AICP Internship Task Week 6

```
In [1]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import seaborn as sns
```

```
In [2]: data = pd.read csv("births.csv")
```

Q1: Add a new column “Decade” by calculating. For example 1969 will be 1960, 1988 will 1980 etc.

```
In [3]: data['Decade'] = (data['year'] // 10) * 10
```

Q2: Show the descriptive statistics of the data.

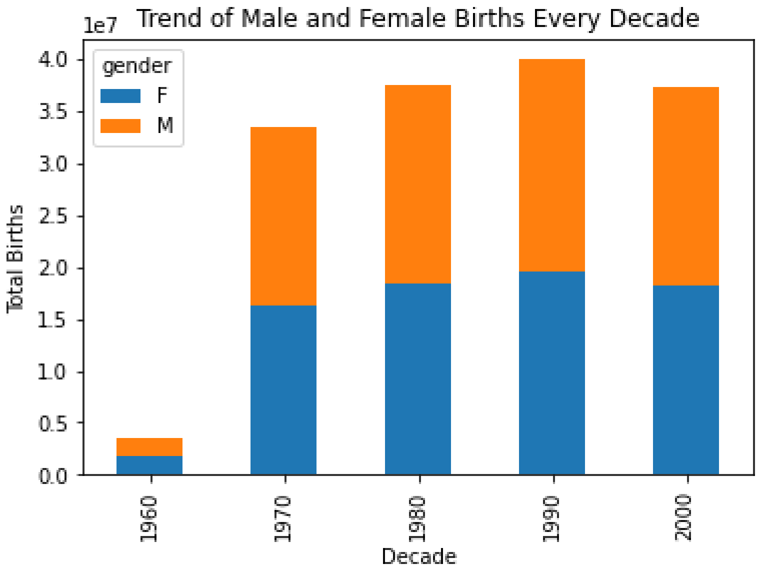
```
In [4]: desc stats = data.describe()
```

Q3: Check if your data contains any missing values.

```
In [5]: missing values = data.isnull().sum()
```

Q4: What is the trend of male & female births every decade?

```
In [6]: trend = data.groupby(['Decade', 'gender'])['births'].sum().unstack(
        trend.plot(kind='bar', stacked=True)
        plt.title('Trend of Male and Female Births Every Decade')
        plt.xlabel('Decade')
        plt.ylabel('Total Births')
        plt.show()
```



Q5: To remove outliers from dataset following techinque to include only those values that fall within 5 standard deviations from the mean. This is a common statistical technique used to focus on the central tendency of the data while excluding extreme values.

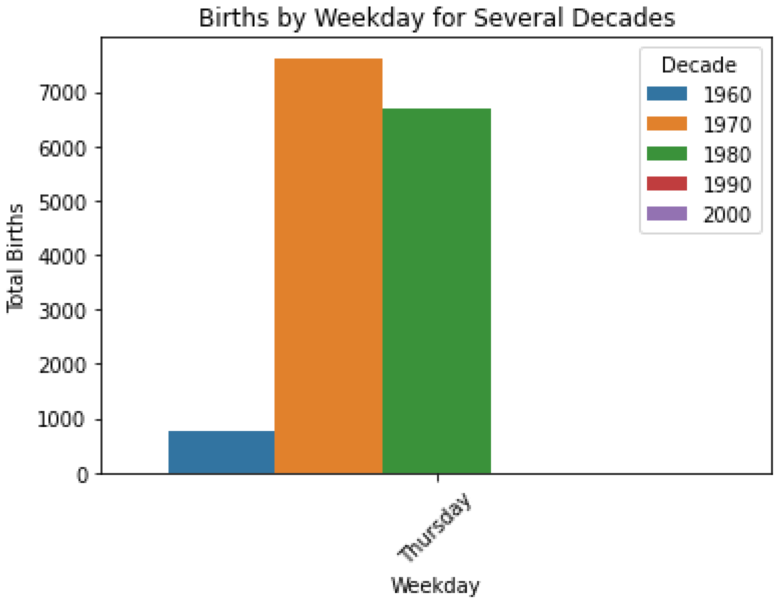
In [7]:
```
def remove_outliers(data, columns):
    z_scores = np.abs((data[columns] - data[columns].mean()) / data
    data_cleaned = data[(z_scores < 5).all(axis=1)]
    return data_cleaned

data_cleaned = remove_outliers(data, ['births'])
```

Use this technique to remove outliers.

Q6: Plot births by weekday for several decades. Write down your observation.

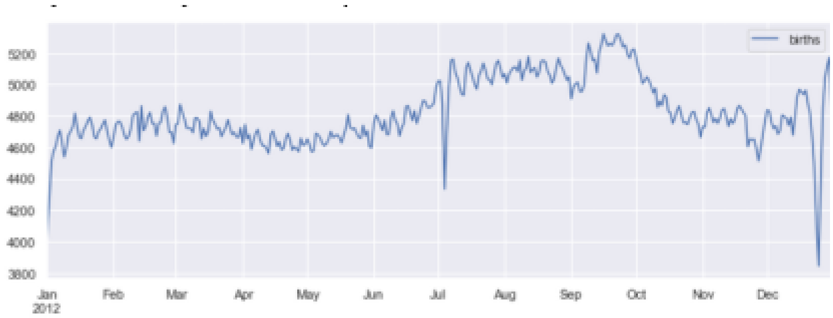
In [8]:
```
data['weekday'] = pd.to_datetime(data['day']).dt.day_name()
sns.countplot(data=data, x='weekday', hue='Decade')
plt.title('Births by Weekday for Several Decades')
plt.xlabel('Weekday')
plt.ylabel('Total Births')
plt.xticks(rotation=45)
plt.show()
```



Q7: Group the data by month and day separately.

In [9]:
```
births_by_month = data.groupby('month')['births'].sum()
births_by_day = data.groupby('day')['births'].sum()
```

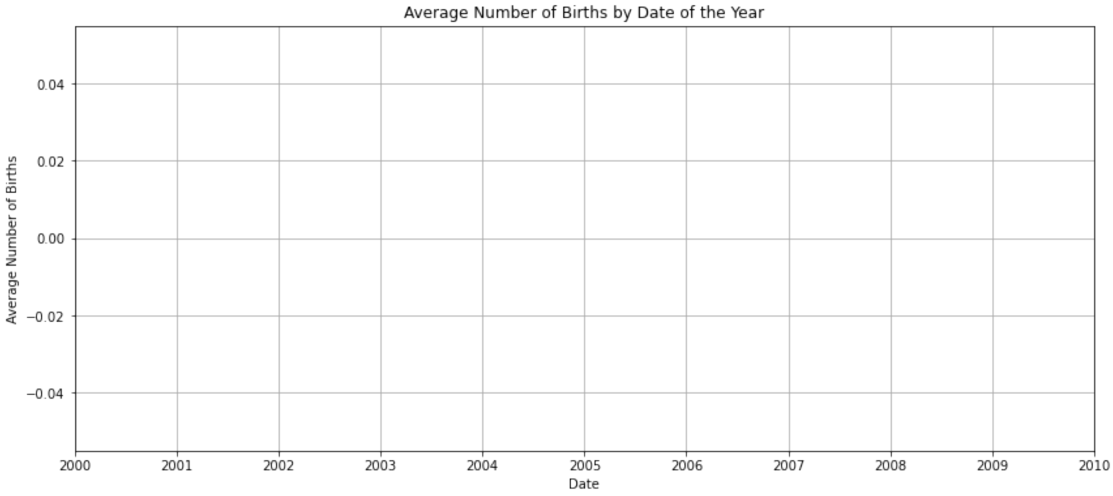
Q.8: Focusing on the month and day only, you have a time series reflecting the average number of births by date of the year. From this, plot the data.



In [11]:
```
average_births_by_date.index = pd.to_datetime(average_births_by_dat

average_births_by_date = average_births_by_date.dropna()

plt.figure(figsize=(14, 6))
plt.plot(average_births_by_date.index, average_births_by_date.value
plt.title('Average Number of Births by Date of the Year')
plt.xlabel('Date')
plt.ylabel('Average Number of Births')
plt.grid(True)
plt.show()
```



Created by: Malik M Shahmeer Rashid