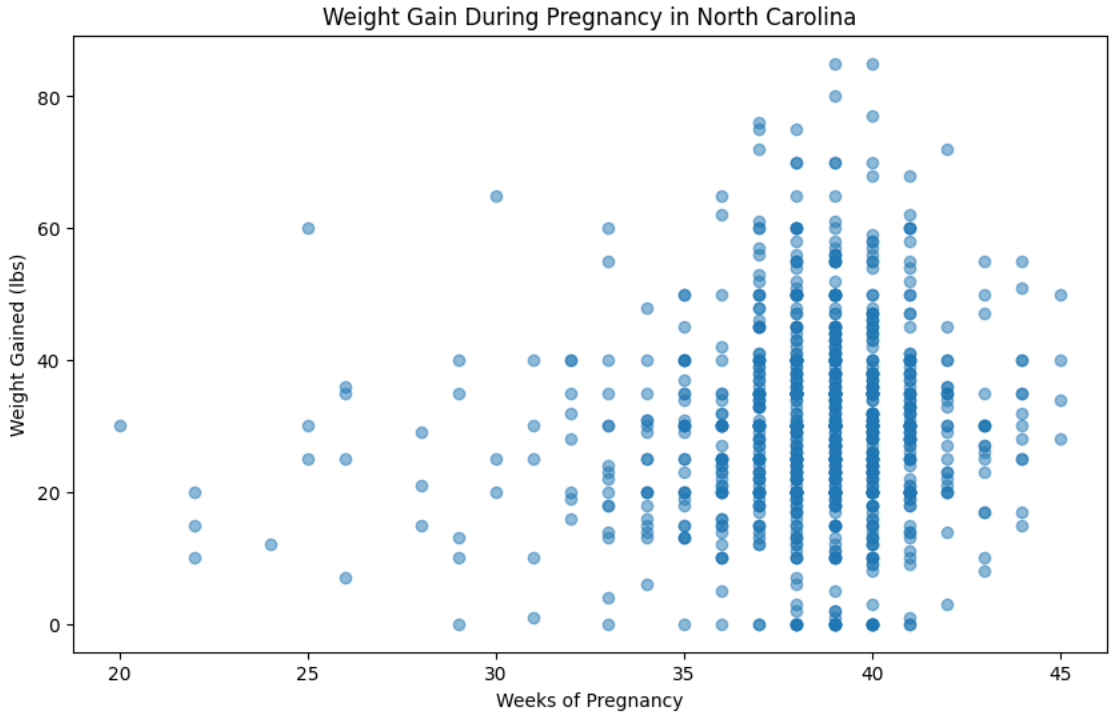
Lab 5 (Title: Feature Distribution and its visualization)

Utkarsh Bhangale
2020802124

i. Make a graph showing weeks on the x-axis and the variable gained on the y-axis:

```
import matplotlib.pyplot as plt
import pandas as pd
df = pd.read_csv("/content/ncbirths.csv")
plt.figure(figsize=(10, 6))
plt.scatter(nc['weeks'], nc['gained'], alpha=0.5)
plt.xlabel('Weeks of Pregnancy')
plt.ylabel('Weight Gained (lbs)')
plt.title('Weight Gain During Pregnancy in North Carolina')
plt.show()
```




```
df.head()
```

	fage	mage	mature	weeks	premie	visits	marital	gained	weight	lowbirthweight	gender
0	NaN	13	younger mom	39.0	full term	10.0	not married	38.0	7.63	not low	male
1	NaN	14	younger mom	42.0	full term	15.0	not married	20.0	7.88	not low	male
2	19.0	15	younger mom	37.0	full term	11.0	not married	38.0	6.63	not low	female

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   fage            829 non-null    float64
 1   mage            1000 non-null   int64
 2   mature          1000 non-null   object
 3   weeks           998 non-null    float64
 4   premie          998 non-null    object
 5   visits          991 non-null    float64
 6   marital         999 non-null    object
 7   gained          973 non-null    float64
 8   weight          1000 non-null   float64
 9   lowbirthweight  1000 non-null   object
 10  gender          1000 non-null   object
 11  habit           999 non-null    object
 12  whitemom        998 non-null    object
```

```
dtypes: float64(5), int64(1), object(7)
memory usage: 101.7+ KB
```

```
df.shape
```

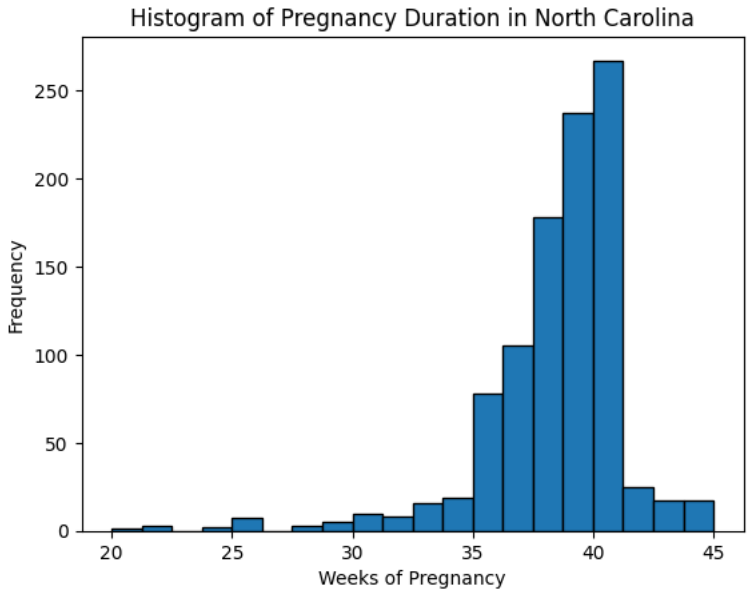
```
(1000, 13)
```

```
df.describe()
```

	fage	mage	weeks	visits	gained	weight
count	829.000000	1000.000000	998.000000	991.000000	973.000000	1000.00000
mean	30.255730	27.000000	38.334669	12.104945	30.325797	7.10100
std	6.763766	6.213583	2.931553	3.954934	14.241297	1.50886
min	14.000000	13.000000	20.000000	0.000000	0.000000	1.00000
25%	25.000000	22.000000	37.000000	10.000000	20.000000	6.38000
50%	30.000000	27.000000	39.000000	12.000000	30.000000	7.31000
75%	35.000000	32.000000	40.000000	15.000000	38.000000	8.06000
max	55.000000	50.000000	45.000000	30.000000	85.000000	11.75000

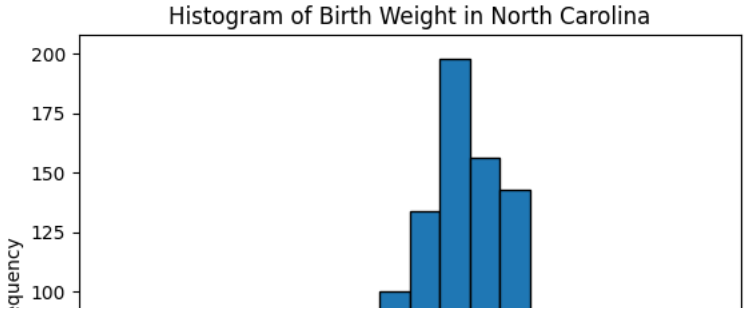
ii. Inspect the histogram of the weeks variable:

```
plt.hist(df['weeks'], bins=20, edgecolor='k')
plt.xlabel('Weeks of Pregnancy')
plt.ylabel('Frequency')
plt.title('Histogram of Pregnancy Duration in North Carolina')
plt.show()
```



iii. Make a histogram of the birth weight of newborns (in lbs)

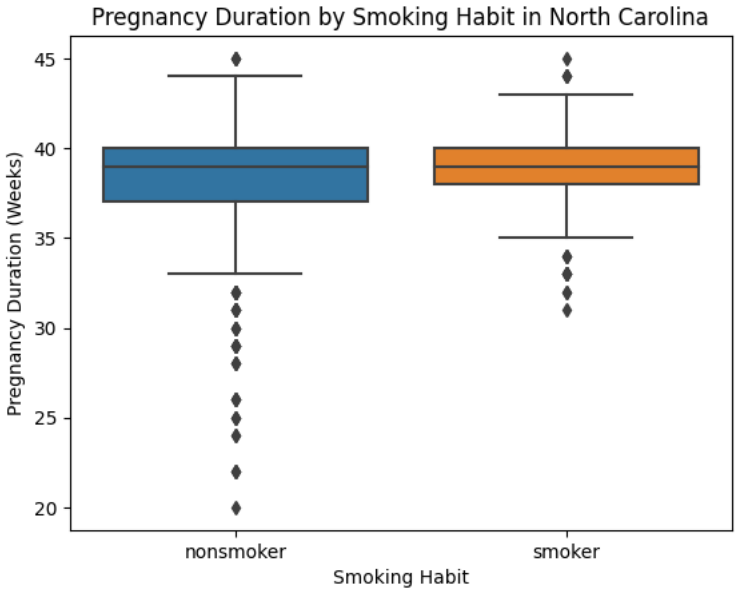
```
plt.hist(df['weight'], bins=20, edgecolor='k')
plt.xlabel('Birth Weight (lbs)')
plt.ylabel('Frequency')
plt.title('Histogram of Birth Weight in North Carolina')
plt.show()
```



iv. Make a boxplot of pregnancy duration in weeks by smoking habit:

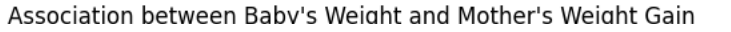
```
import seaborn as sns

sns.boxplot(x='habit', y='weeks', data=nc)
plt.xlabel('Smoking Habit')
plt.ylabel('Pregnancy Duration (Weeks)')
plt.title('Pregnancy Duration by Smoking Habit in North Carolina')
plt.show()
```



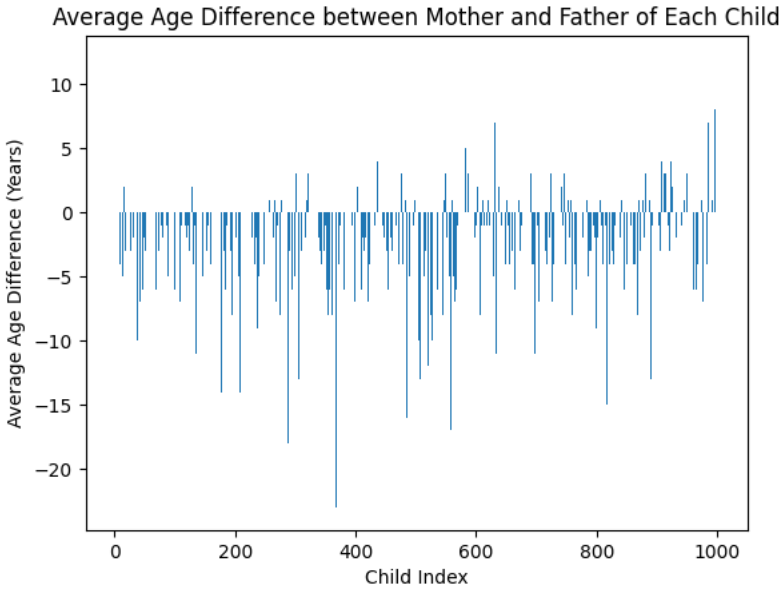
v. Assess the association between baby's weight and mother's weight gain visually using a scatter plot:

```
plt.scatter(df['gained'], df['weight'], alpha=0.5)
plt.xlabel('Weight Gained (lbs)')
plt.ylabel("Baby's Weight (lbs)")
plt.title('Association between Baby\'s Weight and Mother\'s Weight Gain')
plt.show()
```



vi. Assess the average difference in age between mother and father of each child visually using a bar plot:

```
avg_age_diff = df['mage'] - df['fage']
plt.bar(df.index, avg_age_diff)
plt.xlabel('Child Index')
plt.ylabel('Average Age Difference (Years)')
plt.title('Average Age Difference between Mother and Father of Each Child')
plt.show()
```



Conclusion

We have succefully implemented the different types of data visualizations on the this dataset.