Visit

4- Detect Missing or Inconsistent Data

5- Create Visualizations and Dashboards

2 Using Python(Pandas & Matplotlib & Numpy)

Tools & Libraries

- Python using Jupyter

1- Pandas (For data processing and analysis (loading, cleaning, transforming data).

2- Matplotlib & seaborn (For creating beautiful static visualizations and statistical plots)

3- Numpy (For performing mathematical operations and handling arrays)

Reading data :

- load data from a CSV file into a Data Frame for analysis.

[1]:
```
# libraries
import numpy as np
import pandas as pd
from matplotlib import pyplot as plt
import seaborn as sns
```

[2]:
```
df = pd.read_csv('tourism_data.csv')
df.head()
```

[2]:

	Tourist Name	Country	Arrival Date	Departure Date	Age	Gender	Purpose of Visit	Spending (USD)	Feedback
0	Jonathan Rodriguez	France	2025-02-25	03-01-2025	18	Male	BUSINESS	NaN	poor
1	Andrew Woods	MR	2024-06-26	2024-07-06	47	F	Study	NaN	NaN
2	Christopher Powers	Puerto Rico	2025-03-09	2025-03-13	42	f	Business	711	bad
3	Matthew Olson		2025-03-24	2025-04-03	41	male	Holiday	4777	bad
4	Tina Barrera	Egypt	2024-09-22	2024-10-03	NaN	f	Holiday	551.17	excelllent

Rename Column

```
#Rename columns
df.columns = [
    'Name', 'Country', 'Arrival_Date', 'Departure_Date',
    'Age', 'Gender', 'Purpose_of_Visit', 'Spending_USD', 'Feedback'
]
print("Number of rows before cleaning:", df.shape[0])
```

```
Number of rows before cleaning: 100000
```

Data Cleaning and Preprocessing :

```
df.dropna(subset=['Arrival_Date', 'Departure_Date', 'Age', 'Spending_USD'], inplace=True)
```

[6]:
```
print("Number of rows after cleaning:", df.shape[0])
```

```
Number of rows after cleaning: 36109
```

[7]:
```
df.replace(["N/A", "", " ", "n/a", "NA", None], np.nan, inplace=True)

print("Missing value placeholders replaced.")
```

```
Missing value placeholders replaced.
```

Data Cleaning and Preprocessing :

- Fix Date Formats in Arrival and Departure Dates
- Clean Spending Column

[8]:
```
# Fix Date Formats in Arrival and Departure Dates
df['Arrival_Date'] = pd.to_datetime(df['Arrival_Date'], errors='coerce')
df['Departure_Date'] = pd.to_datetime(df['Departure_Date'], errors='coerce')
```

[9]:
```
#Clean Spending Column
def convert_spending(value):
    if pd.isna(value):
        return np.nan
    value = str(value).strip().lower()
    if value == 'one thousand':
        return 1000.0
    elif value.replace('.', '', 1).isdigit():
        return float(value)
    else:
        return np.nan

df['Spending_USD'] = df['Spending_USD'].apply(convert_spending)

print("Spending column cleaned.")
```

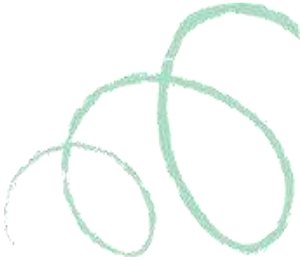
```
Spending column cleaned.
```

- Standardize Gender Column

10]:
```
#Standardize Gender Column
gender_map = {
    'Male': 'Male',
    'male': 'Male',
    'M': 'Male',
    'Female': 'Female',
    'female': 'Female',
    'F': 'Female',
    'MALE': 'Male',
    'Female': 'Female'
}
df['Gender'] = df['Gender'].map(gender_map)

print("Gender Column standardized.")
```

```
Gender Column standardized.
```



- Add Duration of Stay

[13]:
```
# Add Duration of Stay
df['Duration of Stay (days)'] = (df['Departure_Date'] - df['Arrival_Date']).dt.days
```

[14]:
```
#View Cleaned Data
print("First few rows of cleaned data:")
df.head()
```

```
First few rows of cleaned data:
```

- View Cleaned Data

[4]:
```
#View Cleaned Data
print("First few rows of cleaned data:")
df.head()
```

```
First few rows of cleaned data:
```

[4]:

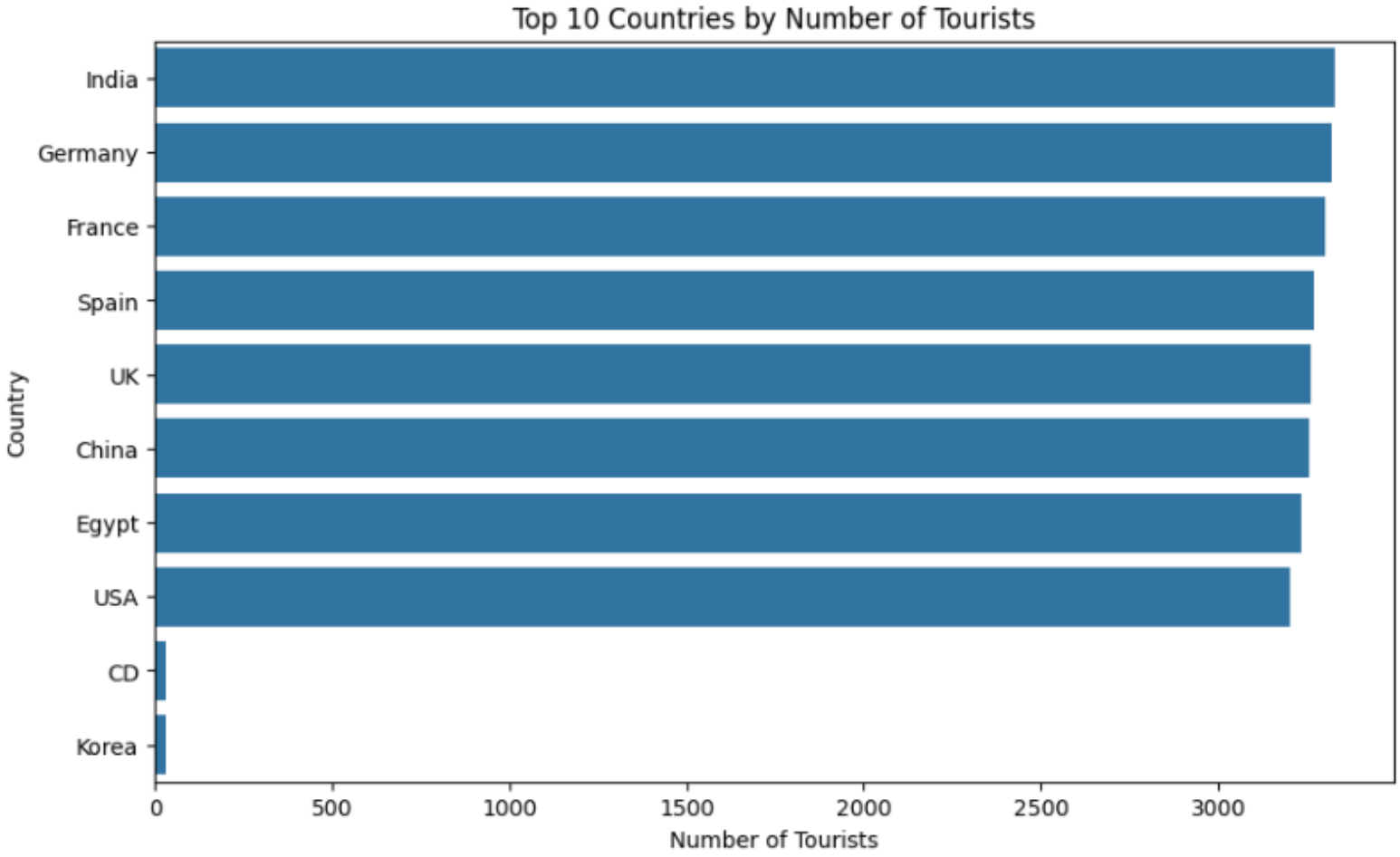
	Name	Country	Arrival_Date	Departure_Date	Age	Gender	Purpose_of_Visit	Spending_USD	Feedback	Duration of Stay (days)
2	Christopher Powers	Puerto Rico	2025-03-09	2025-03-13	42.0	NaN	Business	711.0	bad	4.0
3	Matthew Olson	NaN	2025-03-24	2025-04-03	41.0	Male	Holiday	4777.0	bad	10.0
7	Mr. Rickey Graham	UK	2024-11-22	2024-12-11	38.0	Female	Leisure	3044.0	excelllent	19.0
8	Ariel Kane	China	2025-02-01	2025-02-19	49.0	Male	Holiday	2708.0	excelllent	18.0
11	Jacob Delgado	France	2025-02-28	2025-03-09	47.0	Male	NaN	3918.0	Ok	9.0

Key Analysis Questions(EDA)

1. Which countries have the most visitors?
2. What is the average duration of stay by Purpose_of_Visit?
3. How does feedback vary across travel types?
5. Which countries have the highest average spending?
6. Most Common Reasons for Visiting
7. Average Spending by Purpose of Visit
8. Does gender affect the type of visit preferred?

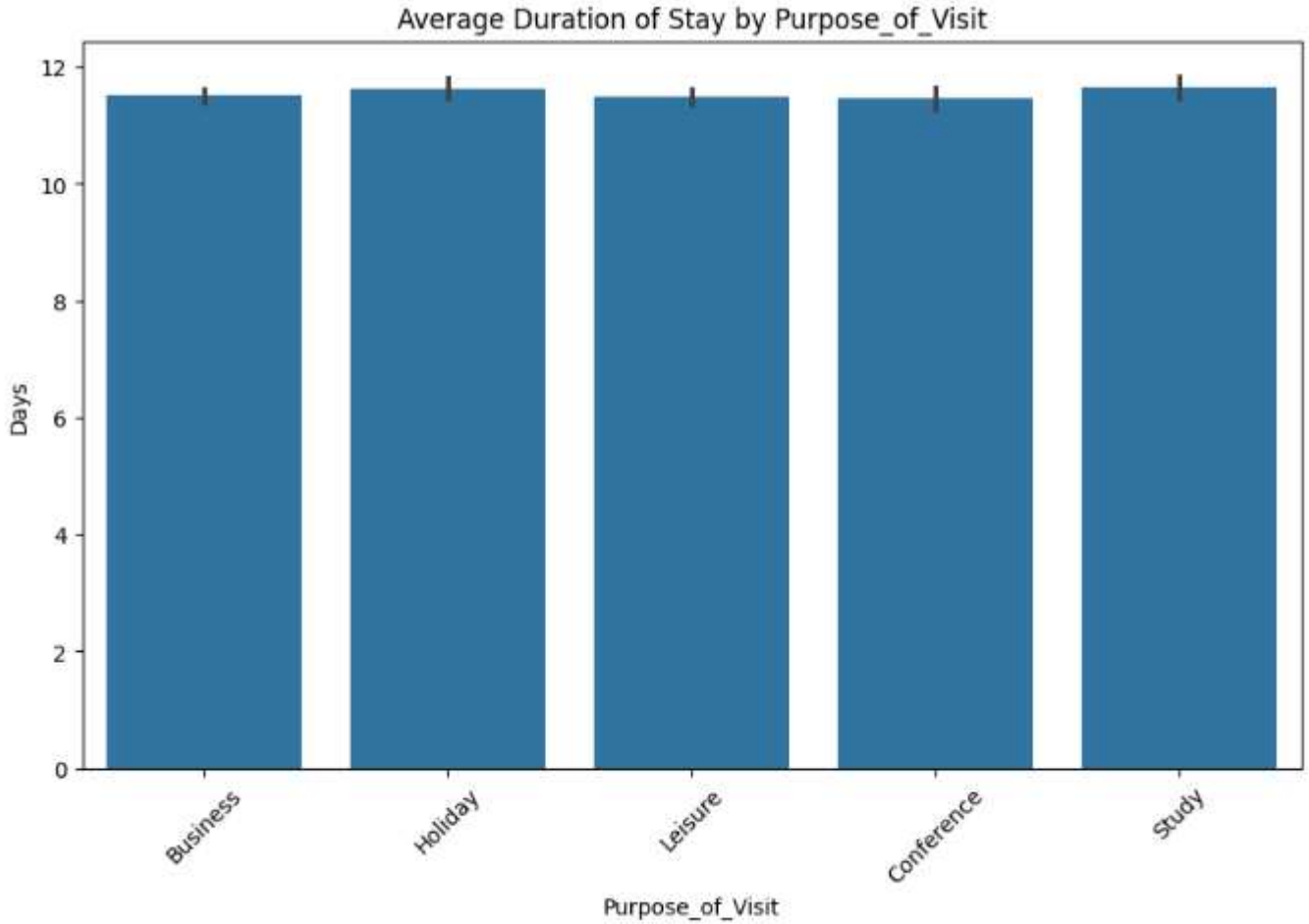
- Q1: Which countries have the most visitors?

[15]:
```
#Q1: Which countries have the most visitors?
top_countries = df['Country'].value_counts(dropna=True).head(10)
plt.figure(figsize=(10,6))
sns.barplot(x=top_countries.values, y=top_countries.index)
plt.title('Top 10 Countries by Number of Tourists')
plt.xlabel('Number of Tourists')
plt.ylabel('Country')
plt.show()
```


Top 10 Countries by Number of Tourists
India
Germany
France
Spain
UK
China
Egypt
USA
CD
Korea
Country
0
500
1000
1500
2000
2500
3000
Number of Tourists

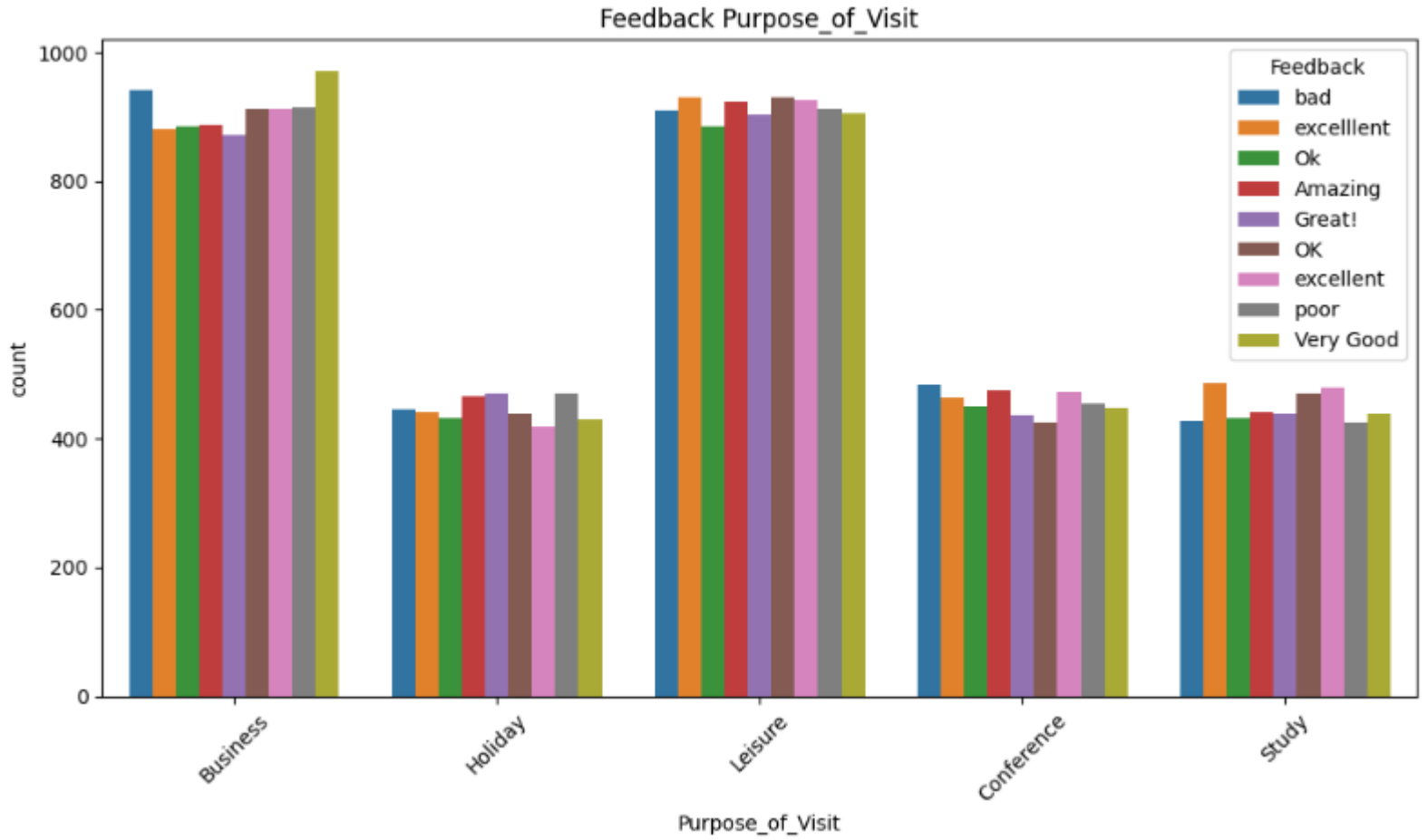
- Q2 : What is the average duration of stay by Purpose_of_Visit?

[38]:
```
#Q2: What is the average duration of stay by Purpose_of_Visit?
plt.figure(figsize=(10,6))
sns.barplot(x='Purpose_of_Visit', y='Duration', data=df, estimator=np.mean)
plt.xticks(rotation=45)
plt.title('Average Duration of Stay by Purpose_of_Visit')
plt.ylabel('Days')
plt.show()
```


Average Duration of Stay by Purpose_of_Visit
Days
0
2
4
6
8
10
12
Business
Holiday
Leisure
Conference
Study
Purpose_of_Visit

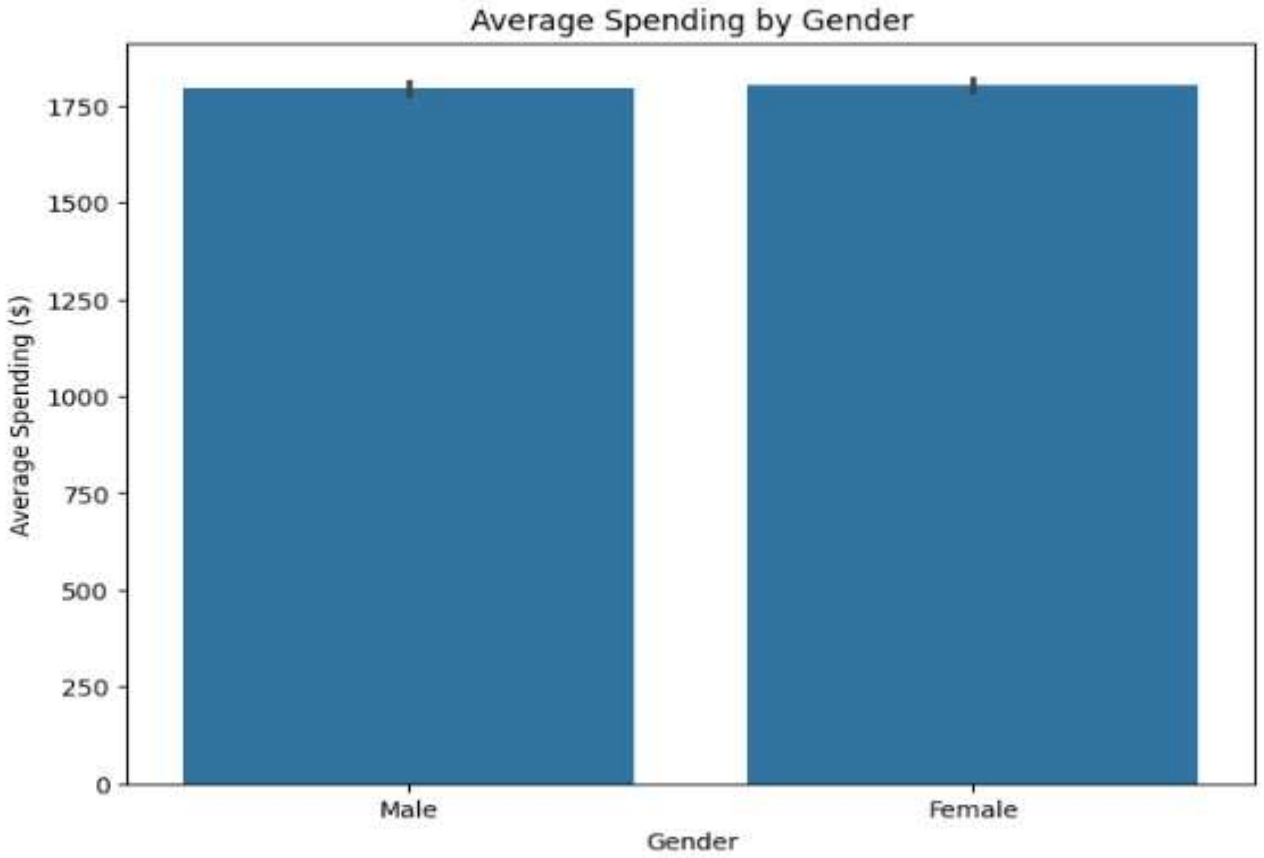
- Q3: How does feedback vary across travel types?

```
39]: #Q3: How does feedback vary across travel types?
     plt.figure(figsize=(10,6))
     sns.countplot(data=df, x='Purpose_of_Visit', hue='Feedback')
     plt.xticks(rotation=45)
     plt.title('Feedback Purpose_of_Visit')
     plt.legend(title='Feedback')
     plt.tight_layout()
     plt.show()
```


Feedback Purpose_of_Visit
Feedback
bad
excelllent
Ok
Amazing
Great!
OK
excellent
poor
Very Good
count
1000
800
600
400
200
0
Business
Holiday
Leisure
Conference
Study
Purpose_of_Visit

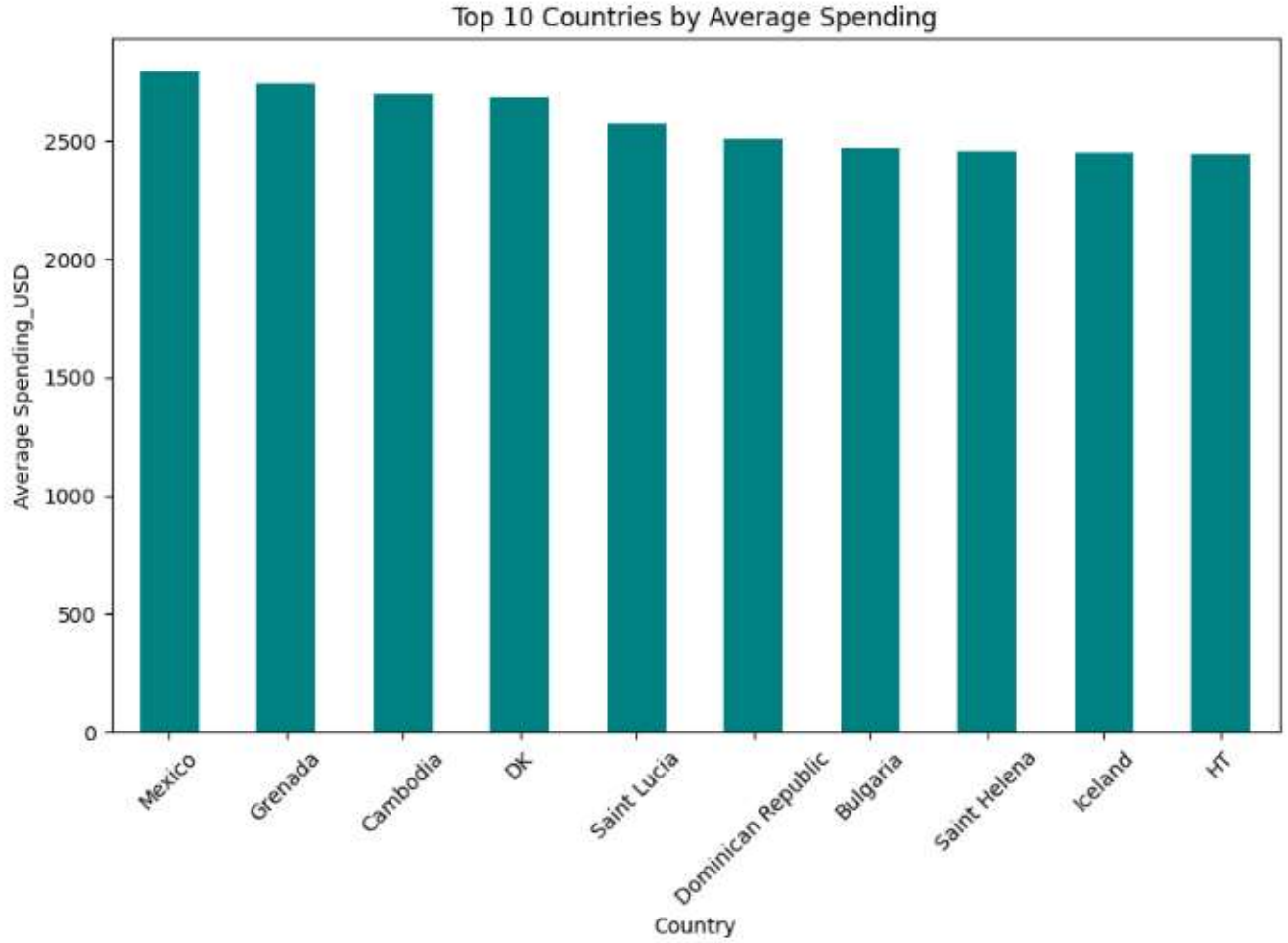
- Q4: What is the average spending by gender?

[23]:
```
# Q4: What is the average spending by gender?
plt.figure(figsize=(8,6))
sns.barplot(x='Gender', y='Spending_USD', data=df, estimator=np.mean)
plt.title('Average Spending by Gender')
plt.ylabel('Average Spending ($)')
plt.show()
```


Average Spending by Gender
Average Spending ($)
1750
1500
1250
1000
750
500
250
0
Male
Female
Gender

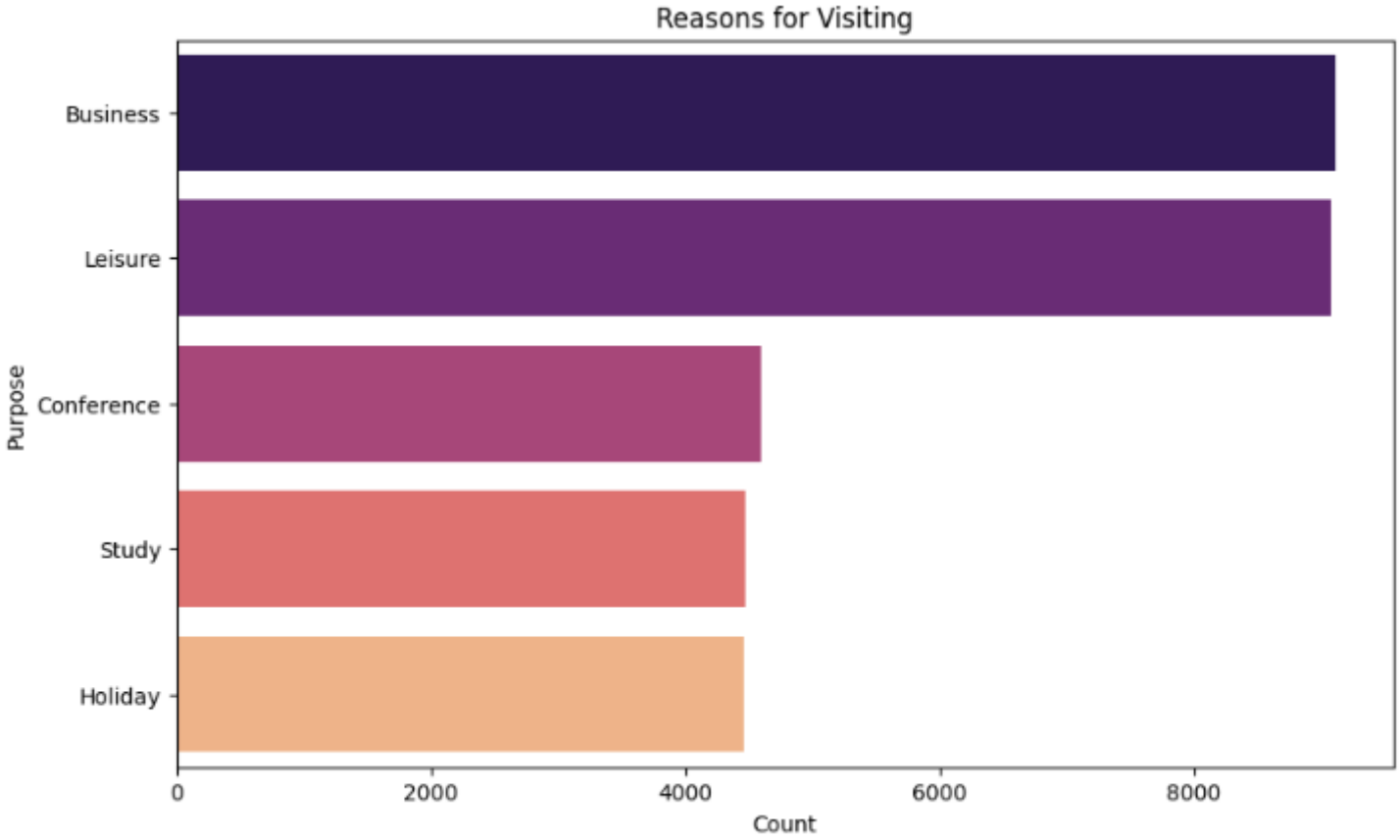
- Q5: Which countries have the highest average spending?

```
#Q5: Which countries have the highest average spending?
avg_spending = df.groupby('Country')['Spending_USD'].mean().sort_values(ascending=False).head(10)
plt.figure(figsize=(10, 6))
avg_spending.plot(kind='bar', color='teal')
plt.title("Top 10 Countries by Average Spending")
plt.xlabel("Country")
plt.ylabel("Average Spending_USD")
plt.xticks(rotation=45)
plt.show()
```


Top 10 Countries by Average Spending
Average Spending_USD
0
500
1000
1500
2000
2500
Mexico
Grenada
Cambodia
DK
Saint Lucia
Dominican Republic
Bulgaria
Saint Helena
Iceland
HT
Country

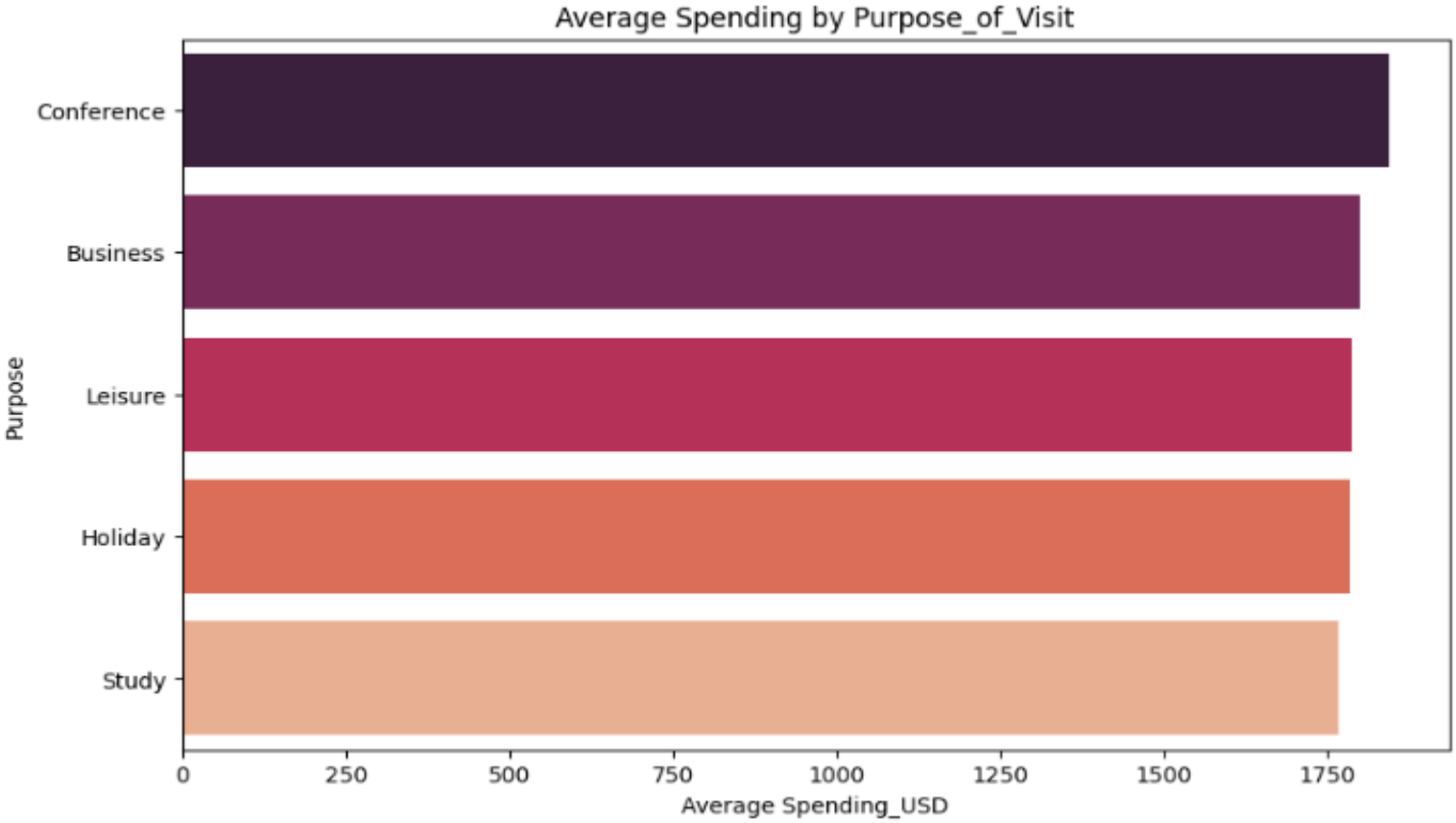
- Q6: Most Common Reasons for Visiting

```
#Q6:Most Common Reasons for Visiting
plt.figure(figsize=(10,6))
purpose_counts = df['Purpose_of_Visit'].value_counts()
sns.barplot(x=purpose_counts.values, y=purpose_counts.index, palette="magma")
plt.title("Reasons for Visiting")
plt.xlabel("Count")
plt.ylabel("Purpose")
plt.show()
```


Reasons for Visiting
Business
Leisure
Conference
Study
Holiday
Purpose
0
2000
4000
6000
8000
Count

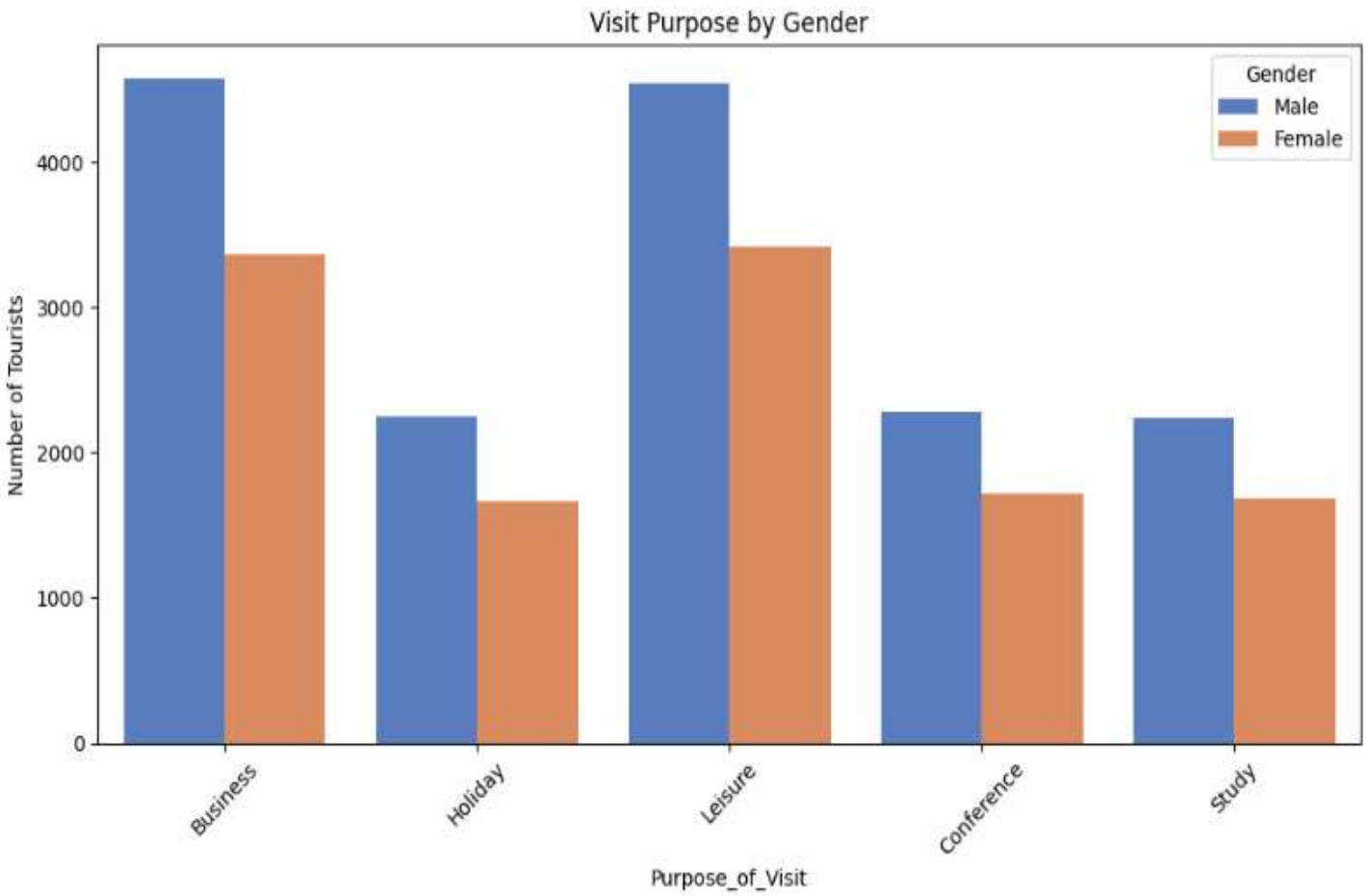
- Q7: Average Spending by Purpose of Visit

```
#Q7 Average Spending by Purpose of Visit
avg_spend_purpose = df.groupby('Purpose_of_Visit')['Spending_USD'].mean().sort_values(ascending=False).head(10)
plt.figure(figsize=(10,6))
sns.barplot(x=avg_spend_purpose.values, y=avg_spend_purpose.index, palette="rocket")
plt.title("Average Spending by Purpose_of_Visit")
plt.xlabel("Average Spending_USD")
plt.ylabel("Purpose")
plt.show()
```


Average Spending by Purpose_of_Visit
Conference
Business
Leisure
Holiday
Study
Purpose
0
250
500
750
1000
1250
1500
1750
Average Spending_USD

- Q8: Does gender affect the type of visit preferred?

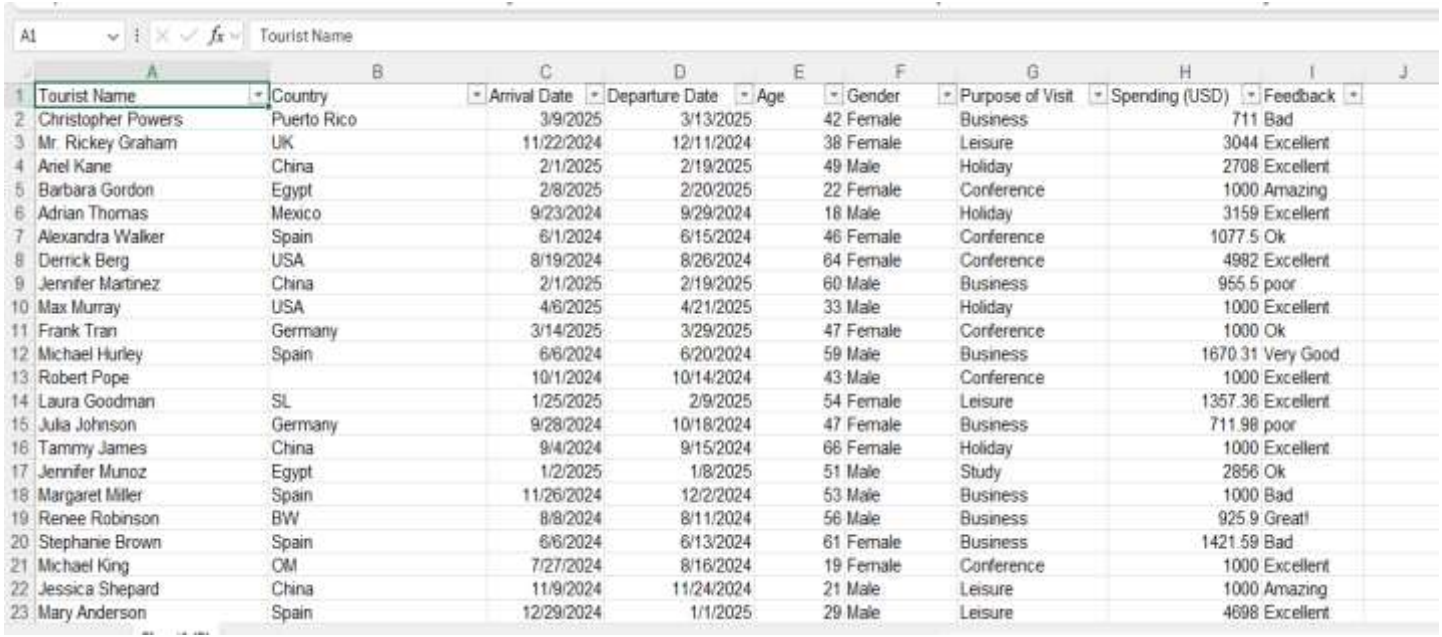
```
#Q8:Does gender affect the type of visit preferred?
plt.figure(figsize=(10, 6))
sns.countplot(x='Purpose_of_Visit', hue='Gender', data=df, palette="muted")
plt.xticks(rotation=45)
plt.title("Visit Purpose by Gender")
plt.xlabel("Purpose_of_Visit")
plt.ylabel("Number of Tourists")
plt.legend(title="Gender")
plt.tight_layout()
plt.show()
```


Visit Purpose by Gender
Gender
Male
Female
Number of Tourists
0
1000
2000
3000
4000
Business
Holiday
Leisure
Conference
Study
Purpose_of_Visit

3 Excel

Data Cleaning & Exploration

- USING POWER QUERY, I CHANGED THE DATA TYPE OF EACH COLUMN AND REMOVED THE NULL AND ERROR DATA.

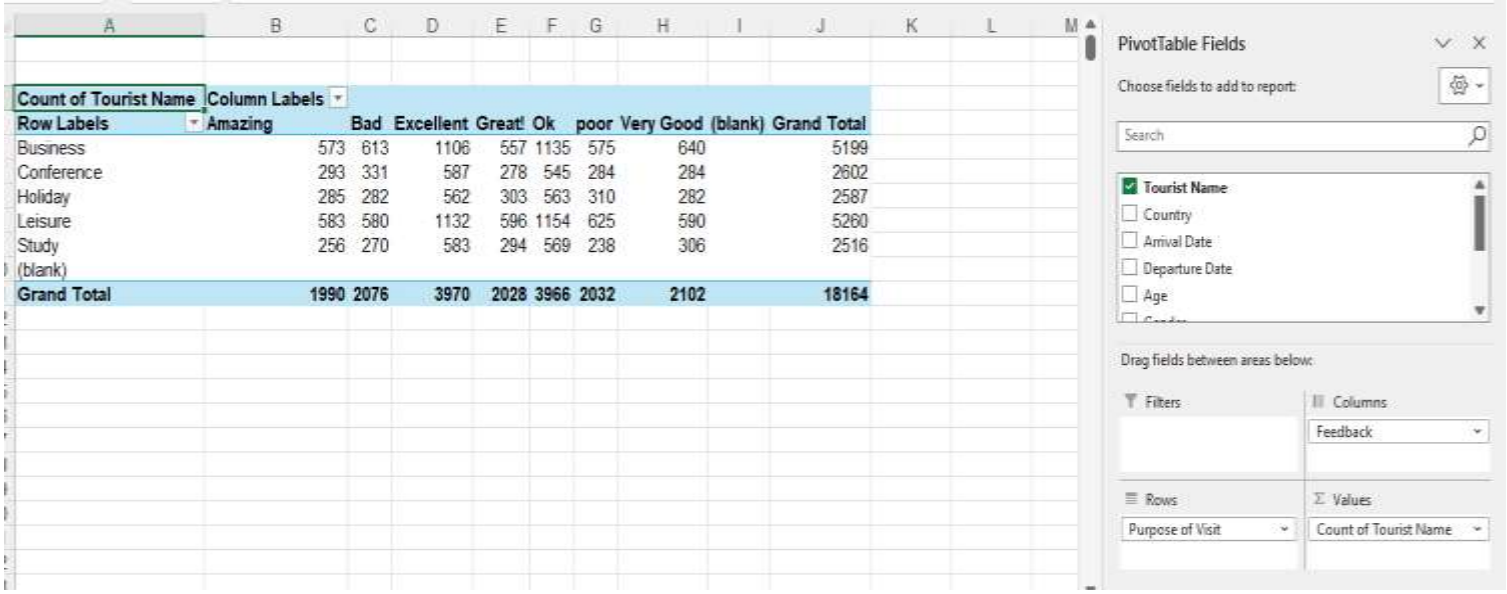


A1 | Tourist Name

	A	B	C	D	E	F	G	H	I	J
1	Tourist Name	Country	Arrival Date	Departure Date	Age	Gender	Purpose of Visit	Spending (USD)	Feedback	
2	Christopher Powers	Puerto Rico	3/9/2025	3/13/2025	42	Female	Business	711	Bad	
3	Mr. Rickey Graham	UK	11/22/2024	12/11/2024	38	Female	Leisure	3044	Excellent	
4	Ariel Kane	China	2/1/2025	2/19/2025	49	Male	Holiday	2708	Excellent	
5	Barbara Gordon	Egypt	2/8/2025	2/20/2025	22	Female	Conference	1000	Amazing	
6	Adrian Thomas	Mexico	9/23/2024	9/29/2024	18	Male	Holiday	3159	Excellent	
7	Alexandra Walker	Spain	6/1/2024	6/15/2024	46	Female	Conference	1077.5	Ok	
8	Derrick Berg	USA	8/19/2024	8/26/2024	64	Female	Conference	4982	Excellent	
9	Jennifer Martinez	China	2/1/2025	2/19/2025	60	Male	Business	955.5	poor	
10	Max Murray	USA	4/6/2025	4/21/2025	33	Male	Holiday	1000	Excellent	
11	Frank Tran	Germany	3/14/2025	3/29/2025	47	Female	Conference	1000	Ok	
12	Michael Hurley	Spain	6/6/2024	6/20/2024	59	Male	Business	1670.31	Very Good	
13	Robert Pope		10/1/2024	10/14/2024	43	Male	Conference	1000	Excellent	
14	Laura Goodman	SL	1/25/2025	2/9/2025	54	Female	Leisure	1357.36	Excellent	
15	Julia Johnson	Germany	9/28/2024	10/18/2024	47	Female	Business	711.98	poor	
16	Tammy James	China	9/4/2024	9/15/2024	66	Female	Holiday	1000	Excellent	
17	Jennifer Munoz	Egypt	1/2/2025	1/8/2025	51	Male	Study	2856	Ok	
18	Margaret Miller	Spain	11/26/2024	12/2/2024	53	Male	Business	1000	Bad	
19	Renee Robinson	BW	8/8/2024	8/11/2024	56	Male	Business	925.9	Great!	
20	Stephanie Brown	Spain	6/6/2024	6/13/2024	61	Female	Business	1421.59	Bad	
21	Michael King	OM	7/27/2024	8/16/2024	19	Female	Conference	1000	Excellent	
22	Jessica Shepard	China	11/9/2024	11/24/2024	21	Male	Leisure	1000	Amazing	
23	Mary Anderson	Spain	12/29/2024	1/1/2025	29	Male	Leisure	4698	Excellent	

Create business-oriented question

- using Pivot Tables
- How does feedback vary by purpose of visit



Count of Tourist Name	Column Labels								
Row Labels	Amazing	Bad	Excellent	Great!	Ok	poor	Very Good	(blank)	Grand Total
Business	573	613	1106	557	1135	575	640		5199
Conference	293	331	587	278	545	284	284		2602
Holiday	285	282	562	303	563	310	282		2587
Leisure	583	580	1132	596	1154	625	590		5260
Study	256	270	583	294	569	238	306		2516
(blank)									
Grand Total	1990	2076	3970	2028	3966	2032	2102		18164

PivotTable Fields

Choose fields to add to report:

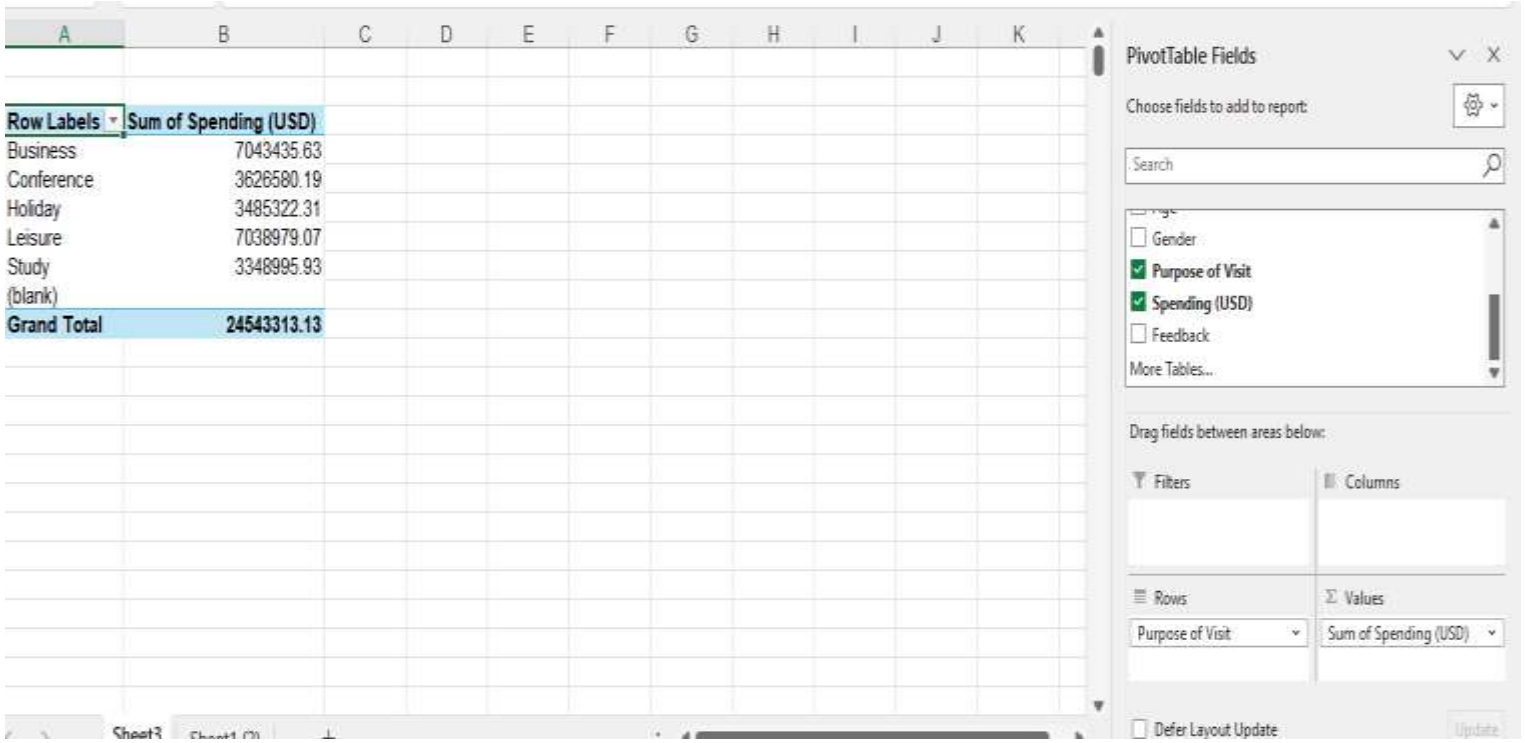
Search

Tourist Name
Country
Arrival Date
Departure Date
Age

Drag fields between areas below:

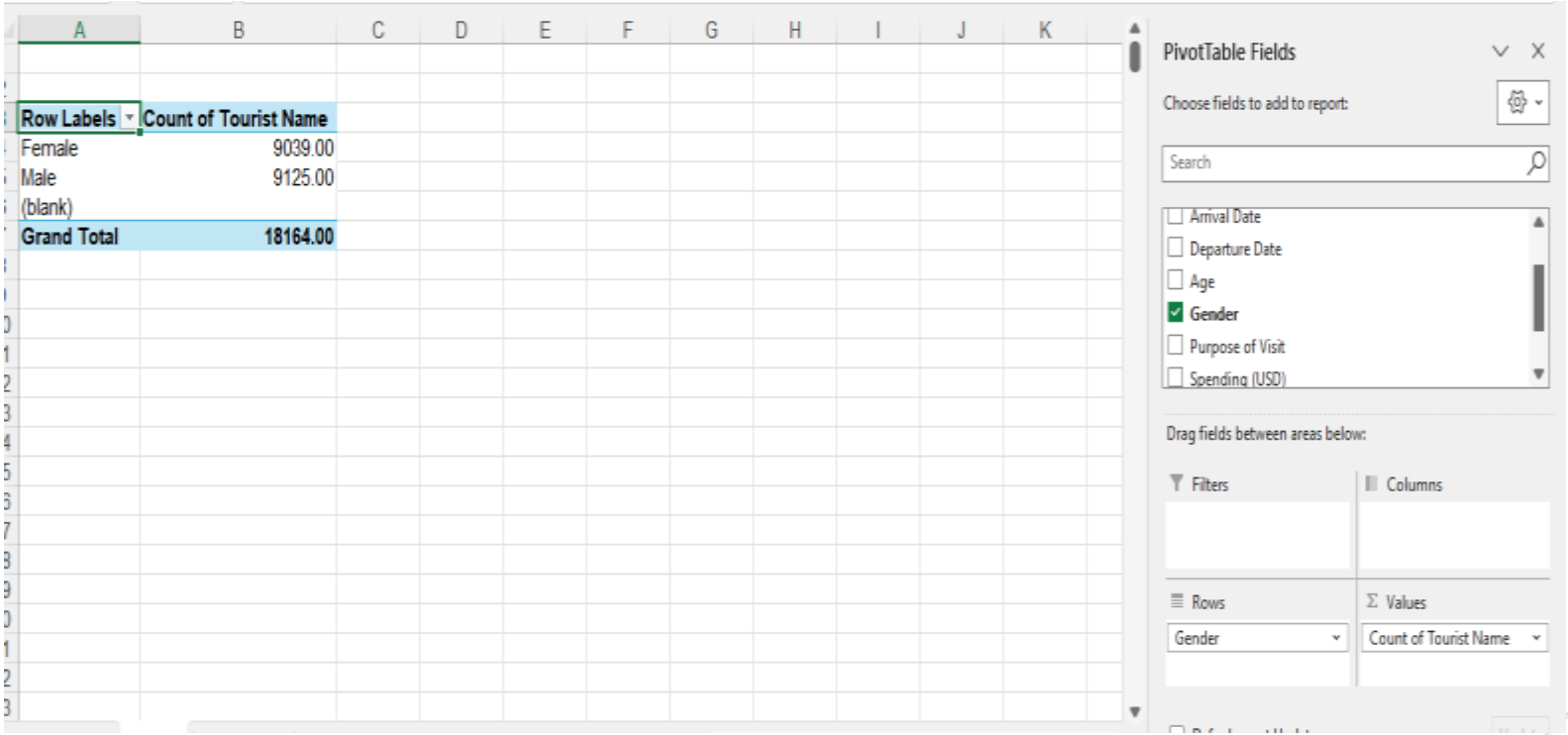
Filters
Columns
Feedback
Rows
Purpose of Visit
Values
Count of Tourist Name

- What is the average spending by purpose of visit?



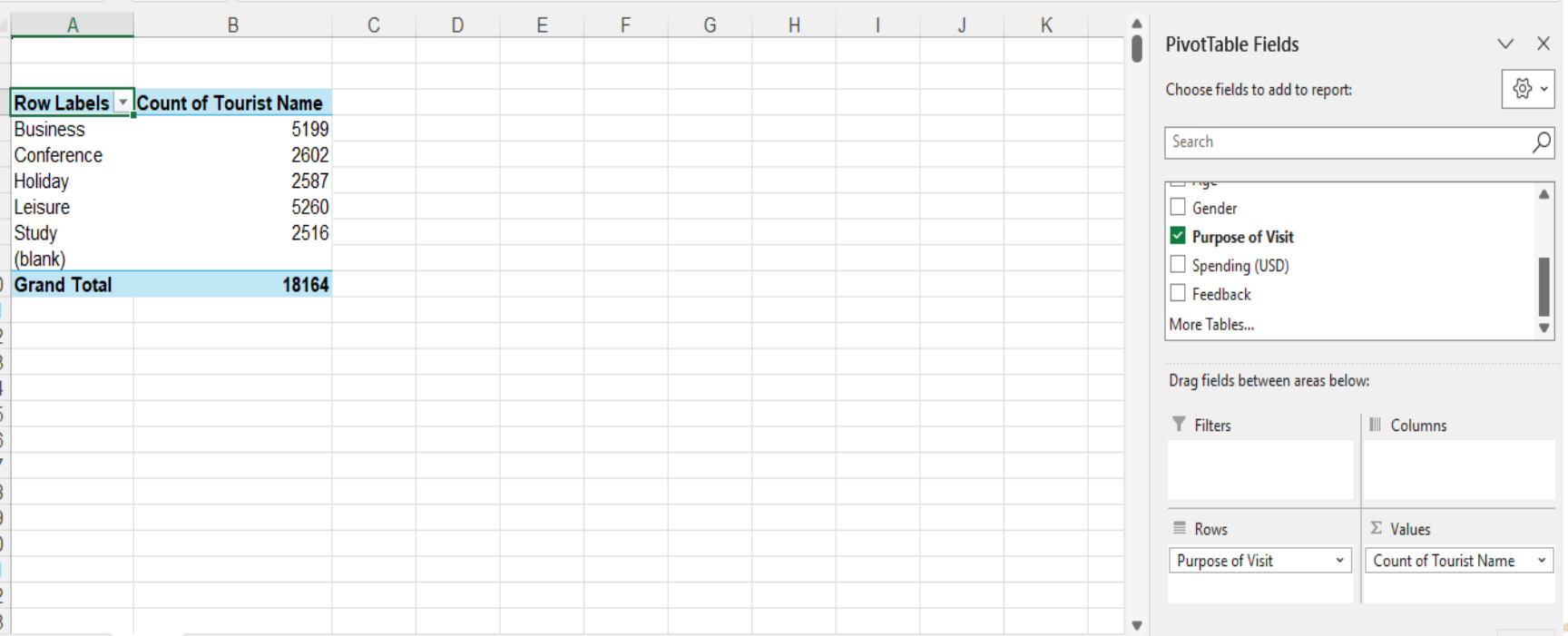
Row Labels	Sum of Spending (USD)
Business	7043435.63
Conference	3626580.19
Holiday	3485322.31
Leisure	7038979.07
Study	3348995.93
(blank)	
Grand Total	**24543313.13**

- What is the distribution of tourists by gender?



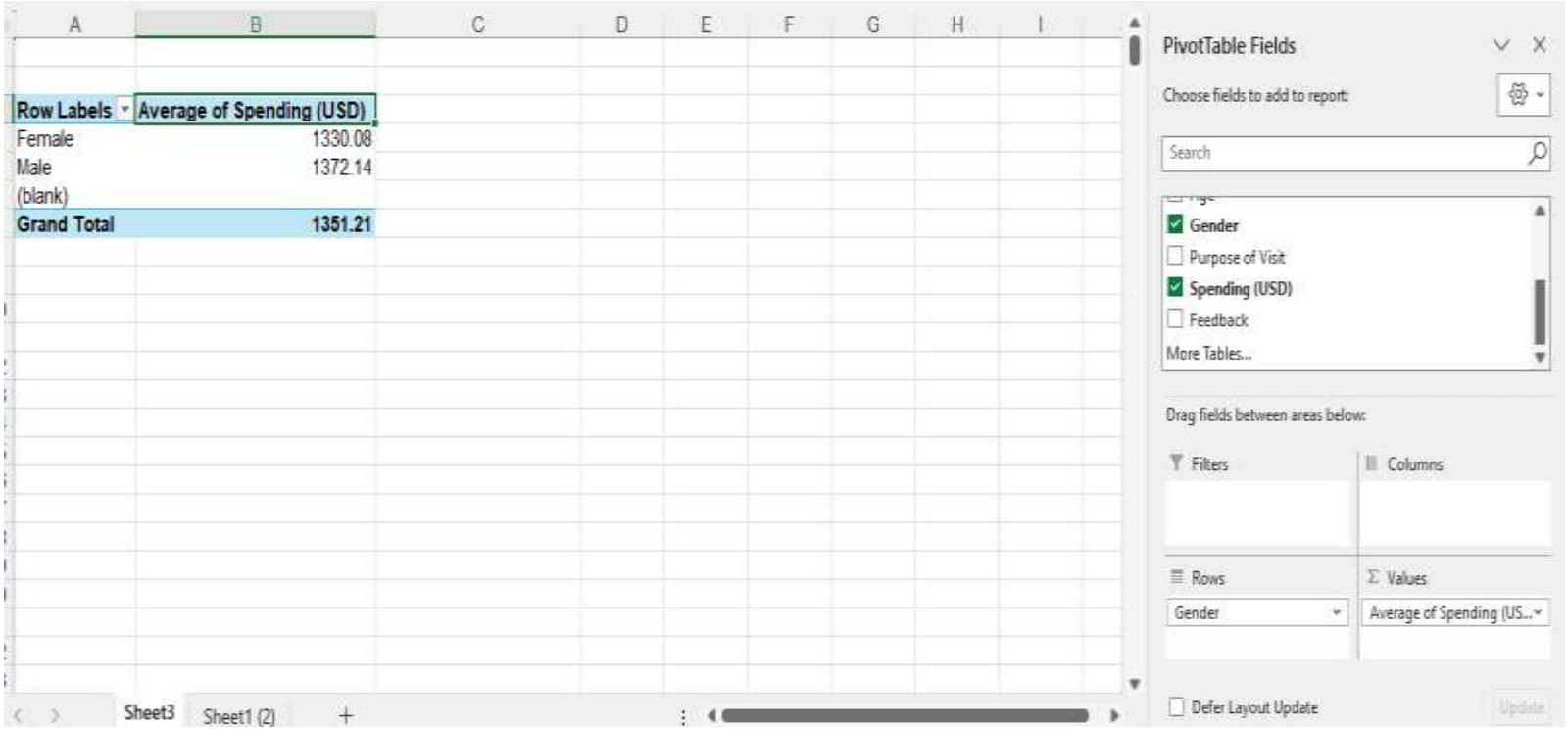
Row Labels	Count of Tourist Name
Female	9039.00
Male	9125.00
(blank)	
Grand Total	**18164.00**

- How many tourists visited by purpose of visit.



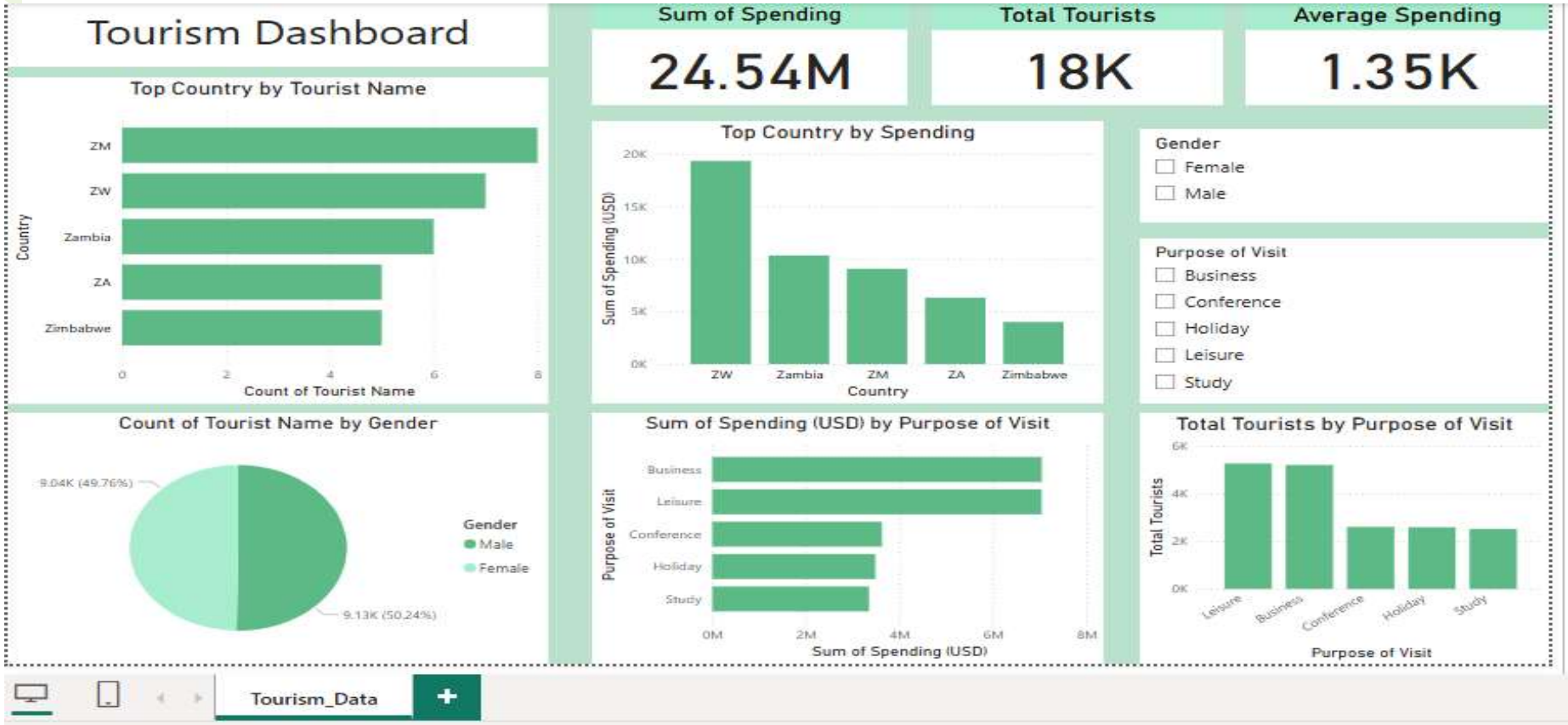
Row Labels	Count of Tourist Name
Business	5199
Conference	2602
Holiday	2587
Leisure	5260
Study	2516
(blank)	
Grand Total	**18164**

- Which gender spends more



4 Dashboard

Dashboard


Tourism Dashboard
Sum of Spending
24.54M
Total Tourists
18K
Average Spending
1.35K
Top Country by Tourist Name
ZM
ZW
Zambia
ZA
Zimbabwe
Country
0
2
4
6
8
Count of Tourist Name
Top Country by Spending
20K
15K
10K
5K
0K
Sum of Spending (USD)
ZW
Zambia
ZM
ZA
Zimbabwe
Country
Gender
Female
Male
Purpose of Visit
Business
Conference
Holiday
Leisure
Study
Count of Tourist Name by Gender
9.04K (49.76%)
9.13K (50.24%)
Gender
Male
Female
Sum of Spending (USD) by Purpose of Visit
Business
Leisure
Conference
Holiday
Study
Purpose of Visit
0M
2M
4M
6M
8M
Sum of Spending (USD)
Total Tourists by Purpose of Visit
6K
4K
2K
0K
Total Tourists
Leisure
Business
Conference
Holiday
Study
Purpose of Visit
Tourism_Data

Summary of Key Points

1. Analysis by Country (Sales Performance by Country):

- Top Countries for Tourists : India, Germany, Spain, Egypt, and France are among the top countries sending tourists.
- Highest Average Spending Per Country : Germany, Egypt, and the USA have the highest average spending.
- Insight :
- Focus on promotional campaigns in these countries.
- Enhance services provided to this category of visitors

Summary of Key Points

2-Purpose of Visit Analysis:

- Most Common Purposes : Business is the primary purpose of visit , Followed by leisure and Study.
- Average Spending by Purpose : Business and conference visitors spend the most.
- Insight : Develop customized services for students.
- special offers for conferences and business events.

Summary of Key Points

3- Spending and Feedback Analysis :

Spending by Gender :

- There is no significant difference in spending between males and females.
- Feedback : Most visitors expressed varying degrees of satisfaction (e.g., Very Good, Excellent).
- Recommendation :
- Improve services that receive poor ratings.
- Encourage visitors to provide feedback after their visit.

Thanks!

Do you have any questions?

Prepared by:
Doha Ali Nasr

Access the Complete
Code From Here :