Homework 2

Problem 1

1. The standard IQR rule assumes symmetry => in skewed distributions the extreme values might be wrongly classified as outliers. Therefore unlike in normal distributions, in non-normal data such as data with extremes, the whiskers are truncated at the dataset's actual min/max values whithin it's range.

2. Boxplots may misrepresent outliers in skewed data by not considering the density variations. We can avoid this problem by scaling the whiskers of the boxplot by skewness.

3. The median represents the central position, while mean is the avarage and therefore is heavily influenced by the skewness of the distribution. Therefore the median is more stable, hence why boxplots priozitize it more, however this can obscure skewness. For instance, in a right-skewed data the median understates the mean, and fails to show the inequality trends which are visible with the mean.

4. Strong right skewness means that the distribution has a long upper tail. Varience Increases due to extreme values, and skewness coefficiend is positive. Models that assume normality, such as linear regression may fail, ~~resulting~~ therefore we might need to use transformations or non-parametric approaches.

5. Boxplots effectively compare central tendency and spread across groups. However, in the case of for example similiar medians but diffrent shapes, the overlapping distributions may obscure difference. Additionally, for small categorical samples, quartiles become unstable.

6. Two few bins may merge the peaks in the data and therefor fail to represent it well, while too many bins create much noise which is not good. In KDE narrow bandwidths overfit, while wide Ones oversmooth. It is crucial to balance the two.

7. Histograms are used for continous data, while bar charts are used for discrete categories. Histograms reflect the frequency density while bar charts reflect the counts with its height. For histograms bin choice is important, meanwhile in bar charts it doesn't matter since all is the same and there are fixed categories.

8. A histogram can distort perception if bins don't align with the data structure. For example, wide bins in a bimodal dataset might merge peaks and resemble unimodal. KDE or violin plots avoid this by smoothing the data or showing full density with all the peaks and gaps.

9. Densidy plots use kerne smoothing to estimate probability density, unlike histogrms that have discrete bins. Meanwhile choosing a kernel also requires balance, since small bandwidths overfit sparse data, while large Ones oversmooth.

10. The orea under the density plot = 1, since it normalized the distribution to a ~~probability~~ PDF, hence probability is 1. This normalization allows direct comparisons of distributions across sample sizes. Unlike in histogram the shape here is emphasized.

Problem 2

1)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
-5	-2	0	3	4	5	5	6	7	7	8	9	9	10	12	15

=> 16 data points

$Ecdf(-5) = \frac{1}{16} = 0{,}0625$

$Ecdf(12) = \frac{15}{16} = 0.9375$

$Esdf(15) = \frac{16}{16} = 1$

$Ecdf(-2) = \frac{2}{16} = \frac{1}{8} = 0.125$

$Ecdf(0) = \frac{3}{16} = 0.1875$

$Ecdf(3) = \frac{4}{16} = 0.25$

$Ecdf(4) = \frac{5}{16} = 0.3125$

$Gddf(5) = \frac{6}{16} = 0.375$

$Ecdf(5) = \frac{7}{16} = 0.4375$

$Ecdf(6) = \frac{8}{16} = 0,5$

$Ecdf(7) = \frac{9}{16} = 0.5625$

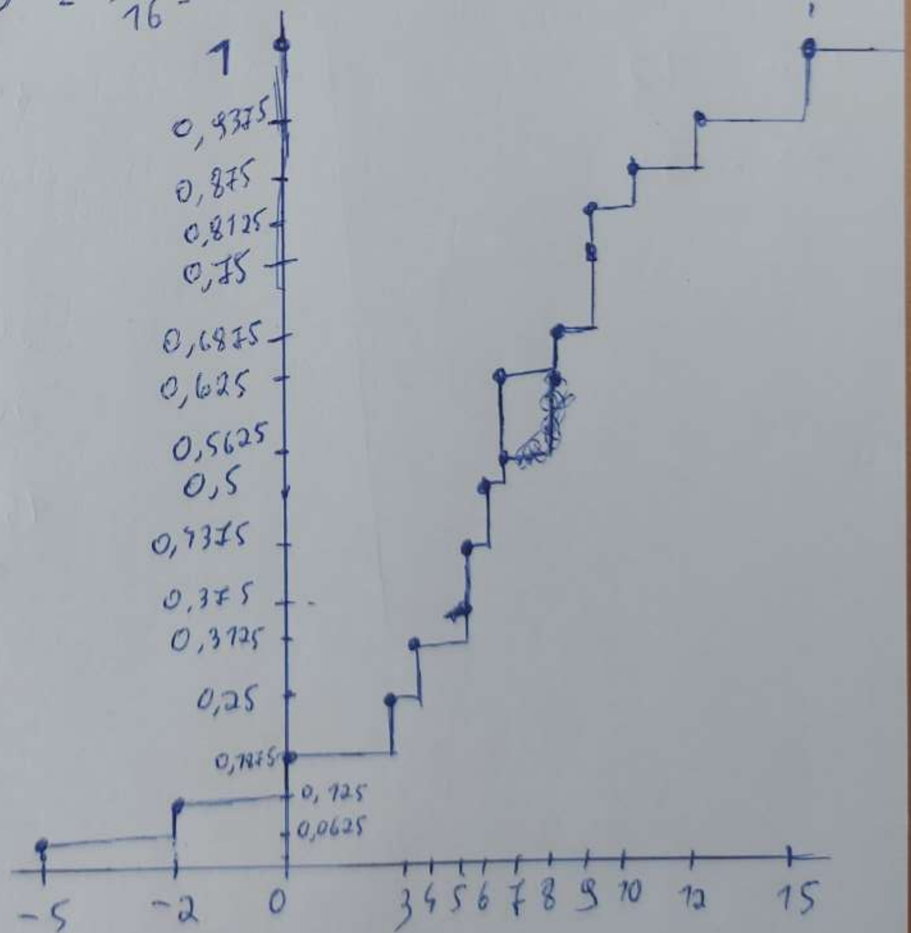
$Ecdf(7) = \frac{10}{16} = 0.625$

$Ecdf(8) = \frac{11}{16} = 0.6875$

$Ecdf(9) = \frac{12}{16} = 0.75$

$Ecdf(9) = \frac{13}{16} = 0.8125$

$Ecdf(10) = \frac{14}{16} = 0.875$



2) $-5 \quad 12 \quad 14 \quad 14 \quad 15 \quad 16. \quad 17 \quad 18. \quad 19 \underset{\substack{\uparrow \\ \text{median}}}{\boxed{20}} \quad 21 \quad 22 \quad 23 \quad 24 \quad 24. \quad 25 \quad 29 \quad 30 \quad 35$

$Q_2 = 20$

$Q_1 = 15$ $\Rightarrow IQR = Q_3 - Q_1 = 24 - 15 = 9$

$Q_3 = 24$

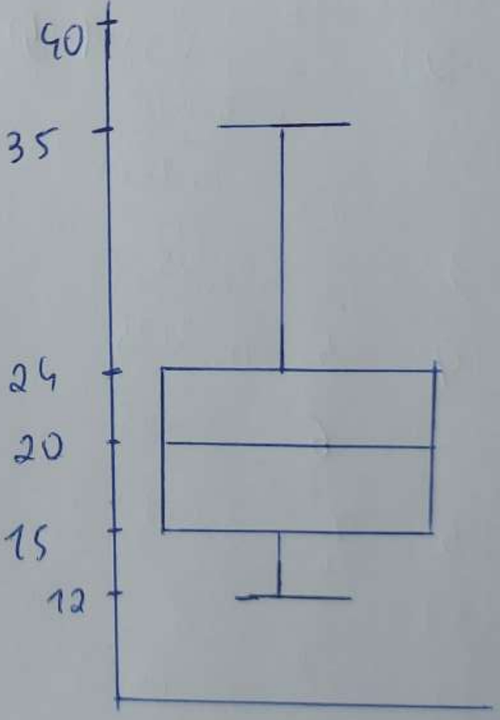
Lower bound $= Q_1 - \frac{3}{2} IQR = 15 - 1.5 \cdot 9 = 1.5$

Upper bound $= Q_3 + \frac{3}{2} IQR = 24 + 1.5 \cdot 9 = 37.5$ $\Rightarrow$

$\Rightarrow$ The only outlier is -5

Min $= 12$

Max $= 35$

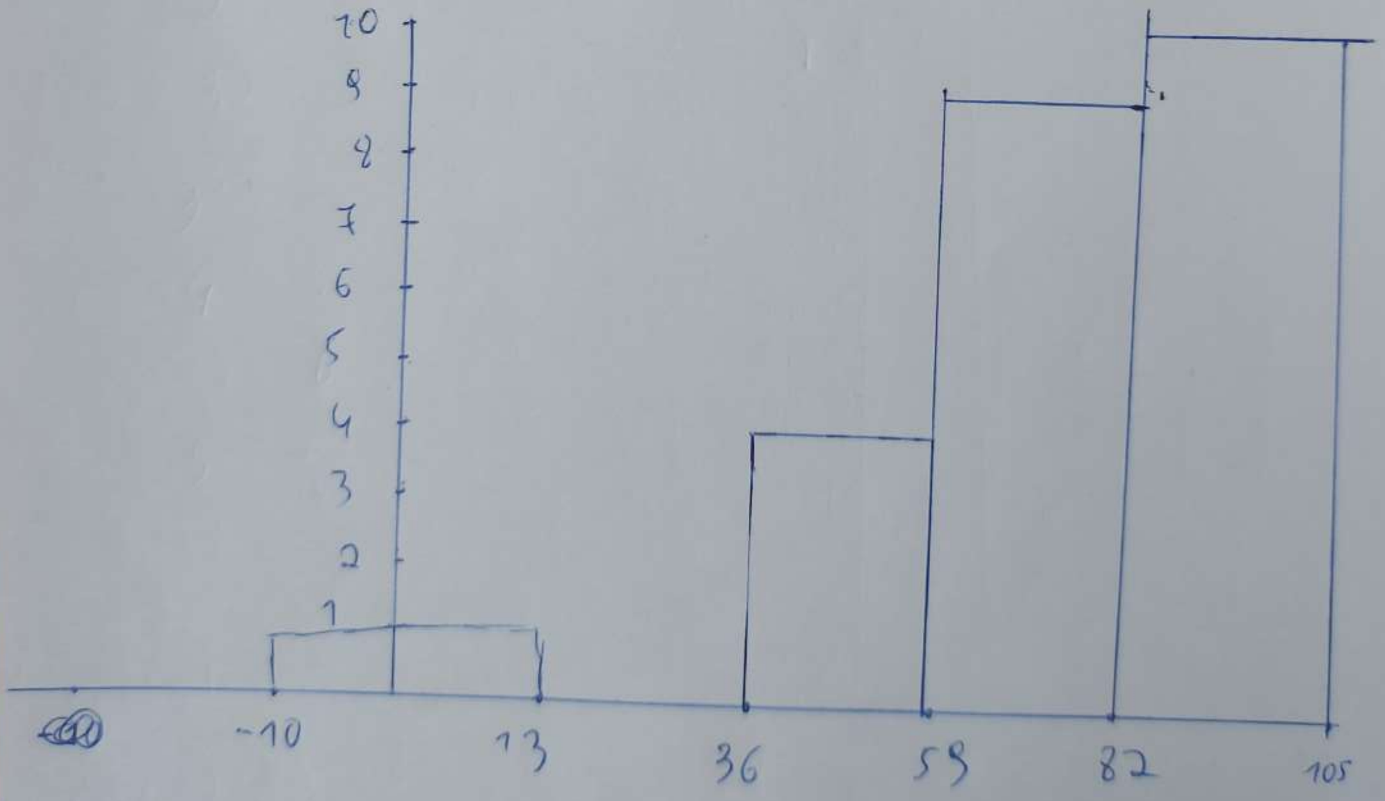


3. -10, 45, 50, 55, 55, 60, 62, 65, 68, 70, 73, 74, 80, 80, 82, 85, 88,

90, 91, 92, 94, 97, 100, 105

Min = -10
Max = 105 |=> range = 105 + 10 = 115 => bin width = $\frac{115}{5}$ = 23

Bin 1: [-10, 13) - 1 value
Bin 2: (13, 36) - 0 value
Bin 3: [36, 59) - 4 values
Bin 4: [59, 82) - 9 values
Bin 5: [82, 105] - 10 values


10
9
8
7
6
5
4
3
2
1
-10
13
36
59
82
105