ANOVA problem solution

1. Running

I use Python2.7 in this homework. With numpy, scipy, matplotlib installed, simply run "python stat.py", it will generate all graphs in img folder and print statistics to the screen.

2. Assumptions of ANOVA

ANOVA assumes:

- Data are randomly sampled. Data samples are indenpent from each other
- The variances of each group are assumed as equal. Empirically, ratio of largest to smallest group standard deviation must be less than 2:1.
- The residues are normally distributed.

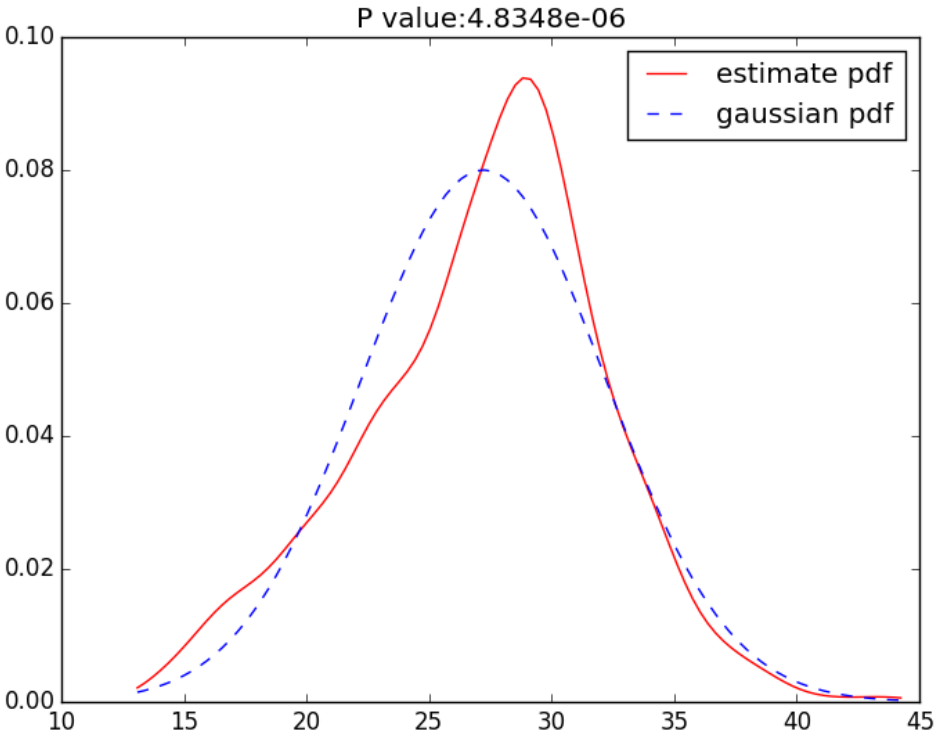
3. Hypothesis

Hypothesis:

- Null hypothesis (H0): Average ages in each category are equal.
- Alternative hypothesis (H1): Not all average ages in different categories are equal

4. ANOVA of average age

a) Estimate probability density function of average age and normality test
Here, we use Kernel Density Estimation with Gaussian kernel to estimate the pdf of average age (red line). To check the normality in a more intuitive way, I draw the pdf of normal distribution (blue line) with the same mean (27.22) and variance(24.85) on the same graph. From the graph, we can see that the pdf is skewed from normal, which imply it's not normal.



To test the normality of average age more convincingly , I use D'Agostino and Pearson's method provided by SciPy, which combines skew test and kurtosis test. This method tests the null hypothesis that a sample comes from a normal distribution. It returns $s^2 + k^2$, where s is the z-score of skew test, and k is the z-score of kurtosis test. The value is chi-square distributed. The corresponding P value indicates the probability of larger chi-square value to be seen, under the hypothesis that the data is normal. The lower P value is, there are more evidence to reject the hypothesis.

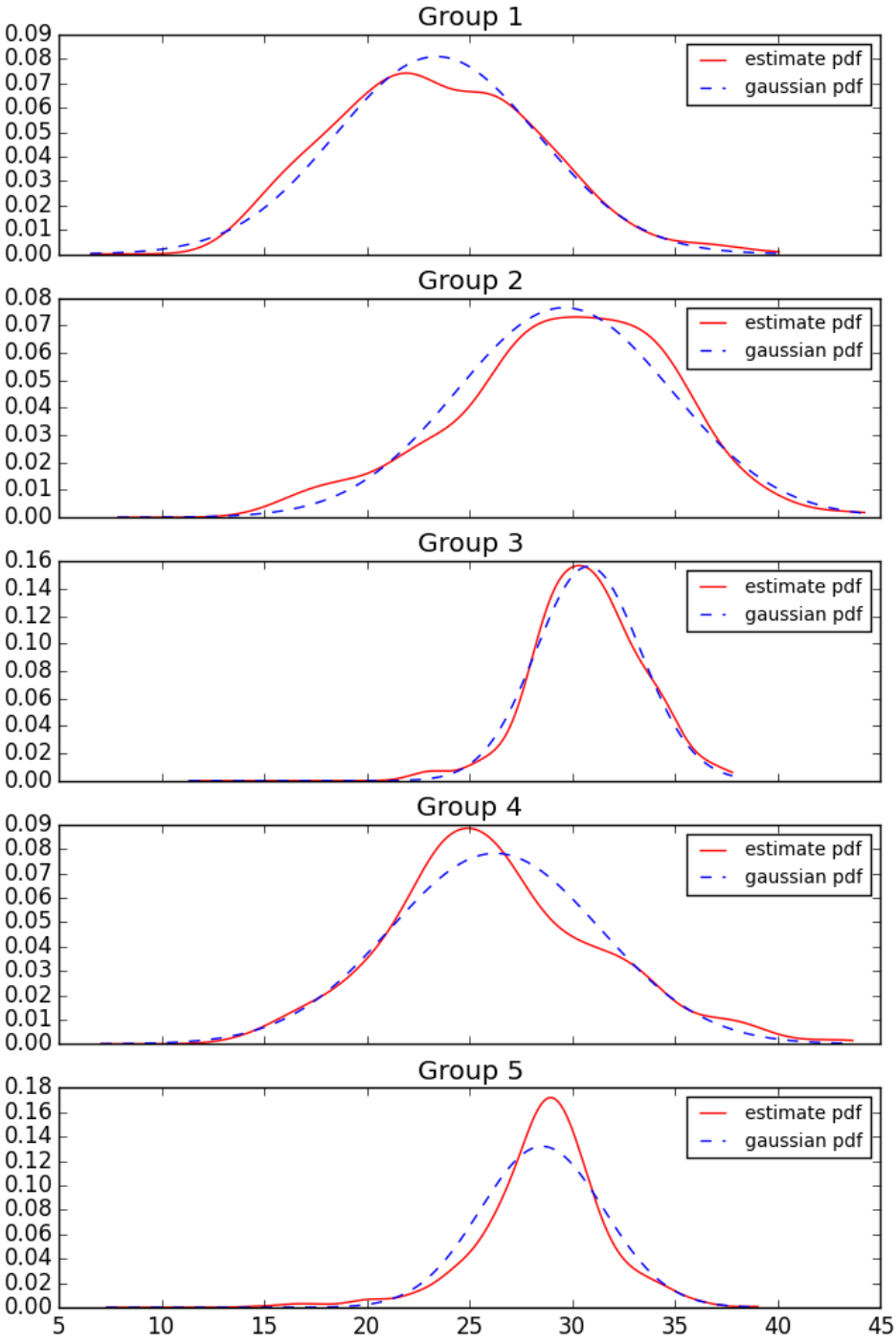
Here, the normality test gives us

$$s^2 + k^2 = 24.48$$

$$p = 4.83 \times 10^{-6}$$

In significance level 0.01, we can say average age of whole data is not normal.

b) Pdf estimation and normality test in each category
Execute the above process for each category, we get



From the graph, we can see the first three groups fits the normal distribution quite well. The last two groups are not. For numerically normality test, check the p value:

	1	2	3	4	5
P value	0. 018686172	0. 063506481	0. 223563684	0. 000531072	2. 23E-20

From the p value of normality test, we can say that average age in category 1,2,3 is normally distributed in significance level 0.01. Category 4 is close to normal, and category 5 is not.

Now we check the homogeneity of variance

	1	2	3	4	5
Variance	24.19	27.13	6.48	25.93	9.10
Std. Dev	4.92	5.21	2.55	5.09	3.02
Max/min stddev ratio	2.04				

Max and Min standard deviation ratio is 2.04. The ratio is slightly larger than 2. We can say the variance is homogeneous, although not strictly.

c) One-way ANOVA for average age.

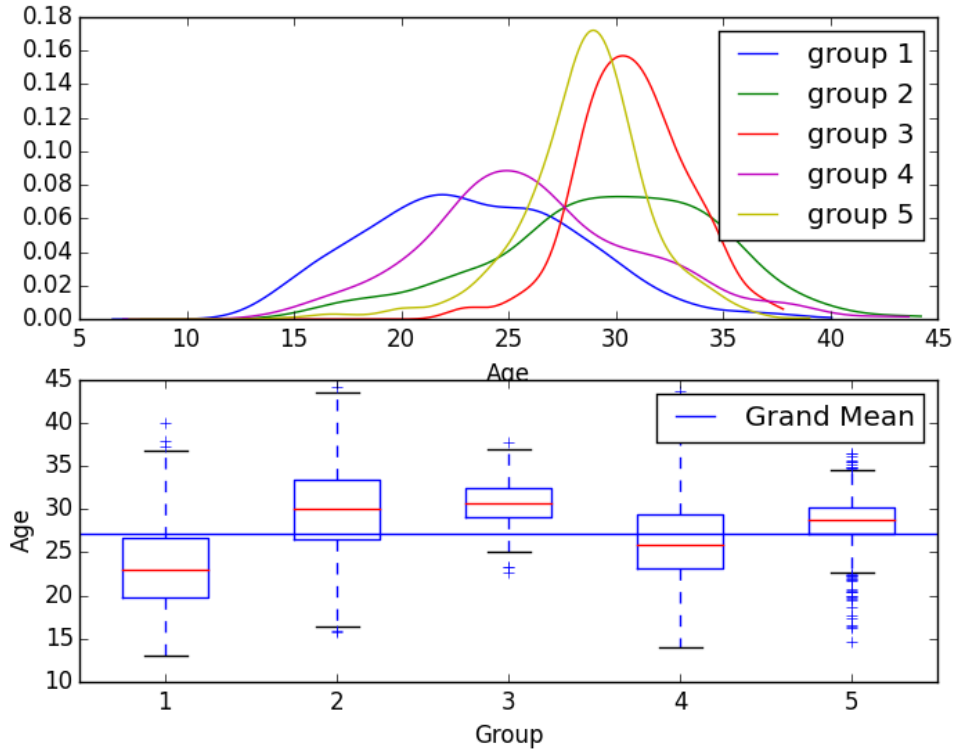
In previous normality test and variance homogeneity test, we see that the data doesn't fully satisfy the assumption of classic ANOVA. But ANOVA is robust against non-normal data and has a small effect on Type-I error. We can still perform ANOVA on this column.

Running ANOVA gives us F statistics and p value

$$F = 171.50$$
$$p = 1.08 \times 10^{-126}$$

The p value is quite small. It's strong evidence to reject the null hypothesis. We can conclude that the not all average age in each category is equal.

To visualize the data, we draw the pdf of each group on the same graph. The distribution of data in different groups is quite different. Further, we draw the box plot of each group to inspire our analysis. The box bottom and top is the first and third quartiles. The dashed line is the median, and the red line the group mean. The blue line is the grand mean.



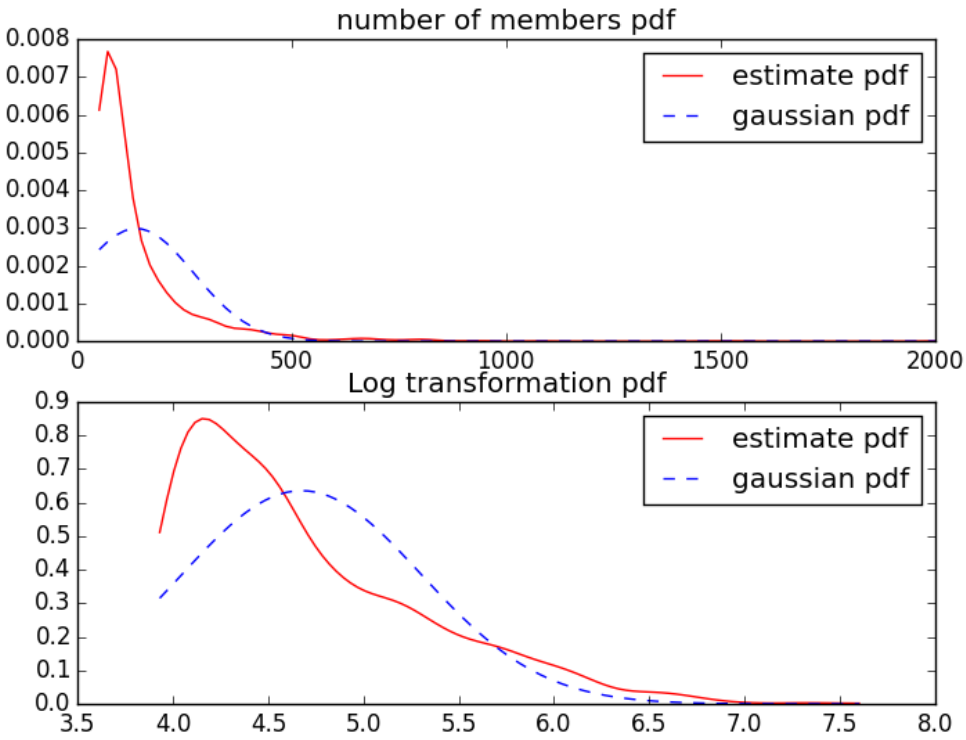
From the box plot, we can see clearly then group 1 has a quite different mean from other groups. The same with group 4. Group 2,3,5 have similar mean.

5. Other columns

We select three columns “number of members 群人数”, “number of messages 消息数”, “number of conversations 会话数”. Execute normality test on them and their log transformation.

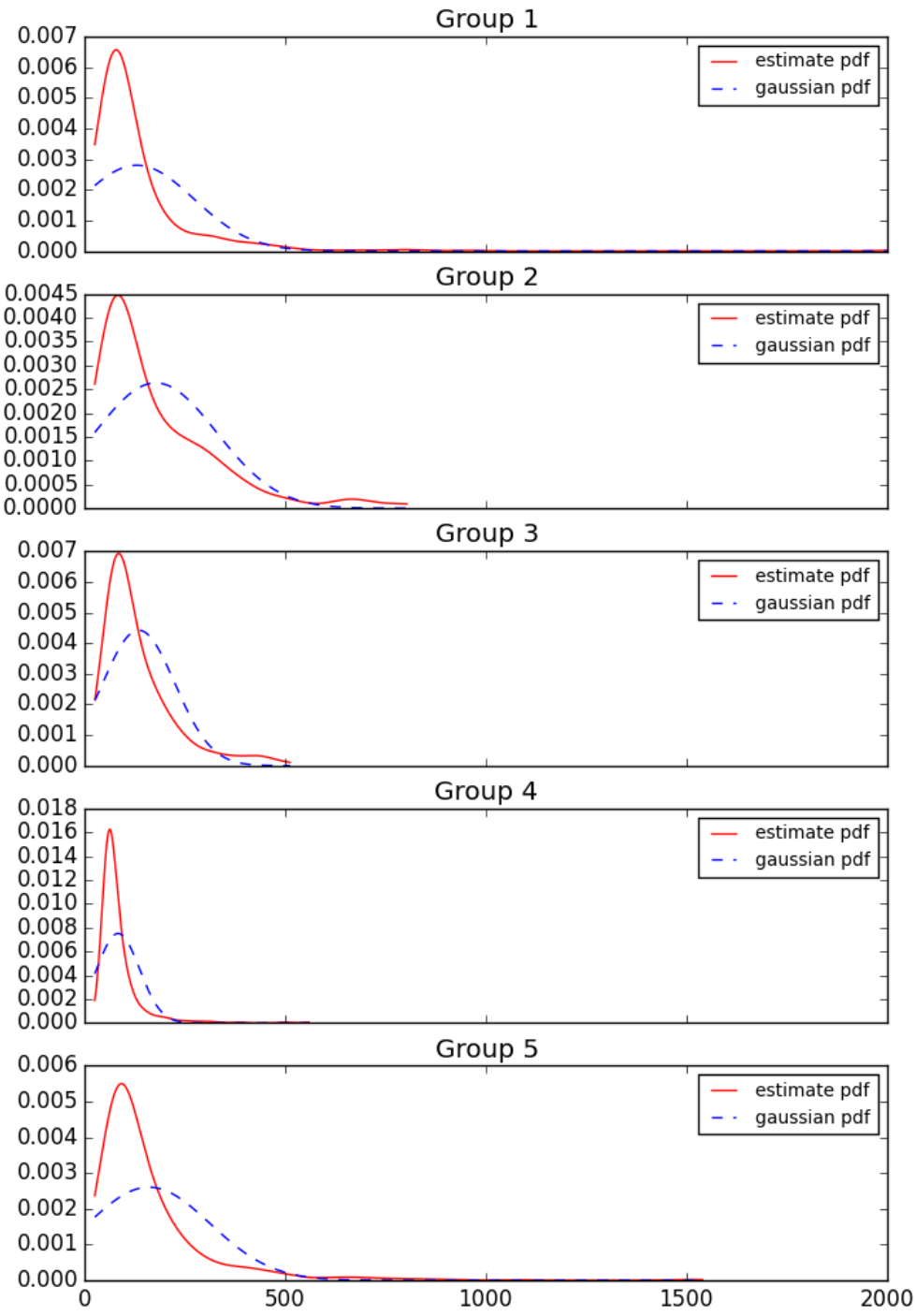
a) Number of members

- Pdf of number of members and its log transformation

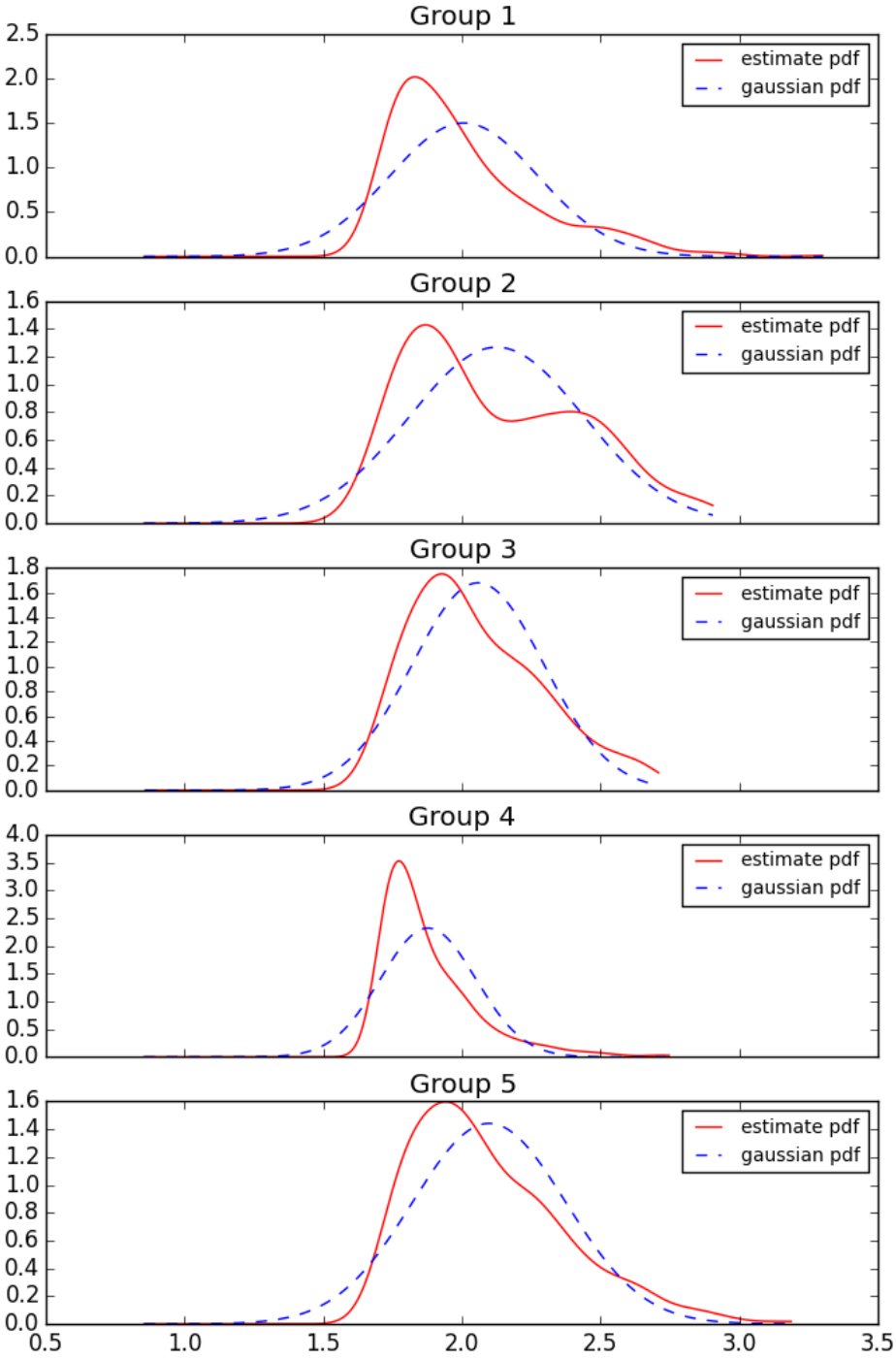


- Pdf of each group and its log transformation

Original:



Log transformation:



- Normality test

	$s^2 + k^2$	p
Number of Member	1959.91	0
Log transformation	312.45	1.41e-68

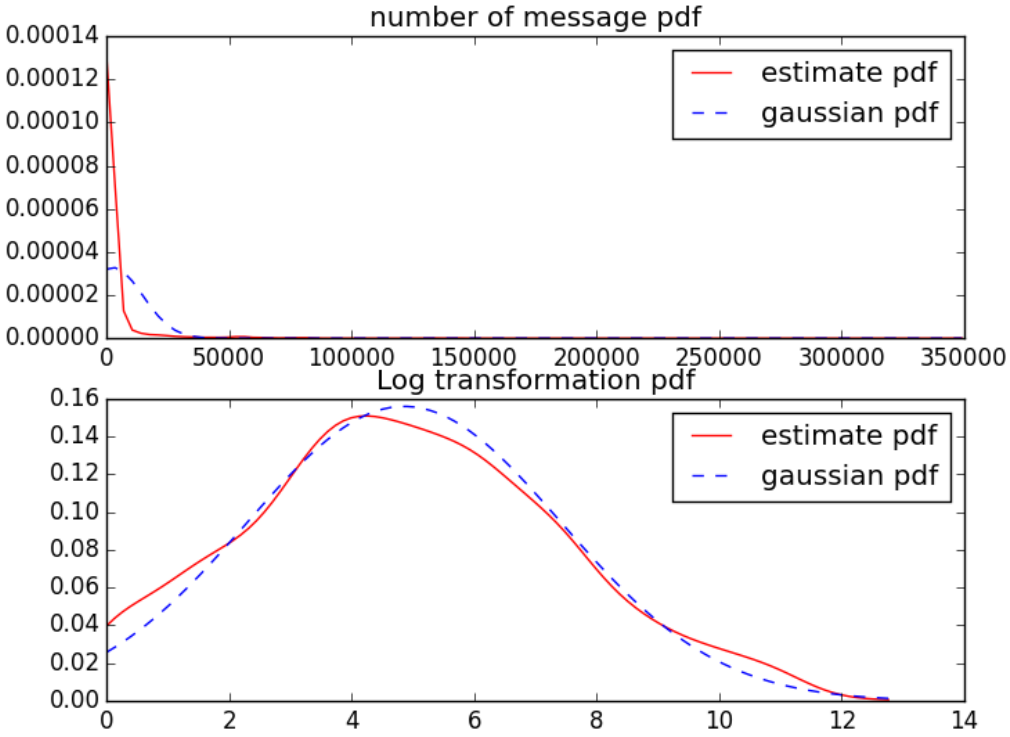
We can see the number of members is quite unlikely to be normal. The log transformation is not, either.
Now we check the normality of each group:

p	Original	Log transformation
1	2.95e-27	4.68e-25
2	4.28e-45	1.64e-10
3	3.31e-19	0.17
4	2.33e-69	1.28e-5
5	1.91e-57	5.80e-6

The original data is far from normal. After log transformation, it's closer to normal. But they still can't be seen as normal.
Because the residue is far from normal, the number of members or its log transformation doesn't satisfy the ANOVA assumption.

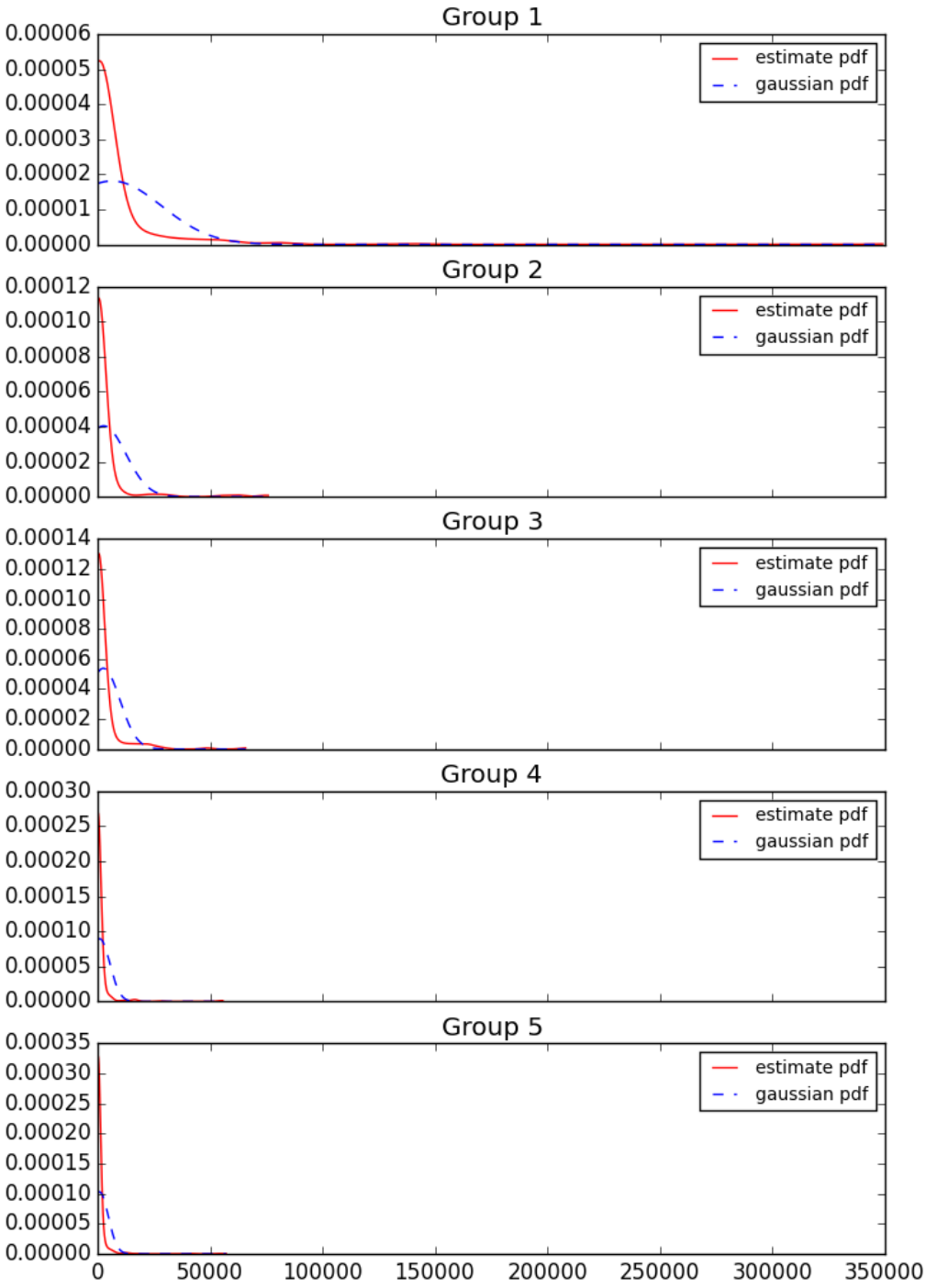
b) Number of messages

- Pdf of number of messages and its log transformation

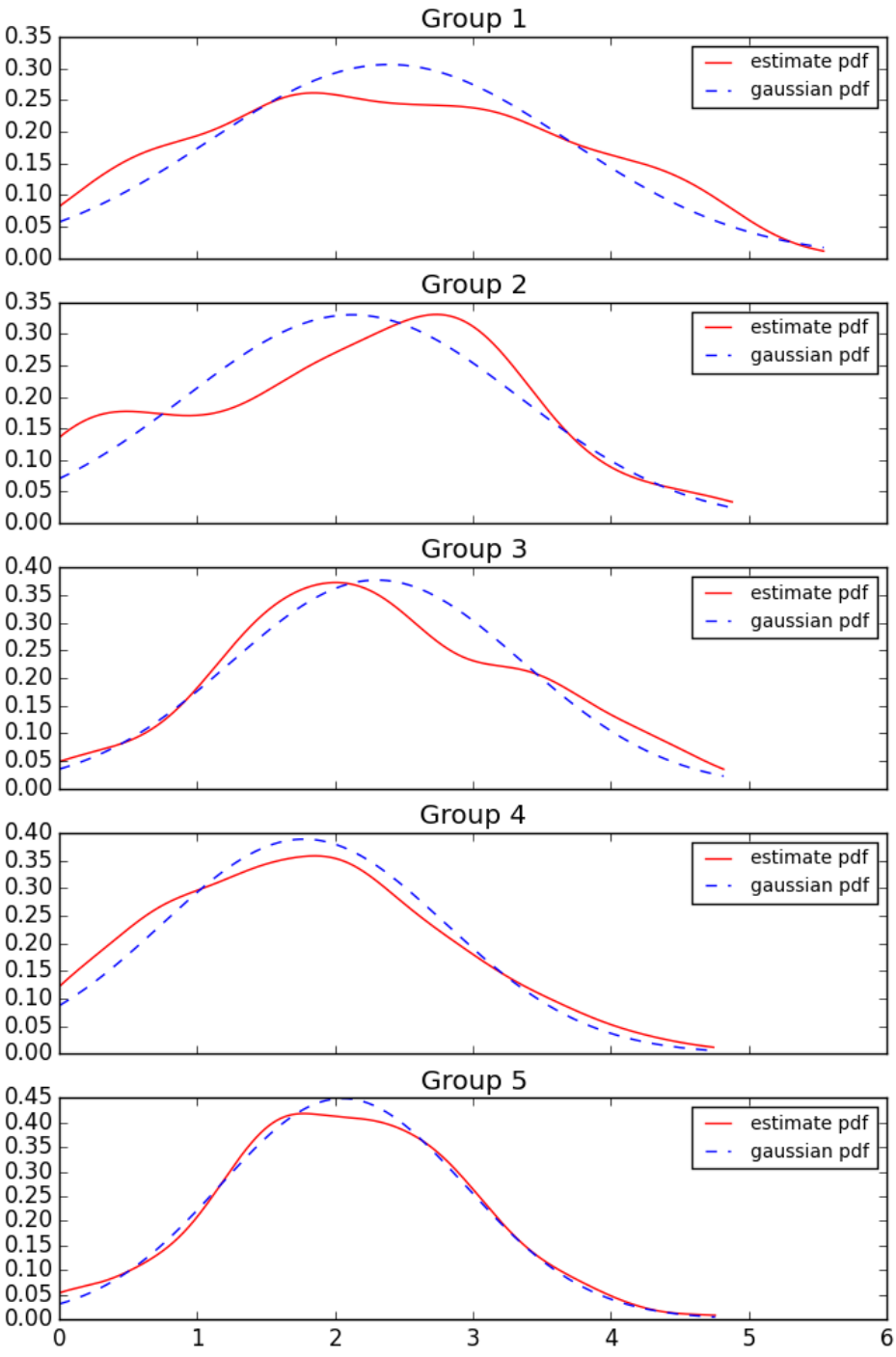


- Pdf of each group and its log transformation:

Original:



Log transformation:



- Normality test for whole data:

	$s^2 + k^2$	p
Number of Member	3947.92	0
Log transformation	32.36	9.39e-8

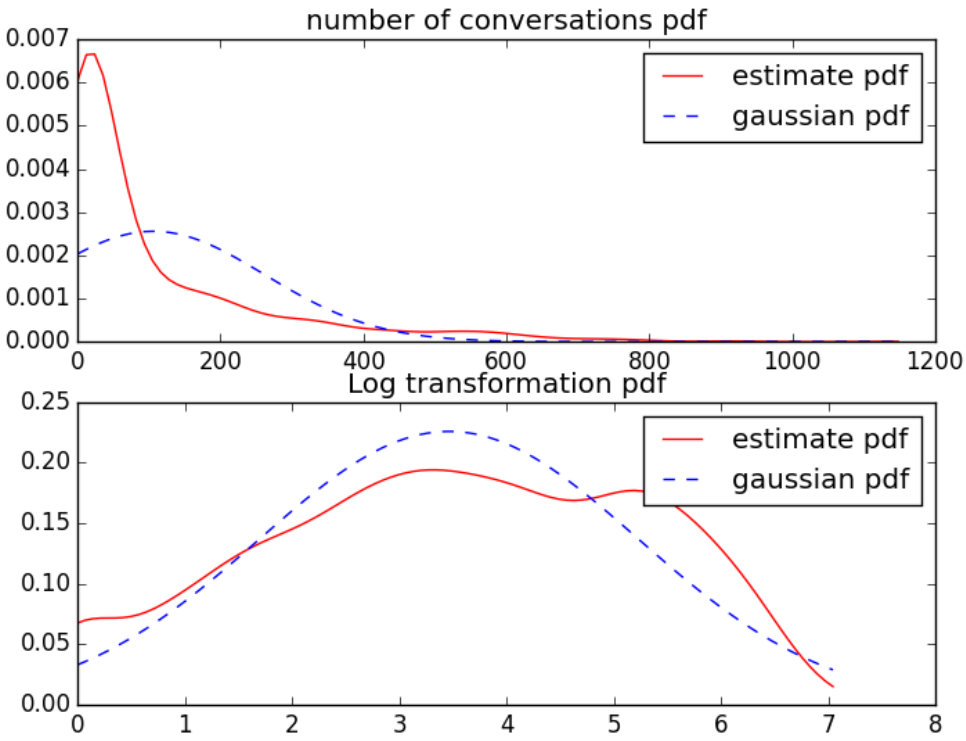
For each group:

p	Original	Log transformation
1	3.06e-172	4.70e-13
2	2.23e-78	0.001
3	4.93e-54	0.24
4	1.16-151	0.006
5	1.37e-240	0.88

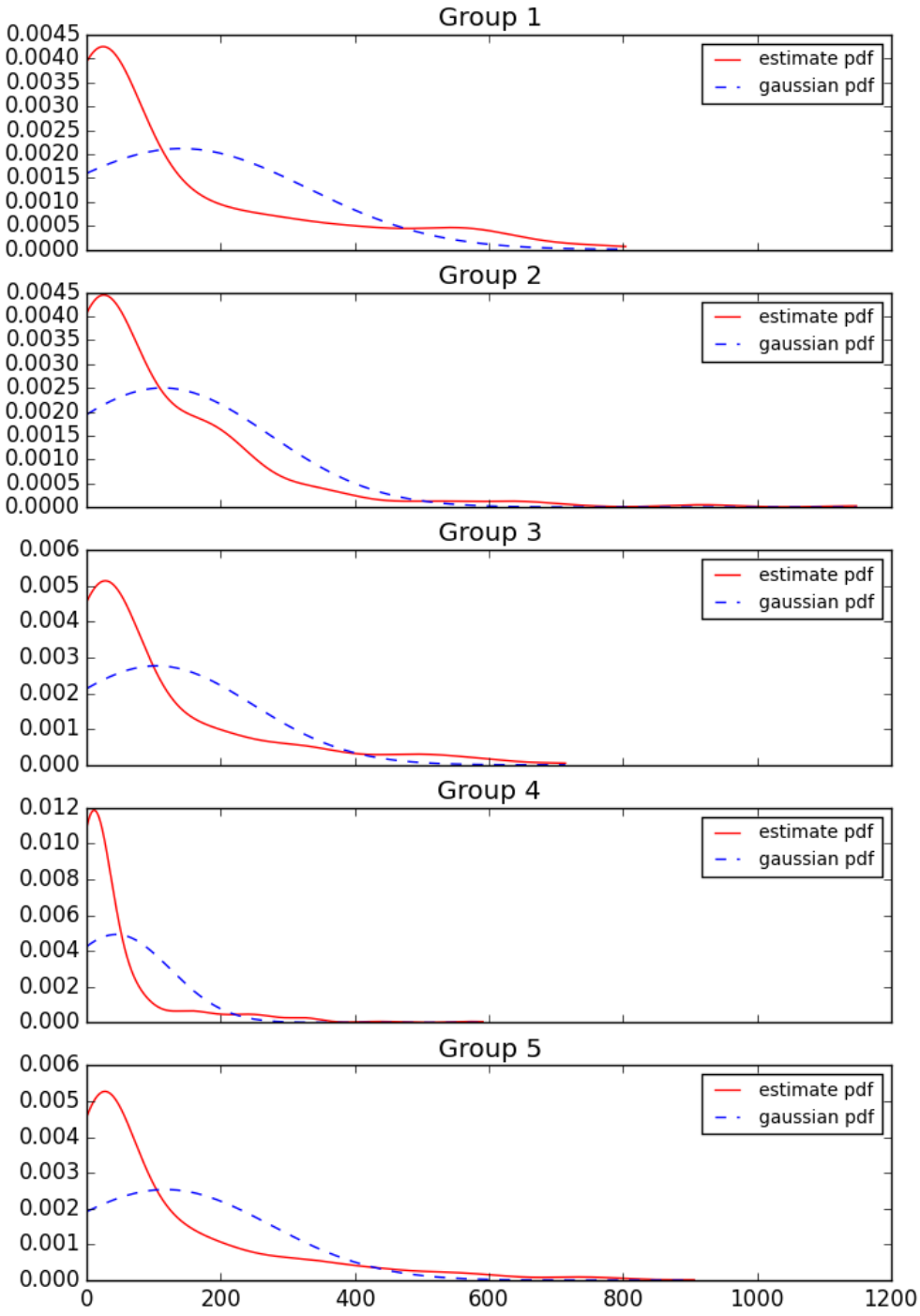
From the normality test, the original number of messages are not normal, failing the ANOVA assumption. But after log transformation, Group3,5 are normal distributed under significance level 0.05 and Group 2,4 are quite close. We further check the the max/min standard deviation ratio is 1.46. So it satisfies the variance homogeneity assumption. So the log transformation of messages roughly satisfies theANOVA assumptions.

c) Number of conversations

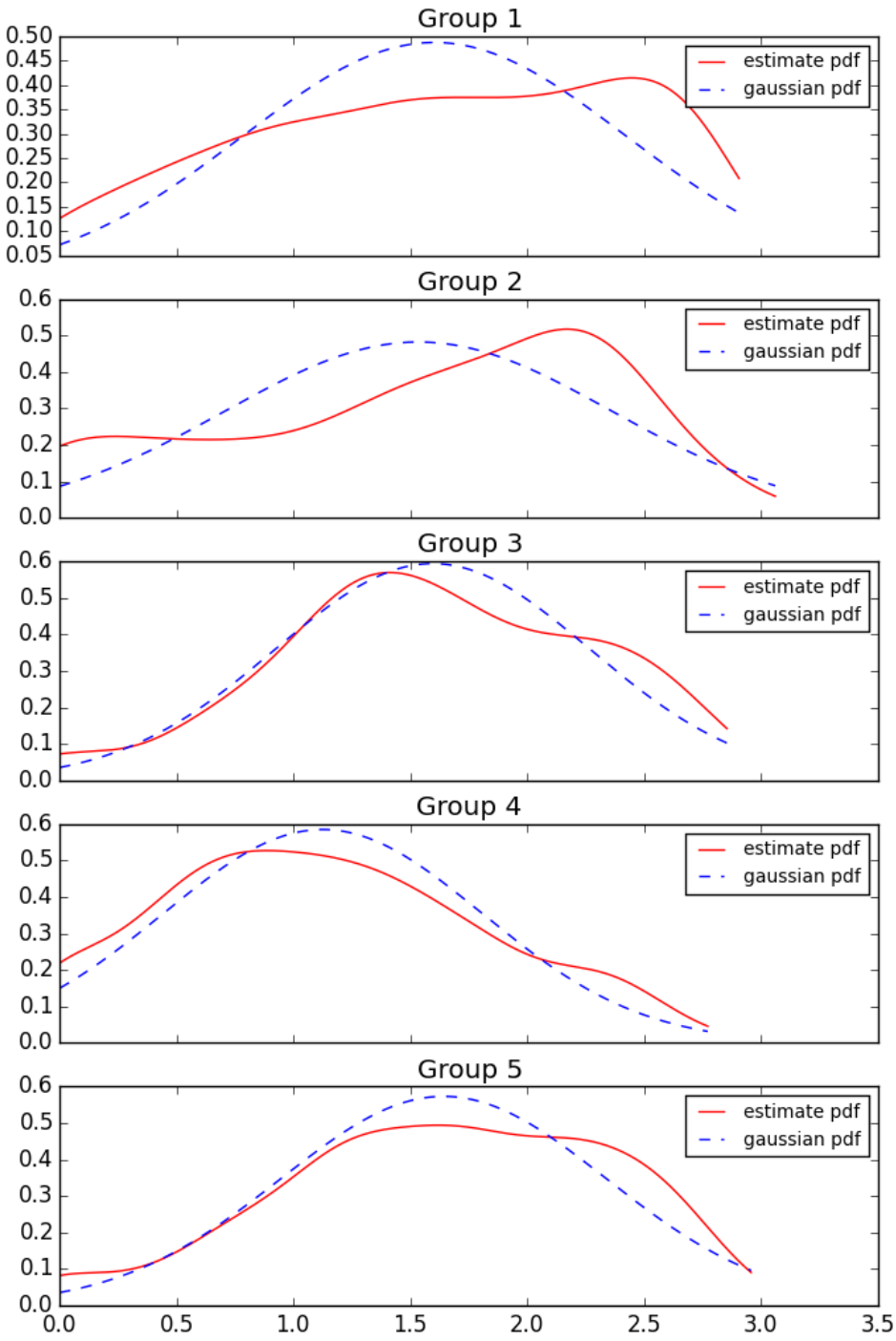
- Pdf of original data and its log transformation:



- Pdf of each group and its log transformation:
  Original:



Log transformation:



- Normality test

Normality test for whole data:

	$s^2 + k^2$	p
Number of Member	913.95	3.44e-199
Log transformation	236.75	3.88e-52

For each group:

p	Original	Log transformation
1	2.95e-27	4.68e-25
2	4.28e-45	1.64e-10
3	3.31e-19	0.18
4	2.33e-69	1.28e-5
5	1.33e-57	5.80e-6

Non of the original data and its log transformation is normal, it doesn't satisfy the ANOVA assumptions.

6. ANOVA for non-normal

a) Method

1, One-way ANOVA is considered robust against the normality assumption, and can usually tolerate data that are non-Gaussian (skewed or kurtotic distributions) with a small effect on the Type I error rate. So we can apply ANOVA on non-normal data.

2, Apply non-parametric Test, including:

- Mann–Whitney U Test (for two groups)
- Kruskal-Wallis H Test (generalization of Mann-Whitney)
- Friedman test
- Kendall's W test (normalization of Friedman test

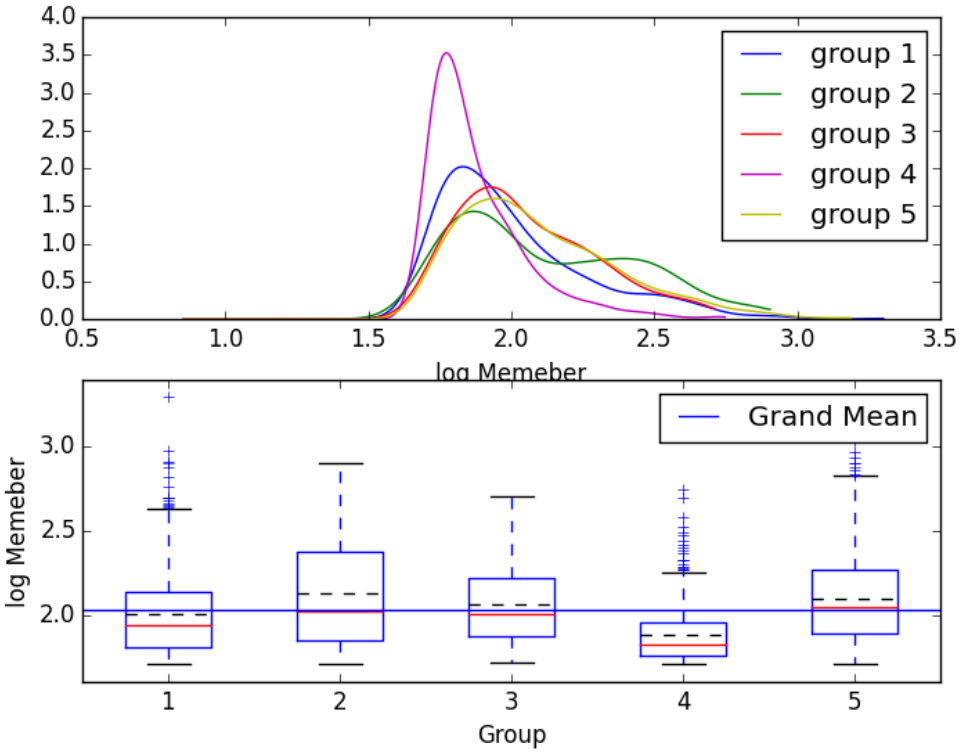
b) ANOVA for non-normal columns

We perform Kruskal-Wallis H Test on the three columns. Because the data magnitude is quite large in the three columns, we visualize its log transformation.

- Number of members

$$H = 256.2$$
$$p = 2.99 \times 10^{-54}$$

Rejecting the null hypothesis in significance level 0.01.

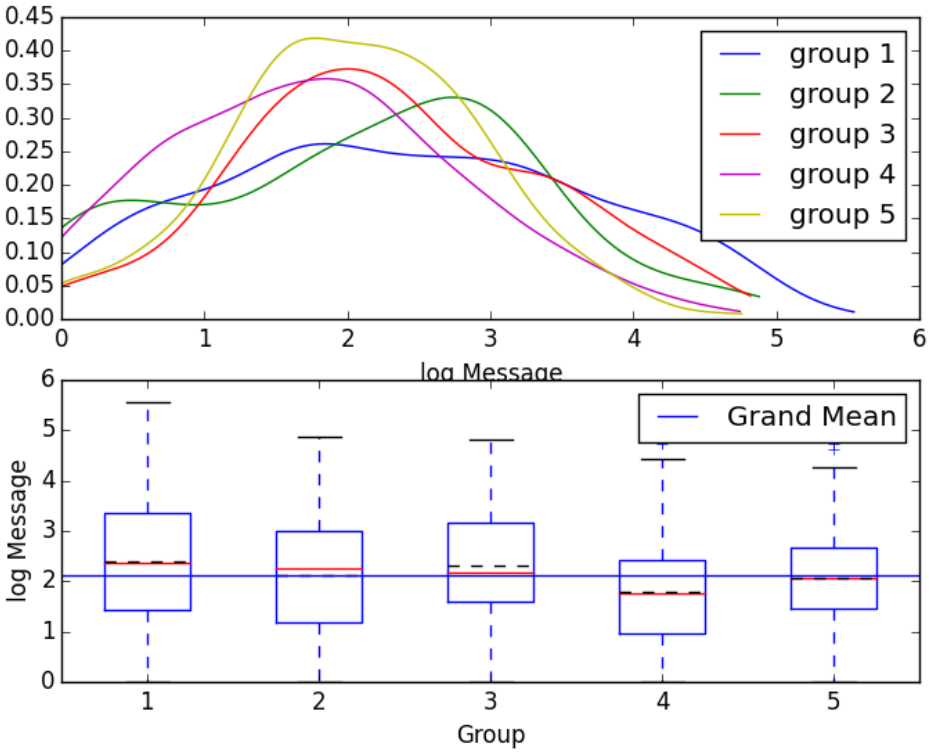


The mean varies significantly in different groups.

- Number of messages

$$H = 66.86$$

$$p = 1.04 \times 10^{-13}$$

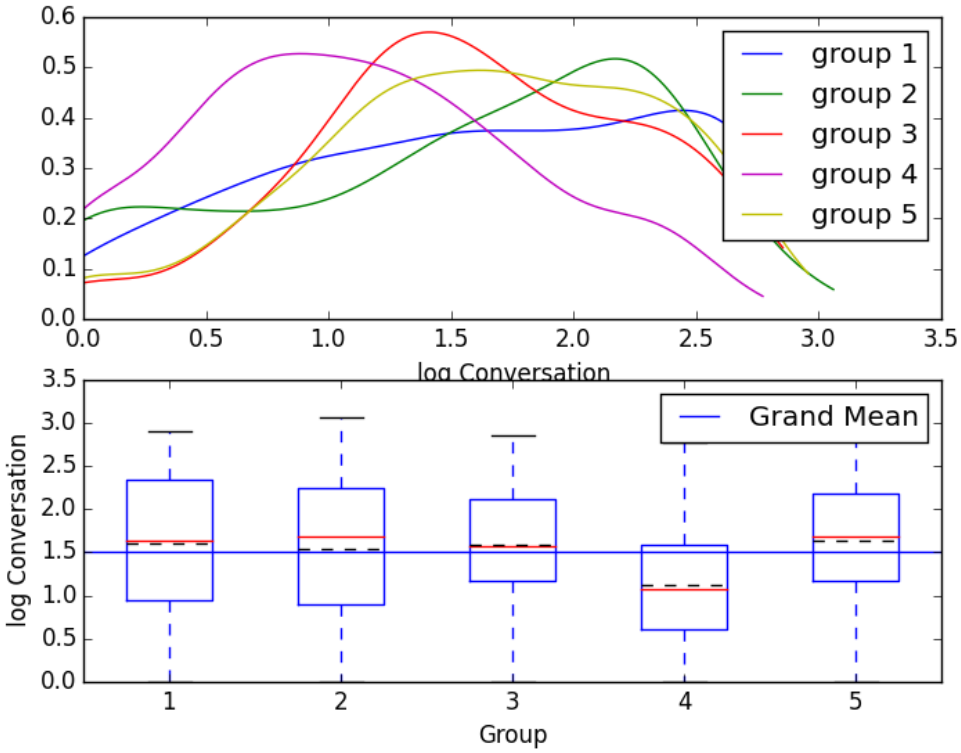


According to the H test, we'll reject the null hypothesis in significance level 0.01. But the p value is quite higher than that of number of member. Just as we can see, the mean of messages varies not so significantly.

- Number of conversations

$$H = 139.29$$

$$p = 4.00 \times 10^{-29}$$



Reject the null hypothesis. And the difference of mean is significant in different groups.