Problem 1. Consider the dataset *housing.arff* which is posted along with this assignment. The dataset has 506 instances and 14 attributes. The 14th attribute in the dataset is *MEDV* (Median value of owner-occupied homes in $1000's). Brief description of all attributes is in the *housing_names.txt* file.
(1). Calculate the mean, median, and standard deviation of the *MEDV* attribute.
(2). Determine Q1, Q2, and Q3, and plot the boxplot of the *MEDV* attribute.
(3). Detect outliers using the IQR method, which we discussed in the class, and show the *MEDV* attribute values of the detected outliers. When detecting outliers, use only the *MEDV* attribute values.

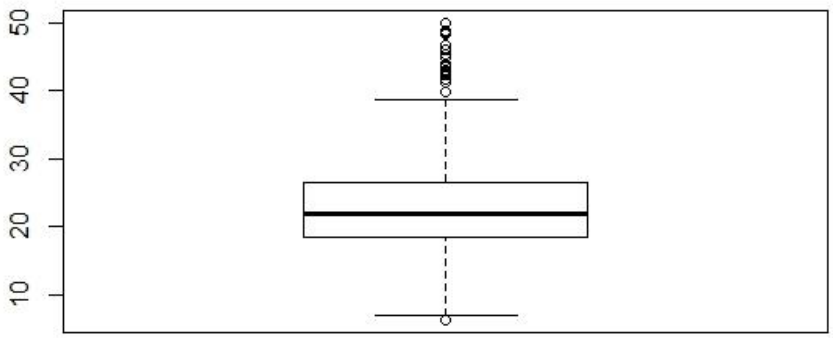
(1)

Mean	**23.75**
Median	**21.95**
Standard Deviation	**8.81**

(2)

MIN	6.3
Q1	18.5
MED	21.9
Q3	26.6
MAX	50

```
Console ~/
> summary(housing1)
      CRIM                ZN              INDUS            CHAS
 Min.   :0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000
 1st Qu.:0.06988   1st Qu.:  0.00   1st Qu.: 4.93   1st Qu.:0.00000
 Median :0.19103   Median :  0.00   Median : 8.14   Median :0.00000
 Mean   :1.42083   Mean   : 12.72   Mean   :10.30   Mean   :0.07743
 3rd Qu.:1.21146   3rd Qu.: 20.00   3rd Qu.:18.10   3rd Qu.:0.00000
 Max.   :9.96654   Max.   :100.00   Max.   :27.74   Max.   :1.00000
      NOX               RM              AGE              DIS              RAD
 Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130   Min.   : 1.000
 1st Qu.:0.4470   1st Qu.:5.927   1st Qu.: 40.95   1st Qu.: 2.355   1st Qu.: 4.000
 Median :0.5190   Median :6.229   Median : 71.80   Median : 3.550   Median : 5.000
 Mean   :0.5408   Mean   :6.344   Mean   : 65.56   Mean   : 4.044   Mean   : 7.823
 3rd Qu.:0.6050   3rd Qu.:6.635   3rd Qu.: 91.62   3rd Qu.: 5.401   3rd Qu.: 7.000
 Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127   Max.   :24.000
      TAX             PTRATIO             B              LSTAT             MEDV
 Min.   :187.0   Min.   :12.60   Min.   :  0.32   Min.   : 1.730   Min.   : 6.30
 1st Qu.:276.8   1st Qu.:16.80   1st Qu.:377.72   1st Qu.: 6.588   1st Qu.:18.50
 Median :307.0   Median :18.60   Median :392.08   Median :10.250   Median :21.95
 Mean   :377.4   Mean   :18.25   Mean   :369.83   Mean   :11.442   Mean   :23.75
 3rd Qu.:411.0   3rd Qu.:20.20   3rd Qu.:396.16   3rd Qu.:15.105   3rd Qu.:26.60
 Max.   :711.0   Max.   :22.00   Max.   :396.90   Max.   :34.410   Max.   :50.00
>
```



(3) Detect outliers using the IQR method
The interquartile range IQR = Q3 — Q1 = 26.6 – 18.5 = 8.1
Upper range: Q3 + 1.5*IQR = 26.6 + 1.5*8.1 = 38.75
Lower range: Q1 - 1.5*IQR = 18.5 – 1.5*8.1 = 6.35
Any value outside these ranges is outlier (Less than lower and more than upper ranges)

Problem 2. Consider the following two objects with their attribute values:

Object	Pressure	Temperature	Flow
O1	29	68	41
O2	32	75	63

Compute the distance between O1 and O2 using (1) Euclidean distance and (2) Manhattan distance.

(1) Euclidean distance

$d(O2, O1) = \sqrt{(32-29)^2 + (75-68)^2 + (63-41)^2} = \sqrt{3^2 + 7^2 + 22^2} = \sqrt{542} = 23.28$

(2) Manhattan distance

$d(O2, O1) = |32-29| + |75-68| + |63-41| = 32$

Problem 3. Consider the following two objects with their attribute values:

Object	Outlook	Temperature	Humidity	Windy
O1	sunny	low	low	True
O2	rainy	high	low	false

Compute the distance between O1 and O2. Here, all attributes are categorical variables.

d(O1,O2) = (all attributes - matches attributes) / all attributes = (4-1)/4 = 0.75

Problem 4. Consider the following two objects with their attribute values:

Object	Fever	Cough	Test1	Test2	Test3	Test4
O1	Y	Y	Negative	Positive	Negative	Negative
O2	Y	N	Negative	Positive	Positive	Positive

Here, all variables are symmetric binary variable. Compute the distance between O1 and O2.

d(O1,O2)= (# mismatches)/(# all attributes) = 3/6 = 0.5

Problem 5 Consider the following dataset (sorted in non-decreasing order):
<11, 18, 22, 22, 31, 34, 37, 58, 59, 64, 68, 68, 72, 74, 79>
(1) Perform the equal width binning on the above data with 3 bins. Note that the bin boundaries are integers in the lecture example (to make the discussion simple). But, for this assignment your bin boundaries may include fractions. For each bin, show the bin interval, data values in the bin, and smoothed values using bin means, bin medians, and bin boundaries.
(2) Repeat the same with equal depth binning with 3 bins.
(3) If you transform the dataset into the interval of [0, 1] using Min-max normalization, what is the new value of 22? What is the new value of 72?
(4) If you transform the dataset using z-score normalization (using the standard deviation), what is the new value of 22? What is the new value of 72?

(1) In equal width partitioning:

$$width_A = (max_A - min_A) / N = (79 - 11)/3 = 22.67$$

	Bin1				Bin2			Bin3							
Original values	11	18	22	22	31	34	37	58	59	64	68	68	72	74	79
Bin Means	18.25				34			67.75							
Values Smoothed by Bin Means	18.25	18.25	18.25	18.25	34	34	34	67.75	67.75	67.75	67.75	67.75	67.75	67.75	67.75
Bin Medians	20				34			68							

Values Smoothed by Bin Medians	20	20	20	20	34	34	34	68	68	68	68	68	68	68	68
Bin Intervals	[11, 22.67)				[22.67, 45.34)			[45.34, 79]							
Values Smoothed by Bin Boundaries	11	22	22	22	31	31	37	58	58	58	58	58	79	79	79

(2) In equal depth partitioning, the number of items in each bin will be 15/3 = 5

	Bin1					**Bin2**					**Bin3**				
Original values	11	18	22	22	31	34	37	58	59	64	68	68	72	74	79
Bin Means	20.8					50.4					72.2				
Values Smoothed by Bin Means	20.8	20.8	20.8	20.8	20.8	50.4	50.4	50.4	50.4	50.4	72.2	72.2	72.2	72.2	72.2
Bin Medians	22					58					72				
Values Smoothed by Bin Medians	22	22	22	22	22	58	58	58	58	58	72	72	72	72	72
Bin Boundaries	(11, 31)					(34, 64)					(68, 79)				
Values Smoothed by Bin Boundaries	11	11	31	31	31	34	34	64	64	64	68	68	68	79	79

(3) Min-Max Normalization:

Values in the range [11 , 79] which have to be normalized to the range [0, 1].

$$v' = new\_minA + [(v - minA) / (maxA - minA)].(new\_maxA - new\_minA)$$

Original values	**Normalized Values**
11	0
18	0.1
22	0.16
22	0.16
31	0.29

34	0.34
37	0.38
58	0.69
59	0.71
64	0.78
68	0.84
68	0.84
72	0.90
74	0.93
79	1
Max Value	79
Min Value	11
New_Max	1
New_min	0

(4) Z-score normalization:

The transformed value v' of the current value v is based on the *mean* (μA) and the *standard deviation* (σA) of the attribute A: $v' = (v - \mu A) / \sigma A$.

Original values	Normalized Values
11	-1.57
18	-1.27
22	-1.10
22	-1.10
31	-0.72
34	-0.59
37	-0.46
58	0.43
59	0.48
64	0.69
68	0.86
68	0.86
72	1.03
74	1.12
79	1.33
μA	47.8
σA	23.48

Problem 6. This problem is a practice of calculating correlations between input attributes (or predictive attributes) and the output attribute (or predictable attribute) in the dataset *red-numeric.arff*. This dataset was downloaded from UCI Machine Learning Lab (https://archive.ics.uci.edu/ml/datasets/Wine+Quality) and modified for this assignment. It has 12 attributes and 1599 tuples. The first 11 attributes are input attributes and the last attribute, *quality*, is the output attribute. Your task is to calculate the correlation between each of the following four input attributes and the output attribute: *density*, *pH*, *sulphates*, and *alcohol*. In other words, you are required to calculate the following four correlations:
correl(*density*, *quality*)
correl(*pH*, *quality*)
correl(*sulphates*, *quality*)
correl(*alcohol*, *quality*)
Here, *correl*(*X*, *Y*) denotes the correlation between *X* and *Y*.
In your submission, include all four correlations, and indicate the attribute which has the strongest correlation with *quality*.

correl(*density*, *quality*): -0.17
correl(*pH*, *quality*): - 0.06
correl(*sulphates*, *quality*): 0.25
correl(*alcohol*, *quality*): 0.48

Alcohol has the strongest correlation with quality because has bigger correlation coefficient than other attributes.