Problem #A

Consider the following data related to the final scores of the four sections of a data science course:

Section	Scores
Clytherin:	95, 95, 65, 70, 65, 65, 75, 75, 75, 80, 65, 95, 75, 65, 75, 80, 65, 95, 75, 65, 95, 75, 65, 80, 65, 70, 70, 65
Jriffindor:	81, 65, 73, 77, 70, 71, 74, 63, 85, 79, 86, 68, 75, 66, 62, 88, 83, 87, 64, 72, 69, 84, 78, 80, 76, 82, 67
Hufflebuff:	84, 77, 77, 61, 77, 84, 75, 89, 66, 75, 61, 66, 80, 61, 70, 66, 73, 89, 73, 70, 73, 70, 80, 84, 89, 80, 75
Ravenklaw:	75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75

Answer the following questions:

A-1: Draw the histograms of scores for each of the above sections. Take bin size as 10 units, starting from score 40. Draw each histogram in one figure. Write the title on each histogram.

A-2: Find the mean, median and mode for each of the above sections.

A-3: Find the population variance and population standard deviation for each of the above sections.

A-4: Find the upper and lower quartile for each of the above sections.

A-5: Draw the box-plot for each of the above sections in one figure. Label x-axis with the name of the section.

A-6: Do hypothesis testing for each section to check if the data follows normal distribution. Assume p-value less than 0.05 or 5% as very small.

A-7: Do pair-wise student's t-test to compare the means of two distribution. Set kwarg alternative='two-sided'. Assume p-value less than 0.05 or 5% as very small.

A-8: Do pair-wise Mann-Whitney U test to compare the underlying distributions. Set kwarg alternative='two-sided'. Assume p-value less than 0.05 or 5% as very small.

In [1]:
```
Data = {
"Jriffindor": [81, 65, 73, 77, 70, 71, 74, 63, 85, 79, 86, 68, 75, 66, 62,
    88, 83, 87, 64, 72, 69, 84, 78, 80, 76, 82, 67],
"Clytherin": [95, 95, 65, 70, 65, 65, 75, 75, 75, 80, 65, 95, 75, 65, 75, 80, 65, 95, 75, 65,
    95, 75, 65, 80, 65, 70, 70, 65],
"Hufflebuff": [84, 77, 77, 61, 77, 84, 75, 89, 66, 75, 61, 66, 80, 61, 70,
    66, 73, 89, 73, 70, 73, 70, 80, 84, 89, 80, 75],
"Ravenklaw": [75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75,
    75, 75, 75, 75, 75, 75, 75]
  }

for section in Data:
  series=Data[section].copy()
```

Problem #B

Consider the data given in CSV file **Bank Info**. Do the following tasks using the data:

B-1: Draw the histograms of all numeric and non-numeric columns.

B-2: Using the histograms generated in *B*-1, provide descriptive comments on the distribution of the following columns:

loan_amount
rate_of_interest
loan_type
property_value

B-3: Draw a plot between **loan_amount** and **Gender**. What can you conclude from the plot?

B-4: Draw a plot between **property_value** and **loan_amount**, differentiated by **Status**. For the defaulted loans, what is the relationship between **property_value** and **loan_amount**. (*Note: A **status** of 1 indicates that the loan has defaulted.*)

B-5: Draw a plot between **property_value** and **loan_amount**, differentiated by **Status** and **Region**. Make the figsize=(9,4), and dpi=200. *Hint: dpi is a keyword argument in plt.figure() method.*

B-6: Display a count plot of **credit_type**, differentiated by **loan_purpose**.

B-7: Draw a plot on the **loan_amount** differentiated by **business_or_commercial** and **occupancy_type**.

B-8: Add two new columns to the data frame with headers: **property value multiple** and **loan multiple**. The values in the new columns can be calculated using the following formulas:

$$\text{property value multiple} = \frac{\text{property\_value}}{\textit{income} \times 12}$$

$$\text{loan multiple} = \frac{\text{loan\_amount}}{\textit{income} \times 12}$$

B-9: Plot the **Credit Score** in ascending order on the x-axis, and the above two new columns on y-axis