import library

In [2]:
```python
# import library
import opendatasets as od
import os
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

Import data from Kaggle

In [3]:
```python
# กำหนด URL ของ dataset บน kaggle
url_data = 'https://www.kaggle.com/datasets/iamsouravbanerjee/airline-dataset'
# download dataset form kaggle โดยต้องระบุ URL ของ dataset ที่ต้องการโหลดก่อน
# ใช้ Kaggle AIP username and key
od.download(url_data)
```

```
Skipping, found downloaded files in ".\airline-dataset" (use force=True to force download)
```

In [4]:
```python
# ระบุ directory ที่จะทำการบันทึก data
data_dir = './airline-dataset'
data_dir
```

Out[4]:
```
'./airline-dataset'
```

In [5]:
```python
# download file ที่อยู่ใน directory ที่ระบุ
# output is file name
os.listdir(data_dir)
```

Out[5]:
```
['Airline Dataset Updated - v2.csv',
 'Airline Dataset Updated.csv',
 'Airline Dataset.csv']
```

Create data frame

In [6]:
```python
# Create data frame
file_path = data_dir + '\Airline Dataset Updated.csv' # file_path = data_dir + 'file name'
df = pd.read_csv(file_path)
df
```

Out[6]:

	Passenger ID	First Name	Last Name	Gender	Age	Nationality	Airport Name	Airport Country Code	Country Name	Airport Continent	Continents	Departure Date	Arrival Airport	Pilot Name	Flight Status
0	ABVWIg	Edithe	Leggis	Female	62	Japan	Coldfoot Airport	US	United States	NAM	North America	6/28/2022	CXF	Edithe Leggis	On Time
1	jkXXAX	Elwood	Catt	Male	62	Nicaragua	Kugluktuk Airport	CA	Canada	NAM	North America	12/26/2022	YCO	Elwood Catt	On Time
2	CdUz2g	Darby	Felgate	Male	67	Russia	Grenoble-Isère Airport	FR	France	EU	Europe	1/18/2022	GNB	Darby Felgate	On Time
3	BRS38V	Dominica	Pyle	Female	71	China	Ottawa / Gatineau Airport	CA	Canada	NAM	North America	9/16/2022	YND	Dominica Pyle	Delayed
4	9kvTLo	Bay	Pencost	Male	21	China	Gillespie Field	US	United States	NAM	North America	2/25/2022	SEE	Bay Pencost	On Time
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
98614	hnGQ62	Gareth	Mugford	Male	85	China	Hasvik Airport	NO	Norway	EU	Europe	12/11/2022	HAA	Gareth Mugford	Cancelled
98615	2omEzh	Kasey	Benedict	Female	19	Russia	Ampampamena Airport	MG	Madagascar	AF	Africa	10/30/2022	IVA	Kasey Benedict	Cancelled
98616	VUPiVG	Darrin	Lucken	Male	65	Indonesia	Albacete-Los Llanos Airport	ES	Spain	EU	Europe	9/10/2022	ABC	Darrin Lucken	On Time
98617	E47NtS	Gayle	Lievesley	Female	34	China	Gagnoa Airport	CI	Côte d'Ivoire	AF	Africa	10/26/2022	GGN	Gayle Lievesley	Cancelled
98618	8JYEcz	Wilhelmine	Touret	Female	10	Poland	Yoshkar-Ola Airport	RU	Russian Federation	EU	Europe	4/16/2022	JOK	Wilhelmine Touret	Delayed

98619 rows × 15 columns

Overview of Data

In [7]:
```python
# show first 5 rows of data
df.head()
```

Out[7]:

	Passenger ID	First Name	Last Name	Gender	Age	Nationality	Airport Name	Airport Country Code	Country Name	Airport Continent	Continents	Departure Date	Arrival Airport	Pilot Name	Flight Status
0	ABVWIg	Edithe	Leggis	Female	62	Japan	Coldfoot Airport	US	United States	NAM	North America	6/28/2022	CXF	Edithe Leggis	On Time
1	jkXXAX	Elwood	Catt	Male	62	Nicaragua	Kugluktuk Airport	CA	Canada	NAM	North America	12/26/2022	YCO	Elwood Catt	On Time
2	CdUz2g	Darby	Felgate	Male	67	Russia	Grenoble-Isère Airport	FR	France	EU	Europe	1/18/2022	GNB	Darby Felgate	On Time
3	BRS38V	Dominica	Pyle	Female	71	China	Ottawa / Gatineau Airport	CA	Canada	NAM	North America	9/16/2022	YND	Dominica Pyle	Delayed
4	9kvTLo	Bay	Pencost	Male	21	China	Gillespie Field	US	United States	NAM	North America	2/25/2022	SEE	Bay Pencost	On Time

In [8]:
```python
# Check information of data as number of rows and columns, all column name, data type and Non-Null Count
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
```

In [9]:
```python
# Check missing value in dataset
df.isnull().sum()
```

Out[9]:
```
Passenger ID            0
First Name              0
Last Name               0
Gender                  0
Age                     0
Nationality             0
Airport Name            0
Airport Country Code    0
Country Name            0
Airport Continent       0
Continents              0
Departure Date          0
Arrival Airport         0
Pilot Name              0
Flight Status           0
dtype: int64
```

In [10]:
```python
# Show statistics about data (show only numeric data)
df.describe()
```

Out[10]:

	Age
count	98619.000000
mean	45.504021
std	25.929849
min	1.000000
25%	23.000000
50%	46.000000
75%	68.000000
max	90.000000

Cleaning data

In [11]:
```python
# Cleaning data
# separate day and month from Departure Date
df['Departure Date'] = pd.to_datetime(df['Departure Date']) # แปลงคอลัม 'Departure Date' เป็น datetime
df['Day_of_week'] = df['Departure Date'].dt.dayofweek # เพิ่มคอลัม 'Day_of_Week' โดยใช้ .dt.dayofweek
df['Month'] = df['Departure Date'].dt.month
```

In [12]:
```python
# Check data after clean
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 98619 entries, 0 to 98618
Data columns (total 17 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  datetime64[ns]
 12  Arrival Airport       98619 non-null  object        
 13  Pilot Name            98619 non-null  object        
 14  Flight Status         98619 non-null  object        
 15  Day_of_week           98619 non-null  int32         
 16  Month                 98619 non-null  int32         
dtypes: datetime64[ns](1), int32(2), int64(1), object(13)
memory usage: 12.0+ MB
```

Visualization

In [13]:
```python
# Create data frame show only Nationality is Thailand
df_th = df.query('Nationality == "Thailand"').reset_index().drop('index', axis = 1 )
df_th
```

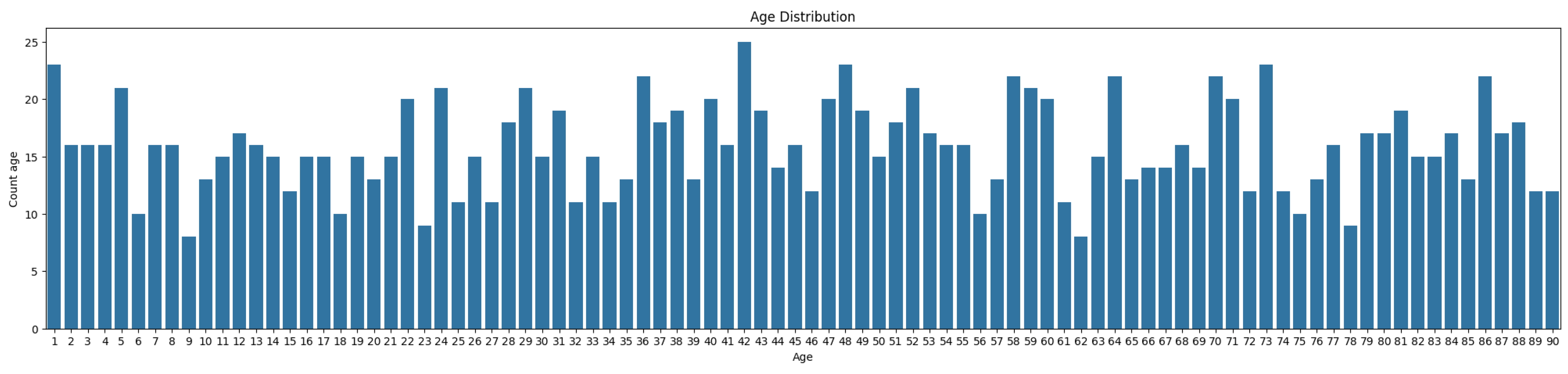
Out[13]:

	Passenger ID	First Name	Last Name	Gender	Age	Nationality	Airport Name	Airport Country Code	Country Name	Airport Continent	Continents	Departure Date	Arrival Airport	Pilot Name	Flight Status	Day_of_week	Month
0	3jmudz	Burlie	Schustl	Male	13	Thailand	Vermilion Airport	CA	Canada	NAM	North America	2022-04-06	YVG	Burlie Schustl	On Time	2	4
1	x6rnf2	Aprilette	Veysey	Female	13	Thailand	Trona Airport	US	United States	NAM	North America	2022-05-28	TRH	Aprilette Veysey	Cancelled	5	5
2	vGOv3o	Neal	Kulver	Male	69	Thailand	Kaitaia Airport	NZ	New Zealand	OC	Oceania	2022-04-24	KAT	Neal Kulver	Delayed	6	4
3	bXrWFO	Jody	Philpin	Male	34	Thailand	Punta Colorada Airport	MX	Mexico	NAM	North America	2022-10-14	PCO	Jody Philpin	On Time	4	10
4	4t9AmH	Alisa	Dillintone	Female	70	Thailand	Crossville Memorial Whitson Field	US	United States	NAM	North America	2022-03-07	CSV	Alisa Dillintone	On Time	0	3
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1421	5pPhx9	Lyndsey	Dewer	Female	52	Thailand	Honinabi Airport	PG	Papua New Guinea	OC	Oceania	2022-03-12	HNN	Lyndsey Dewer	On Time	5	3
1422	UUCLbD	Aymer	Pregel	Male	27	Thailand	Dywer Airbase	AF	Afghanistan	AS	Asia	2022-04-22	DWR	Aymer Pregel	Cancelled	4	4
1423	Au5ftj	Arvie	Brewster	Male	37	Thailand	Tarakbits Airport	PG	Papua New Guinea	OC	Oceania	2022-09-04	TRJ	Arvie Brewster	Delayed	6	9
1424	g8QB5o	Torie	Measor	Female	80	Thailand	Kokkola-Pietarsaari Airport	FI	Finland	EU	Europe	2022-05-25	KOK	Torie Measor	Cancelled	2	5
1425	rqUg8x	Rosalie	Waldock	Female	6	Thailand	Trombetas Airport	BR	Brazil	SAM	South America	2022-12-08	TMT	Rosalie Waldock	On Time	3	12

1426 rows × 17 columns

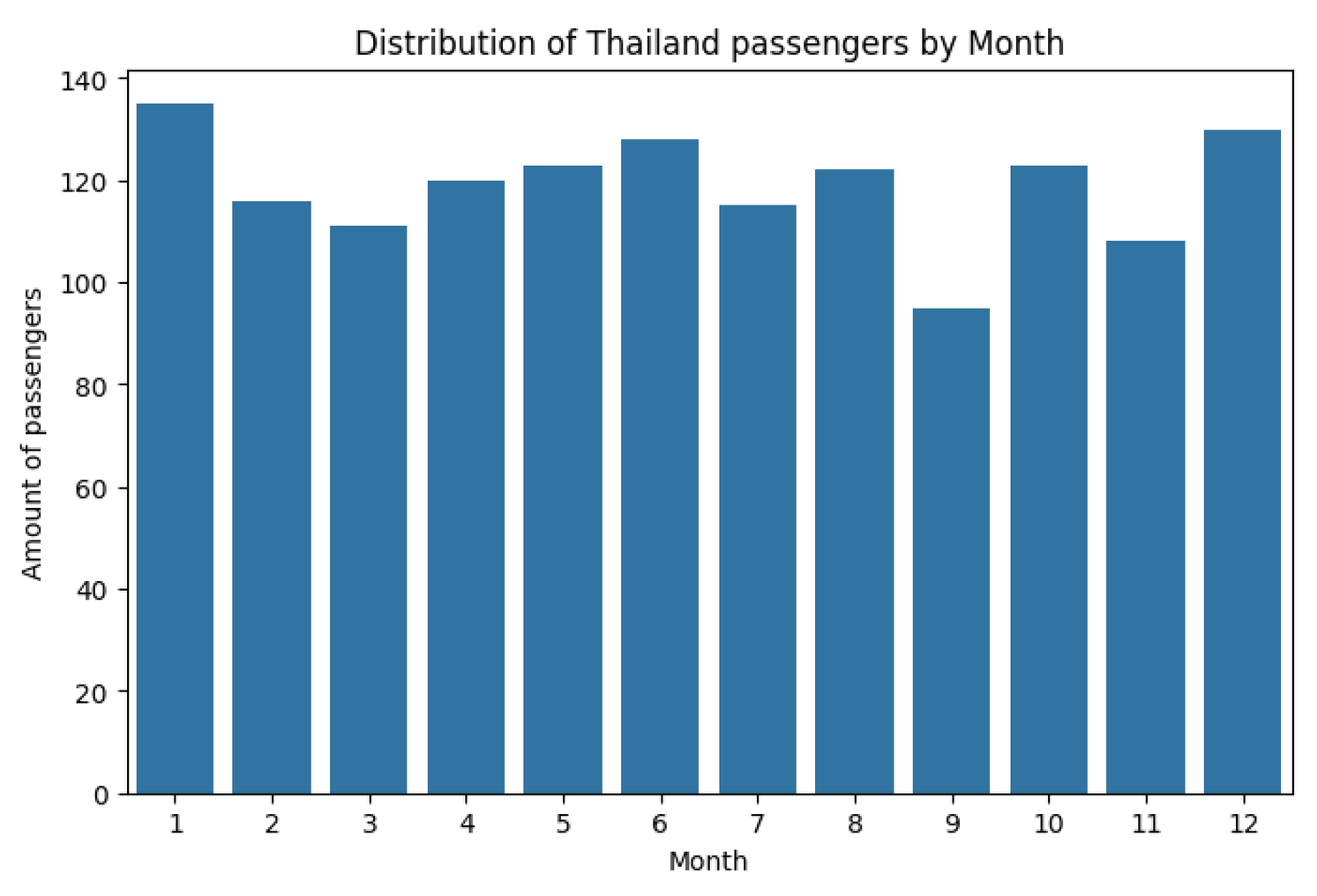
In [14]:
```python
# Create Chart
plt.figure(figsize=(25,5))
age_counts = df_th['Age'].value_counts()
sns.barplot(x = age_counts.index, y=age_counts.values)
plt.title('Age Distribution')
plt.xlabel('Age')
plt.ylabel('Count age')
```

Out[14]:
```
Text(0, 0.5, 'Count age')
```



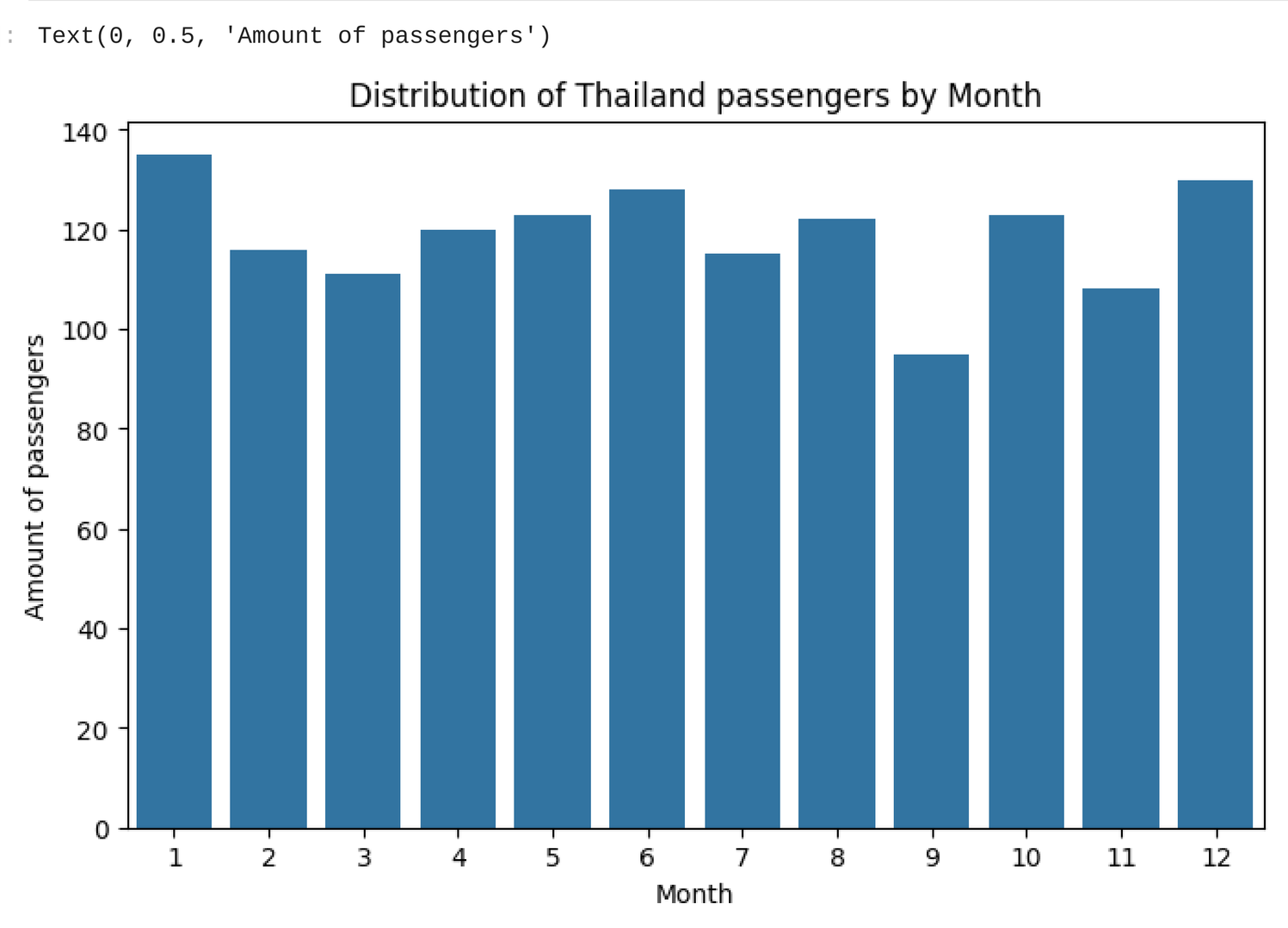
In [19]:
```python
# Create Chart
plt.figure(figsize=(8,5))
sns.countplot(x= 'Month', data = df_th)
plt.title('Distribution of Thailand passengers by Month')
plt.xlabel('Month')
plt.ylabel('Amount of passengers')
```

Out[19]:
```
Text(0, 0.5, 'Amount of passengers')
```



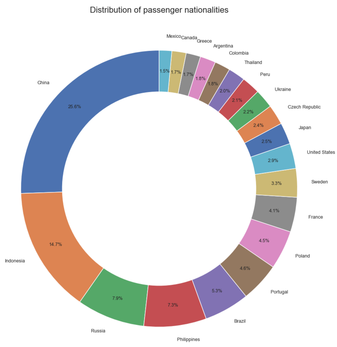
In [17]:
```python
# Create Chart
plt.figure(figsize=(8,5))
sns.countplot(x= 'Month', data = df_th)
plt.title('Distribution of Thailand passengers by Month')
plt.xlabel('Month')
plt.ylabel('Amount of passengers')
```

Out[17]:
```
Text(0, 0.5, 'Amount of passengers')
```



In [20]:
```python
top_20 = df['Nationality'].value_counts().nlargest(20)

plt.figure(figsize=(12, 12))
sns.set(style="whitegrid")

ax = top_20.plot.pie(autopct='%1.1f%%', pctdistance=0.85, startangle= 90, textprops={'fontsize': 9})

centre_circle = plt.Circle((0,0),0.70,fc='white')
fig = plt.gcf()
fig.gca().add_artist(centre_circle)
plt.ylabel('')

plt.title('Distribution of passenger nationalities', fontsize=16)
```

Out[20]:
```
Text(0.5, 1.0, 'Distribution of passenger nationalities')
```


Distribution of passenger nationalities
Mexico
Canada
Greece
Argentina
Colombia
Thailand
Peru
Ukraine
Czech Republic
Japan
United States
Sweden
France
Poland
Portugal
Brazil
Philippines
Russia
Indonesia
China
1.5%
1.7%
1.7%
2.0%
2.0%
2.1%
2.2%
2.5%
2.9%
3.3%
4.1%
4.5%
4.6%
5.3%
7.3%
7.9%
14.7%
25.4%

In [ ]: