csszgnn8l

September 7, 2024

1 Aim of the Analysis:

2 "Understanding Gender Distribution Across Age Groups to Optimize Pricing and Marketing Strategies for airline company"

[ ]:

[1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

[2]:
```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

[6]:
```
df = pd.read_csv('/content/drive/My Drive/busines case dsml/airline_data/
 ↪Airline Dataset Updated .csv')
```

[7]:
```
df.head()
```

[7]:

	Passenger ID	First Name	Last Name	Gender	Age	Nationality	Airport Name	Airport Country Code	Country Name
0	ABVWIg	Edithe	Leggis	Female	62	Japan	Coldfoot Airport	US	United States
1	jkXXAX	Elwood	Catt	Male	62	Nicaragua	Kugluktuk Airport	CA	Canada
2	CdUz2g	Darby	Felgate	Male	67	Russia	Grenoble-Isère Airport	FR	France
3	BRS38V	Dominica	Pyle	Female	71	China	Ottawa / Gatineau Airport	CA	Canada
4	9kvTLo	Bay	Pencost	Male	21	China	Gillespie Field	US	United States

```
  Airport Continent     Continents Departure Date Arrival Airport  \
0               NAM  North America      6/28/2022             CXF
1               NAM  North America     12/26/2022             YCO
2                EU         Europe      1/18/2022             GNB
3               NAM  North America      9/16/2022             YND
4               NAM  North America      2/25/2022             SEE

            Pilot Name Flight Status
0  Fransisco Hazeldine       On Time
1      Marla Parsonage       On Time
2         Rhonda Amber       On Time
3       Kacie Commucci       Delayed
4           Ebonee Tree       On Time
```

[8]:
```
df.shape
```

[8]:
```
(98619, 15)
```

[9]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 98619 entries, 0 to 98618
Data columns (total 15 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Passenger ID          98619 non-null  object
 1   First Name            98619 non-null  object
 2   Last Name             98619 non-null  object
 3   Gender                98619 non-null  object
 4   Age                   98619 non-null  int64
 5   Nationality           98619 non-null  object
 6   Airport Name          98619 non-null  object
 7   Airport Country Code  98619 non-null  object
 8   Country Name          98619 non-null  object
 9   Airport Continent     98619 non-null  object
 10  Continents            98619 non-null  object
 11  Departure Date        98619 non-null  object
 12  Arrival Airport       98619 non-null  object
 13  Pilot Name            98619 non-null  object
 14  Flight Status         98619 non-null  object
dtypes: int64(1), object(14)
memory usage: 11.3+ MB
```

[10]:
```
df.nunique()
```

[10]:
```
Passenger ID            98619
First Name               8437
Last Name               41658
Gender                      2
Age                        90
Nationality               240
Airport Name             9062
Airport Country Code      235
Country Name              235
Airport Continent           6
Continents                  6
Departure Date            364
Arrival Airport          9024
Pilot Name              98605
Flight Status               3
dtype: int64
```

[29]:
```
# Taking Below columns from data set 'df'
df_f_travel = df[['Passenger ID', 'Gender','Age', "Departure Date"]]
df_f_travel = pd.DataFrame(df_f_travel)
```

[41]:
```
df_f_travel.drop("Gender code", axis=1, inplace= True)
```

[43]:
```
df_f_travel.head()
```

[43]:

	Passenger ID	Gender	Age	Departure Date	Gender_code
0	ABVWIg	Female	62	6/28/2022	0
1	jkXXAX	Male	62	12/26/2022	1
2	CdUz2g	Male	67	1/18/2022	1
3	BRS38V	Female	71	9/16/2022	0
4	9kvTLo	Male	21	2/25/2022	1

[44]:
```
df_f_travel.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 98619 entries, 0 to 98618
Data columns (total 5 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Passenger ID    98619 non-null  object
 1   Gender          98619 non-null  object
 2   Age             98619 non-null  int64
 3   Departure Date  98619 non-null  object
 4   Gender_code     98619 non-null  int64
dtypes: int64(2), object(3)
memory usage: 3.8+ MB
```

[47]:
```
df_f_travel['Gender_code'].value_counts()
```

[47]:
```
Gender_code
1    49598
0    49021
Name: count, dtype: int64
```

[52]:
```
df_f_travel.groupby("Gender")["Age"].agg(['min', "max", "count", "mean"])
```

[52]:

Gender	min	max	count	mean
Female	1	90	49021	45.51943
Male	1	90	49598	45.48879

[54]:
```
# Define bin and lables
bins = [0, 20, 40, 60, 90]

labels= ['1-20', '20-40', '40-60', '60-90']
```

[59]:
```
# create a new column "Age_Bin" with the binned categories
df_f_travel['Age_Bin'] = pd.cut(df_f_travel['Age'], bins=bins, labels=labels,␣
 ↪right=False)

result = df_f_travel.groupby(['Gender', "Age_Bin"])['Age'].agg(['min', 'max',␣
 ↪'count', 'max'])
print(result)


df_age = df_f_travel.groupby(['Gender', 'Age_Bin'])['Age'].count()
```

Gender	Age_Bin	min	max	count	max
Female	1-20	1	19	10267	19
	20-40	20	39	10943	39
	40-60	40	59	10981	59
	60-90	60	89	16305	89
Male	1-20	1	19	10464	19
	20-40	20	39	10975	39
	40-60	40	59	11117	59
	60-90	60	89	16491	89

[60]:
```
df_age
```

[60]:

Gender	Age_Bin	
Female	1-20	10267
	20-40	10943
	40-60	10981

```
        60-90      16305
Male    1-20       10464
        20-40      10975
        40-60      11117
        60-90      16491
Name: Age, dtype: int64
```

[66]:
```
df_age_pivot = df_f_travel.groupby(['Gender', 'Age_Bin'])['Age'].count().
 ↪unstack(fill_value=0).reset_index()
```

[69]:
```
df_age_pivot
```

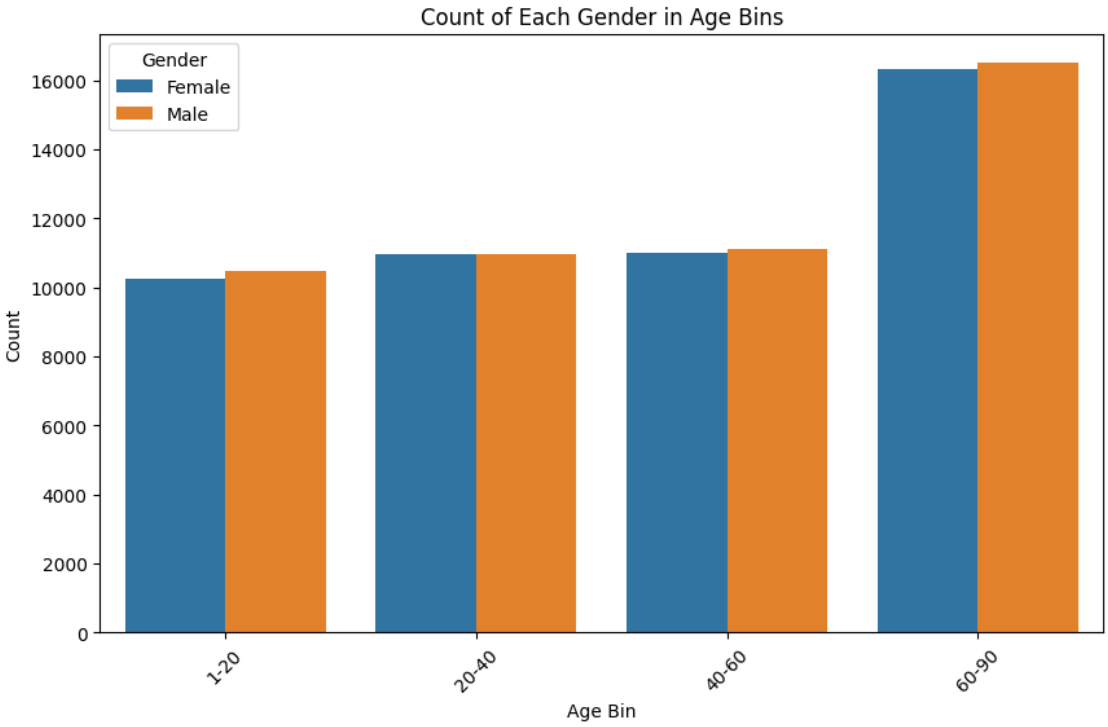
[69]:
```
Age_Bin  Gender   1-20  20-40  40-60  60-90
0        Female  10267  10943  10981  16305
1          Male  10464  10975  11117  16491
```

[71]:
```
df_age_pivot= pd.DataFrame(df_age_pivot)
```

[76]:
```
# Melt the DataFrame to long format for sns.barplot
df_age_melted = df_age_pivot.melt(id_vars='Gender', var_name='Age_Bin',␣
 ↪value_name='Count')

# Create the bar plot
plt.figure(figsize=(10, 6))
sns.barplot(data=df_age_melted, x='Age_Bin', y='Count', hue='Gender')
# Add labels and title
plt.xlabel('Age Bin')
plt.ylabel('Count')
plt.title('Count of Each Gender in Age Bins')
plt.legend(title='Gender')
plt.xticks(rotation=45)  # Rotate x labels if needed for better readability

plt.show()
```



[75]:
```
df_age_melted
```

[75]:

	Gender	Age_Bin	Count
0	Female	1-20	10267
1	Male	1-20	10464
2	Female	20-40	10943
3	Male	20-40	10975
4	Female	40-60	10981
5	Male	40-60	11117
6	Female	60-90	16305
7	Male	60-90	16491

3 Analysis Summary:

Our analysis of gender distribution across different age groups for the airline company reveals an intriguing trend. The bar plot indicates that the age group between 20 and 40 shows a notably balanced number of males and females. This suggests that many travelers in this age range are likely couples.

4 Recommendation:

To capitalize on this insight, we recommend the following strategies:

Enhanced Pricing for Couples and Groups:

Introduce attractive pricing options for group bookings and couples. This will not only incentivize bookings from pairs but also attract more customers in the 20-40 age range who are likely traveling together. Promote Honeymoon Destinations:

Increase promotional efforts for honeymoon destinations. This targeted approach will appeal to the significant number of couples in this age group and potentially boost bookings for romantic getaways.

[ ]: