Jason Yeoh
A20457826

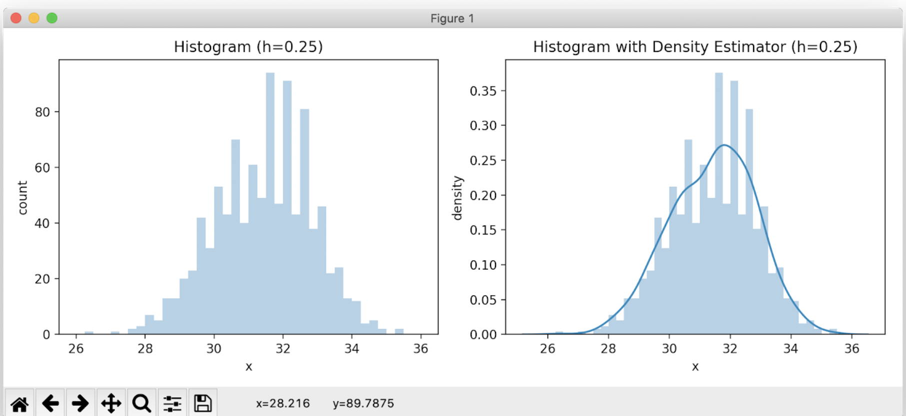
CS584 HW1
Spring 2020

QUESTION 1:

a) According to Izenman (1991) method, what is the recommended bin-width for the histogram of x is **$h = 2(IQR)N^{-1/3}$** where IQR = interquartile range.
The minimum value of x is **26.3** and the maximum value of x is **35.4**.

b) a = **26,** b = **36**

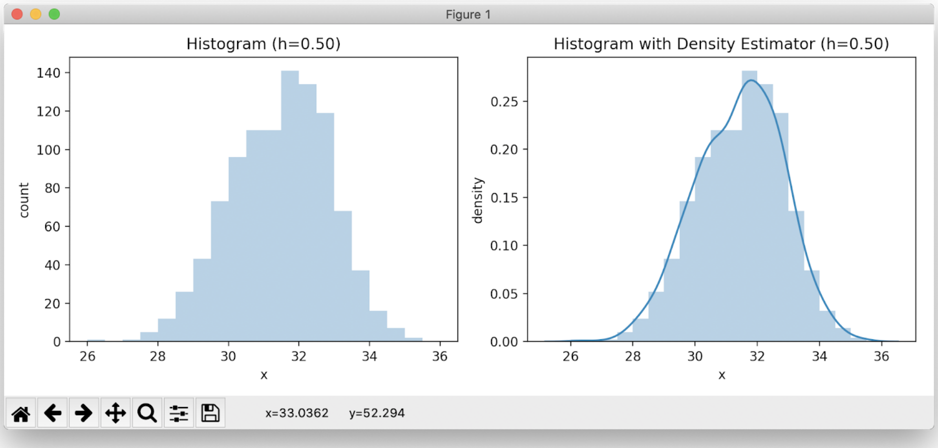
c) Histogram at h = 0.25:



Coordinates – m_i, $p(m_i)$

(26.125, 0.000)
(26.375, 0.004)
(26.625, 0.000)
(26.875, 0.000)
(27.125, 0.004)
(27.375, 0.000)
(27.625, 0.008)
(27.875, 0.016)
(28.125, 0.024)
(28.375, 0.036)
(28.625, 0.036)
(28.875, 0.072)
(29.125, 0.060)
(29.375, 0.148)
(29.625, 0.112)
(29.875, 0.188)
(30.125, 0.148)
(30.375, 0.268)
(30.625, 0.184)
(30.875, 0.228)
(31.125, 0.176)
(31.375, 0.332)
(31.625, 0.240)
(31.875, 0.324)
(32.125, 0.228)
(32.375, 0.284)
(32.625, 0.212)
(32.875, 0.228)
(33.125, 0.108)
(33.375, 0.132)
(33.625, 0.052)
(33.875, 0.064)
(34.125, 0.036)
(34.375, 0.024)
(34.625, 0.012)
(34.875, 0.008)
(35.125, 0.000)
(35.375, 0.008)
(35.625, 0.000)
(35.875, 0.000)
(36.125, 0.000)

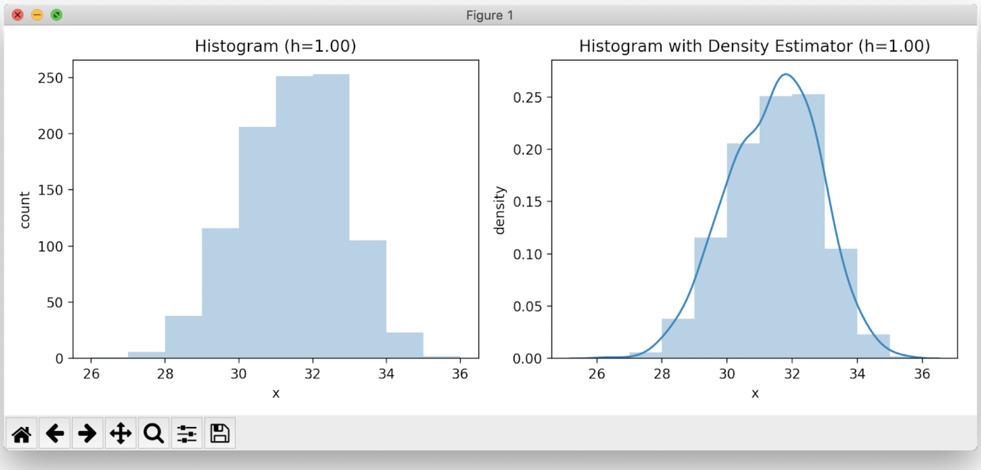
d) Histogram at h = 0.50



Coordinates – m_i, $p(m_i)$

(26.250, 0.002)
(26.750, 0.000)
(27.250, 0.002)
(27.750, 0.012)
(28.250, 0.030)
(28.750, 0.054)
(29.250, 0.104)
(29.750, 0.150)
(30.250, 0.208)
(30.750, 0.206)
(31.250, 0.254)
(31.750, 0.282)
(32.250, 0.256)
(32.750, 0.220)
(33.250, 0.120)
(33.750, 0.058)
(34.250, 0.030)
(34.750, 0.010)
(35.250, 0.004)
(35.750, 0.000)
(36.250, 0.000)

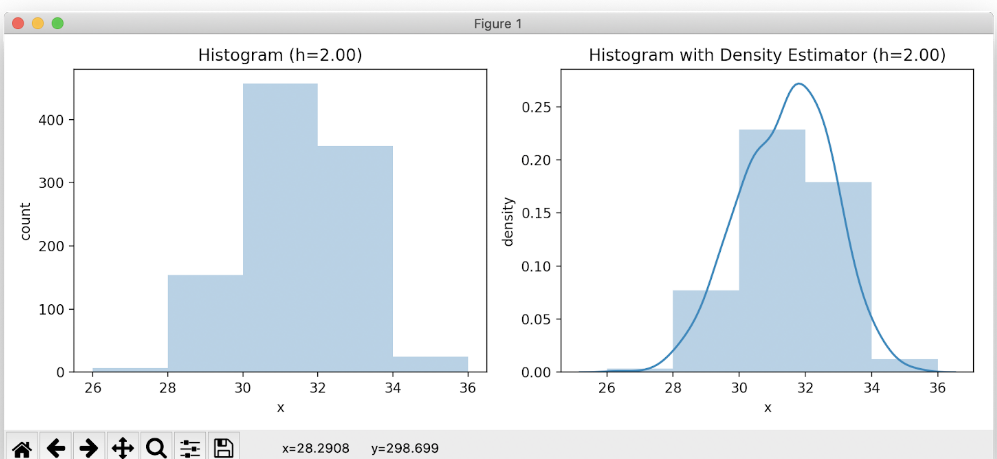
e) Histogram at h = 1



Coordinates – m_i, $p(m_i)$

(26.500, 0.001)
(27.500, 0.007)
(28.500, 0.042)
(29.500, 0.127)
(30.500, 0.207)
(31.500, 0.268)
(32.500, 0.238)
(33.500, 0.089)
(34.500, 0.020)
(35.500, 0.002)
(36.500, 0.000)

f) Histogram at h = 2



Coordinates – m_i, $p(m_i)$

(27.000, 0.004)
(29.000, 0.084)
(31.000, 0.237)
(33.000, 0.163)
(35.000, 0.011)
(37.000, 0.000)

g) Personally, I would **histogram at h = 0.5** (at part e) as it portrays the shape and spread of the distribution of the data very well. The histogram at h = 0.25 provides an overly specific distribution of the data in which does not give an overall insight on the data. While, histograms at h = 1 and h = 2 projects the appropriate shape and spread of the data. These two histograms lack on specificity on the data as it gives a generalized trend over the data.

QUESTION 2

a) Five-number summary of x:

Min	26.3
Q1	30.4
Median	31.5
Q3	32.4
Max	35.4

IQR: 2.0
1.5IQR Whiskers: [27.4 - 35.4]

b) Five-number summary of x on group 0:

Min	26.3
Q1	29.4
Median	30.0
Q3	30.6
Max	32.2

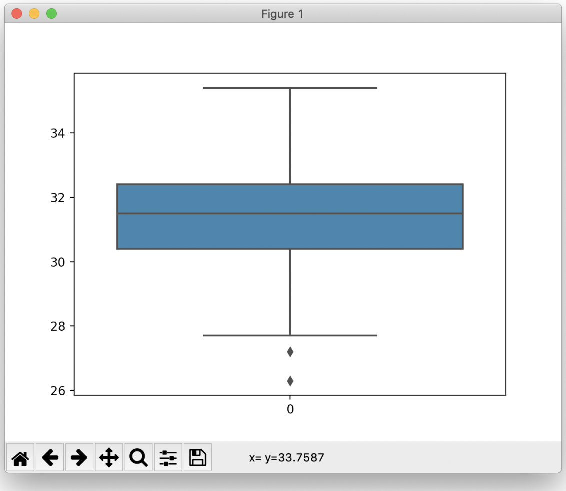
IQR: 1.2
1.5IQR Whiskers: [27.6 - 32.4]

Five-number summary of x on group 1:

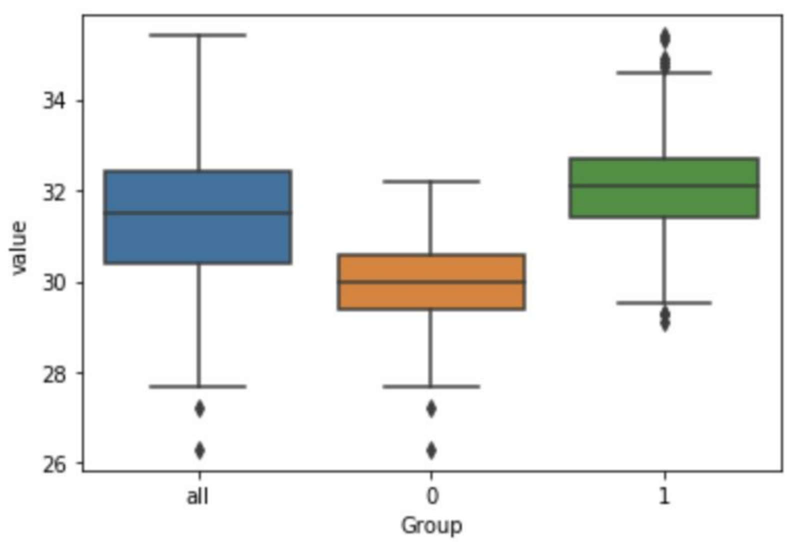
Min	29.1
Q1	31.4
Median	32.1
Q3	32.7
Max	35.4

IQR: 1.3
1.5IQR Whiskers: [29.45 - 32.65]

c) **Yes,** because it correctly represents two low outliers (<27.4) and no high outliers (>35.4).



d) **Box plots** (entire data, group 0 data, group 1 data)



OUTLIERS FOR THE ENTIRE DATA – (position, value)
70 27.2
295 26.3

OUTLIERS FOR GROUP 0 – (position, value)
70 27.2
295 26.3

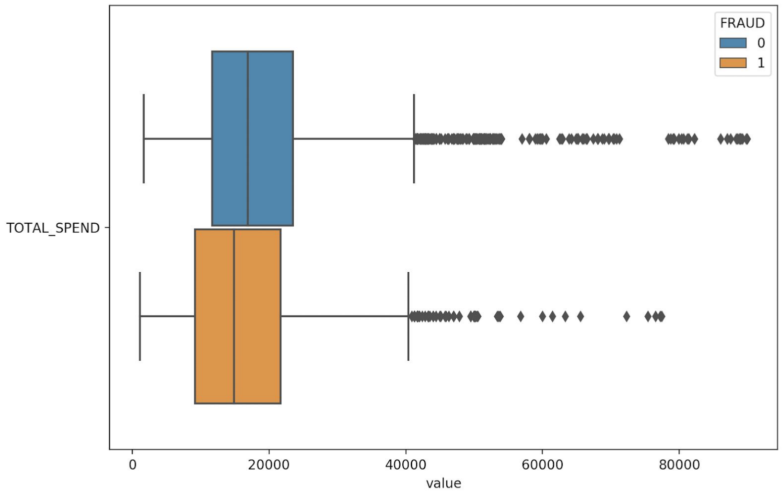
OUTLIERS FOR GROUP 1 – (position, value)

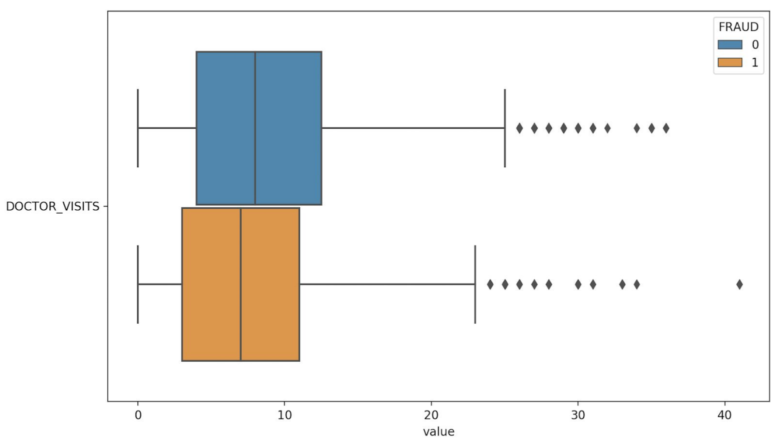
30	35.3	846	34.7
107	29.3	907	34.8
297	35.4	938	29.3
812	34.9	975	29.1

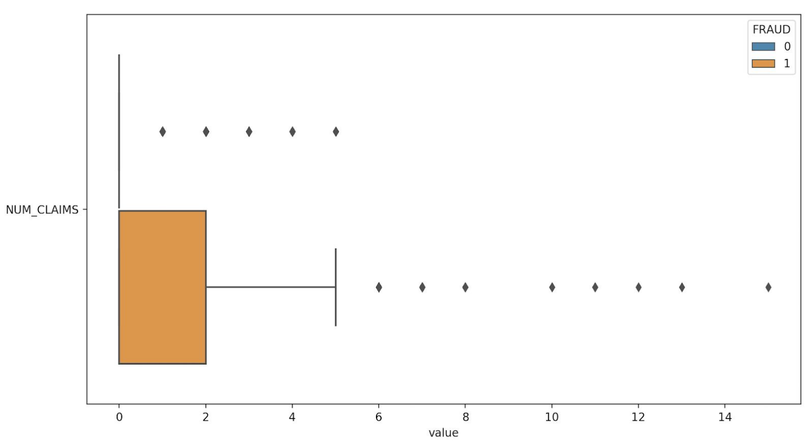
QUESTION 3

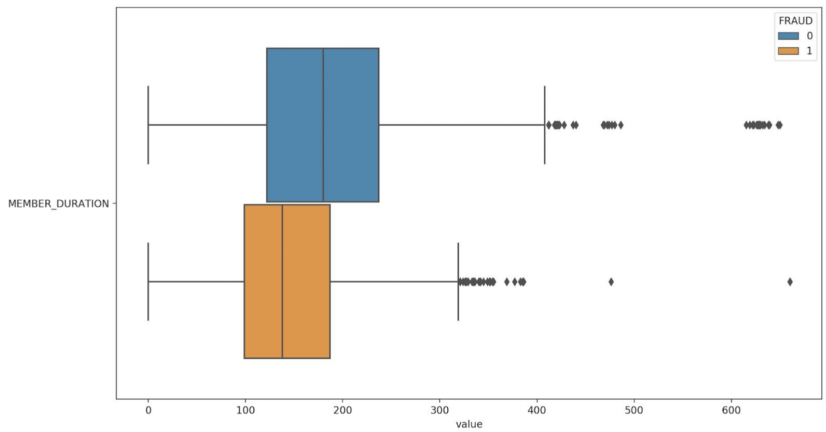
a) % of fraudulent investigations = **0.1995**

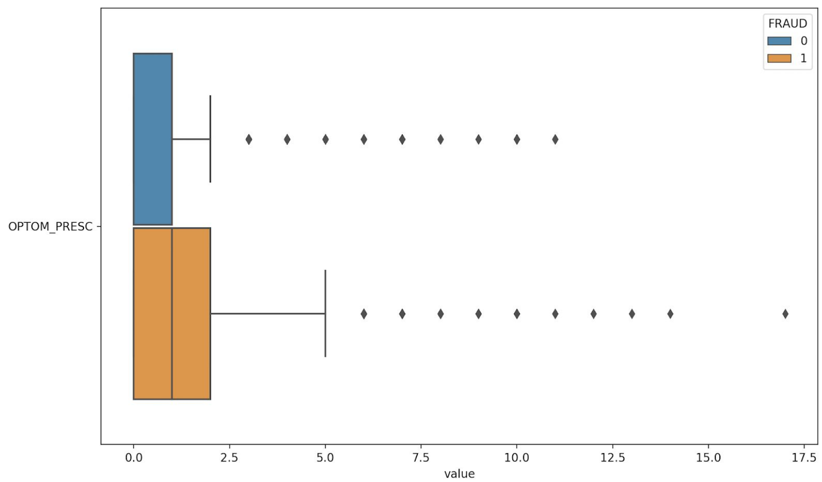
b) **Boxplots** (0 = non-fraud, 1 = fraud)

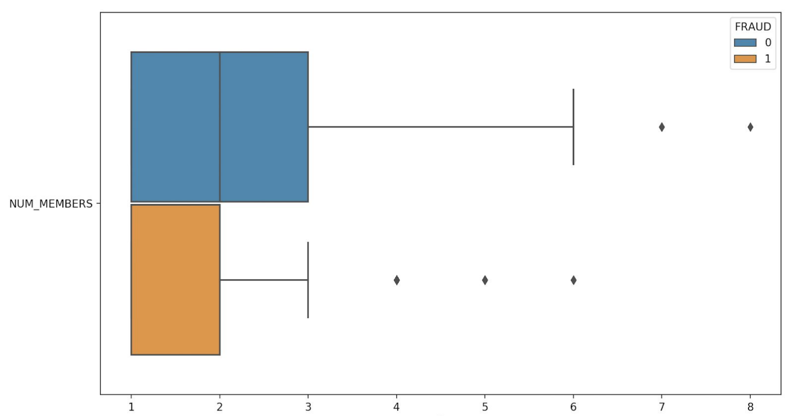












c) Orthonormalization
   i) There were **six dimensions** used.

```
Eigenvalues of x =
 [6.84728061e+03 8.38798104e+03 1.80639631e+04 3.15839942e+05
 8.44539131e+07 2.81233324e+12]
```

ii) Transformation matrix:

```
Transformation Matrix =
 [[-6.49862374e-08 -2.41194689e-07  2.69941036e-07 -2.42525871e-07
  -7.90492750e-07  5.96286732e-07]
 [ 7.31656633e-05 -2.94741983e-04  9.48855536e-05  1.77761538e-03
   3.51604254e-06  2.20559915e-10]
 [-1.18697179e-02  1.70828329e-03 -7.68683456e-04  2.03673350e-05
   1.76401304e-07  9.09938972e-12]
 [ 1.92524315e-06 -5.37085514e-05  2.32038406e-05 -5.78327741e-05
   1.08753133e-04  4.32672436e-09]
 [ 8.34989734e-04 -2.29964514e-03 -7.25509934e-03  1.11508242e-05
   2.39238772e-07  2.85768709e-11]
 [ 2.10964750e-03  1.05319439e-02 -1.45669326e-03  4.85837631e-05
   6.76601477e-07  4.66565230e-11]]
```

The resulting variable is **ORTHONORMAL** since xtx provides an **identity matrix**.

```
Expect an Identity Matrix =
 [[ 1.00000000e+00 -3.00432422e-16 -4.61219604e-16  5.45323877e-15
   1.20996962e-15 -1.28911638e-16]
 [-3.00432422e-16  1.00000000e+00 -6.44449771e-16 -2.76820667e-14
  -1.23512311e-15  7.78890841e-16]
 [-4.61219604e-16 -6.44449771e-16  1.00000000e+00  3.49546780e-15
   1.21430643e-16 -2.39391840e-16]
 [ 5.45323877e-15 -2.76820667e-14  3.49546780e-15  1.00000000e+00
   1.14968798e-14 -3.47812057e-15]
 [ 1.20996962e-15 -1.23512311e-15  1.21430643e-16  1.14968798e-14
   1.00000000e+00 -6.31439345e-16]
 [-1.28911638e-16  7.78890841e-16 -2.39391840e-16 -3.47812057e-15
  -6.31439345e-16  1.00000000e+00]]
```

(see code for the resulting variable)

d) Score function
   i) The score function returned a value of **~81.96%** (0.8196308724832215). I used 80/20 train-test data split. Without any train-test data split, the score function returns **~87.79% (**0.8778523489932886).

   ii) The score function returns the mean (average) accuracy on the given test data and labels.

e) Nearest neighbors (of transformed matrix)
   i) [553, 3870,   31, 1030, 2748, 2173],
   ii) [425, 3160, 5673, 1245, 2173, 2748],
   iii) [1101, 5202, 4228, 1231, 2224, 776],
   iv) [457, 4968, 5416, 733, 776, 2224],
   v) [1232, 176, 4776, 369, 1893, 478]

f) The predicted probability of fraudulent investigation is right around **10%** (0.10067114093959731). The observation at (e) would be **misclassified** as it was wrongfully classified as **non-fraud** even though it is a fraudulent investigation.