Analysis of Complaint Patterns at Airports

November 3, 2024

- Sauda Haywood

- DSC640 Weeks 9 & 10 Exercise - TSA Complaints

1 Introduction

This analysis investigates patterns and insights from four datasets related to complaint records. Each dataset covers different aspects of complaint data, including airport-specific complaints, complaint categories, subcategories, and general distribution across geographic locations. The primary audience for this analysis is researchers and analysts, with the goal of identifying trends, prevalent issues, and temporal patterns that may inform quality improvements and operational enhancements. Visualizations throughout the notebook provide a clear and data-driven narrative, supporting decision-making in areas with high complaint volumes.

2 Exploratory Data Analysis

[3]:
```
import pandas as pd

# Load all datasets
data1 = pd.read_csv('iata-icao.csv')
data2 = pd.read_csv('complaints-by-subcategory.csv')
data3 = pd.read_csv('complaints-by-category.csv')
data4 = pd.read_csv('complaints-by-airport.csv')
```

[7]:
```
# Summarize the datasets
data1.head(5)
```

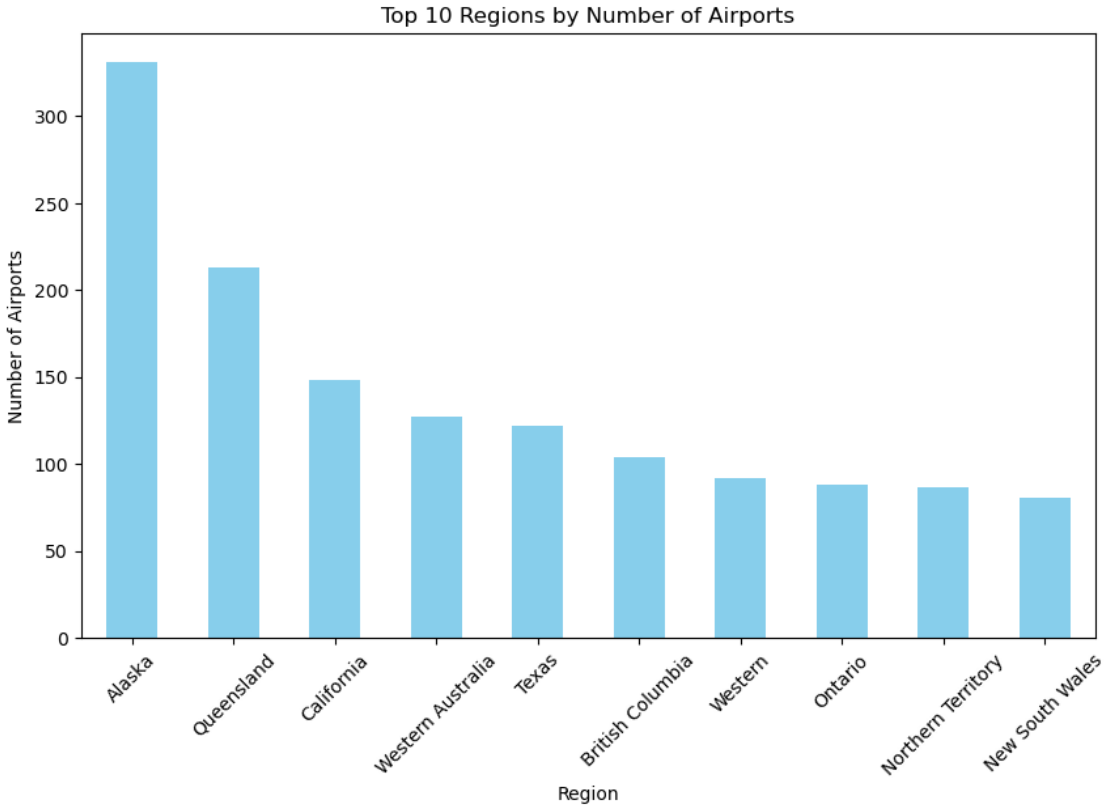
[7]:
```
  country_code region_name iata  icao                          airport  \
0           AE    Abu Zaby  AAN  OMAL     Al Ain International Airport
1           AE    Abu Zaby  AUH  OMAA  Abu Dhabi International Airport
2           AE    Abu Zaby  AYM   NaN         Yas Island Seaplane Base
3           AE    Abu Zaby  AZI  OMAD      Al Bateen Executive Airport
4           AE    Abu Zaby  DHF  OMAM               Al Dhafra Air Base

   latitude  longitude
0   24.2617    55.6092
1   24.4330    54.6511
2   24.4670    54.6103
```

```
3   24.4283    54.4581
4   24.2482    54.5477
```

[8]:
```
print(data1.info())
print(data1.describe())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8937 entries, 0 to 8936
Data columns (total 7 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   country_code  8905 non-null   object
 1   region_name   8937 non-null   object
 2   iata          8937 non-null   object
 3   icao          7794 non-null   object
 4   airport       8937 non-null   object
 5   latitude      8937 non-null   float64
 6   longitude     8937 non-null   float64
dtypes: float64(2), object(5)
memory usage: 488.9+ KB
None
          latitude    longitude
count  8937.000000  8937.000000
mean     20.151642     1.906493
std      28.610367    95.945375
min     -62.190800  -179.877000
25%      -4.196630   -82.783800
50%      26.760600     5.750000
75%      42.795500    92.819600
max      82.517800   179.976000
```

[19]:
```
import pandas as pd
import matplotlib.pyplot as plt

# Bar Plot: Top 10 regions by Number of Airports
top_countries = data1['region_name'].value_counts().head(10)

# Plotting
plt.figure(figsize=(10, 6))
top_countries.plot(kind='bar', color='skyblue')
plt.title('Top 10 Regions by Number of Airports')
plt.xlabel('Region')
plt.ylabel('Number of Airports')
plt.xticks(rotation=45)
plt.show()
```

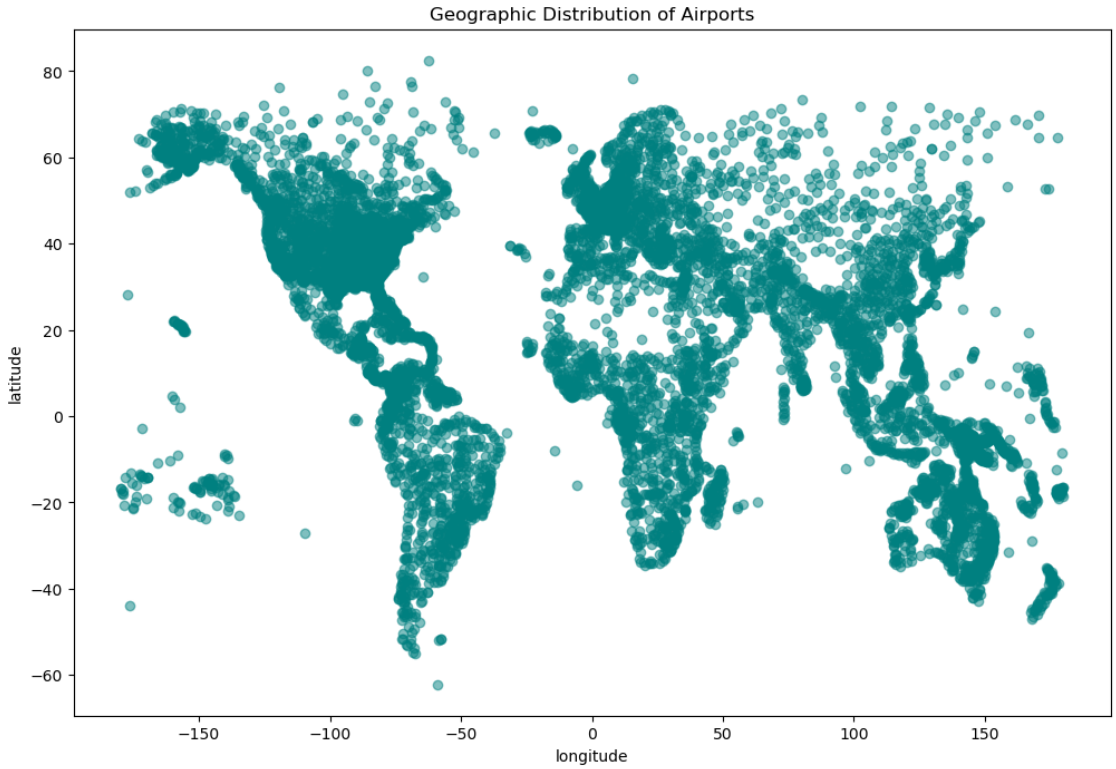


- This bar chart highlights the top 10 regions with the highest number of airports.
- It provides a quick glance at which regions may have high air traffic or aviation infrastructure.

Call to Action:

- Regions with the most airports may have either large geographic areas, high population density, or a well-developed transportation network.
- This could point toward a need for better airport management and resources in these countries due to potentially higher traffic.

[22]:
```
# Scatter Plot: Geographic Distribution of Airports by Longitude and Latitude
plt.figure(figsize=(12, 8))
plt.scatter(data1['longitude'], data1['latitude'], alpha=0.5, color='teal')
plt.title('Geographic Distribution of Airports')
plt.xlabel('longitude')
plt.ylabel('latitude')
plt.show()
```



- This scatter plot provides a visual representation of airport distribution worldwide.
- Each point represents an airport, and clustering can reveal high-density regions.

Call to Action:

- Dense clusters could represent highly connected regions with strong aviation networks.
- Sparse regions may indicate areas with limited or no airport coverage, potentially useful for market expansion or infrastructure development analysis.

[9]:
```
# Summarize the dataset 2
data2.head(5)
```

[9]:

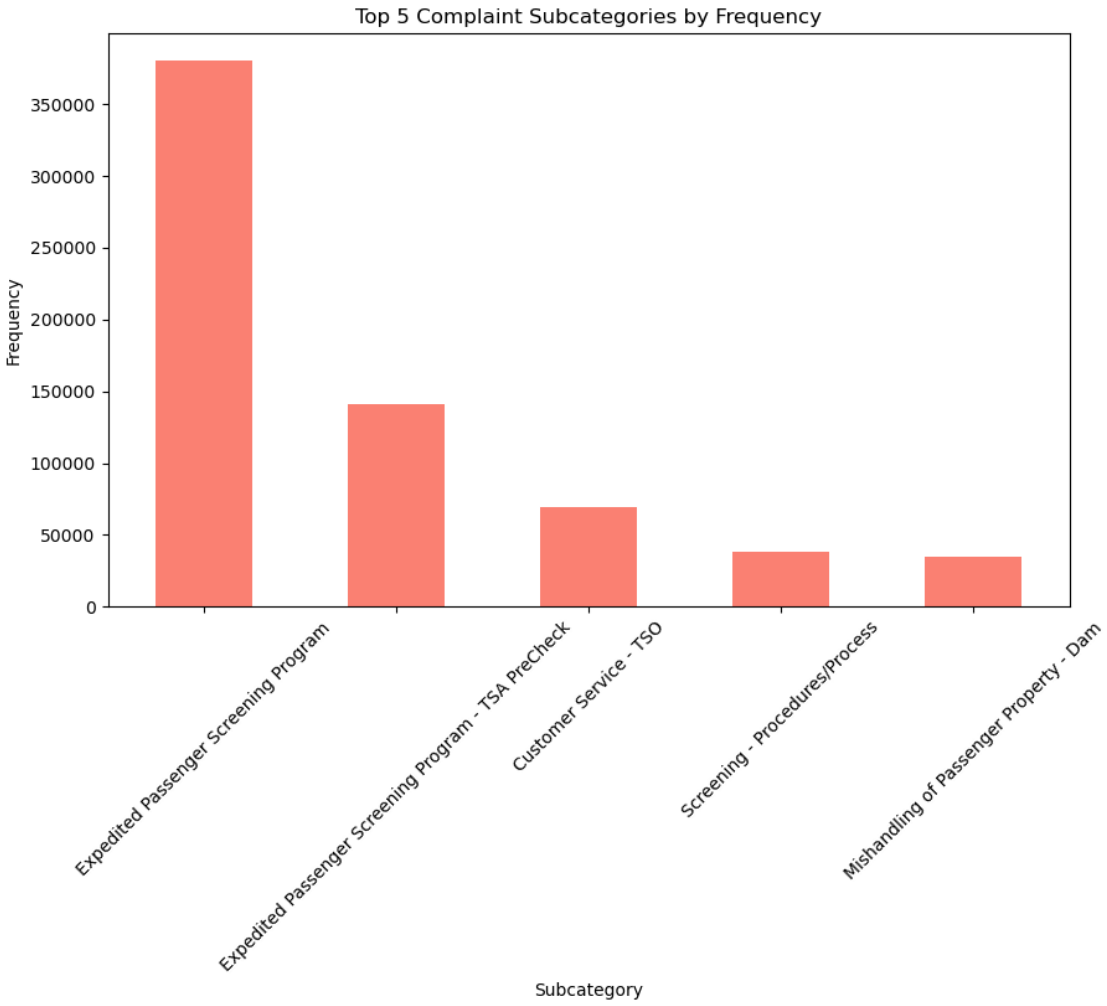
	pdf_report_date	airport	category	subcategory	year_month	count
0	2019-02	ABE	Hazardous Materials Safety	General	2015-01	0
1	2019-02	ABE	Mishandling of Passenger Property	Damaged/Missing Items--Checked Baggage	2015-01	0
2	2019-02	ABE	Hazardous Materials Safety	General	2015-02	0
3	2019-02	ABE	Mishandling of Passenger Property	Damaged/Missing Items--Checked Baggage	2015-02	0
4	2019-02	ABE	Hazardous Materials Safety			

```
4                                 General    2015-03      0

                          clean_cat                              clean_subcat  \
0         Hazardous Materials Safety                                   General
1  Mishandling of Passenger Property  *Damaged/Missing Items--Checked Baggage
2         Hazardous Materials Safety                                   General
3  Mishandling of Passenger Property  *Damaged/Missing Items--Checked Baggage
4         Hazardous Materials Safety                                   General

  clean_cat_status clean_subcat_status  is_category_prefix_removed
0         original            original                       False
1         original            original                       False
2         original            original                       False
3         original            original                       False
4         original            original                       False
```

[15]:
```
# Summarize the dataset 2
print(data2.info())
print(data2.describe())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 504512 entries, 0 to 504511
Data columns (total 11 columns):
 #   Column                      Non-Null Count   Dtype
---  ------                      --------------   -----
 0   pdf_report_date             504512 non-null  object
 1   airport                     491048 non-null  object
 2   category                    504512 non-null  object
 3   subcategory                 504512 non-null  object
 4   year_month                  504512 non-null  object
 5   count                       504512 non-null  int64
 6   clean_cat                   504512 non-null  object
 7   clean_subcat                504512 non-null  object
 8   clean_cat_status            504512 non-null  object
 9   clean_subcat_status         504512 non-null  object
 10  is_category_prefix_removed  504512 non-null  bool
dtypes: bool(1), int64(1), object(9)
memory usage: 39.0+ MB
None
               count
count  504512.000000
mean        2.024295
std        46.414187
min         0.000000
25%         0.000000
50%         0.000000
75%         1.000000
max      4588.000000
```

[27]:
```
# Top 5 Complaint Subcategories by Frequency
top_complaints = data2.groupby('subcategory')['count'].sum().
 ↪sort_values(ascending=False).head(5)

# Plotting
plt.figure(figsize=(10, 6))
top_complaints.plot(kind='bar', color='salmon')
plt.title('Top 5 Complaint Subcategories by Frequency')
plt.xlabel('Subcategory')
plt.ylabel('Frequency')
plt.xticks(rotation=45)
plt.show()
```

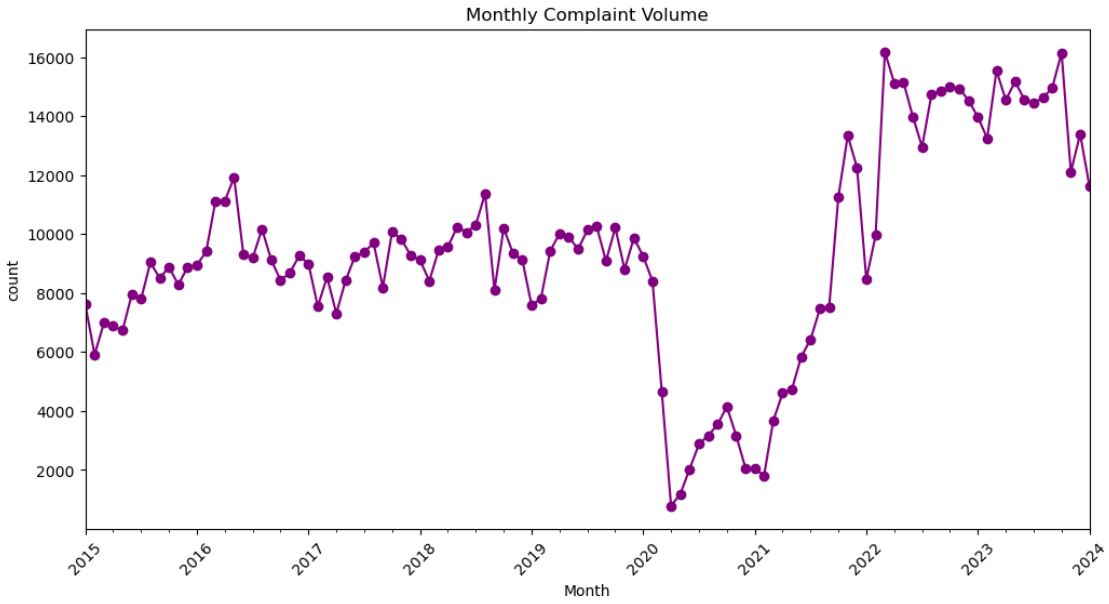


- This bar plot highlights the most frequent complaint subcategories.
- It allows easy identification of the types of issues that occur most frequently.

Call to Action:

- The most common subcategories of complaints can guide areas of focus for quality improvement or customer service enhancement.
- Analyzing these categories can help prioritize resources to address the most frequent issues.

[37]:
```
# Trend Line,  Monthly Complaint Volume
data2['year_month'] = pd.to_datetime(data2['year_month'])
monthly_trend = data2.groupby(data2['year_month'].dt.to_period('M'))['count'].
 ↪sum()

# Plotting
plt.figure(figsize=(12, 6))
monthly_trend.plot(kind='line', marker='o', color='purple')
plt.title('Monthly Complaint Volume')
plt.xlabel('Month')
plt.ylabel('count')
plt.xticks(rotation=45)
plt.show()
```



- The line plot shows complaint frequency over time on a monthly basis.

Call to Action:

- Spikes in certain months may indicate specific events, such as product launches or policy changes, that lead to higher complaint volumes.
- Consistent trends could reveal cyclical or seasonal issues, which could help in proactive issue management.

[40]:
```
# Summarize the dataset 3
data3.head(5)
```

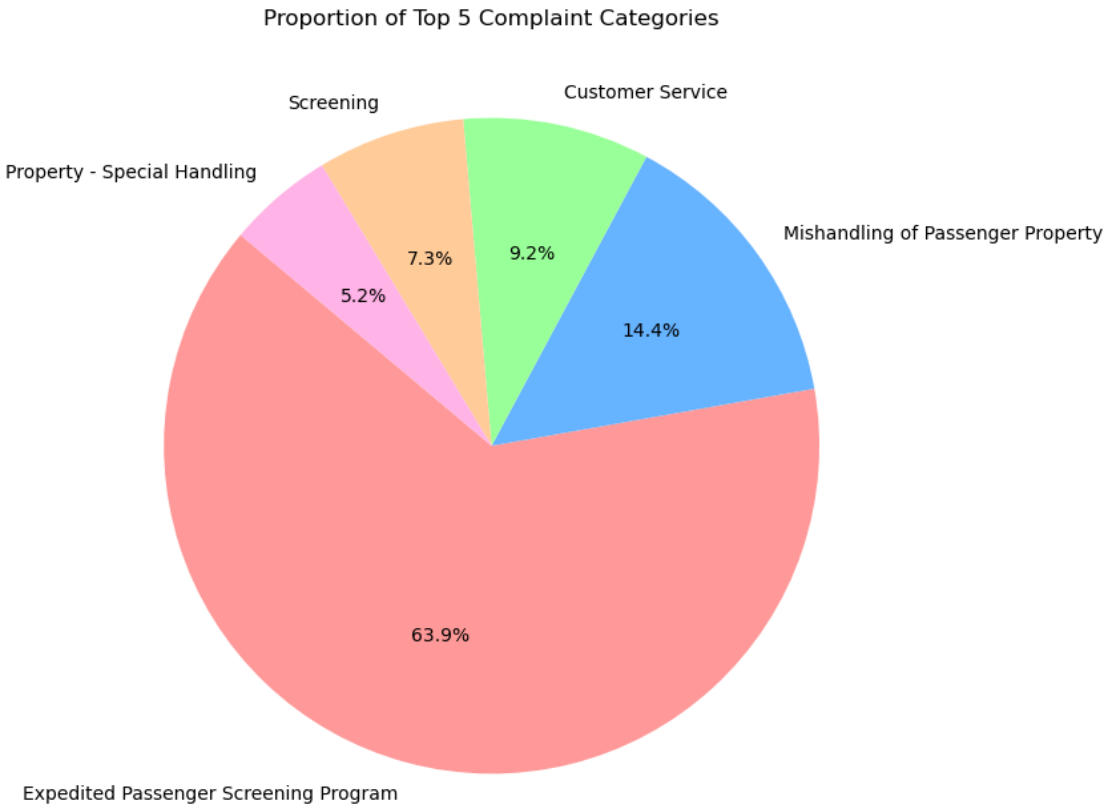
[40]:
```
  pdf_report_date airport                          category year_month  \
0         2019-02     ABE         Hazardous Materials Safety    2015-01
1         2019-02     ABE  Mishandling of Passenger Property    2015-01
2         2019-02     ABE         Hazardous Materials Safety    2015-02
3         2019-02     ABE  Mishandling of Passenger Property    2015-02
4         2019-02     ABE         Hazardous Materials Safety    2015-03

   count                          clean_cat clean_cat_status
0      0         Hazardous Materials Safety         original
1      0  Mishandling of Passenger Property         original
2      0         Hazardous Materials Safety         original
3      0  Mishandling of Passenger Property         original
4      0         Hazardous Materials Safety         original
```

[12]:
```
# Summarize the dataset 3
print(data3.info())
print(data3.describe())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 241588 entries, 0 to 241587
Data columns (total 7 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   pdf_report_date   241588 non-null  object
 1   airport           237546 non-null  object
 2   category          241588 non-null  object
 3   year_month        241588 non-null  object
 4   count             241588 non-null  int64
 5   clean_cat         241588 non-null  object
 6   clean_cat_status  241588 non-null  object
dtypes: int64(1), object(6)
memory usage: 12.9+ MB
None
               count
count  241588.000000
mean        4.227366
std        74.506112
min         0.000000
25%         0.000000
50%         0.000000
75%         1.000000
max      5953.000000
```

[42]:
```
# Proportion of Top 5 Complaint Categories
# Calculate total complaint counts by category and select the top 5
top_5_categories = data3.groupby('category')['count'].sum().nlargest(5)

# Plotting
```

```
plt.figure(figsize=(8, 8))
top_5_categories.plot(kind='pie', autopct='%1.1f%%', startangle=140,␣
↪colors=['#ff9999','#66b3ff','#99ff99','#ffcc99','#ffb3e6'])
plt.title('Proportion of Top 5 Complaint Categories')
plt.ylabel('')
plt.show()
```



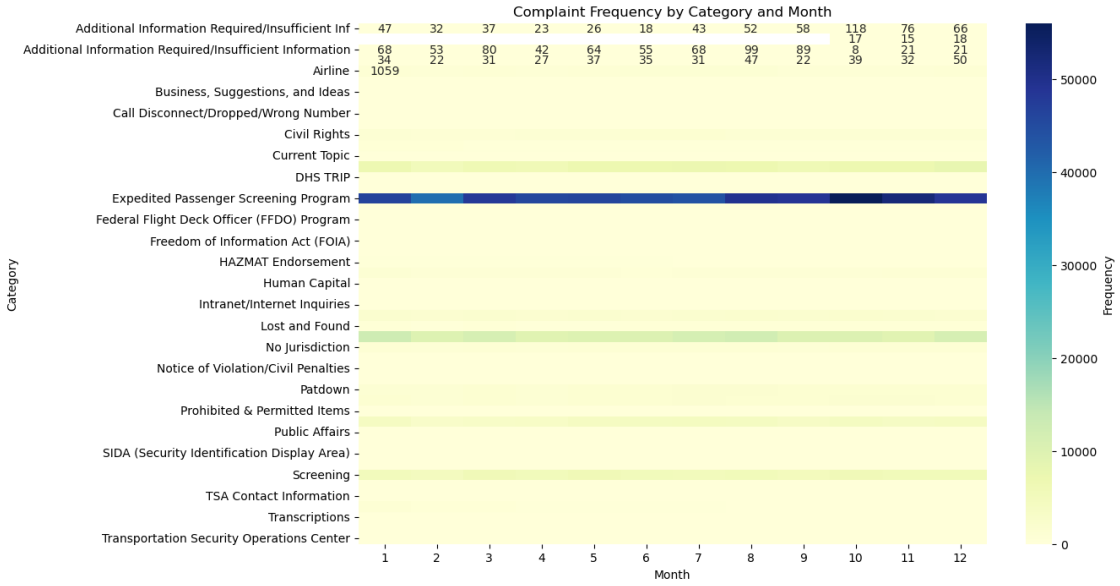
- This pie provides a focused view of the most significant complaint categories, making it easier to identify where most issues occur.

Call to Action:

- This view helps prioritize efforts to address the most critical categories affecting customer experience, potentially leading to a more significant impact on satisfaction levels.

[47]:
```
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd
import warnings
warnings.filterwarnings('ignore')
```

```
# Convert 'Date' column to datetime if it exists
data3['year_month'] = pd.to_datetime(data3['year_month'])

# Extract month and year for grouping
data3['Month'] = data3['year_month'].dt.month
data3['Year'] = data3['year_month'].dt.year

# Group by Category and Month to get total complaints in each combination
category_month_data = data3.groupby(['category', 'Month'])['count'].sum().
 ↪unstack()

# Plotting heat map
plt.figure(figsize=(12, 8))
sns.heatmap(category_month_data, annot=True, fmt='g', cmap="YlGnBu",␣
 ↪cbar_kws={'label': 'Frequency'})
plt.title('Complaint Frequency by Category and Month')
plt.xlabel('Month')
plt.ylabel('Category')
plt.show()
```



- It highlights if certain complaint categories have increasing or decreasing trends.

Call to Action: * Categories with increasing complaint volumes may indicate worsening issues that require attention.

[13]:
```
# Summarize the dataset 4
data4.head(5)
```

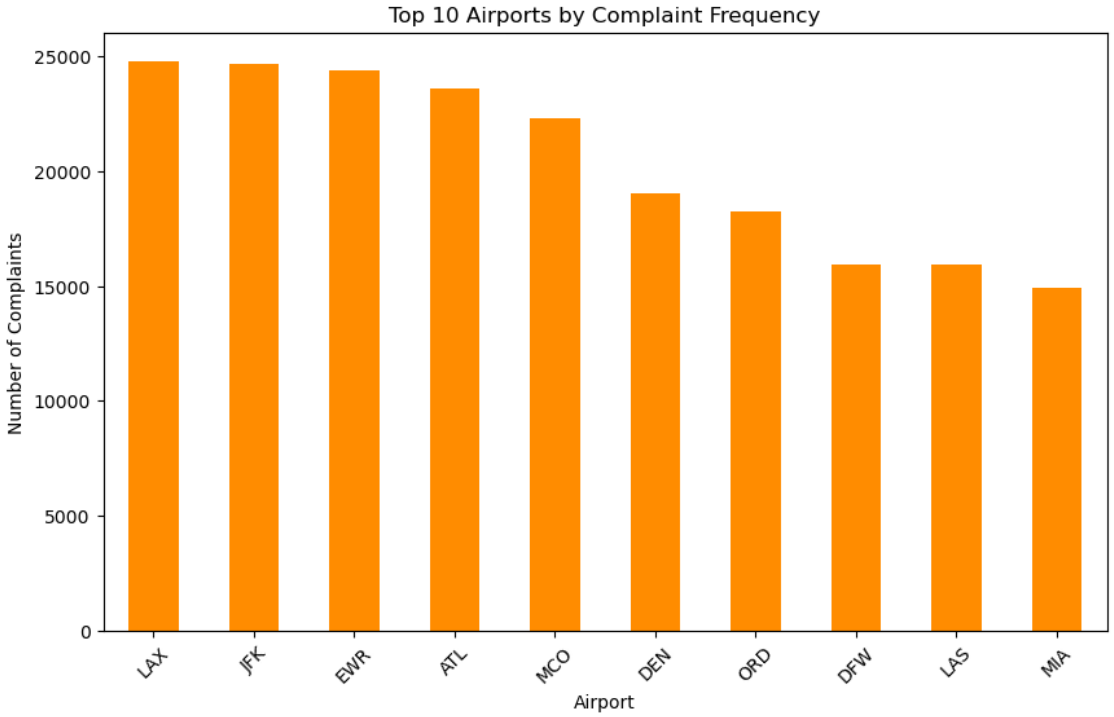
[13]:

	pdf_report_date	airport	year_month	count
0	2019-02	ABE	2015-01	0
1	2019-02	ABE	2015-02	0
2	2019-02	ABE	2015-03	0
3	2019-02	ABE	2015-04	0
4	2019-02	ABE	2015-05	2

[14]:
```
# Summarize the dataset 4
print(data4.info())
print(data4.describe())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41721 entries, 0 to 41720
Data columns (total 4 columns):
 #   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
 0   pdf_report_date  41721 non-null  object
 1   airport          41612 non-null  object
 2   year_month       41721 non-null  object
 3   count            41721 non-null  int64 
dtypes: int64(1), object(3)
memory usage: 1.3+ MB
None
              count
count  41721.000000
mean      24.478824
std      220.301338
min        0.000000
25%        0.000000
50%        1.000000
75%        6.000000
max     6604.000000
```

[45]:
```
# Top 10 Airports by Complaint Frequency
top_airports = data4.groupby('airport')['count'].sum().
 ↪sort_values(ascending=False).head(10)

# Plotting
plt.figure(figsize=(10, 6))
top_airports.plot(kind='bar', color='darkorange')
plt.title('Top 10 Airports by Complaint Frequency')
plt.xlabel('Airport')
plt.ylabel('Number of Complaints')
plt.xticks(rotation=45)
plt.show()
```



- This bar plot shows the airports with the highest complaint frequencies.
- It provides an easy way to identify which airports have the most reported issues.

Call to Action:

- Airports with a high volume of complaints may have operational, customer service, or infrastructure challenges that need attention.
- Prioritizing improvements at these airports could enhance customer satisfaction and reduce complaints.

2.1 Conclusion

This analysis provides valuable insights into complaint patterns across multiple dimensions, including airports, complaint types, and temporal trends. By examining the frequency and distribution of complaints, we identified key airports and categories that experience higher volumes of issues, suggesting areas where targeted improvements could enhance customer satisfaction. Monthly trends also revealed potential peak times for complaints, enabling proactive planning and resource allocation during high-traffic periods. These findings equip researchers and analysts with a deeper understanding of complaint dynamics, highlighting critical areas that may benefit from operational adjustments or service enhancements.

[ ]: