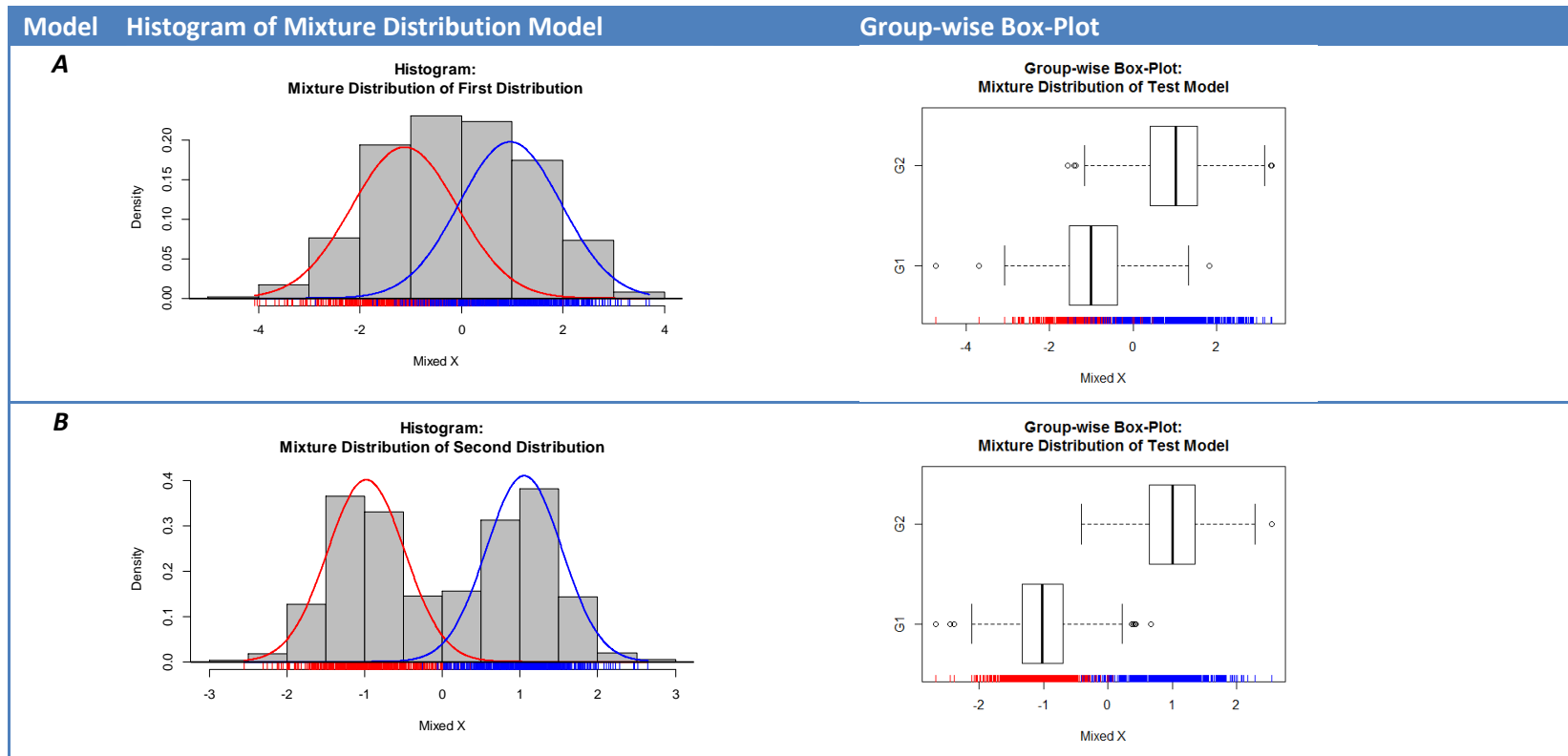
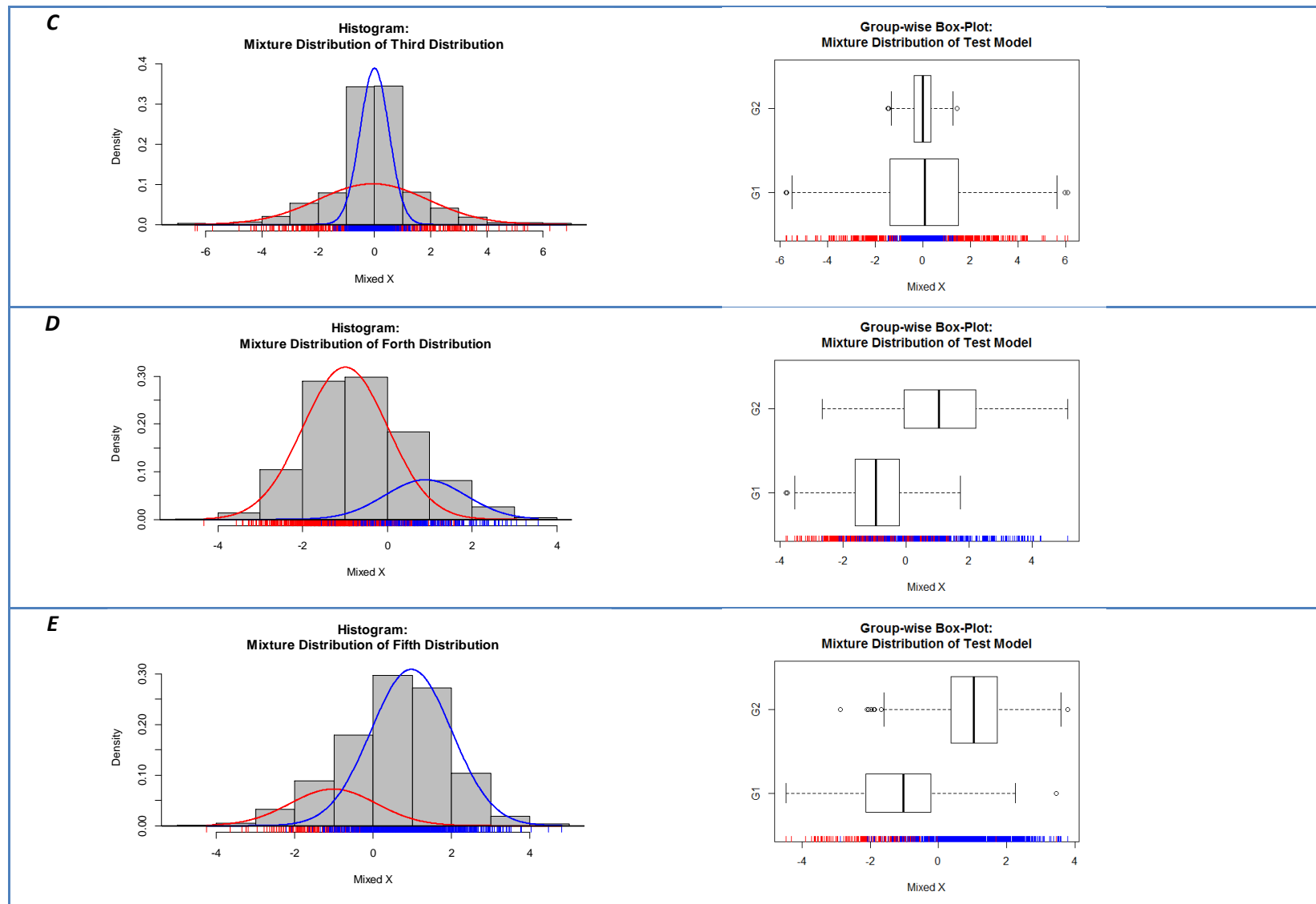


Sample Answer Lab 04: Univariate Descriptive Statistics

Task 1: Descriptive Statistics and Relation to the Mixture Distributions of Random Variables (2.5 points)





[b] Report your selected parameters of each of the mixture distributions models. (0.5 points)

Model	Mixture Proportion	μ_1	σ_1	μ_2	σ_2
-------	--------------------	---------	------------	---------	------------

A	0.5	-1	1	1	1
B	0.5	-1	0.6	1	0.6
C	0.5	0	2	0	0.5
D	0.75	-1	1	1	1.1
E	0.25	-1	1	1	1

[c] Report in a table the estimated statistics describing the joint distribution of each of the five mixture distribution models. (0.5 points)

Model	Mean	Winsorized Mean at 10%	Standard Deviation	Skewness	Kurtosis	Bimodality Index
A	-0.047	-0.061	1.389	0.091	-0.503	0.402
B	0.016	0.020	1.163	-0.014	-1.141	0.522
C	0.047	0.035	1.518	0.207	2.676	0.183
D	-0.464	-0.522	1.307	0.401	0.203	0.362
E	0.515	0.549	1.339	-0.257	-0.044	0.360

[d] Explain how the skewness, kurtosis and bimodality index relate to the selected parameters that you have selected for the five mixture distribution models? (0.5 points)

Model A: Unimodal and symmetric with negative kurtosis. Make two means close enough to strengthen the center of the joint distribution.

Model B: Bimodal and symmetric. Same proportion and standard deviation. Separate two means far enough.

Model C: Unimodal and symmetric with positive kurtosis. Make two means the same because the mixture distribution should be unimodal and symmetric. Increase the standard deviation of the first parent distribution to fatten the both tails of the joint distribution.

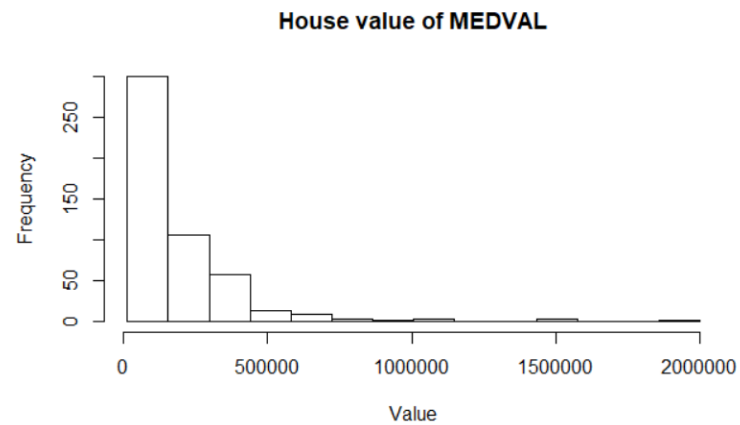
Model D: Unimodal with positive skewness. Make two means close enough. The second distribution has a larger standard deviation.

Model E: Unimodal with negative skewness. Make two means close enough. The first distribution has a larger standard deviation.

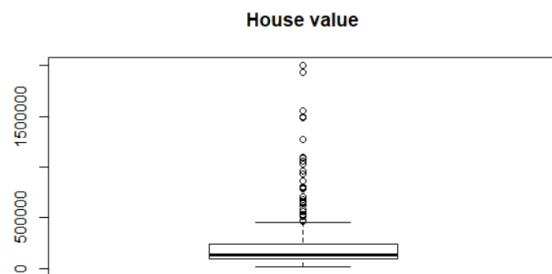
Task 2: Aggregation – Merger – Weighted Mean (1.5 point)

[a] Generate a professionally labeled histogram and box-plot of the variable **MEDVALHOME** (caution: there are missing values) as well as a side-by-side box-plot for **MEDVALHOME** broken down by the factor **CITYPERI**. **Describe** the distribution of the median home values for the sectors of Dallas County and provide **meaningful** summary statistics of the median home values within each sector. (0.3 points)

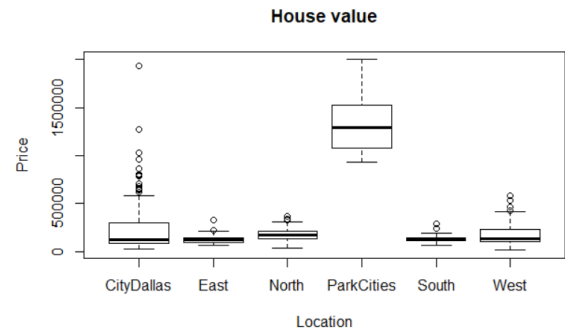
```
hist(na.omit(tract$MEDVALHOME),  
     ,breaks = seq(12500,2000500,by = 142000),xlab = "Value",  
     main="House value of MEDVAL")
```



```
boxplot(na.omit(tract$MEDVALHOME), main="House value")
```



```
boxplot(tract$MEDVALHOME~tract$CITYPERI, main="House value",xlab = "Location",ylab="Price")
```



```
tapply(tract$MEDVALHOME, tract$CITYPERI, summary)
```

```
$CityDallas
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
29800	83100	128200	219270	297950	1932700	19

```
$East
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
68400	94975	119800	126225	144900	327300	2

```
$North
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
35400	137200	170200	182496	213500	365300	2

```
$ParkCities
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
936400	1082550	1292350	1340313	1511700	2000001

```
$South
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
69300	111700	131100	135021	144900	286100

```
$West
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
12600	106100	136100	181778	229000	577600	4

Comment: Based on the histogram, boxplot, and summary statistics, we can observe the overall home values have a negative skewed distribution. The mean home value is larger than the median home value. The boxplot indicates many outliers (Very high home value) are existed. Among the six neighborhoods, Park City has the highest mean home value.


[b] Aggregate the data-frame by the factor **CITYPERI** into a new data-frame with the aggregated statistics **mean**, and **sd** for the variable **MEDVALHOME** as well as number of census tracts in each sector (use the **length()** -function). Show your code and the aggregated data-frame with your calculated statistics. Name the variables properly. (0.4 points)

```
DF <- aggregate (MEDVALHOME~CITYPERI,data = tract,function(x){ c(mean = mean(x),sd =sd(x), Number = length(x)) })
```

CITYPERI	MEDVALHOME.mean	MEDVALHOME.sd	MEDVALHOME.Number
CityDallas	219270.34	221320.93	263.00
East	126225.00	43002.67	72.00
North	182495.92	68611.73	49.00
ParkCities	1340312.62	356946.92	8.00
South	135021.21	42983.43	33.00
West	181777.92	120756.12	77.00

[c] Compare the regular mean of the median home values based on the census tracts with the **weighted mean** based on the aggregated sector means. Use the number of census tracts in each sector as weight. Justify verbally why both means do not differ. (0.4 points)

```
RegularMean <- mean(tract$MEDVALHOME)
208911.8
WeightMean <- weighted.mean(DF$MEDVALHOME[,1], DF$MEDVALHOME[,3])
208911.8
```

[d] Merge the aggregated information to your census tract data-frame. Show your code and check that the merger was performed properly by showing the first six records (see the  function **head()**). (0.4 points)

```
m <- merge(DF, tract, by="CITYPERI")
head(m)
```