Experiments

1. (**5 points**) Recall and write down the **assumptions** which **one-way ANOVA** are based on.
    - 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 residuals are normally distributed (not skewed or partial)
2. (**5 points**) Focus on two columns: **Category** (Col[2]) and **Average Age** (Col[7]). Taking feature Average Age as an example, we want to measure whether the average age varied significantly across the categories. Clearly state the null (H0) and the alternative (H1) hypotheses for this task.

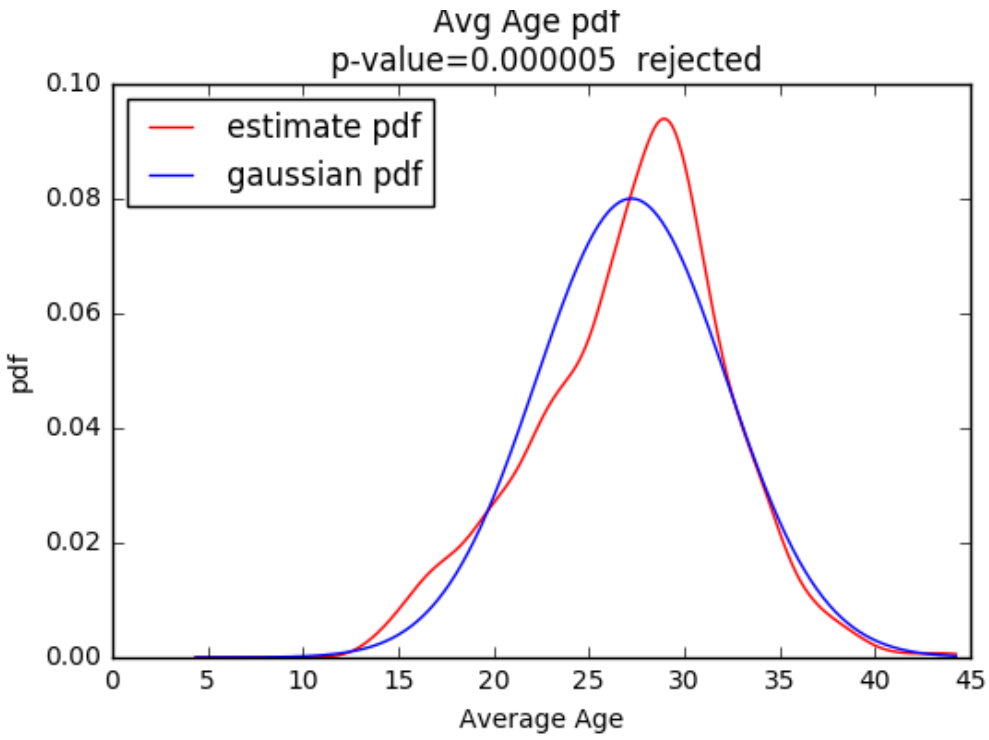
    the null (H0)： Average ages in each category are equal.

    the alternative (H1)： Not all average ages in different categories are equal
3. Use your favorite statistics analysis software, like Matlab, R, Excel, SPSS or ...

    In below practice, I use the jupyter(IPython notebook), and many statistic analysis tool, such as sciPy, numpy,matplotlib,pandas, sklearn, etc.

    1. (**10 points**) Draw the **empirical probability density function** of Col[7], i.e. the empirical pdf of average age. Does the data in this dimension follow **Gaussian distribution**? Test **normality** of Col[7].

        the empirical pdf of average age:

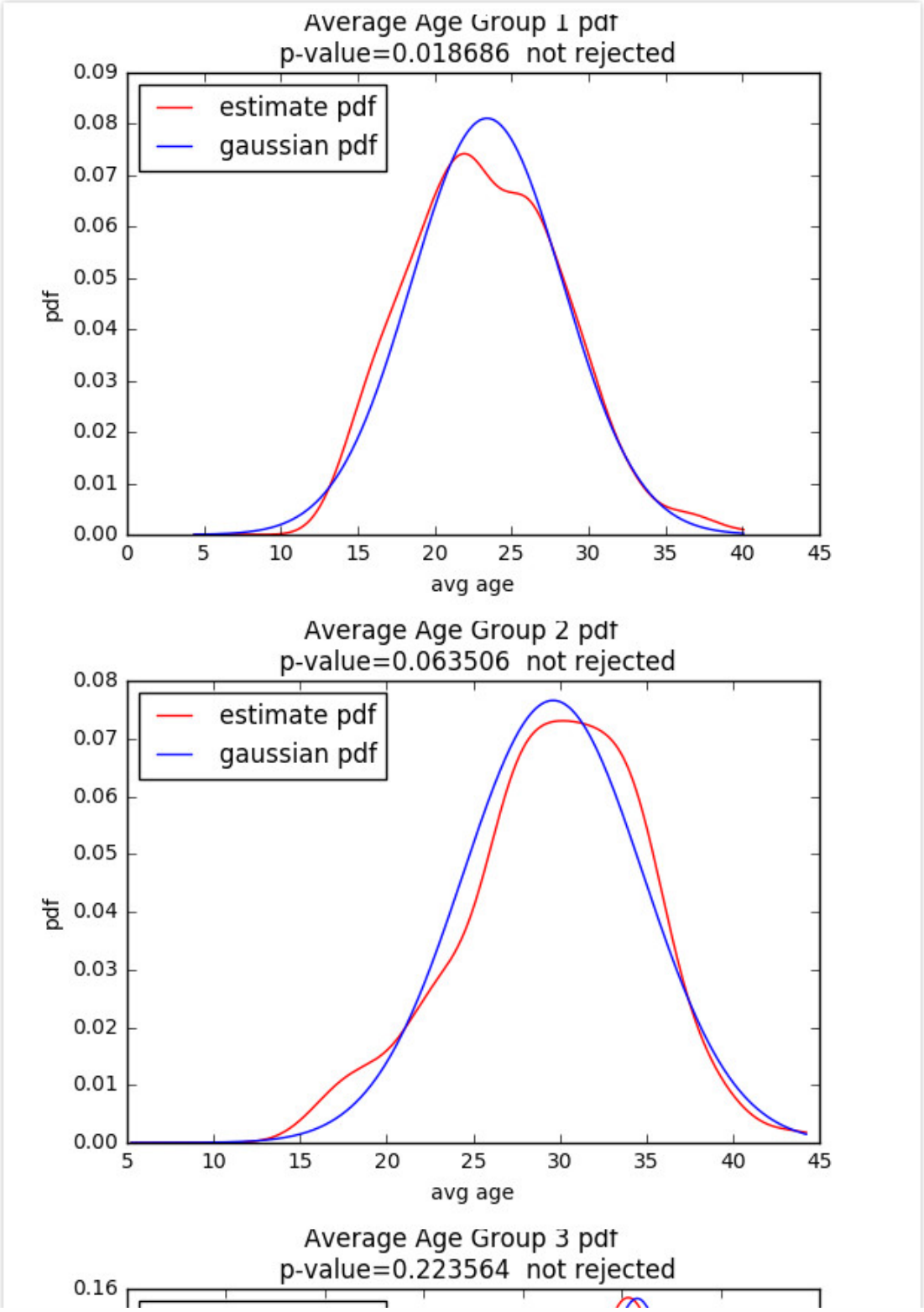
        

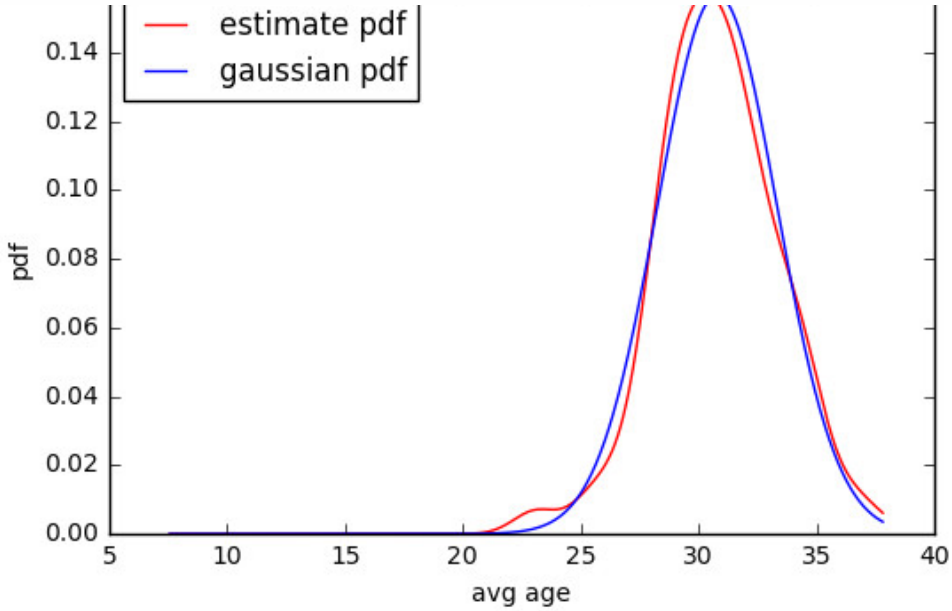
        the red line represents estimate probability density function. The blue one represents gaussian pdf which has the same mean and variance with estimate pdf. In this part, I use the function ***gaussian_kde*** provided by sciPy to draw pdf, which combines skew test and kurtosis test. I used ***spicy.stats.normaltest*** to get P value, 4.8348e-6, which is

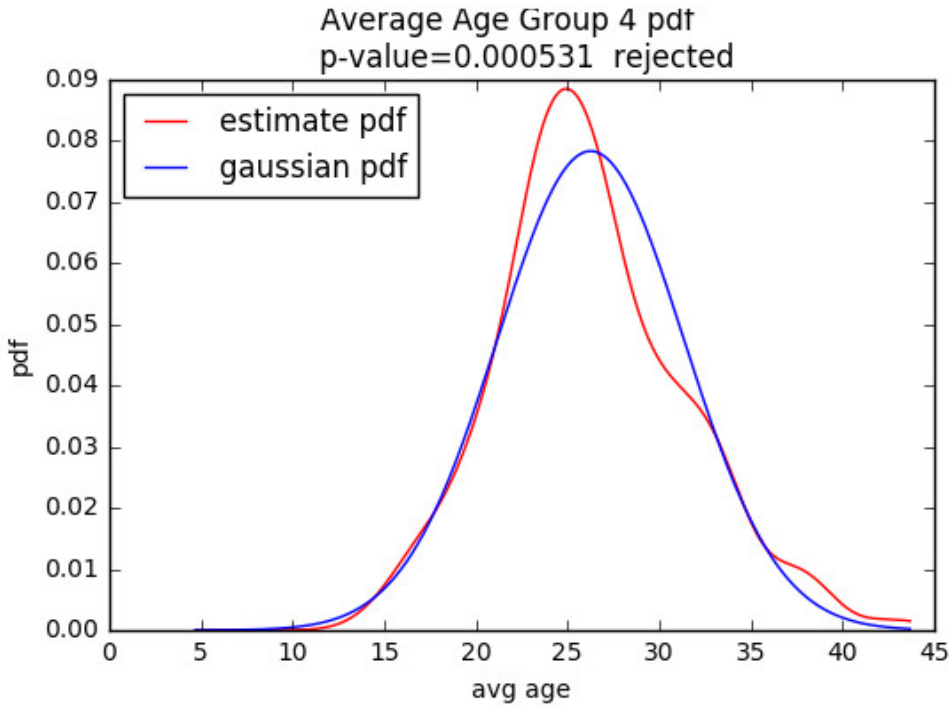
less than 0.01,the significance level. So we can imply that it has enough evidence to reject the null hypotheses H0(the data follow Gaussian distribution) under the significance level 0.01.

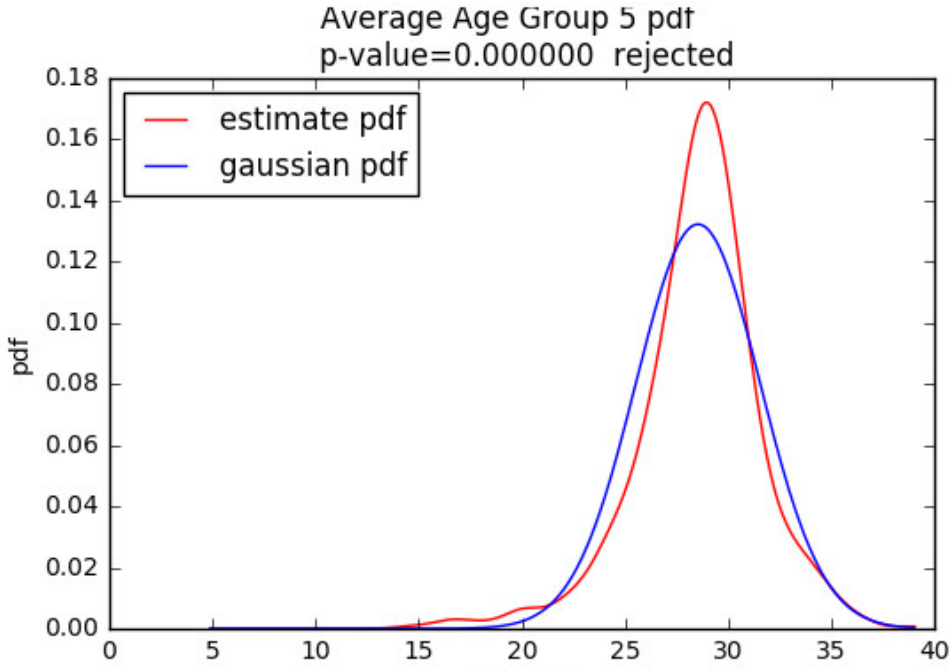
Mean	Variance	p	Significance level
27.22	24.85	4.8348e-6	0.01

2. (**10 points**) In Col[7], there are 5 components divided by category labels. We denote the data in Col[7] with category i (where i = 1,...,5) as Col[7|categoty=i]. Test **the normality of each components** and test **the homogeneity of variances**.




estimate pdf
gaussian pdf
pdf
avg age


Average Age Group 4 pdf
p-value=0.000531 rejected
estimate pdf
gaussian pdf
pdf
avg age


Average Age Group 5 pdf
p-value=0.000000 rejected
estimate pdf
gaussian pdf
pdf

avg age

	Group1	Group2	Group3	Group4	Group5
P-value	0.0186862	0.0635065	0.223564	0.000531072	2.22673e-20
Is_rejected	not rejected	not rejected	not rejected	Rejected	Rejected

We get P-value of 5 groups in the average age. There are 3 groups follow the Gaussian distribution, that is , Group1 Group2 Group3 have good normality.

Next, we need to get the homogeneity of variances.

we test the homogeneity of variances through avg_age_std.max()/avg_age_std.min().

if avg_age_std.max()/avg_age_std.min()>2, that is, , ratio of largest to smallest group standard deviation is more than 2:1, it doesn't follow the homogeneity of variances.

```
avg_age_std.max()/avg_age_std.min() = 2.04551837158
```

Although 2.04551837158 is slightly larger than 2, the distribution of average age follows the homogeneity of variances.

3. (**20 points**) Do **the one-way ANOVA test** for Col[7] with categories in Col[2]. Write down your **conclusion**, supporting statistics, and **visualize your data** which inspire the process.

Although the data doesn't satisfy the assumptions of ANOVA strictly, fortunately, ANOVA is not very sensitive to moderate deviations from normality. We still use ANOVA to analysis the data in the column average age.

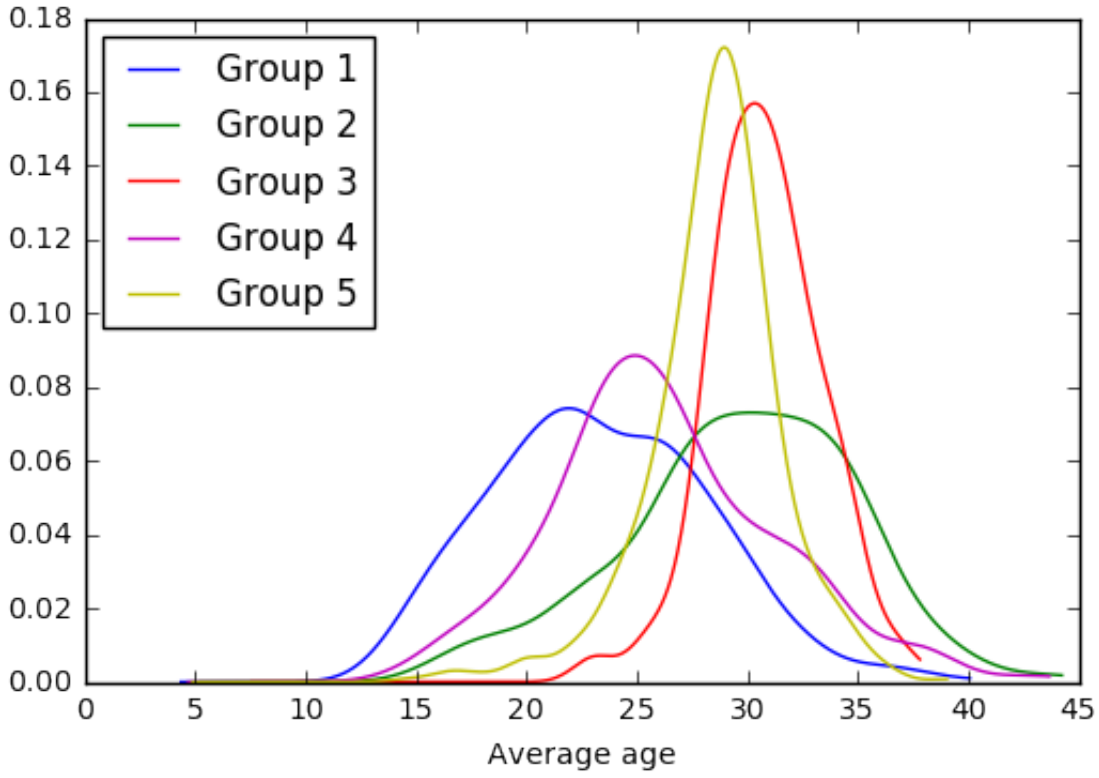
I use ***stats.f_oneway(*group)*** to get F value and P value.

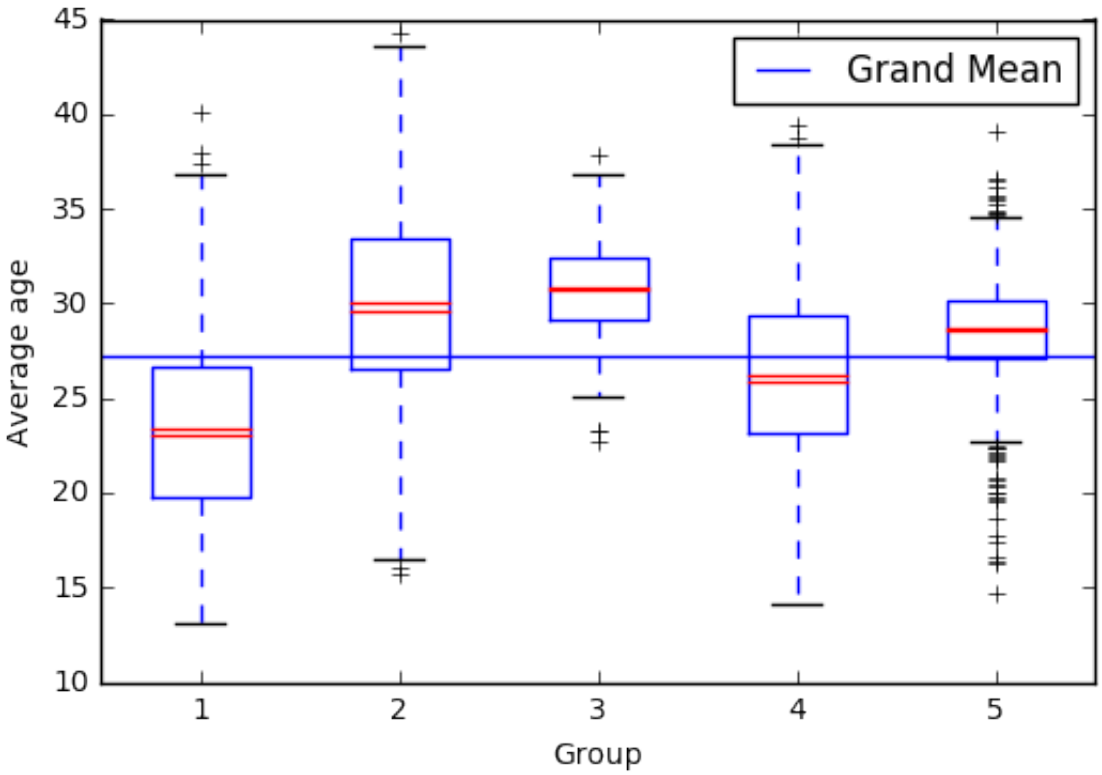
F value	**171.507032707**
P value	**1.08209160648e-126**

P-value<<0.01. we can imply that the null hypothesis is rejected.

The **conclusion** is that not all average age in each category is equal.

Next part , I will display the pdf of each group to show how different group differs from each other.



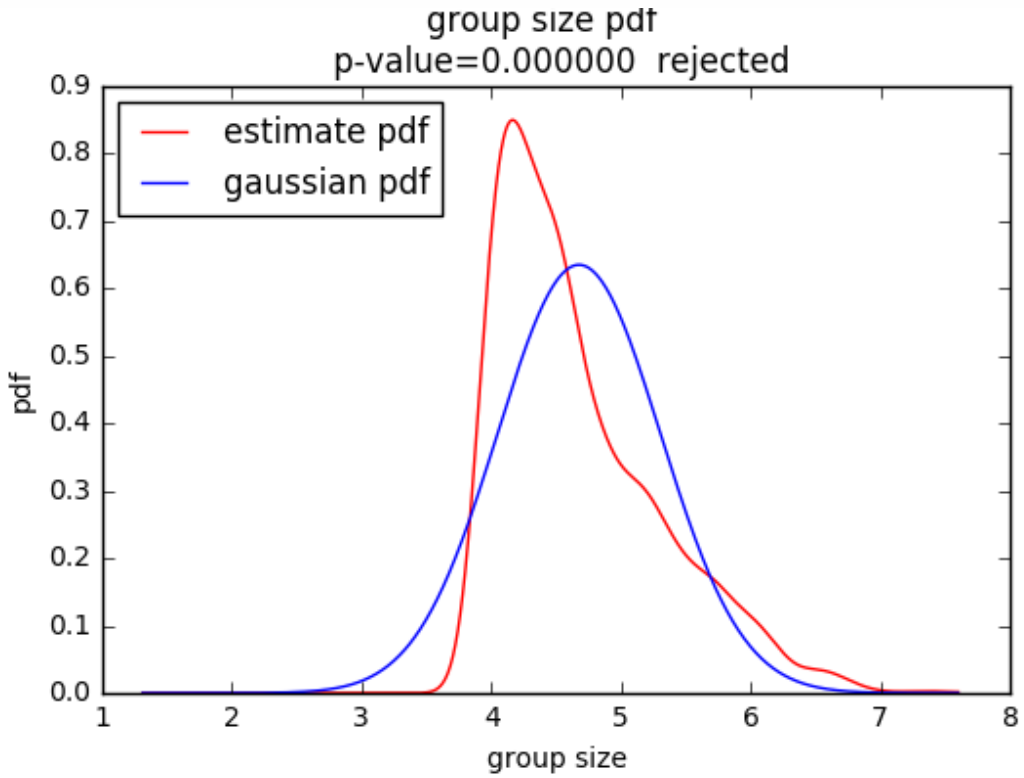


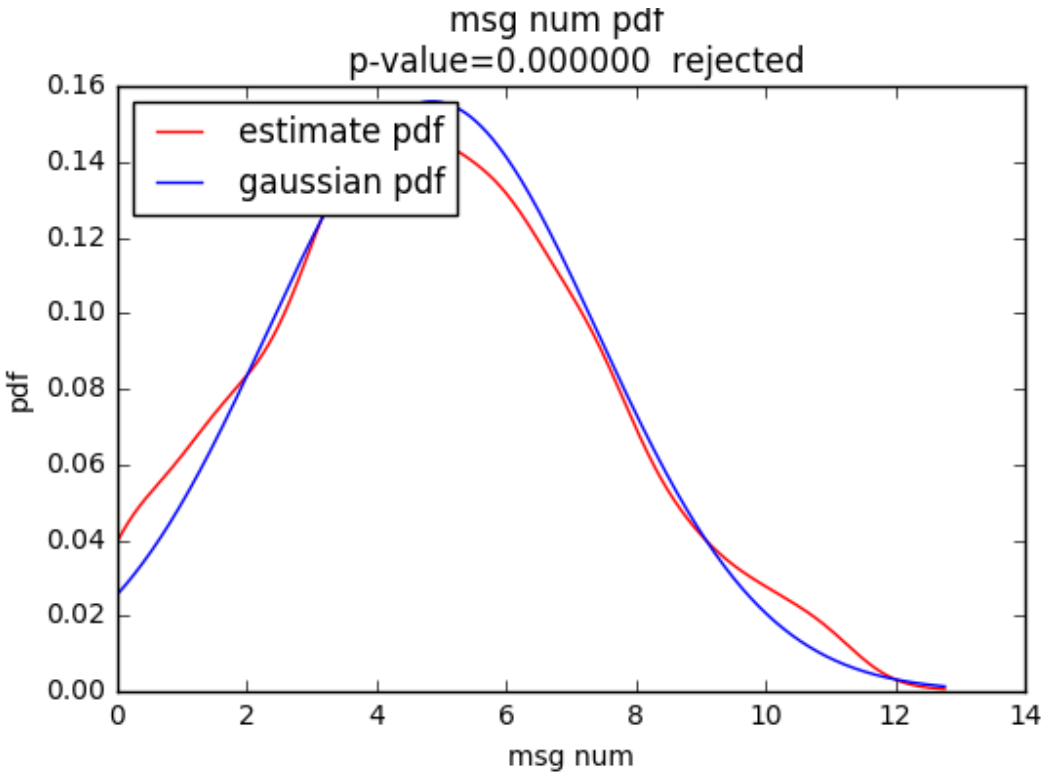
The second graph is a boxplot, which shows the data of different groups' mean, median number's position, etc. In conclusion, we can get the basic statistic information of the data.

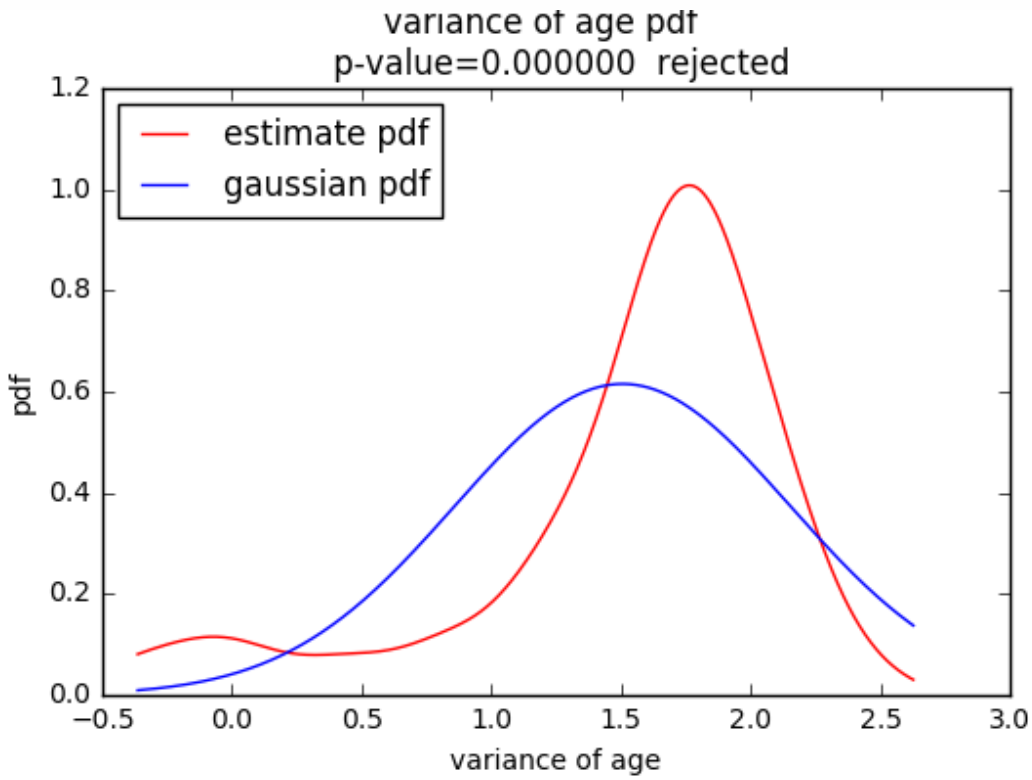
4. (**15 points**) Choose another 3 columns, draw the empirical pdf of each feature columns and test which column follows these assumptions in question 1? How about their corresponding log transformation?

   I choose **group size**, **number of message**, **variance of age**.

   First, I draw the empirical pdf of each feature columns in whole data.


group size pdf
p-value=0.000000 rejected
estimate pdf
gaussian pdf
pdf
group size
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
2
3
4
5
6
7
8


msg num pdf
p-value=0.000000 rejected
estimate pdf
gaussian pdf
pdf
msg num
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
0
2
4
6
8
10
12
14



From the pictures above, obviously, group size and variance don't follow the Gaussian distribution.

In the following normality test table, it displays the p-value and the result of whether the null hypothesis is rejected.

	P-value	**is rejected**
Group size	0	Rejected
number of message	0	Rejected
Variance of age	0.331845	not rejected

the log transformation can import the normality well.

we can conclude it from the following graphs.

next we need to test the assumptions

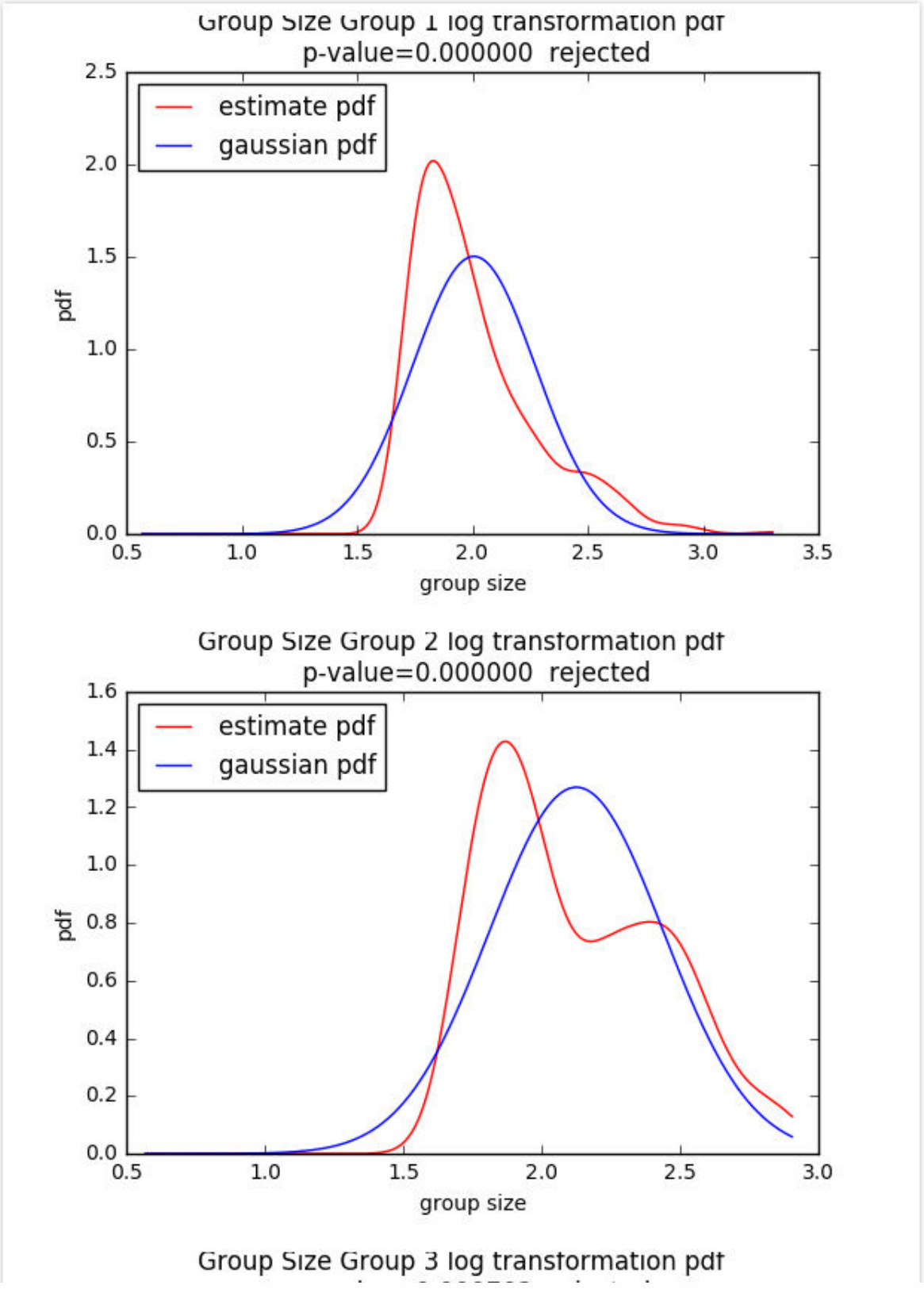
as for the second assumption:

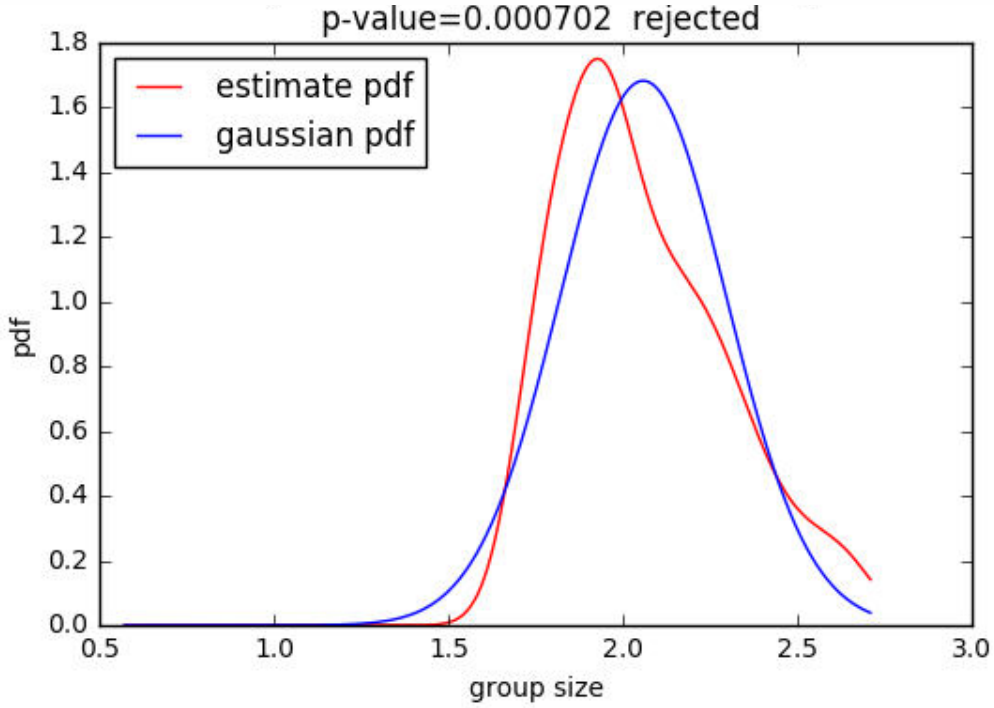
- The variances of each group are assumed as equal. Empirically, ratio of largest to smallest group standard deviation must be less than 2:1.
- The residuals are normally distributed (not skewed or partial)

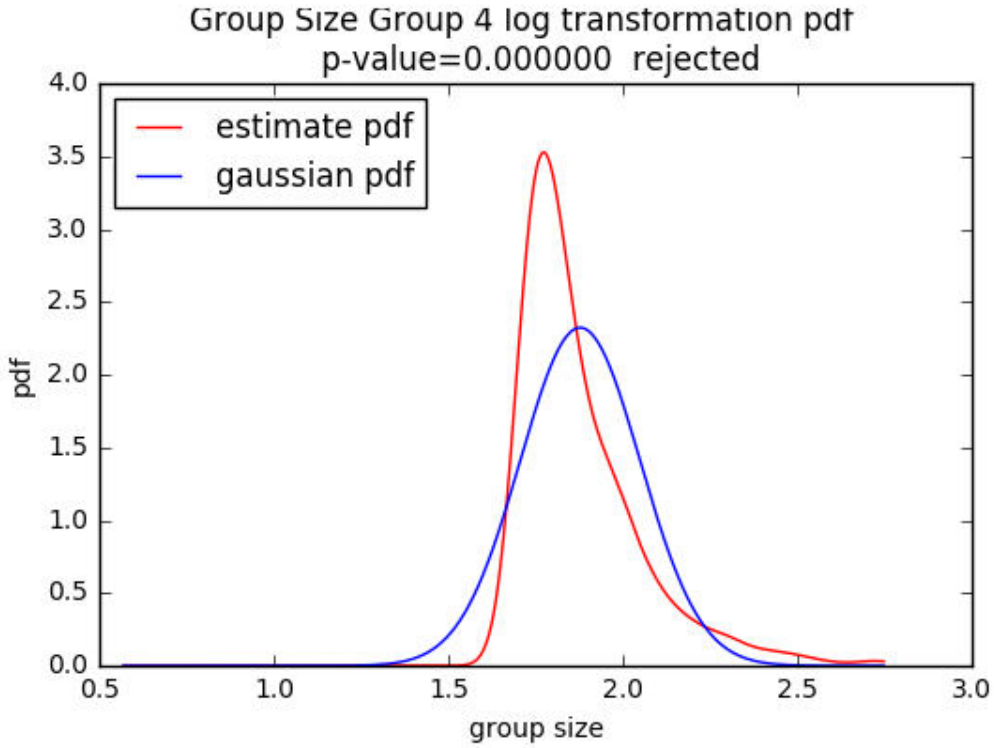
  we use **stats.skewtest** to get the degree of skew.

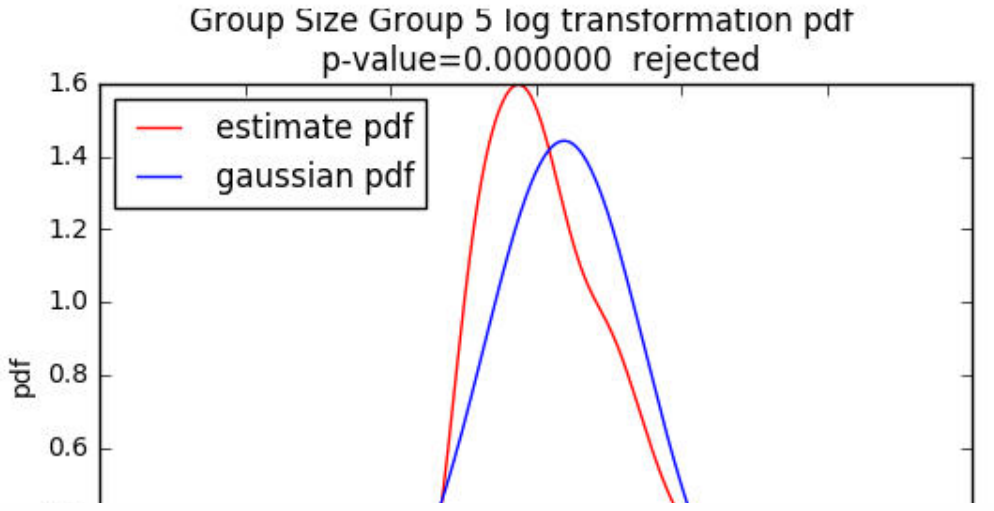
	Statistic	P-value
Group size	36.433162486099576	1.2712362807765213e-290
msg number	54.260089202222915	0.0
Variance of age	1.4750544033746911	0.14019791769383663

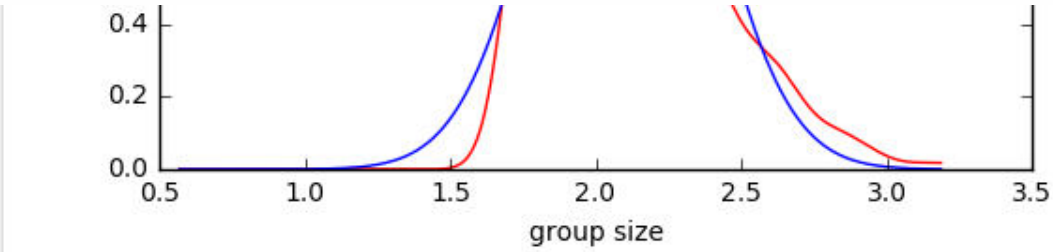
Group size log transformation



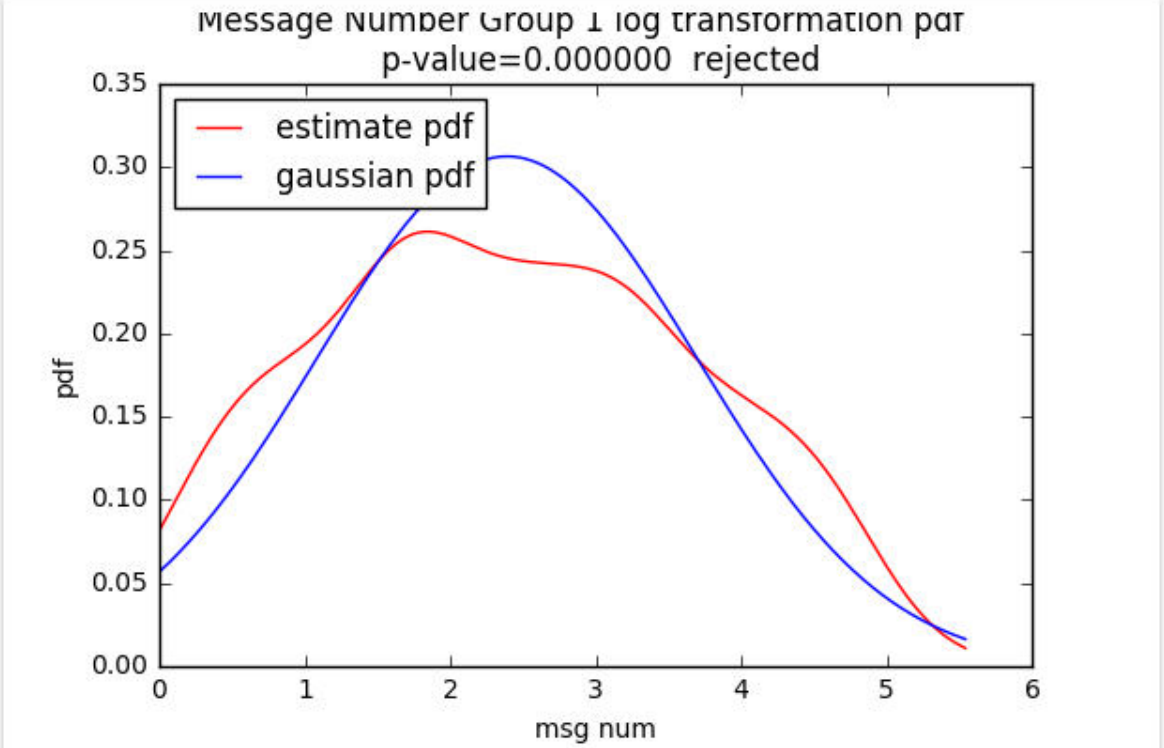

p-value=0.000702 rejected
estimate pdf
gaussian pdf
pdf
group size

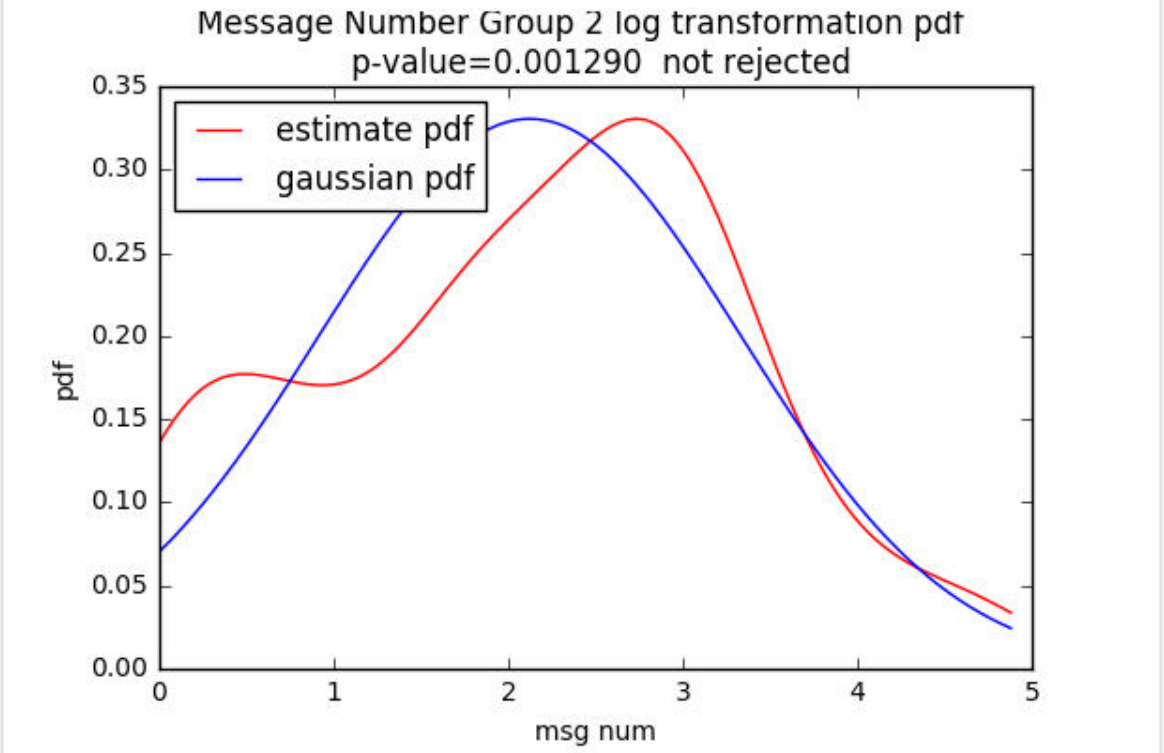

Group Size Group 4 log transformation pdf
p-value=0.000000 rejected
estimate pdf
gaussian pdf
pdf
group size

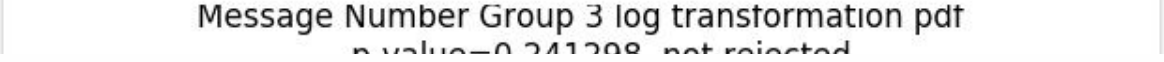

Group Size Group 5 log transformation pdf
p-value=0.000000 rejected
estimate pdf
gaussian pdf
pdf

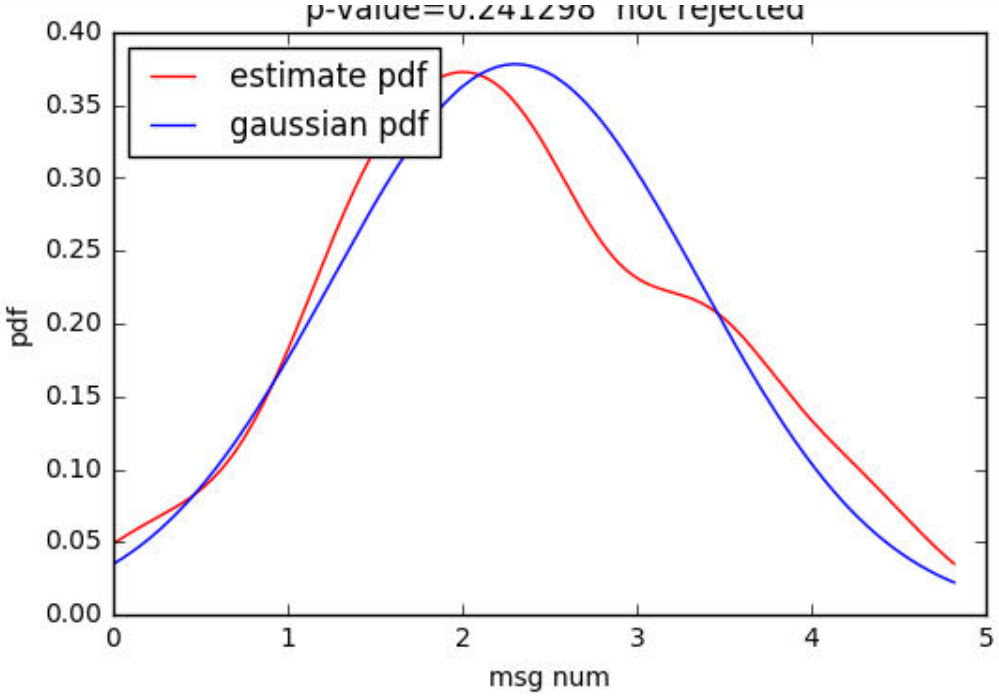


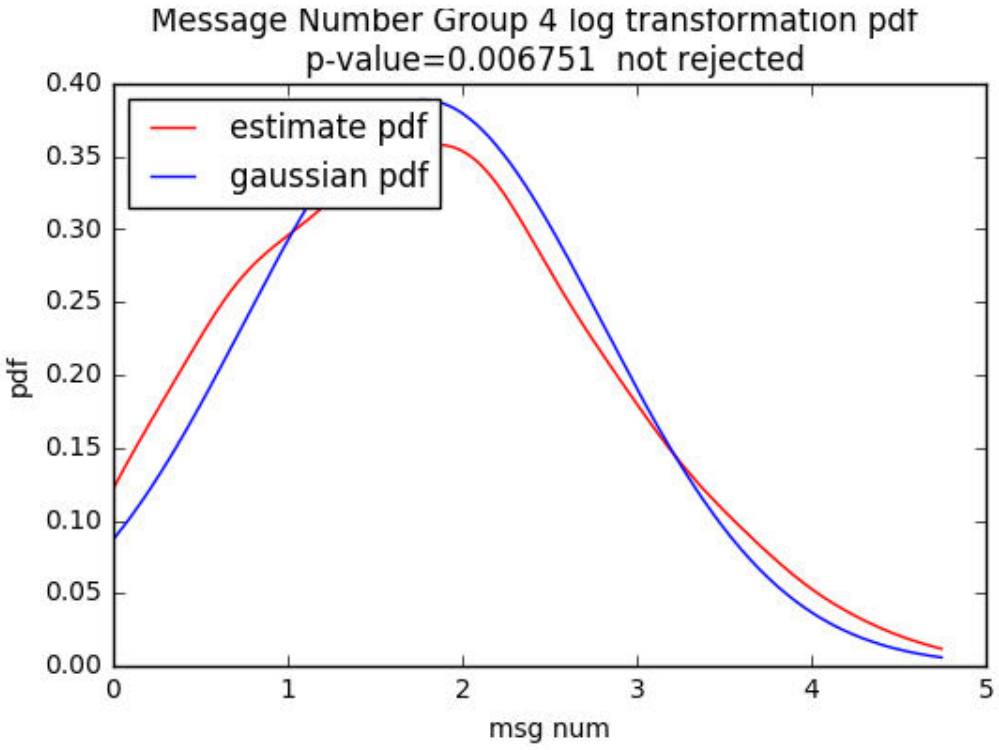
Message number log transformation

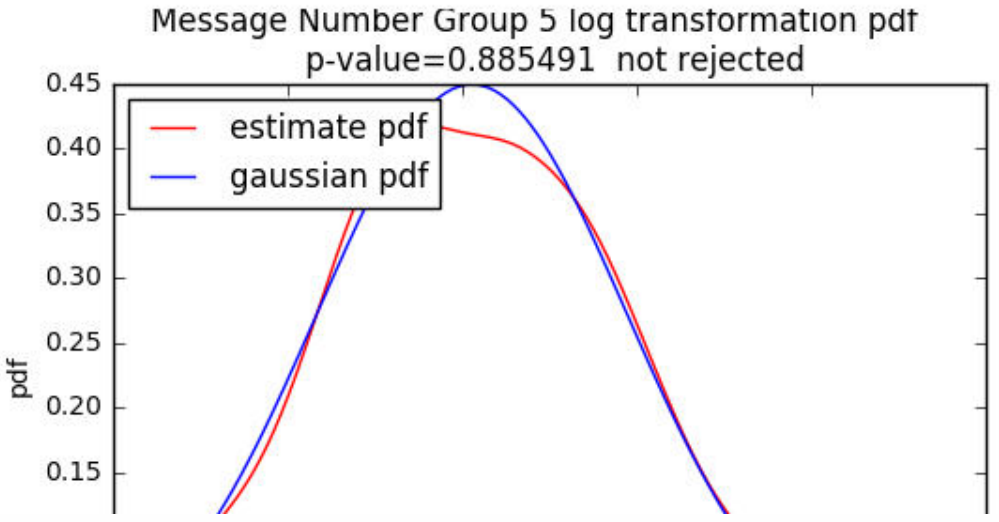


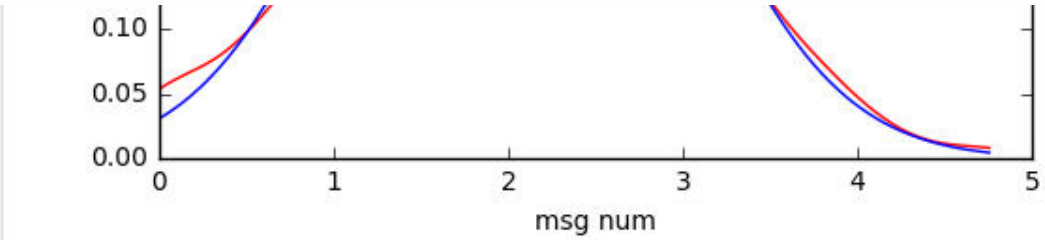




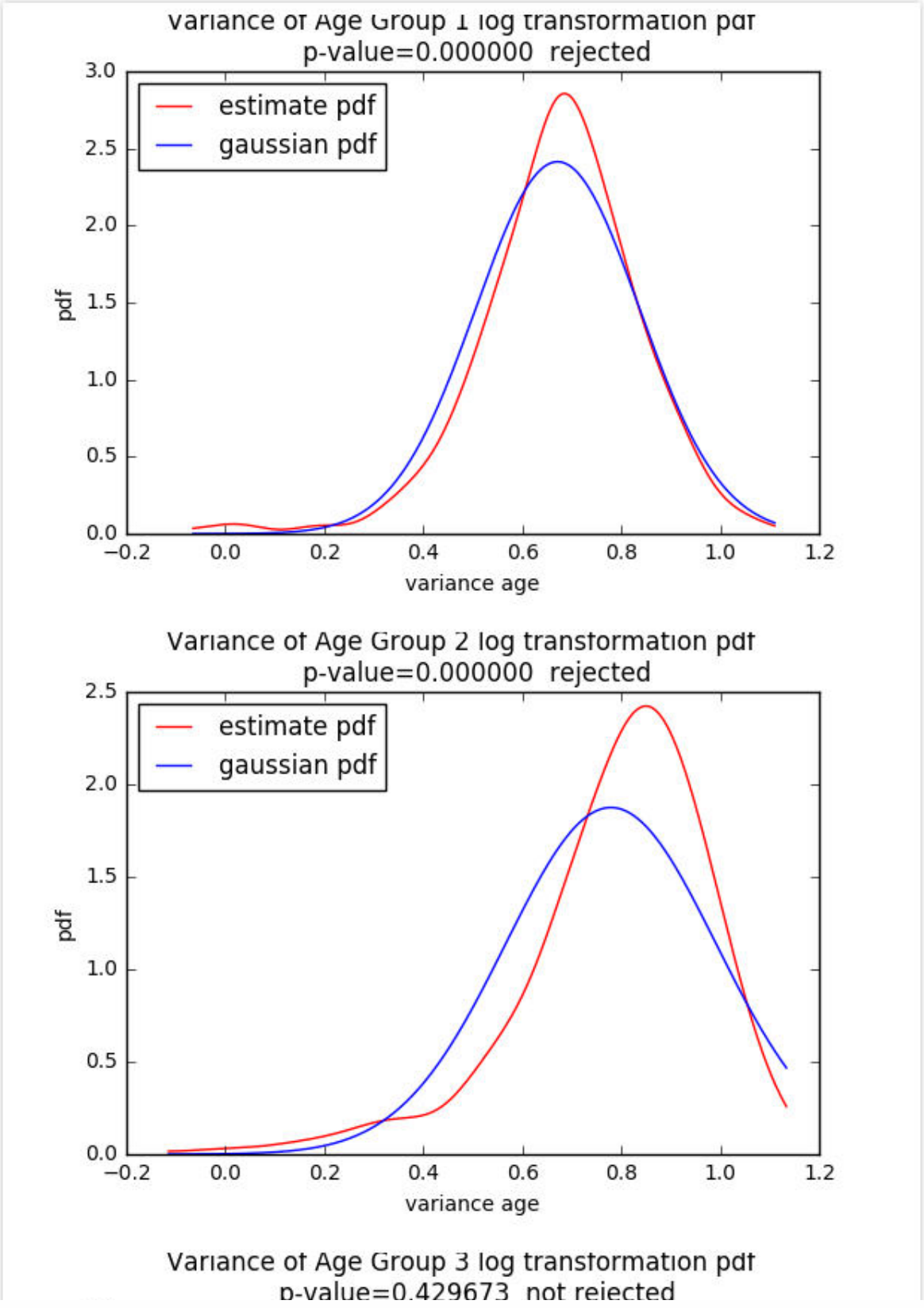

p-value=0.241298 not rejected
estimate pdf
gaussian pdf
pdf
msg num
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0
1
2
3
4
5

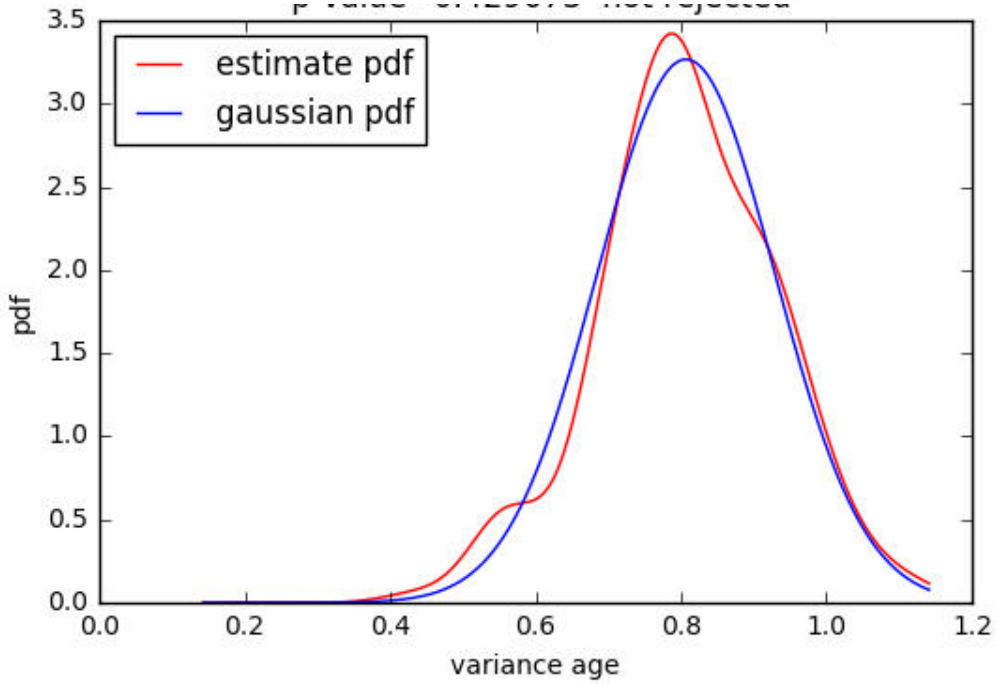

Message Number Group 4 log transformation pdf
p-value=0.006751 not rejected
estimate pdf
gaussian pdf
pdf
msg num
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0
1
2
3
4
5

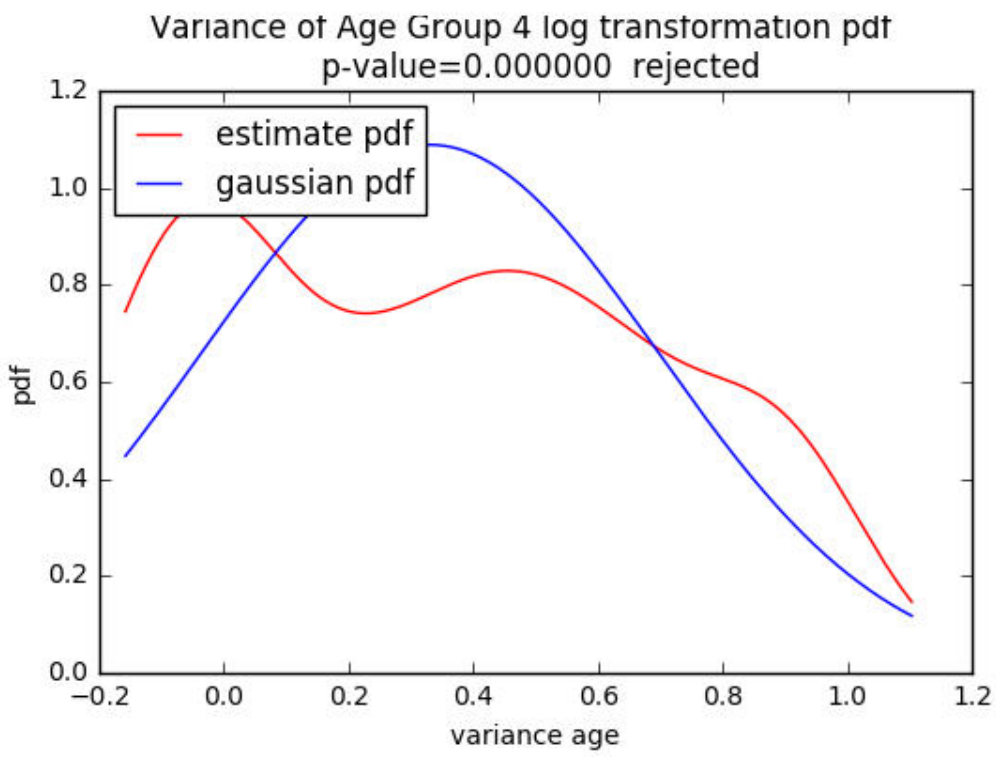

Message Number Group 5 log transformation pdf
p-value=0.885491 not rejected
estimate pdf
gaussian pdf
pdf
0.15
0.20
0.25
0.30
0.35
0.40
0.45

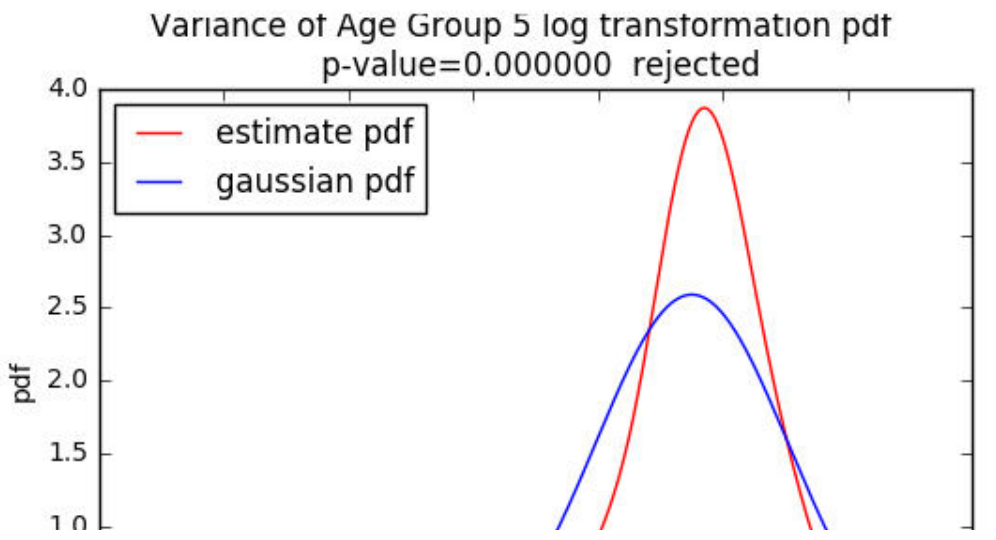


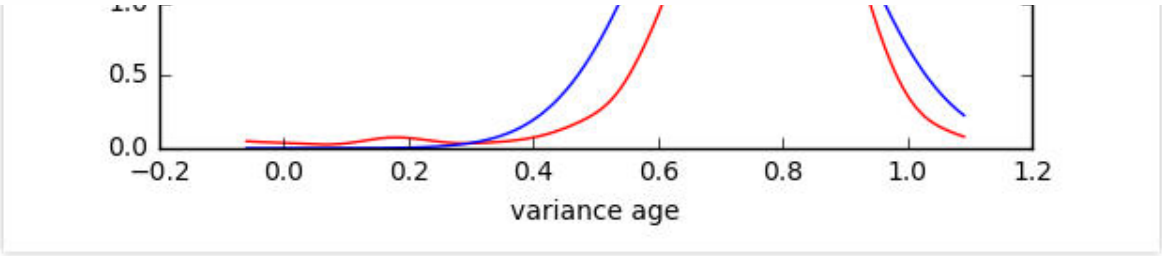
variance of age log transformation




estimate pdf
gaussian pdf
pdf
variance age


Variance of Age Group 4 log transformation pdf
p-value=0.000000 rejected
estimate pdf
gaussian pdf
pdf
variance age

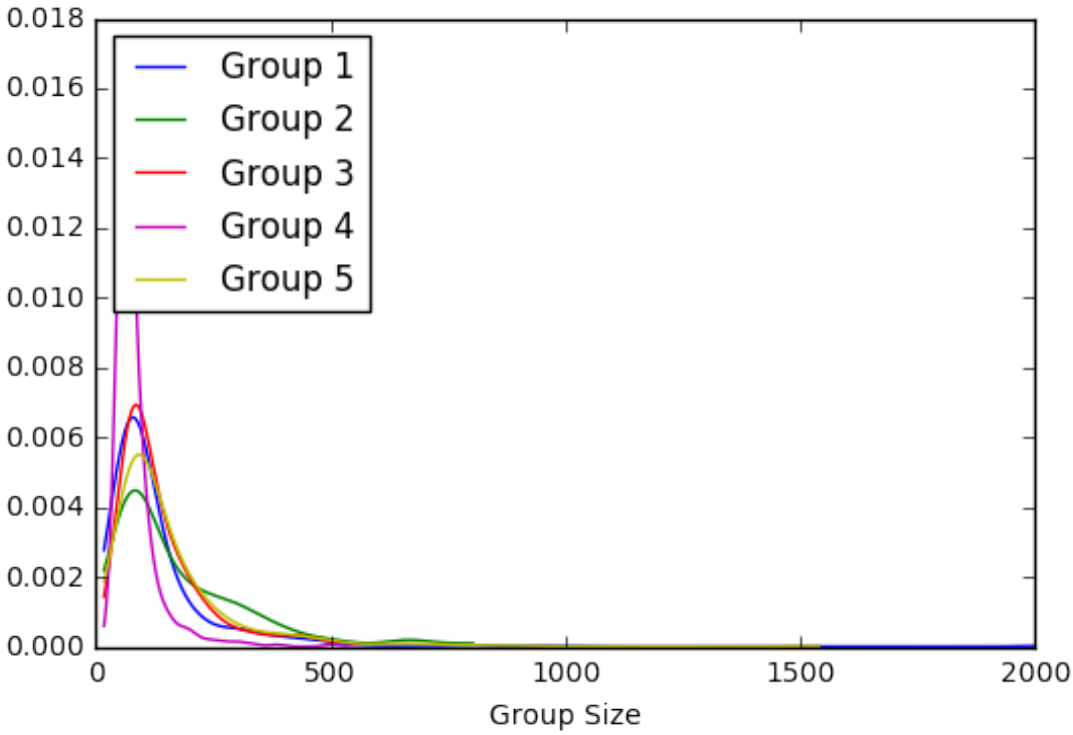

Variance of Age Group 5 log transformation pdf
p-value=0.000000 rejected
estimate pdf
gaussian pdf
pdf

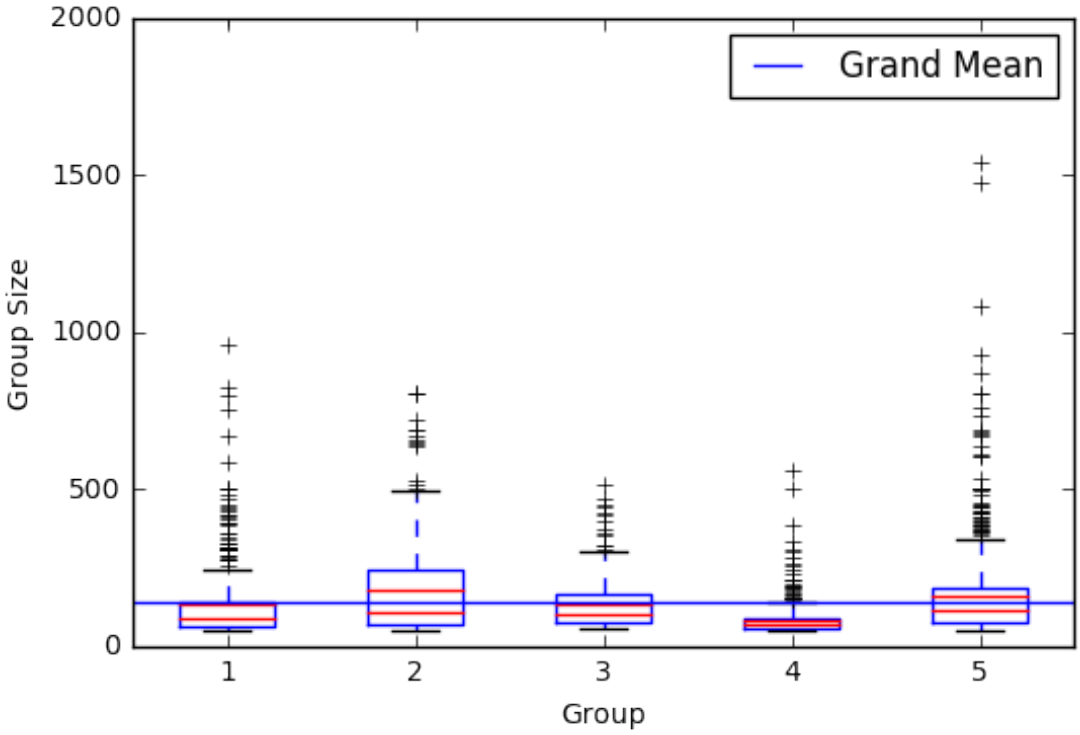


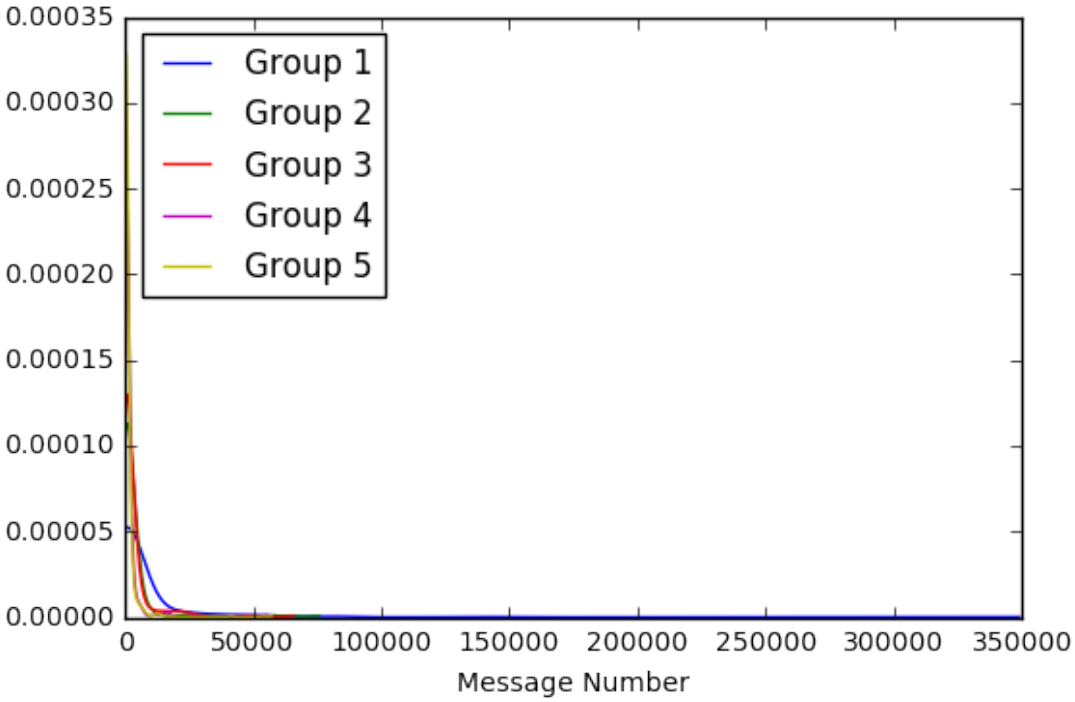
5. How to do one-way ANOVA with the non-normal data?
    1. (**10 points**) Find and list the **possible solutions set**.
        1. **ANOVA**: many data sets that are significantly non-normal would be perfectly appropriate for an anovaor other parametric test. Fortunately, an anova is not very sensitive to moderate deviations from normality; simulation studies, using a variety of non-normal distributions, have shown that the false positive rate is not affected very much by this violation of the assumption (Glass et al. 1972, Harwell et al. 1992, Lix et al. 1996).
            1. **Welch's anova**: If the data show a lot of heteroscedasticity (different groups have different standard deviations), the one-way anova can yield an inaccurate *P* value; the probability of a false positive may be much higher than 5%. In that case, you should use Welch's anova.
        2. **data transformations**: If your histogram looks like a normal distribution that has been pushed to one side, like the sulphate data above, you should try different data transformationsto see if any of them make the histogram look more normal.
        3. **non-parametric test**: parametric tests are not very sensitive to deviations from normality.every parametric statistical test has a non-parametric substitute.
            1. **Kruskal–Wallis test**: instead of a one-way anova
            2. **Wilcoxon signed-rank test**:instead of a paired *t*–test
            3. **Spearman rank correlation**: instead of linear regression/correlation
    2. (**15 points**) Do the one-way ANOVA on the 3 columns you choose. Do these feature columns vary significantly? Visualize the results.

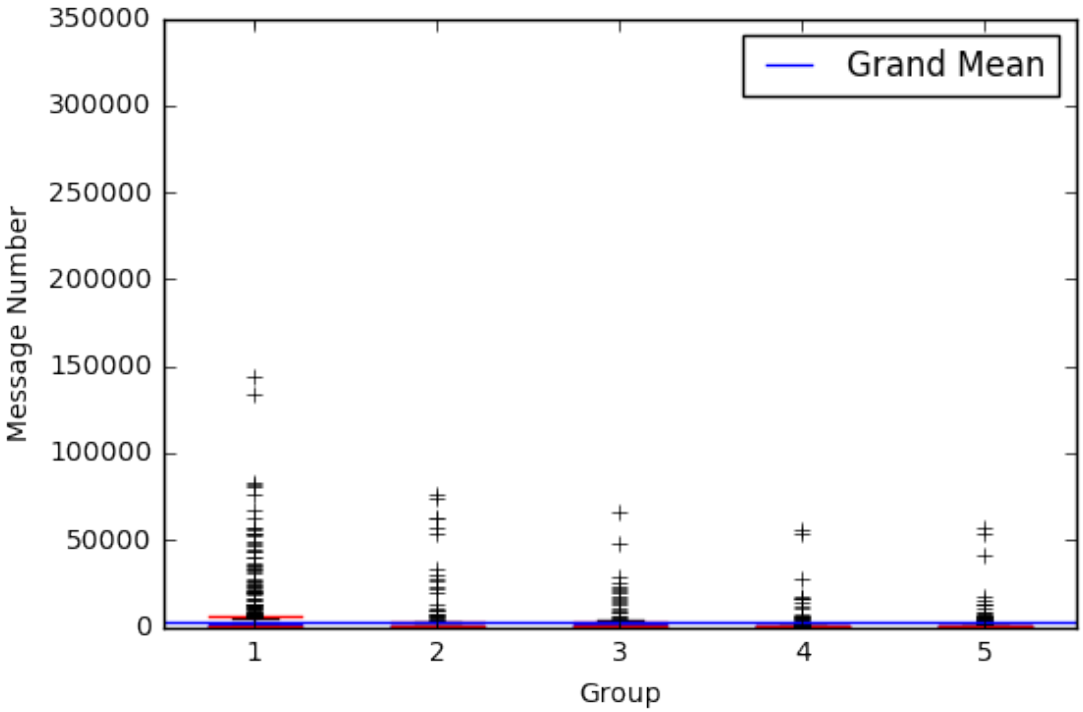
        the null (H0)： {group size|massage number|variance of age} in each category are equal.

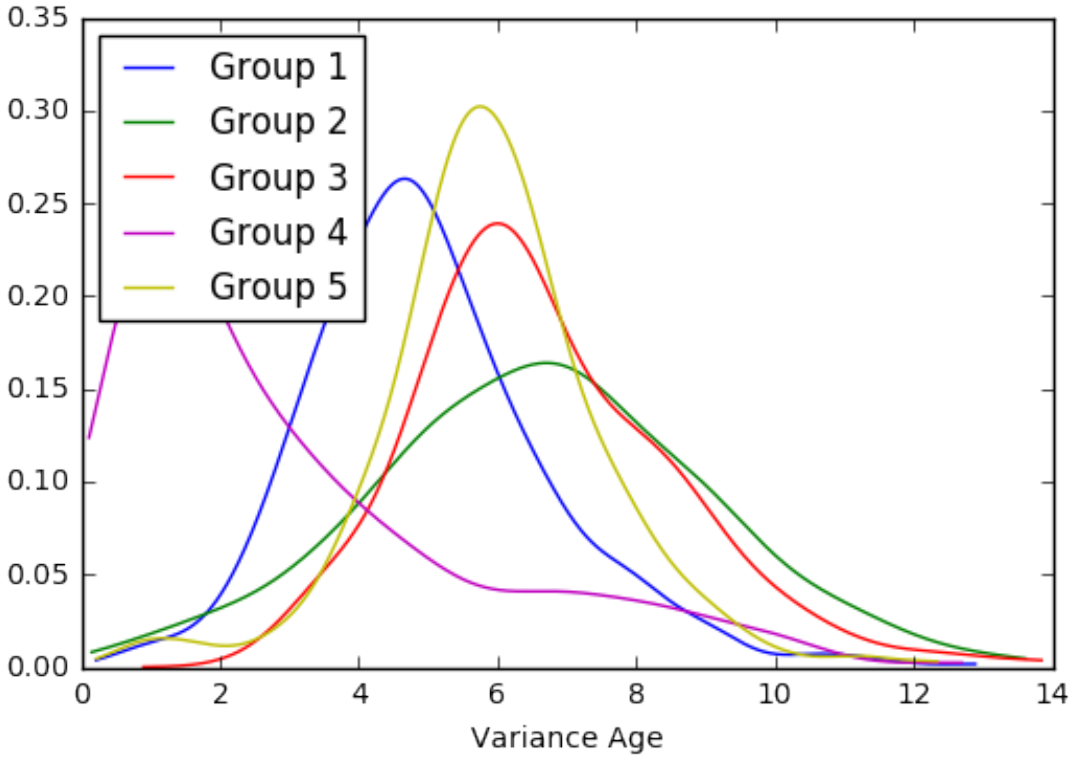
        the alternative (H1)： Not all {group size|massage number|variance of age} in different categories are equal

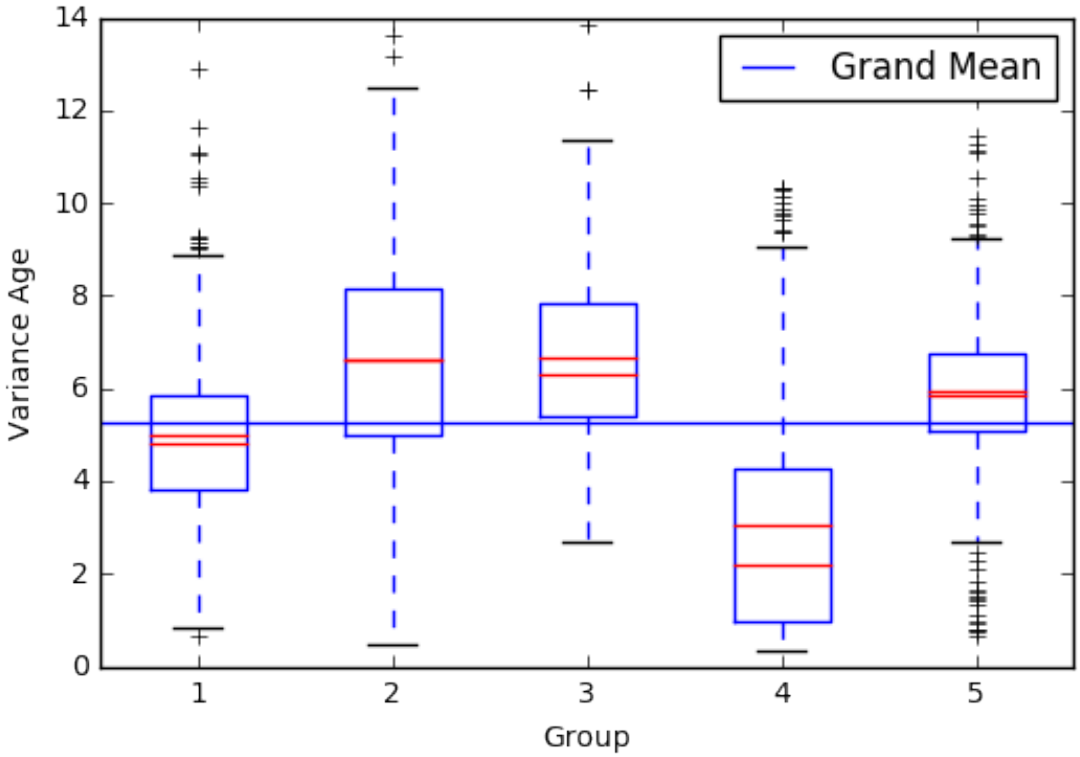

Group 1
Group 2
Group 3
Group 4
Group 5
0.018
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
0
500
1000
1500
2000
Group Size


Grand Mean
2000
1500
1000
500
0
Group Size
1
2
3
4
5
Group


0.00035
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
Group 1
Group 2
Group 3
Group 4
Group 5
0
50000
100000
150000
200000
250000
300000
350000
Message Number


350000
300000
250000
200000
150000
100000
50000
0
Message Number
Grand Mean
1
2
3
4
5
Group





6. (**10 points**) Choose any two categories, and classify them by logistical regression, or you can try multi-class classification on all categories.

I choose category 1 and 2. classify them by logistical regression.

I use **sklearn. LogisticRegression** to train the classifier.

```
precision:  0.778245835579 [ 0.80769231  0.75308642  0.78666667  0.76
0.78378378]
recall:  0.834585289515    [ 0.875       0.84722222  0.83098592  0.8028169
0.81690141]
cross_val_score 0.805285343361 [ 0.84    0.79738562  0.80821918  0.78082192
0.8       ]
```