

```
from google.colab import files
upload= files.upload()
```



Choose Files GlobalLand...Country.csv

- **GlobalLandTemperaturesByCountry.csv**(text/csv) - 22680393 bytes, last modified: 1/11/2025 - 100% done
Saving GlobalLandTemperaturesByCountry.csv to GlobalLandTemperaturesByCountry.csv

```
import pandas as pd
```

```
df = pd.read_csv('GlobalLandTemperaturesByCountry.csv')
print(df.to_string())
```



577404	2008-12-01	23.938	0.352	Zimbabwe
577405	2009-01-01	23.898	0.488	Zimbabwe
577406	2009-02-01	23.269	0.317	Zimbabwe
577407	2009-03-01	22.141	0.406	Zimbabwe
577408	2009-04-01	20.472	0.529	Zimbabwe
577409	2009-05-01	19.286	0.465	Zimbabwe
577410	2009-06-01	17.770	0.520	Zimbabwe
577411	2009-07-01	15.234	0.414	Zimbabwe
577412	2009-08-01	18.465	0.489	Zimbabwe
577413	2009-09-01	22.831	0.329	Zimbabwe
577414	2009-10-01	24.544	0.429	Zimbabwe
577415	2009-11-01	23.886	0.421	Zimbabwe
577416	2009-12-01	24.731	0.430	Zimbabwe
577417	2010-01-01	24.660	0.459	Zimbabwe
577418	2010-02-01	24.297	0.565	Zimbabwe
577419	2010-03-01	23.499	0.249	Zimbabwe
577420	2010-04-01	21.907	0.293	Zimbabwe
577421	2010-05-01	20.147	0.534	Zimbabwe
577422	2010-06-01	16.377	0.476	Zimbabwe
577423	2010-07-01	16.668	0.410	Zimbabwe
577424	2010-08-01	18.260	0.447	Zimbabwe
577425	2010-09-01	23.109	0.225	Zimbabwe
577426	2010-10-01	25.943	0.369	Zimbabwe
577427	2010-11-01	25.211	0.356	Zimbabwe
577428	2010-12-01	23.757	0.533	Zimbabwe
577429	2011-01-01	22.982	0.460	Zimbabwe
577430	2011-02-01	23.166	0.260	Zimbabwe
577431	2011-03-01	23.668	0.183	Zimbabwe
577432	2011-04-01	21.759	0.319	Zimbabwe
577433	2011-05-01	19.627	0.398	Zimbabwe
577434	2011-06-01	16.939	0.599	Zimbabwe
577435	2011-07-01	15.803	0.447	Zimbabwe
577436	2011-08-01	17.883	0.328	Zimbabwe
577437	2011-09-01	22.902	0.612	Zimbabwe
577438	2011-10-01	24.966	0.441	Zimbabwe
577439	2011-11-01	25.521	0.318	Zimbabwe
577440	2011-12-01	24.013	0.356	Zimbabwe
577441	2012-01-01	23.872	0.247	Zimbabwe
577442	2012-02-01	24.294	0.305	Zimbabwe
577443	2012-03-01	23.596	0.354	Zimbabwe
577444	2012-04-01	20.349	0.462	Zimbabwe
577445	2012-05-01	19.712	0.312	Zimbabwe
577446	2012-06-01	16.631	0.277	Zimbabwe
577447	2012-07-01	16.048	0.783	Zimbabwe
577448	2012-08-01	18.946	1.127	Zimbabwe
577449	2012-09-01	22.609	0.643	Zimbabwe
577450	2012-10-01	23.482	0.574	Zimbabwe
577451	2012-11-01	24.606	0.532	Zimbabwe
577452	2012-12-01	24.111	0.846	Zimbabwe
577453	2013-01-01	23.812	1.218	Zimbabwe
577454	2013-02-01	24.075	1.286	Zimbabwe
577455	2013-03-01	23.226	0.564	Zimbabwe
577456	2013-04-01	21.142	0.495	Zimbabwe
577457	2013-05-01	19.059	1.022	Zimbabwe
577458	2013-06-01	17.613	0.473	Zimbabwe
577459	2013-07-01	17.000	0.453	Zimbabwe
577460	2013-08-01	19.759	0.717	Zimbabwe
577461	2013-09-01	NaN	NaN	Zimbabwe

```
print(df.info())
```

```
print(df.describe())
```



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 577462 entries, 0 to 577461
Data columns (total 4 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   dt                                    577462 non-null  object
1   AverageTemperature                  544811 non-null  float64
2   AverageTemperatureUncertainty      545550 non-null  float64
```

```

3    Country                    577462 non-null    object
dtypes: float64(2), object(2)
memory usage: 17.6+ MB
None
AverageTemperature  AverageTemperatureUncertainty
count      544811.000000      545550.000000
mean         17.193354         1.019057
std          10.953966         1.201930
min         -37.658000         0.052000
25%          10.025000         0.323000
50%          20.901000         0.571000
75%          25.814000         1.206000
max          38.842000         15.003000

```

```
print(df.head(1000))
```

```

↗
   dt  AverageTemperature  AverageTemperatureUncertainty  Country
0  1743-11-01            4.384                        2.294    Åland
1  1743-12-01            NaN                        NaN    Åland
2  1744-01-01            NaN                        NaN    Åland
3  1744-02-01            NaN                        NaN    Åland
4  1744-03-01            NaN                        NaN    Åland
..  ...                ...                        ...    ...
995 1826-10-01           8.292                      2.658    Åland
996 1826-11-01           3.104                      0.510    Åland
997 1826-12-01           1.290                      2.905    Åland
998 1827-01-01          -4.002                      3.275    Åland
999 1827-02-01          -6.543                      1.744    Åland

```

```
[1000 rows x 4 columns]
```

✓ Data Cleaning

```
df.dropna(inplace=True)
print(df)
```

```
↗ Show hidden output
```

```
print(df.info())
```

```
↗ Show hidden output
```

```

# Remove duplicates
df = df.drop_duplicates()
print(df)

```

```
↗ Show hidden output
```

```

# Handle missing values
print(df.isnull().sum())

```

```

↗
dt                0
AverageTemperature  0
AverageTemperatureUncertainty  0
Country            0
dtype: int64

```

```

# Handle wrong data (convert to correct data type)
df['dt'] = pd.to_datetime(df['dt'], errors='coerce')
print(df)

```

```

↗
   dt  AverageTemperature  AverageTemperatureUncertainty  Country
0  1743-11-01            4.384                        2.294    Åland
5  1744-04-01            1.530                        4.680    Åland
6  1744-05-01            6.702                        1.789    Åland
7  1744-06-01           11.609                        1.577    Åland
8  1744-07-01           15.342                        1.410    Åland
...  ...                ...                        ...    ...
577456 2013-04-01          21.142                      0.495    Zimbabwe
577457 2013-05-01          19.059                      1.022    Zimbabwe
577458 2013-06-01          17.613                      0.473    Zimbabwe
577459 2013-07-01          17.000                      0.453    Zimbabwe
577460 2013-08-01          19.759                      0.717    Zimbabwe

```

```
[544811 rows x 4 columns]
```

```
# 4. Convert columns to appropriate data types
df['Country'] = df['Country'].astype('category')
print(df)
```

```

dt      AverageTemperature  AverageTemperatureUncertainty  Country
0      1743-11-01          4.384                        2.294    Åland
5      1744-04-01          1.530                        4.680    Åland
6      1744-05-01          6.702                        1.789    Åland
7      1744-06-01         11.609                        1.577    Åland
8      1744-07-01         15.342                        1.410    Åland
...      ...              ...                          ...      ...
577456 2013-04-01         21.142                        0.495    Zimbabwe
577457 2013-05-01         19.059                        1.022    Zimbabwe
577458 2013-06-01         17.613                        0.473    Zimbabwe
577459 2013-07-01         17.000                        0.453    Zimbabwe
577460 2013-08-01         19.759                        0.717    Zimbabwe
```

[544811 rows x 4 columns]

```
# Handle conversion errors (for invalid data)
df['AverageTemperature'] = pd.to_numeric(df['AverageTemperature'], errors='coerce')
print(df)
```

Show hidden output

```
# 7. Rename columns
```

```
df = df.rename(columns={'dt': 'Date', 'AverageTemperature': 'AvgTemperature'})
print(df)
```

```

Date      AvgTemperature  AverageTemperatureUncertainty  Country
0      1743-11-01          4.384                        2.294    Åland
5      1744-04-01          1.530                        4.680    Åland
6      1744-05-01          6.702                        1.789    Åland
7      1744-06-01         11.609                        1.577    Åland
8      1744-07-01         15.342                        1.410    Åland
...      ...              ...                          ...      ...
577456 2013-04-01         21.142                        0.495    Zimbabwe
577457 2013-05-01         19.059                        1.022    Zimbabwe
577458 2013-06-01         17.613                        0.473    Zimbabwe
577459 2013-07-01         17.000                        0.453    Zimbabwe
577460 2013-08-01         19.759                        0.717    Zimbabwe
```

[544811 rows x 4 columns]

```
# 8. Replace specific values
```

```
df['Country'] = df['Country'].replace('unknown', None)
print(df)
```

```

Date      AvgTemperature  AverageTemperatureUncertainty  Country
0      1743-11-01          4.384                        2.294    Åland
5      1744-04-01          1.530                        4.680    Åland
6      1744-05-01          6.702                        1.789    Åland
7      1744-06-01         11.609                        1.577    Åland
8      1744-07-01         15.342                        1.410    Åland
...      ...              ...                          ...      ...
577456 2013-04-01         21.142                        0.495    Zimbabwe
577457 2013-05-01         19.059                        1.022    Zimbabwe
577458 2013-06-01         17.613                        0.473    Zimbabwe
577459 2013-07-01         17.000                        0.453    Zimbabwe
577460 2013-08-01         19.759                        0.717    Zimbabwe
```

[544811 rows x 4 columns]

```
# Split columns
```

```
df['Date'] = pd.to_datetime(df['Date'])
df['Year'] = df['Date'].dt.year
print(df)
```

```

Date      AvgTemperature  AverageTemperatureUncertainty  Country \
0      1743-11-01          4.384                        2.294    Åland
5      1744-04-01          1.530                        4.680    Åland
6      1744-05-01          6.702                        1.789    Åland
7      1744-06-01         11.609                        1.577    Åland
8      1744-07-01         15.342                        1.410    Åland
...      ...              ...                          ...      ...
577456 2013-04-01         21.142                        0.495    Zimbabwe
577457 2013-05-01         19.059                        1.022    Zimbabwe
```

```

577458 2013-06-01      17.613      0.473  Zimbabwe
577459 2013-07-01      17.000      0.453  Zimbabwe
577460 2013-08-01      19.759      0.717  Zimbabwe

```

```

      Year
0      1743
5      1744
6      1744
7      1744
8      1744
...      ...
577456 2013
577457 2013
577458 2013
577459 2013
577460 2013

```

```
[544811 rows x 5 columns]
```

```
# Display the cleaned DataFrame
print(df.head())
```

```

Date AvgTemperature AverageTemperatureUncertainty Country Year
0 1743-11-01      4.384      2.294 Åland 1743
5 1744-04-01      1.530      4.680 Åland 1744
6 1744-05-01      6.702      1.789 Åland 1744
7 1744-06-01     11.609      1.577 Åland 1744
8 1744-07-01     15.342      1.410 Åland 1744

```

```
print(df.info())
```

```

<class 'pandas.core.frame.DataFrame'>
Index: 544811 entries, 0 to 577460
Data columns (total 5 columns):
#   Column              Non-Null Count  Dtype
---  -
0   Date                 544811 non-null  datetime64[ns]
1   AvgTemperature       544811 non-null  float64
2   AverageTemperatureUncertainty  544811 non-null  float64
3   Country              544811 non-null  category
4   Year                 544811 non-null  int32
dtypes: category(1), datetime64[ns](1), float64(2), int32(1)
memory usage: 19.8 MB
None

```

```

# extract year from column dt
df['Country'] = df['Country'].astype('string')
df = df[df['Year']>=1800]
print(df.info())
print(df.describe().to_string())

```

```

<class 'pandas.core.frame.DataFrame'>
Index: 510679 entries, 674 to 577460
Data columns (total 5 columns):
#   Column              Non-Null Count  Dtype
---  -
0   Date                 510679 non-null  datetime64[ns]
1   AvgTemperature       510679 non-null  float64
2   AverageTemperatureUncertainty  510679 non-null  float64
3   Country              510679 non-null  string
4   Year                 510679 non-null  int32
dtypes: datetime64[ns](1), float64(2), int32(1), string(1)
memory usage: 21.4 MB
None

```

	Date	AvgTemperature	AverageTemperatureUncertainty	Year
count	510679	510679.000000	510679.000000	510679.000000
mean	1922-10-26 22:42:22.295258112	17.737920	0.819909	1922.362028
min	1800-01-01 00:00:00	-37.658000	0.052000	1800.000000
25%	1880-05-01 00:00:00	11.100500	0.312000	1880.000000
50%	1925-07-01 00:00:00	21.860000	0.527000	1925.000000
75%	1969-09-01 00:00:00	25.970000	1.020000	1969.000000
max	2013-09-01 00:00:00	38.842000	9.366000	2013.000000
std	NaN	10.873421	0.801990	55.197808

```

# Rearrange the columns
df= df[['Year', 'Country', 'AvgTemperature', 'AverageTemperatureUncertainty']]
df.sort_values(by=['Year'], inplace=True)
df.reset_index(drop= True, inplace=True)
print(df.head(10).to_string())

```

```

Year Country AvgTemperature AverageTemperatureUncertainty
0 1800 Åland -5.567 2.118
1 1800 Luxembourg 11.725 2.093

```

```

2 1800 Luxembourg 15.259 1.546
3 1800 Luxembourg 13.890 1.490
4 1800 Luxembourg 17.915 1.890
5 1800 Luxembourg 18.666 3.348
6 1800 Luxembourg 14.764 3.261
7 1800 Luxembourg 8.507 3.202
8 1800 Luxembourg 5.847 3.141
9 1800 Luxembourg 1.231 2.727

```

```
<ipython-input-26-5e3d4eb9c5e5>:3: SettingWithCopyWarning:
```

```
A value is trying to be set on a copy of a slice from a DataFrame
```

See the caveats in the documentation: [https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-df.sort_values\(by=\['Year'\],inplace=True\)](https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-df.sort_values(by=['Year'],inplace=True))

```
# Average temperature in country per year
```

```

AvgTemp_Year= df.groupby(['Year']).agg(
    Country=('Country', 'first'),
    AvgTemp_C=('AvgTemperature', lambda x:round(x.mean(),2)),
    AvgTempUncertainty_C=('AverageTemperatureUncertainty', lambda x:round(x.mean(),2)),

```

```

).sort_values(by=['Year']).reset_index()
print(AvgTemp_Year.head(5).to_string())
print(AvgTemp_Year.info())

```

```

↗
   Year      Country  AvgTemp_C  AvgTempUncertainty_C
0 1800      Åland      10.80      2.71
1 1801        Malta      11.03      2.41
2 1802  United Kingdom (Europe)  10.98      2.55
3 1803      Netherlands      10.94      2.63
4 1804        Cyprus      11.72      3.05

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 214 entries, 0 to 213
Data columns (total 4 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Year                  214 non-null  int32
1   Country                214 non-null  string
2   AvgTemp_C              214 non-null  float64
3   AvgTempUncertainty_C  214 non-null  float64
dtypes: float64(2), int32(1), string(1)
memory usage: 6.0 KB
None

```

```
# Cleaned data
```

```

Cleaned_Data= AvgTemp_Year
print(Cleaned_Data.head(5).to_string())

```

```

↗
   Year      Country  AvgTemp_C  AvgTempUncertainty_C
0 1800      Åland      10.80      2.71
1 1801        Malta      11.03      2.41
2 1802  United Kingdom (Europe)  10.98      2.55
3 1803      Netherlands      10.94      2.63
4 1804        Cyprus      11.72      3.05

```

✓ Univariate Analysis

```
# Cleaned data info
```

```

print(Cleaned_Data.describe())
print()
print(Cleaned_Data.info())

```

```

↗
   count  Year  AvgTemp_C  AvgTempUncertainty_C
count  214.000000  214.000000  214.000000
mean    1906.500000  16.828505  1.074159
std       61.920648  3.064192  0.864530
min     1800.000000  8.400000  0.310000
25%     1853.250000  16.132500  0.380000
50%     1906.500000  18.395000  0.655000
75%     1959.750000  18.807500  1.497500
max     2013.000000  19.880000  3.560000

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 214 entries, 0 to 213
Data columns (total 4 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Year                  214 non-null  int32
1   Country                214 non-null  string
2   AvgTemp_C              214 non-null  float64
3   AvgTempUncertainty_C  214 non-null  float64

```

```
dtypes: float64(2), int32(1), string(1)
memory usage: 6.0 KB
None
```

```
# visualization of temp distribution
import matplotlib.pyplot as plt
import seaborn as sns
```

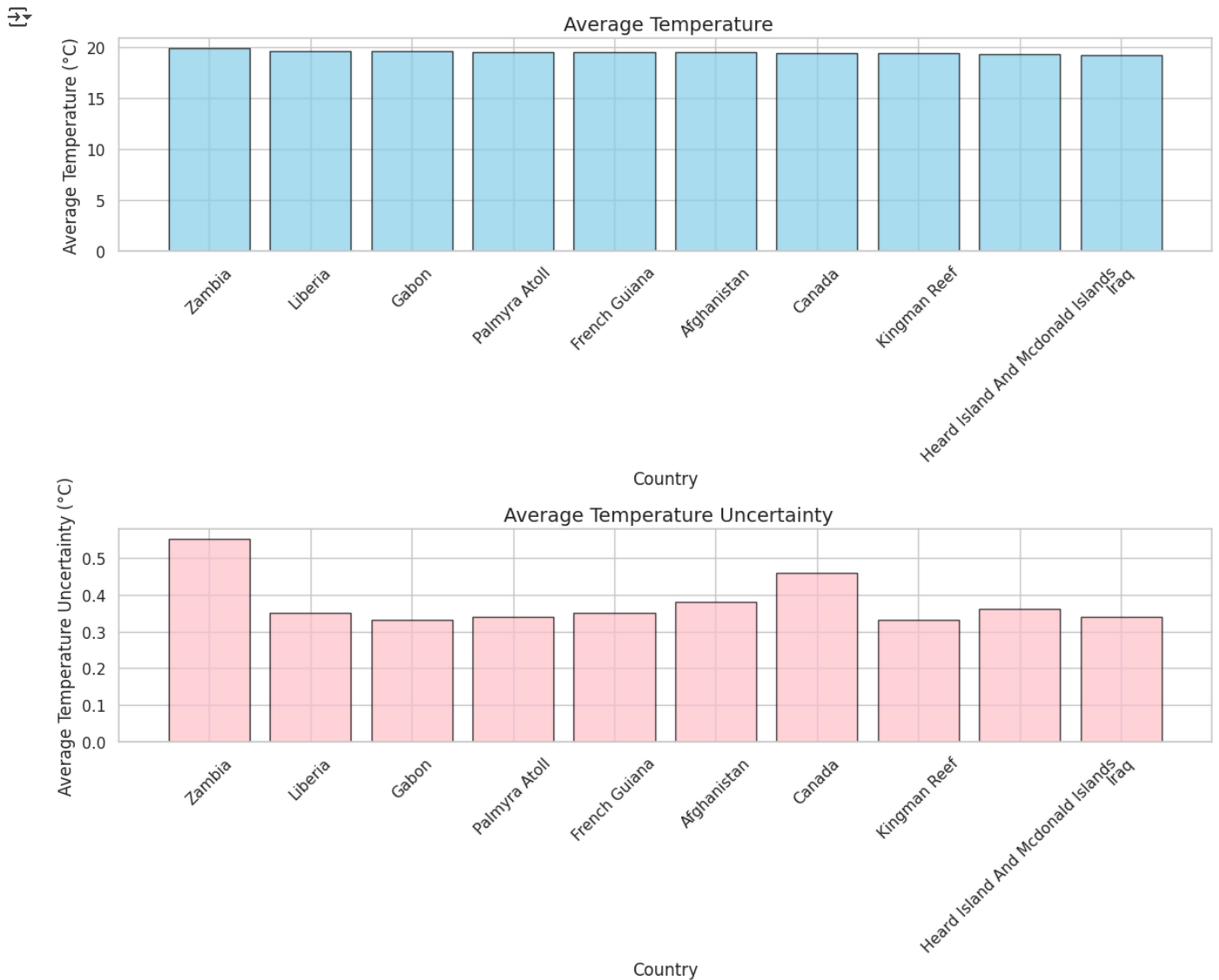
```
sns.set_theme(style='whitegrid')
plt.figure(figsize=(12, 10))
```

```
avg_temp_by_country = Cleaned_Data.groupby('Country')['AvgTemp_C'].mean().sort_values(ascending=False).head(10)
avg_uncertainty_by_country = Cleaned_Data.groupby('Country')['AvgTempUncertainty_C'].mean().loc[avg_temp_by_country.index]
```

```
plt.subplot(2, 1, 1)
plt.bar(avg_temp_by_country.index, avg_temp_by_country.values, color='skyblue', edgecolor='black', alpha=0.7)
plt.title(' Average Temperature', fontsize=14)
plt.xlabel('Country')
plt.ylabel('Average Temperature (°C)')
plt.xticks(rotation=45)
```

```
plt.subplot(2, 1, 2)
plt.bar(avg_uncertainty_by_country.index, avg_uncertainty_by_country.values, color='pink', edgecolor='black', alpha=0.7)
plt.title(' Average Temperature Uncertainty', fontsize=14)
plt.xlabel('Country')
plt.ylabel('Average Temperature Uncertainty (°C)')
plt.xticks(rotation=45)
```

```
plt.tight_layout()
plt.show()
```



```
#pie chart
import matplotlib.pyplot as plt
import seaborn as sns

sns.set_theme(style='whitegrid')
plt.figure(figsize=(12, 10))

avg_temp_by_country = Cleaned_Data.groupby('Country')['AvgTemp_C'].mean().sort_values(ascending=False).head(10)
avg_uncertainty_by_country = Cleaned_Data.groupby('Country')['AvgTempUncertainty_C'].mean().loc[avg_temp_by_country.index]

plt.subplot(2, 1, 1)
plt.pie(avg_temp_by_country.values, labels=avg_temp_by_country.index, autopct='%1.1f%%',
        startangle=140, wedgeprops={'edgecolor': 'magenta'})
plt.title(' Average Temperature', fontsize=14)

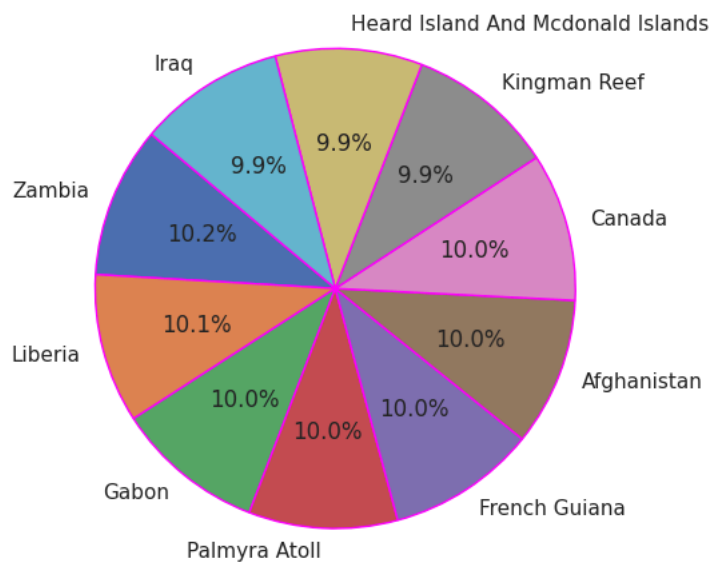
plt.subplot(2, 1, 2)
plt.pie(avg_uncertainty_by_country.values, labels=avg_uncertainty_by_country.index, autopct='%1.1f%%',
        startangle=140, wedgeprops={'edgecolor': 'black'})
plt.title(' Average Temperature Uncertainty', fontsize=14)

plt.tight_layout()
```

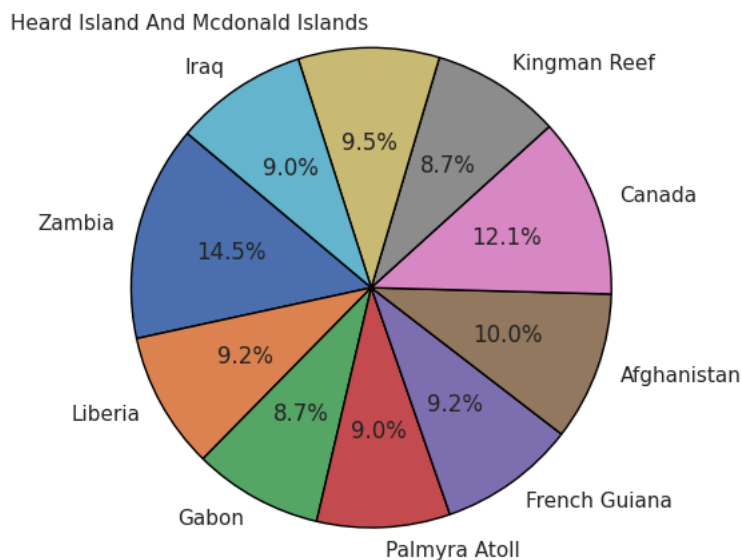
```
plt.show()
```



Average Temperature



Average Temperature Uncertainty



```
# line chart
```

```
import matplotlib.pyplot as plt
```

```
avg_temp_by_country = Cleaned_Data.groupby('Country')['AvgTemp_C'].mean().sort_values(ascending=False).head(10)
```

```
avg_uncertainty_by_country = Cleaned_Data.groupby('Country')['AvgTempUncertainty_C'].mean().loc[avg_temp_by_country.index]
```

```
plt.figure(figsize=(12, 6))
```

```
plt.plot(avg_temp_by_country.index, avg_temp_by_country.values, marker='o', label='Average Temperature (°C)', color='b')
```

```
plt.xticks(rotation=45)
```


```
plt.plot(avg_uncertainty_by_country.index, avg_uncertainty_by_country.values, marker='o', label='Temperature Uncertainty (°C)', color='r')
```

```
plt.xticks(rotation=45)
```

```
plt.title('Trends in Average Temperature and Uncertainty ', fontsize=14)
```

```
plt.xlabel('Country', fontsize=12)
```

```
plt.ylabel('Value (°C)', fontsize=12)
```


 Text(0, 0.5, 'Value (°C)')

