Aviation Data Set

Final Project Submission
Please fill out:
Student name: Charity Nguru
Student pace: part time
Scheduled project review date/time: 21/11/24
Instructor name: NOAH KANDIE

Business Understanding

Overview:
Exploring what the business expects to gain from the
project.Using the data in this project ,we will be able to identify the various potential risks in aviation industry.By identifying the various risks involved, such as financial risks,operational risks,safety and security risks,market and competitive risks among other risks that are associatedto aviation. With the data(Civil Aviation Accidents) we will focus more on safety and security risks, where we will analyze the number of accidents that occur and which routes they wereusing.This will help us identify the safest route to use which will have lower risks.

BUSINESS OBJECTIVES .

To gain profits and financial stability . Offer quality services to customersfor customer satisfaction . expand our market and staying competitive by adapting the market
trends . ensuring passenger and crew safety and security(main issue)

BUSINESS CRITERIA . To meet market demand . financial stability . regulatory compliance . operational efficiency

STATEMENT OF THE PROBLEM

Our company plans to invest in the aviation industry by
developing our own airstrip. To ensure safety and security, a critical component of this project will involve analyzing aviation accident data to identify patterns and key risk factors, such as routes or environmental conditions associated with higher accident probabilities. By leveraging these insights, we aim to make data-driven decisions to design safer operational procedures, select optimal routes, and implement strategies that mitigate safety risks effectively.

1. Business Questions
   Below are some of the business questions our project aims to address:

a) Safety and Security Risks:
What are the most significant factors contributing to aviation accidents in our target region?
Which routes, environmental conditions, or operational practiceshas the highest safety risks?
b)Operational Planning:
How can we optimize route planning to
minimize accident risk while maintaining operational efficiency?
What are the critical infrastructure or design requirements to enhance airstrip safety?
C) Risk Mitigation:
What strategies or technologies can be implemented to reduce accident risks?
How can we monitor and predict potential safety risks in real-time?
d)Financial Viability: What are the potential cost implications of implementing robust safety measures, and how do they compare to the costs ofaccidents or liabilities?

1. Any other business requirements

a) Access to historical aviation accident data (e.g., frequency, location, and causes of accidents).

b) Environmental and weather data for the proposed airstrip location.

c) Data on airstrip design, traffic volume, and regional aviationregulations.

2. expected benefits

a) Operational efficiency

b) Excellent reputation of the airstrip

c)Enhanced safety

d) financial savings as a result of reduced cost incurred from accidents

BUSINESS CRITERIA

1. Safety and Security Performance

-Accident Rate Reduction: Success will be measured by achieving a significant reduction in the projected accident rates compared to industry averages in similar conditions.

-Implementation of Risk Mitigation Strategies: Effective implementation of safety protocols, technologies, and operational practices designed to mitigate risks.

-Compliance with Safety Standards: Full adherence to local and internationalaviation safety regulations.

2. Financial stability Cost Efficiency: The project remains within the allocated budget, including the costs of safety measures, infrastructure, and operational processes. Revenue Targets: Achieving projected revenue from airstrip operations within the first year of launch.
3. Customer and Stakeholder Satisfaction Airline and Passenger Confidence: Positive feedback from airlines, passengers, and other stakeholders regarding safety and reliability. Partnership Development: Establishing strong relationships with local governments, airlines, and aviation authorities.

What sort of data are available for the project?

a) Existing Data:

Aviation Incident Data: Aviationaccident data from 1962–2023.
Variables: location, flight phases, weather conditions, accident causes, time of day, aircrafttype.
Environmental Data: Weather patterns, terrain details, and visibility conditions.
Variables:Temperature, wind speed, storms, icing, turbulence.
Operational Data: Airstrip infrastructure and flight management systems.
Variables: Runway lengths, lighting systems, navigation aids, airtraffic density.

Do we have the necessary resources to complete the project?

a) Human Resources:

DomainExperts: Aviation safety experts, pilots, and engineers for contextual insights.
Data Analysts/Scientists: Skilled in data cleaning, visualization, and statistical modeling.

Developers: For building dashboards or systems to operationalize the findings.

b) Technical Resources:

Hardware: High-performance computers for data processing and modeling.

Software: Tools foranalysis (e.g., Python, R, Tableau) and data storage (e.g., SQL databases).

Access to Data: Agreements or licenses for proprietary datasets.

c) Financial Resources: Budget for acquiring data, tools, and expertise. Allocation for additional resources, like cloud computing or training.

What are the risks involved?

a) Data Risks:

Incomplete Data: Missing critical variables may skewthe analysis.

Impact: Reduces the reliability of models and insights.

Data Privacy: Handling sensitive information (e.g., pilot data or accident reports) may pose legal issues.

Impact: Potential delays due to compliance requirements.

b) Operational Risks:

Resource Shortage: Lack of qualified personnel or tools.

Impact: Project delays or subpar outcomes.

Stakeholder Misalignment: Differing priorities or unclear goals among teams.

Impact: Inefficient execution orscope creep.

c) Analytical Risks:

Bias in Models: Historical biases may lead to flawed recommendations.

Impact: Misleading results and poor decisions.

Complexity: Over complicatedmodels may be difficult to implement.

Do we have a contingency plan for each risk?

a) Data Risks:

Mitigation for Incomplete Data: Collect additional data from alternative sources (e.g., international databases).

Use imputation techniques to handle missing values.

b) Operational Risks:

Resource Shortage: Outsource tasks to third-party experts if necessary.

Train internal teams to fill skill gaps.

Stakeholder Misalignment: Conduct frequentmeetings to align on goals and progress.

Document clear objectives .

c) Analytical Risks:

Bias in Models: Conduct sensitivity analyses and validate models with diversedatasets.

Complexity: Start with simpler models and gradually build complexity if needed.

In [2]:
```
#1. Inspect the Dataset
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [3]:
```
aviation_data = pd.read_csv('data/Aviation_Data.csv')

print(aviation_data.head())
```

```
C:\Users\Murugi\Anaconda3\envs\learn-env\lib\site-packages\IPython\core\interactivesh
ell.py:3145: DtypeWarning: Columns (6,7,28) have mixed types.Specify dtype option on
import or set low_memory=False.
  has_raised = await self.run_ast_nodes(code_ast.body, cell_name,
        Event.Id Investigation.Type Accident.Number  Event.Date  \
0  20001218X45444           Accident      SEA87LA080  1948-10-24
1  20001218X45447           Accident      LAX94LA336  1962-07-19
2  20061025X01555           Accident      NYC07LA005  1974-08-30
3  20001218X45448           Accident      LAX96LA321  1977-06-19
4  20041105X01764           Accident      CHI79FA064  1979-08-02

          Location        Country Latitude Longitude Airport.Code  \
0  MOOSE CREEK, ID  United States      NaN       NaN          NaN
1   BRIDGEPORT, CA  United States      NaN       NaN          NaN
2    Saltville, VA  United States  36.9222  -81.8781          NaN
3       EUREKA, CA  United States      NaN       NaN          NaN
4       Canton, OH  United States      NaN       NaN          NaN

  Airport.Name  ... Purpose.of.flight Air.carrier Total.Fatal.Injuries  \
0          NaN  ...          Personal         NaN                  2.0
1          NaN  ...          Personal         NaN                  4.0
2          NaN  ...          Personal         NaN                  3.0
3          NaN  ...          Personal         NaN                  2.0
4          NaN  ...          Personal         NaN                  1.0

  Total.Serious.Injuries Total.Minor.Injuries  Total.Uninjured  \
0                    0.0                  0.0              0.0
1                    0.0                  0.0              0.0
2                    NaN                  NaN              NaN
3                    0.0                  0.0              0.0
4                    2.0                  NaN              0.0

  Weather.Condition  Broad.phase.of.flight   Report.Status  Publication.Date
0               UNK                 Cruise  Probable Cause              NaN
1               UNK               Unknown  Probable Cause       19-09-1996
2               IMC                 Cruise  Probable Cause       26-02-2007
3               IMC                 Cruise  Probable Cause       12-09-2000
4               VMC              Approach  Probable Cause       16-04-1980

[5 rows x 31 columns]
```

In [4]:
```
print(aviation_data.info())
```

```
<class 'pandas.core.frame.DataFrame'>
Range Index: 90348 entries, 0 to
90347 Data columns (total 31
columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Event.Id                88889 non-null  object
 1   Investigation.Type      90348 non-null  object
 2   Accident.Number         88889 non-null  object
 3   Event.Date              88889 non-null  object
 4   Location                88837 non-null  object
 5   Country                 88663 non-null  object
 6   Latitude                34382 non-null  object
 7   Longitude               34373 non-null  object
 8   Airport.Code            50249 non-null  object
```

```
 9   Airport.Name            52790 non-null  object
 10  Injury.Severity         87889 non-null  object
 11  Aircraft.damage         85695 non-null  object
 12  Aircraft.Category       32287 non-null  object
 13  Registration.Number     87572 non-null  object
 14  Make                    88826 non-null  object
 15  Model                   88797 non-null  object
 16  Amateur.Built           88787 non-null  object
 17  Number.of.Engines       82805 non-null  float64
 18  Engine.Type             81812 non-null  object
 19  FAR.Description         32023 non-null  object
 20  Schedule                12582 non-null  object
 21  Purpose.of.flight       82697 non-null  object
 22  Air.carrier             16648 non-null  object
 23  Total.Fatal.Injuries    77488 non-null  float64
 24  Total.Serious.Injuries  76379 non-null  float64
 25  Total.Minor.Injuries    76956 non-null  float64
 26  Total.Uninjured         82977 non-null  float64
 27  Weather.Condition       84397 non-null  object
 28  Broad.phase.of.flight   61724 non-null  object
 29  Report.Status           82508 non-null  object
 30  Publication.Date        73659 non-null  object
dtypes: float64(5), object(26)
memory usage: 21.4+ MB
None
```

In [5]:
```
missing_values = (aviation_data.isnull().sum() / len(aviation_data))
print(missing_values.sort_values(ascending=False))
```

```
Schedule                0.860738
Air.carrier             0.815735
FAR.Description         0.645559
Aircraft.Category       0.642637
Longitude               0.619549
Latitude                0.619449
Airport.Code            0.443828
Airport.Name            0.415704
Broad.phase.of.flight   0.316819
Publication.Date        0.184719
Total.Serious.Injuries  0.154613
Total.Minor.Injuries    0.148227
Total.Fatal.Injuries    0.142339
Engine.Type             0.094479
Report.Status           0.086776
Purpose.of.flight       0.084684
Number.of.Engines       0.083488
Total.Uninjured         0.081585
Weather.Condition       0.065868
Aircraft.damage         0.051501
Registration.Number     0.030726
Injury.Severity         0.027217
Country                 0.018650
Amateur.Built           0.017278
Model                   0.017167
Make                    0.016846
Location                0.016724
Event.Date              0.016149
Accident.Number         0.016149
Event.Id                0.016149
Investigation.Type      0.000000
dtype: float64
```

In [6]:
```
print(aviation_data.columns)
```

```
Index(['Event.Id', 'Investigation.Type', 'Accident.Number', 'Event.Date',
       'Location', 'Country', 'Latitude', 'Longitude', 'Airport.Code',
       'Airport.Name', 'Injury.Severity', 'Aircraft.damage',
       'Aircraft.Category', 'Registration.Number', 'Make', 'Model',
       'Amateur.Built', 'Number.of.Engines', 'Engine.Type', 'FAR.Description',
```

```
       'Schedule', 'Purpose.of.flight', 'Air.carrier', 'Total.Fatal.Injuries',
       'Total.Serious.Injuries', 'Total.Minor.Injuries', 'Total.Uninjured',
       'Weather.Condition', 'Broad.phase.of.flight', 'Report.Status',
       'Publication.Date'],
      dtype='object')
```

In [7]:
```
print(aviation_data.isnull().sum())
```

```
Event.Id                 1459
Investigation.Type          0
Accident.Number          1459
Event.Date               1459
Location                 1511
Country                  1685
Latitude                55966
Longitude               55975
Airport.Code            40099
Airport.Name            37558
Injury.Severity          2459
Aircraft.damage          4653
Aircraft.Category       58061
Registration.Number      2776
Make                     1522
Model                    1551
Amateur.Built            1561
Number.of.Engines        7543
Engine.Type              8536
FAR.Description         58325
Schedule                77766
Purpose.of.flight        7651
Air.carrier             73700
Total.Fatal.Injuries    12860
Total.Serious.Injuries  13969
Total.Minor.Injuries    13392
Total.Uninjured          7371
Weather.Condition        5951
Broad.phase.of.flight   28624
Report.Status            7840
Publication.Date        16689
dtype: int64
```

In [8]:
```
# categorical columns with the mode (most frequent value)
aviation_data['Location']  =  aviation_data['Location'].fillna(aviation_data['Location
aviation_data['Country']  =  aviation_data['Country'].fillna(aviation_data['Country'].
aviation_data['Injury.Severity']  =  aviation_data['Injury.Severity'].fillna(aviation_
aviation_data['Aircraft.damage']  =  aviation_data['Aircraft.damage'].fillna(aviation_
aviation_data['Make']  =  aviation_data['Make'].fillna(aviation_data['Make'].mode()[0]
aviation_data['Model']  =  aviation_data['Model'].fillna(aviation_data['Model'].mode()
aviation_data['Amateur.Built']  =  aviation_data['Amateur.Built'].fillna(aviation_data

#  numeric columns with the median
aviation_data['Total.Minor.Injuries']  =  aviation_data['Total.Minor.Injuries'].fillna
aviation_data['Total.Uninjured']  =  aviation_data['Total.Uninjured'].fillna(aviation_
```

In [9]:
```
print(aviation_data.isnull().sum())
```

```
Event.Id                 1459
Investigation.Type          0
Accident.Number          1459
Event.Date               1459
Location                    0
Country                     0
Latitude                55966
Longitude               55975
Airport.Code            40099
Airport.Name            37558
Injury.Severity             0
Aircraft.damage             0
```

```
Aircraft.Category         58061
Registration.Number        2776
Make                          0
Model                         0
Amateur.Built                 0
Number.of.Engines          7543
Engine.Type                8536
FAR.Description           58325
Schedule                  77766
Purpose.of.flight          7651
Air.carrier               73700
Total.Fatal.Injuries      12860
Total.Serious.Injuries    13969
Total.Minor.Injuries          0
Total.Uninjured               0
Weather.Condition          5951
Broad.phase.of.flight     28624
Report.Status              7840
Publication.Date          16689
dtype: int64
```

In [10]:
```
aviation_data = aviation_data.dropna(subset=['Event.Date'],['Latitude'],['Longitude'
print(aviation_data.isnull().sum())
```

```
  File "<ipython-input-10-69c860e5953d>", line 1
    aviation_data = aviation_data.dropna(subset=['Event.Date'],['Latitude'],['Longitu
de'],['Airport.Code'],['Airport.Name']  ,['Aircraft.Category'],['Registration.Numbe
r'],['FAR.Description'],['Schedule'],['Air.Carrier'] ,['Total.Fatal.Injuries'],['Tota
l.Serious.Injuries'],['Weather.Condition'],['Broad.phase.of.flight'],['Report.Statu
s'],['Publication.Date'])
                                                                                    ^
SyntaxError: positional argument follows keyword argument
```

In [11]:
```
aviation_data = aviation_data.dropna(subset=[
    'Event.Date', 'Latitude', 'Longitude', 'Airport.Code', 'Airport.Name',
    'Aircraft.Category', 'Registration.Number', 'FAR.Description', 'Schedule',
    'Air.carrier', 'Total.Fatal.Injuries', 'Total.Serious.Injuries',
    'Weather.Condition', 'Broad.phase.of.flight', 'Report.Status', 'Publication.Date
])
```

In [13]:
```
print(aviation_data.isnull().sum())
```

```
Event.Id                  0
Investigation.Type        0
Accident.Number           0
Event.Date                0
Location                  0
Country                   0
Latitude                  0
Longitude                 0
Airport.Code              0
Airport.Name              0
Injury.Severity           0
Aircraft.damage           0
Aircraft.Category         0
Registration.Number       0
Make                      0
Model                     0
Amateur.Built             0
Number.of.Engines         0
Engine.Type               0
FAR.Description           0
Schedule                  0
Purpose.of.flight         6
Air.carrier               0
Total.Fatal.Injuries      0
Total.Serious.Injuries    0
Total.Minor.Injuries      0
```

```
Total.Uninjured           0
Weather.Condition         0
Broad.phase.of.flight     0
Report.Status             0
Publication.Date          0
dtype: int64
```

AFTER CLEANING ALL THE MISSING DATA, ANALYZE THE DATA

What are the most significant factors contributing to aviation accidents in our target region?

I'll consider factors like; environment factors i.e weather.conditions, Latitude and Longitude ,Event.
Date aircraft characteristics i.e Aircraft.category, Registration.Number, Human factors i.e Broad.Phase.of.light, Report.status, Total.Fatal.Injuries, Total.Serious.Injuries

1. What are the most significant factors contributing to aviation accidents in our target region?
   ...to answer this statement problem, I'll consider significant factors that include:

   1) weatherconditions-Poor visibility, storms, or icing can lead to accidents.

   Variables: visibility, wind speed, temperature, etc

   2) Human Error- Pilot, air traffic control, or maintenance crew errors.
   Variables: pilot experience, crew fatigue,training hours

   3) Flight Phase

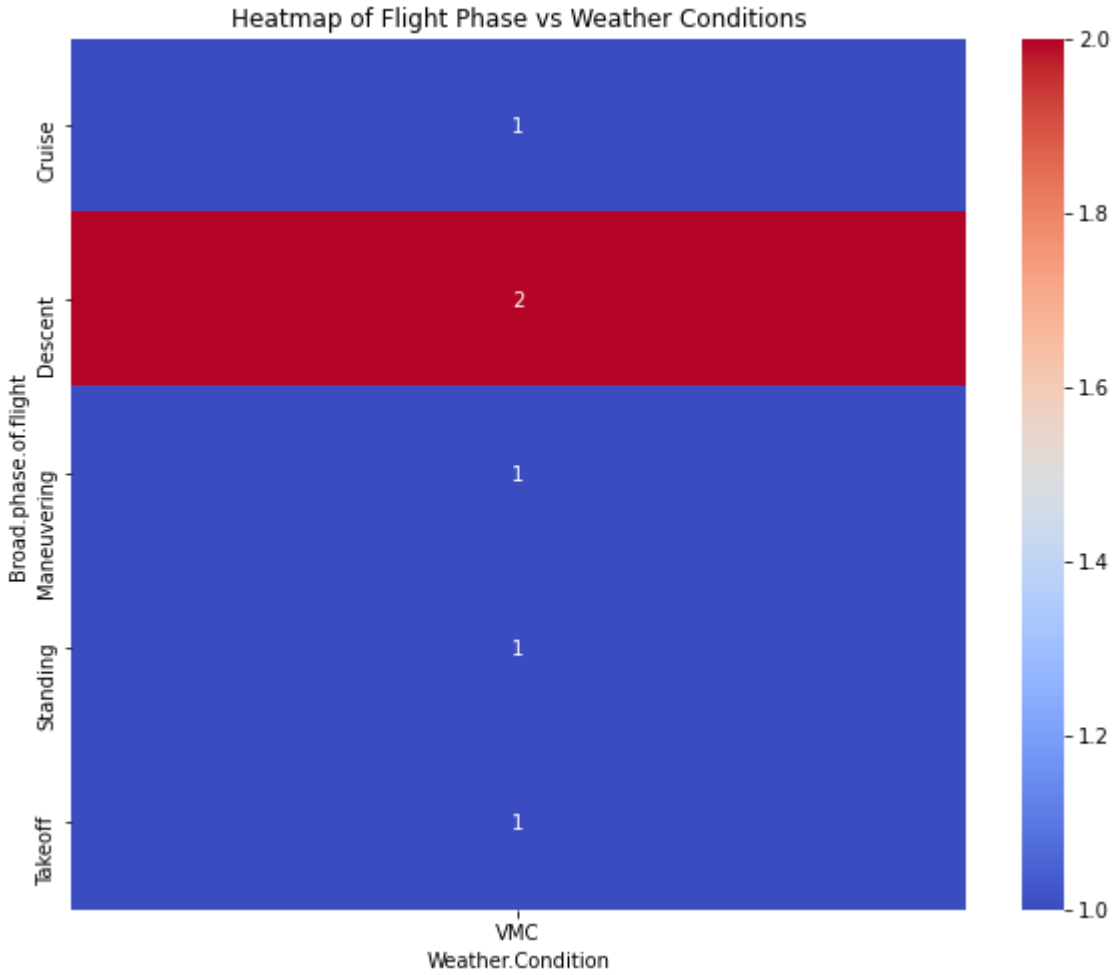
   Accidents often occur during takeoff, landing, or approach. Variables: flight phase, altitude, speed.

   4) Aircraft Technical Issues

   Mechanical failures or maintenance oversights. Variables: aircraft age, maintenance frequency, system failures.

Below is a heat map to analyze the relationship between flight phase and weather condition. Ideally weather conditions have a high influence on accident occurrence during any phase of flight.

In [20]:
```
#  Heatmap of flight phase vs weather conditions
heatmap_data = aviation_data.pivot_table(index='Broad.phase.of.flight', columns='Wea
plt.figure(figsize=(10, 8))
sns.heatmap(heatmap_data, annot=True, cmap='coolwarm', fmt='.0f')
plt.title('Heatmap of Flight Phase vs Weather Conditions')
plt.show()
```



In [ ]:
```
INTERPRATATION:

1. General Layout
X-axis: Represents the Weather Condition (in this case, "VMC").
Y-axis: Represents the Broad Phase of Flight (e.g., "Cruise," "Descent," etc.).
Cells: The numbers in the cells indicate the count of accidents under each combinati
```

. Observations From the heatmap:

Most flight phases (like "Cruise," "Landing," "Standing," and "Takeoff") have 1 accident under the "VMC" weather condition. The "Descent" phase has 2 accidents, which is highlighted by the red color indicating a higher value.

High-risk phase: The "Descent" phase under "VMC" (Visual Meteorological Conditions) shows a higher frequency of accidents compared to other phases.
This might suggest that there are Operational challenges during descent, even in clear weather caused by Possible issues like pilot error,communication, or procedural difficulties.

Low-risk phases: Other phases like "Takeoff" and "Cruise" show fewer accidents under the same weather condition.

1. Actionable Steps Based on the analysis ,our company should :

Focus on "VMC" Training: Since these accidents occur under Visual Meteorological Conditions, this might indicate human error, so targeted pilot training could help.

second statement problem:

Which routes, environmental conditions, or operational practices has the highest safety risks?

I'll consider using the following factors :

a) Routes:

Variables: Origin and destination airports, flight paths, regional accident density.
Focus: Identifyaccident-prone routes or areas with higher risk (e.g., mountainous terrain, high air traffic regions).

B)Environmental Conditions:

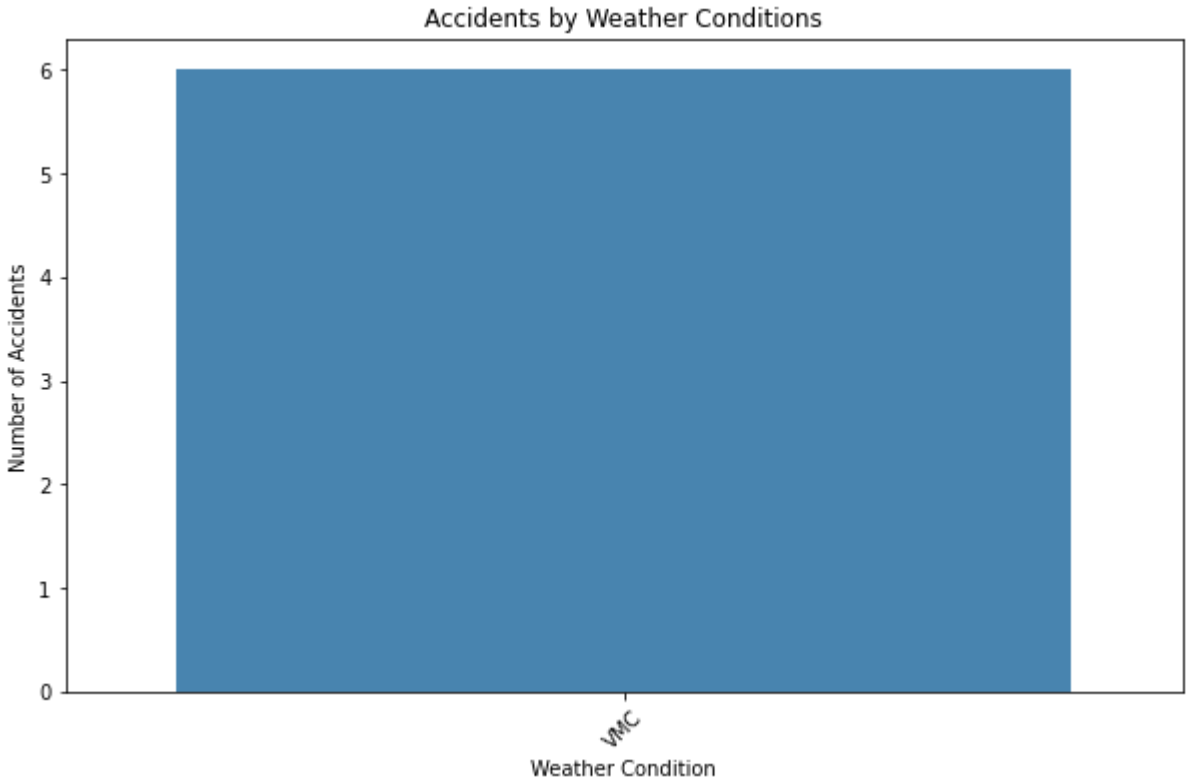
Variables: Weather conditions (e.g., VMC, IMC, turbulence, icing, wind).
Focus: To analyze accidentsunder adverse weather conditions like storms, poor visibility, or high wind speeds.

C)Operational Practices:

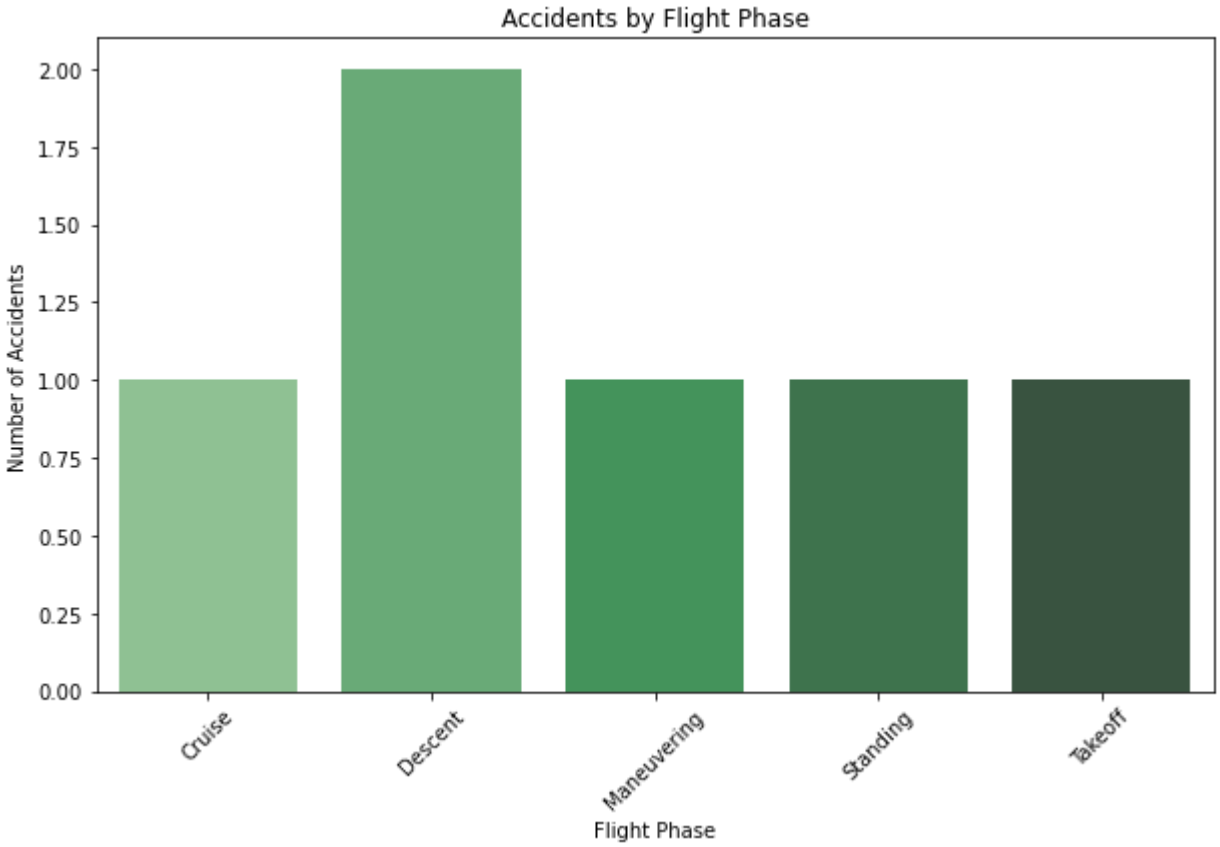
Variables: Flight phase (e.g., takeoff, landing, cruise), maintenance records, pilot experience.
Focus: Identify flight phases or practices with higher accident frequencies (e.g., improper procedures, fatigue).

we will identify High-Risk Categories by determine which routes, conditions, or practices have the most accidents or highest severity.

In [24]:
```
import matplotlib.pyplot as plt
import seaborn as sns

# Group accidents by Weather Condition
weather_data = aviation_data.groupby('Weather.Condition')['Accident.Number'].count()
weather_data.rename(columns={'Accident.Number': 'Accident Count'}, inplace=True)

# Plot Bar Graph
plt.figure(figsize=(10, 6))
sns.barplot(data=weather_data, x='Weather.Condition', y='Accident Count', palette='B
plt.title('Accidents by Weather Conditions')
plt.xlabel('Weather Condition')
plt.ylabel('Number of Accidents')
plt.xticks(rotation=45)
plt.show()
```



In [25]:
```
# Group accidents by Flight Phase
flight_phase_data = aviation_data.groupby('Broad.phase.of.flight')['Accident.Number'
flight_phase_data.rename(columns={'Accident.Number': 'Accident Count'}, inplace=True

# Plot Bar Graph
plt.figure(figsize=(10, 6))
sns.barplot(data=flight_phase_data, x='Broad.phase.of.flight', y='Accident Count', p
plt.title('Accidents by Flight Phase')
plt.xlabel('Flight Phase')
plt.ylabel('Number of Accidents')
plt.xticks(rotation=45)
plt.show()
```



ADDRESSING HIGH-RISK FACTORS

Environmental Conditions: Implement stricter weather-related takeoff and landing policy Train pilots for handling adverse weather scenarios.

Operational Practices: Increase oversight on maintenance and pre-flight checks. Introduce more advanced flight simulators to train pilots for risky phases (e.g., landing, descent).

Operational Planning

1. How can we optimize route planning to minimize accident risk while maintaining operational efficiency?

To optimize route planning while balancing safety and efficiency,consider these strategies:

A)Risk-Based Route Prioritization By Analyzing Historical Accident Data:

B)Optimize Routes for Environmental Conditions:

Avoid routes prone to severe weather conditions (e.g., turbulence, icing). Implement real-time weather monitoring and dynamic route adjustments based on forecasts.

Flight Path Optimization:

Favor direct routes to reduce time in air while maintaining safe separation from obstacles. Use predictive analytic to identify and mitigate potential risks on planned routes.

C)Enhance Pilot and Crew Decision-Making Provide Detailed Navigation Data by Equipping pilots with accurate terrain maps, weather overlays, and predictive hazard warnings.

D)Route Categorization: Define routes as "High Risk" or "Low Risk" based on safety metrics, andallocate experienced crews to higher-risk routes.

E)Integrate Technology Advanced Flight Management Systems: Use automated systems to calculate the safest and most fuel-efficient routes.

F)Traffic Flow Management: Optimize air traffic control to reduce congestion on high-risk routes and alternate routes during peak hours.

1.What are the critical infrastructure or design requirements to enhance airstrip safety?

To enhance airstrip safety, focus on both infrastructure and operational design:

a) Infrastructure Requirements.

Runway Design and Maintenance: To Ensure runways are of sufficient length, width, and strength to handle the largest aircraft using the airstrip. Regularly inspect and repair runway surfaces to avoid debris or wear-related accidents.

Lighting and Markings: Install high-visibility runway lighting (e.g., LED lights) for night and low-visibility operations.

Use clear and standardized markings for runways, taxiways, and aprons.

Navigation Aids: Install Instrument Landing Systems (ILS), GPS-based navigation aids, and radar for improved landing accuracy. Enhance airstrip weather monitoring systems for real-time updates to pilots and controllers.

Emergency Services: Ensure proximity to fire services, medical facilities, and emergency response teams.I.e Equipping airstrips with quick-response fire extinguishing systems and crash rescue vehicles.

Perimeter Security: Prevent wildlife intrusion or unauthorized access through robust fencing and surveillance systems.

b) Operational Design Requirements

Air Traffic Management (ATM): Upgrade communication systems for seamless coordination between pilots and air traffic control. Implement surface movement guidance systems to prevent ground collisions.

Standardized Safety Protocols: Regularly update airstrip operating procedures to meet international aviation standards (e.g., ICAO). Conduct frequent drills to ensure readiness for emergencies.

Capacity Management: Ensure airstrip capacity aligns with traffic demand to avoid congestion. Use advanced scheduling software to minimize ground delays and optimize runway use.

The above are solutions to our various business questions that had arise earlier. The company needs to implement the above for smooth running of the business as well as accomplishing business goals.

Name: Charity Nguru
Email: charitymurugi55@gmail.com
Linkedin: https://www.linkedin.com/in/charity-murugi-070bb831b/

In [ ]: