PSTAT 105 HW 6

Matthew Xu (5752811)

26 Feburary 2021

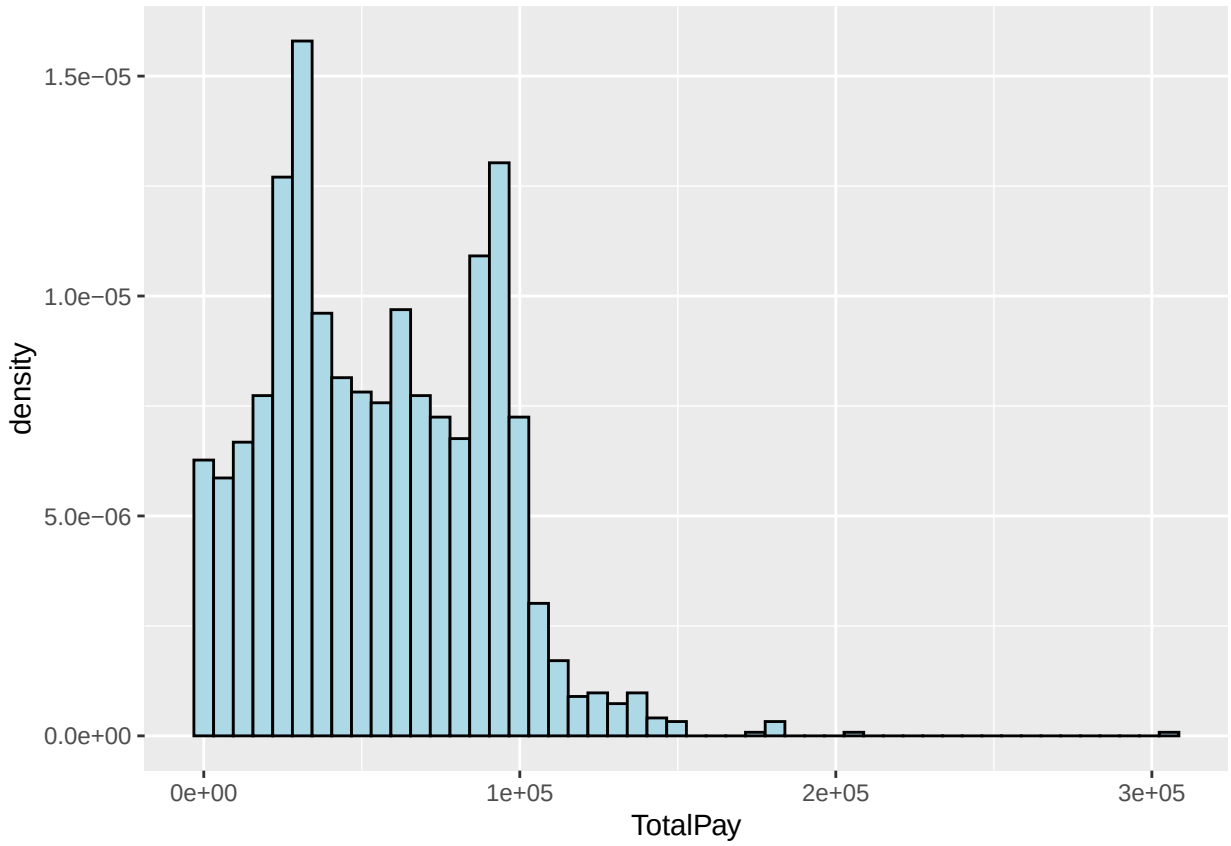
```
# libraries
library(tidyverse)
library(nortest)

# scan in SB unif salaries data
salaries <- read.table("SBUnifSalaries.txt", sep = " ", header = TRUE)
```

Please complete these questions and submit the answers on GauchoSpace.

In California, public employees' salaries are considered public knowledge and are published online. I downloaded all the salaries for people working for the Santa Barbara Unified School District and included the relevant salary information in the file SBUnifSalaries.txt. We are going to analyze the pay data in the column TotalPay.

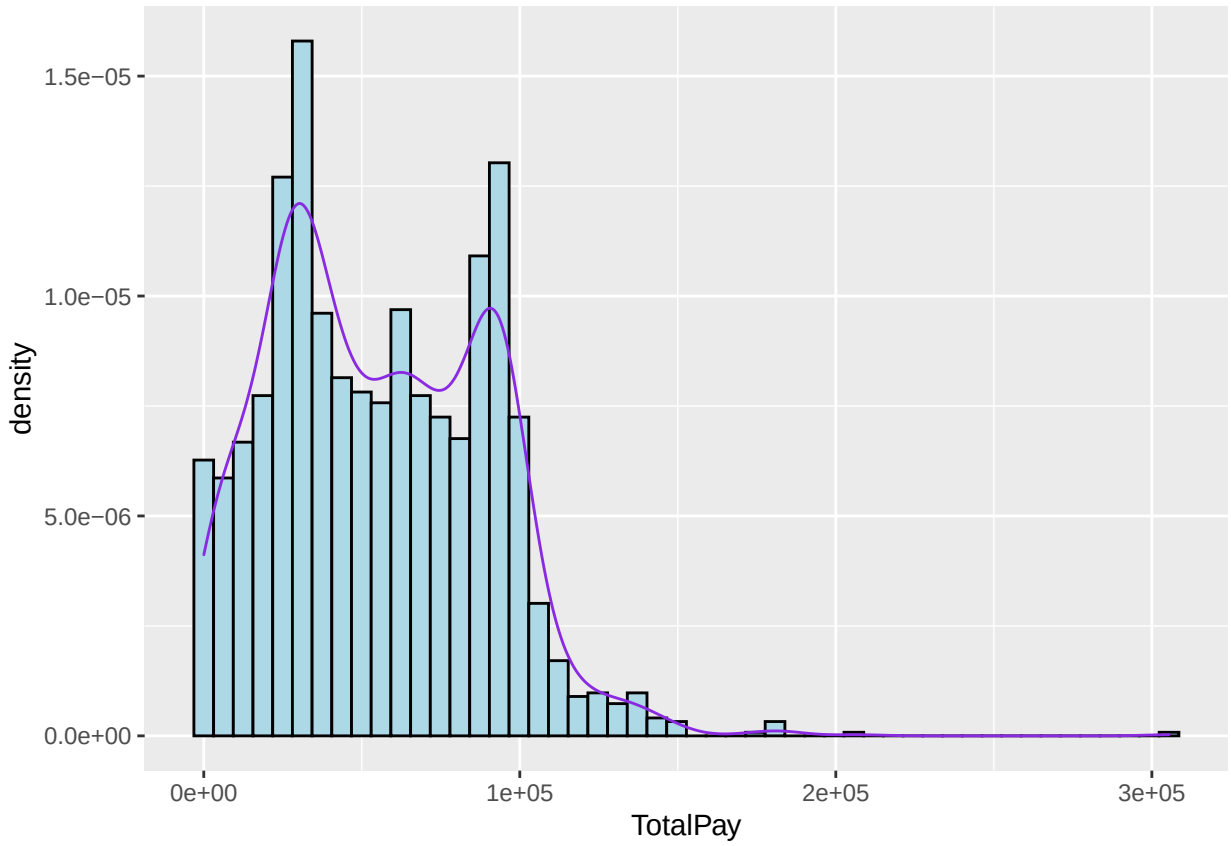
1. Create a histogram of this data with a reasonable number of bins that allows you to see the characteristics of the data.

```
# analysze column TotalPay
salaries_hist <- ggplot(salaries, aes(TotalPay)) + geom_histogram(bins = 50, aes(y = ..density
    fill = "lightblue", col = "black")

salaries_hist
```



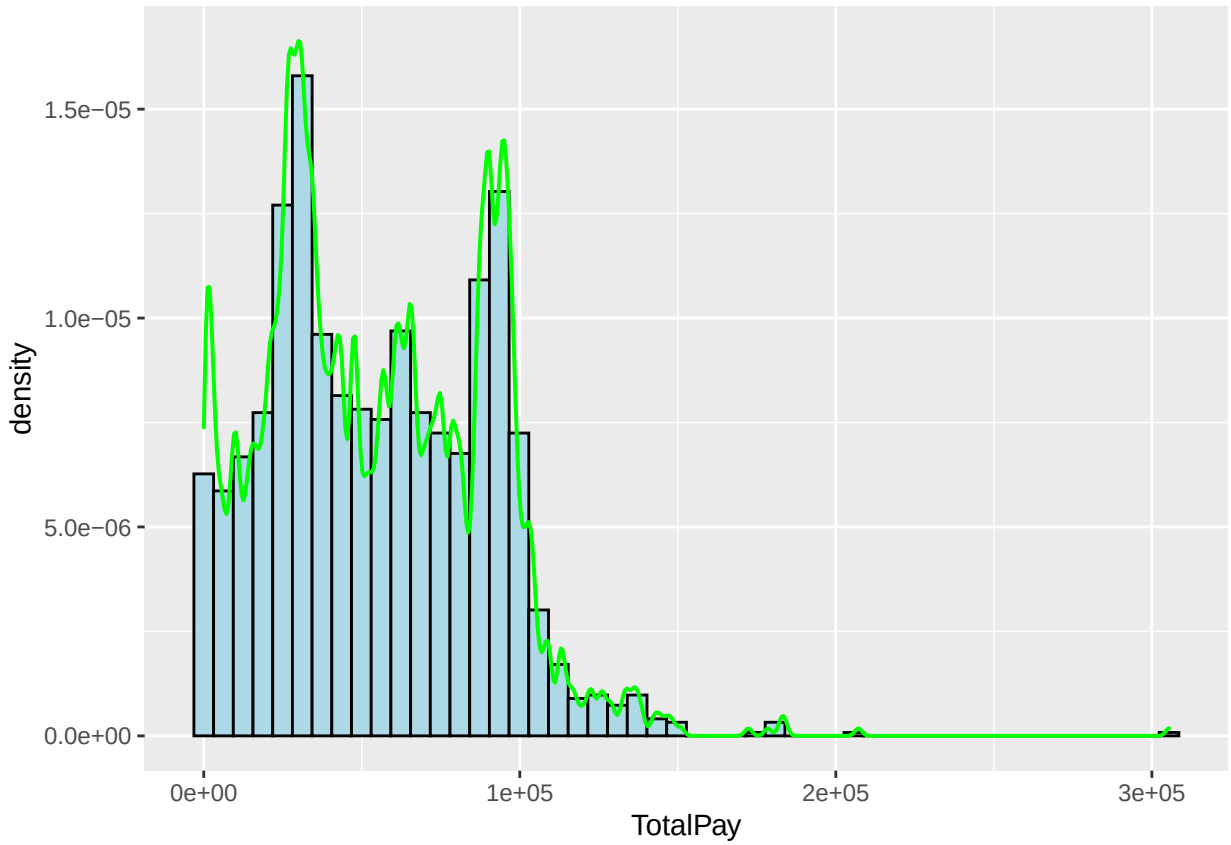
2. Draw a kernel density estimate using the normal reference density algorithm for choosing the bandwidth.

```
salaries_hist + geom_density(bw = "nrd", kernel = "gaussian", col = "blueviolet",
    size = 0.5)
```



3. Draw a kernel density estimate using the cross validation algorithm, "ucv." Which bandwidth do you think is better?

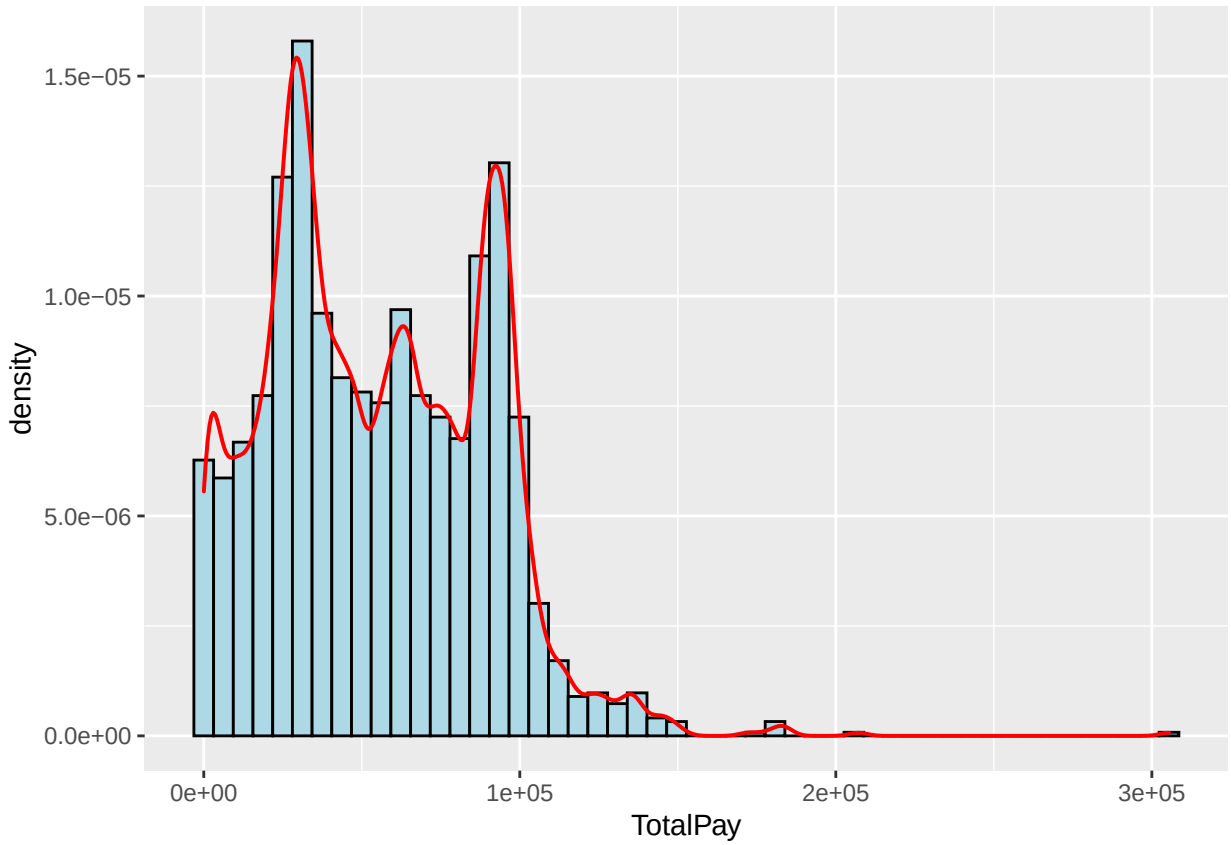
```
salaries_hist + geom_density(bw = "ucv", kernel = "gaussian", col = "green", size = 0.7)
```



The nrd bandwidth estimator seems the best so far, as it is smoother and has less peaks than uvc at certain peaks (uvc sometimes has 2 peaks at one point). Ucv also has a much larger to overestimate the distribution more often than the slight underestimation from nrd.

4. Sheather's paper recommends using a different algorithm which gets a more refined version of the second derivative to plug into the AIMSE. This is implemented via the bandwidth algorithm "SJ" in the density function. Use this option to find a bandwidth and comment on whether you agree that this method is working better for this data.

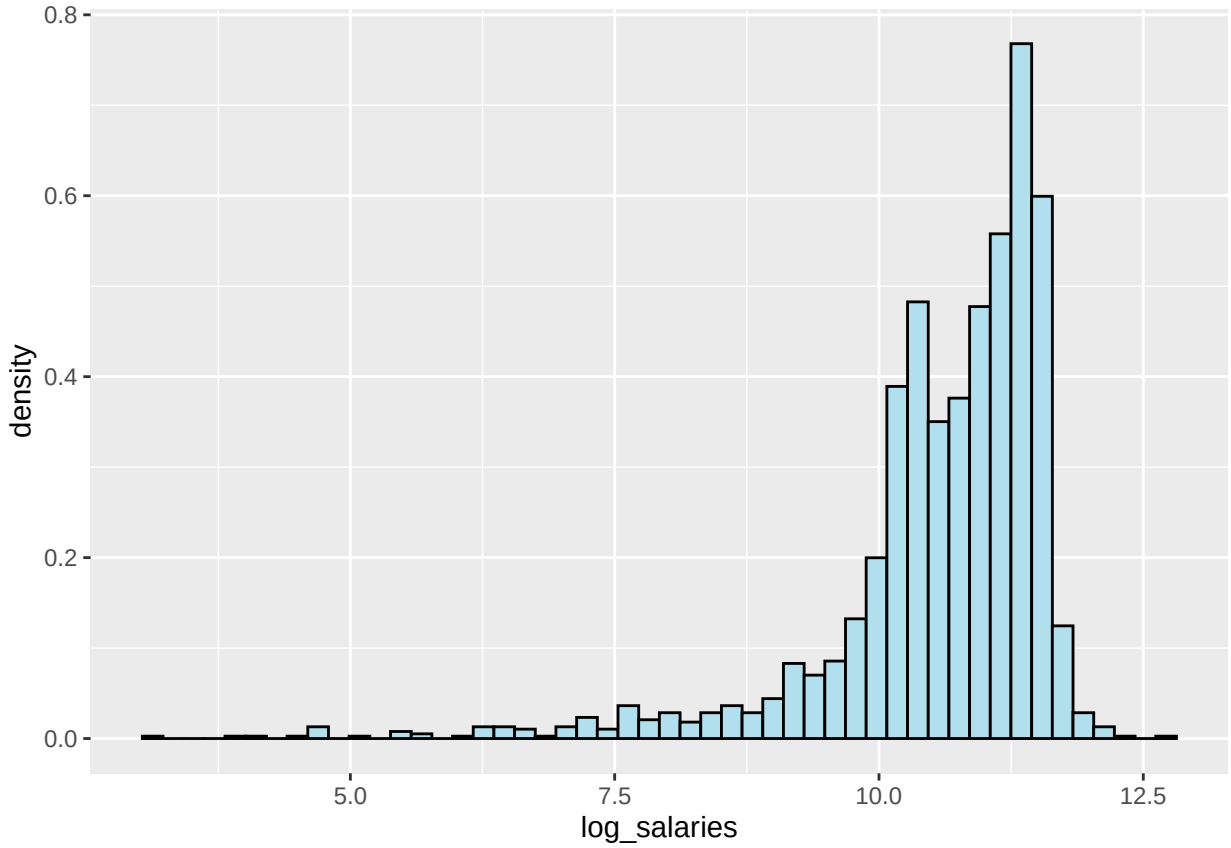
```
salaries_hist + geom_density(bw = "SJ", kernel = "gaussian", col = "red", size = 0.7)
```



This bandwidth algorithm SJ is the best one used so far, agreeing with the statement, as it lies closest to the distrubution depicted by the histogram. Each peak lies closely to the corresponding peak of the histogram, and the lows do as well, making SJ very accuracte. It is also quite smooth, with no ridges and edges at the peak areas.

5. Salary data like this data set is often skewed towards higher values. One way to make the data look more symmetric is to apply a log transform. Draw a histogram of the logs of the salaries.

```
# taking log for transformation
log_salaries = log(salaries$TotalPay)

log_hist = ggplot(data.frame(log_salaries), aes(log_salaries)) + geom_histogram(bins = 50,
    aes(y = ..density..), fill = "lightblue2", col = "black")

log_hist
```



6. Perform a Lilliefors test on the log-transformed data. Is it reasonable to assume that this data is normally distributed?

```
lillie.test(log_salaries)
```

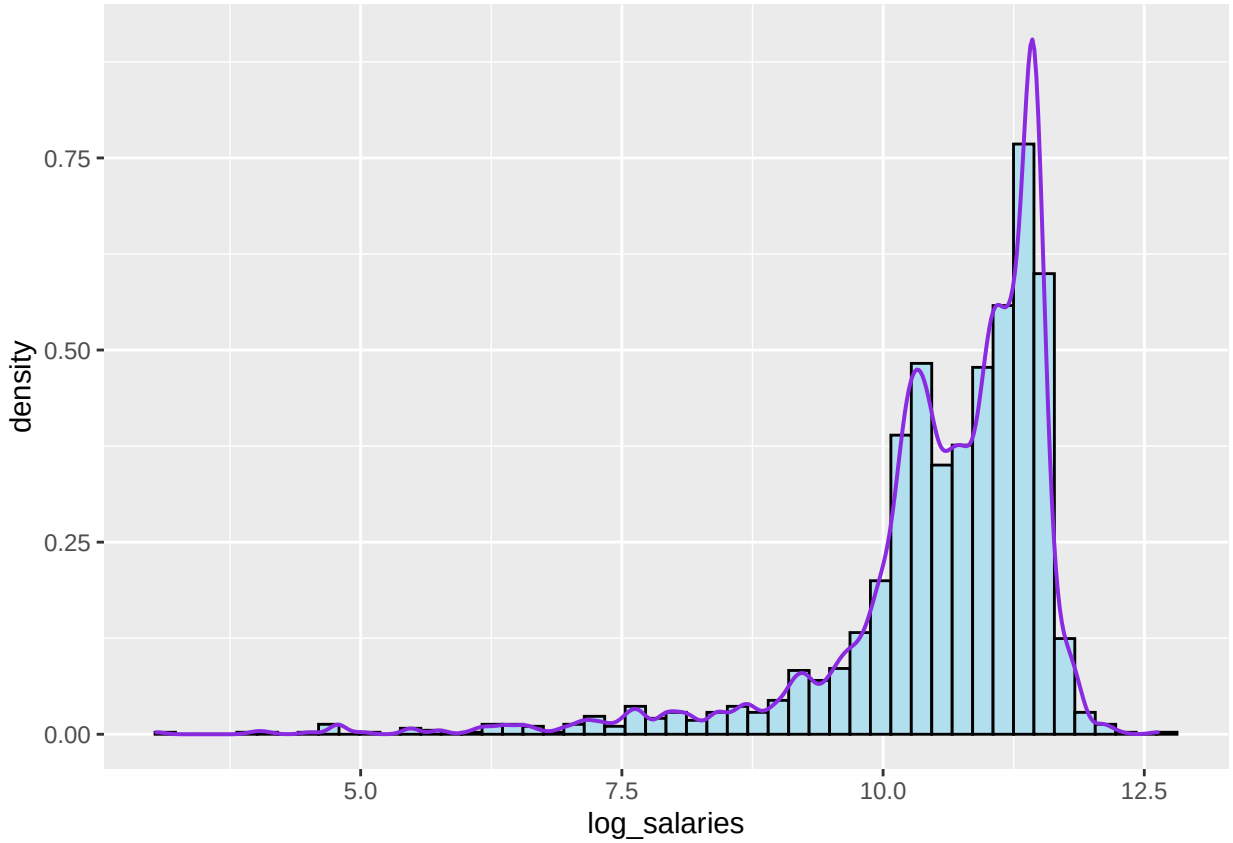
```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  log_salaries
## D = 0.13609, p-value < 2.2e-16
```

H0: The log of salaries data is normally distributed Ha: The log of the salaries data is not normally distributed

Because the p value of $<$ 2.2e-16 is less than the significance level of 0.05, we reject the null hypothesis. There is sufficient evidence to state that the log of the salaries data is not normally distributed and that it is stastically significant.

7. Draw an appropriate density estimate on your histogram of the transformed data, and comment on ways in which it appears non-Gaussian.

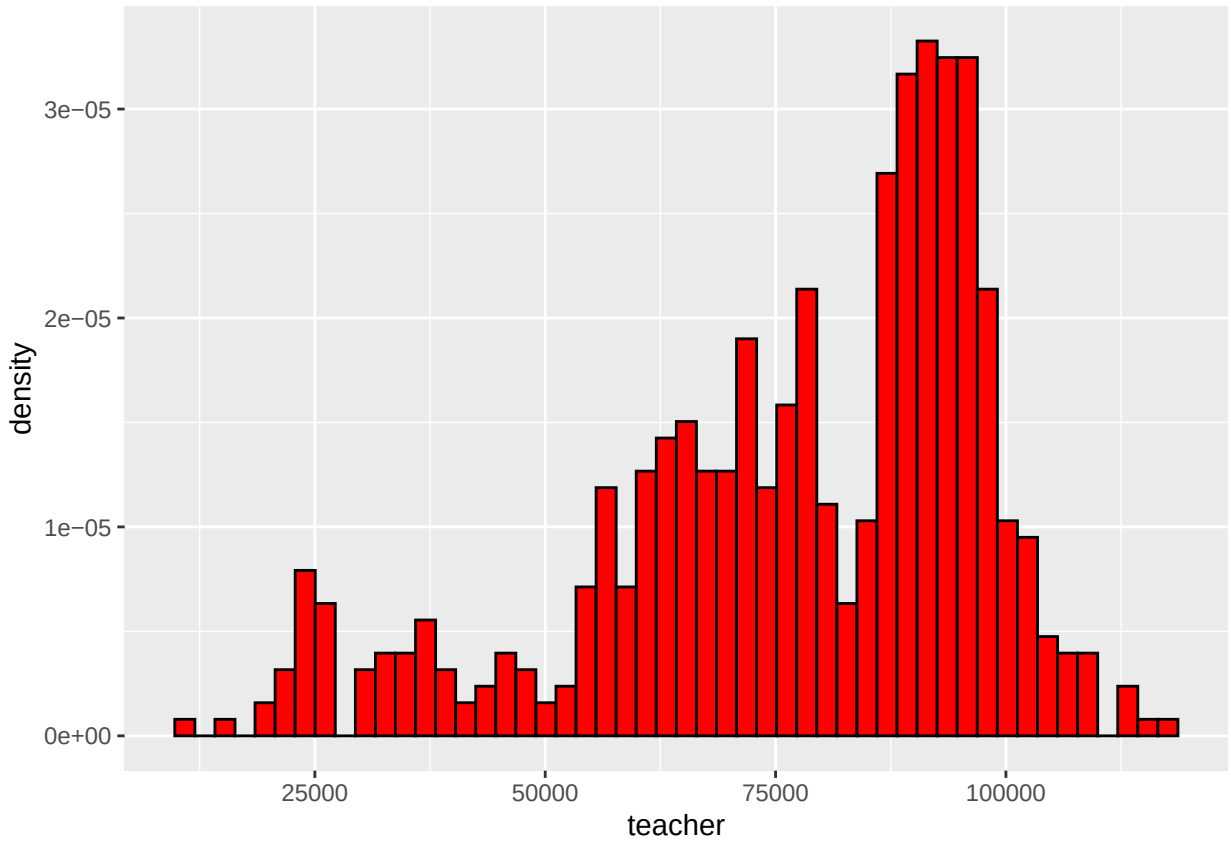
```
log_hist + geom_density(bw = 0.08, kernel = "gaussian", col = "blueviolet", size = 0.7)
```



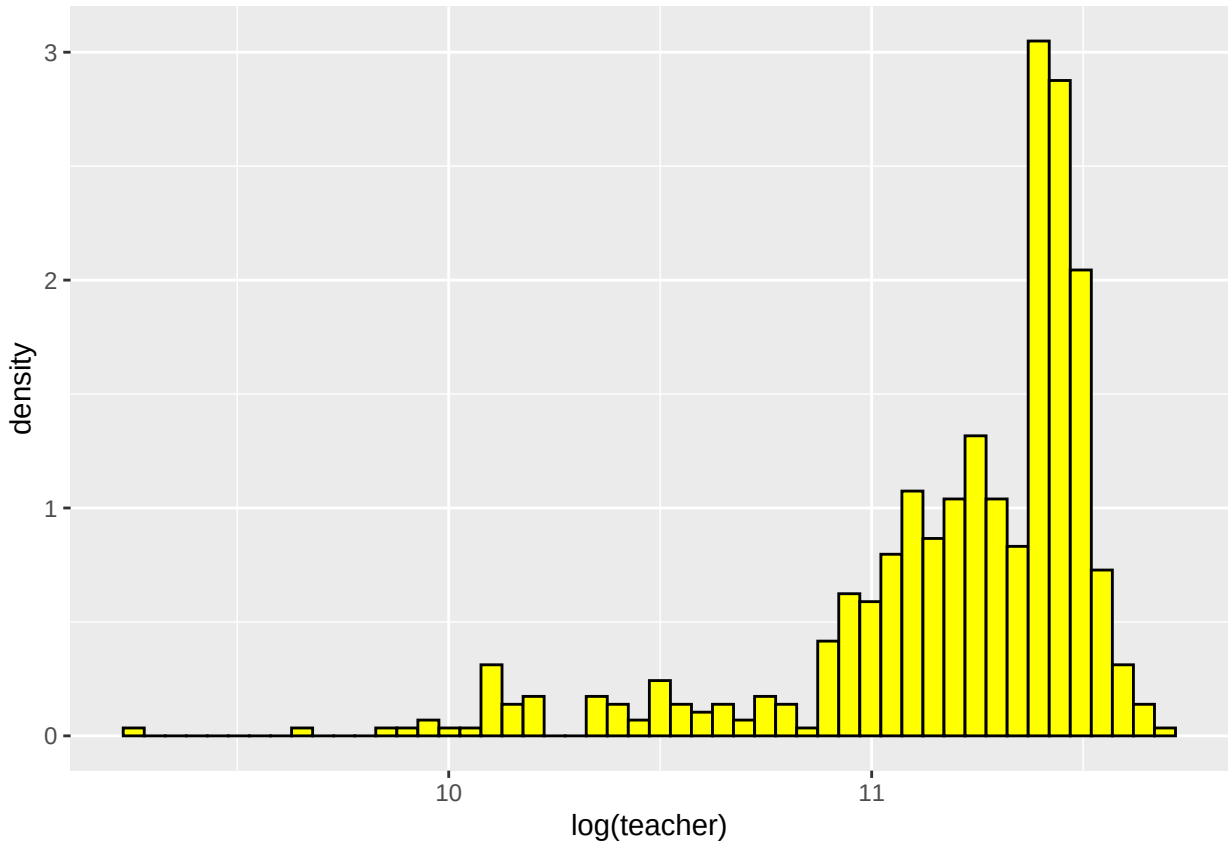
The plot depicted above is non-gaussian as it only contains non negative values, and appears to be more than one peak, which is unlikely. The distribution is also skewed, not symmetric, making it non gaussian.

8. One issue with the data is that it includes employees with many different sorts of jobs from Substitute Teachers to the Superintendent. Separate out just the folks that are Teachers and plot a histogram of their salaries. Would a log transform be appropriate to make the distribution of this data look more symmetric?

```
# extract totalpay of the salaires with job as teacher
teacher = salaries[which(salaries$JobTitle == "TEACHER"), ]$TotalPay

teacher_hist = ggplot(data.frame(teacher), aes(teacher)) + geom_histogram(bins = 50,
    aes(y = ..density..), fill = "red", col = "black")

teacher_hist
```



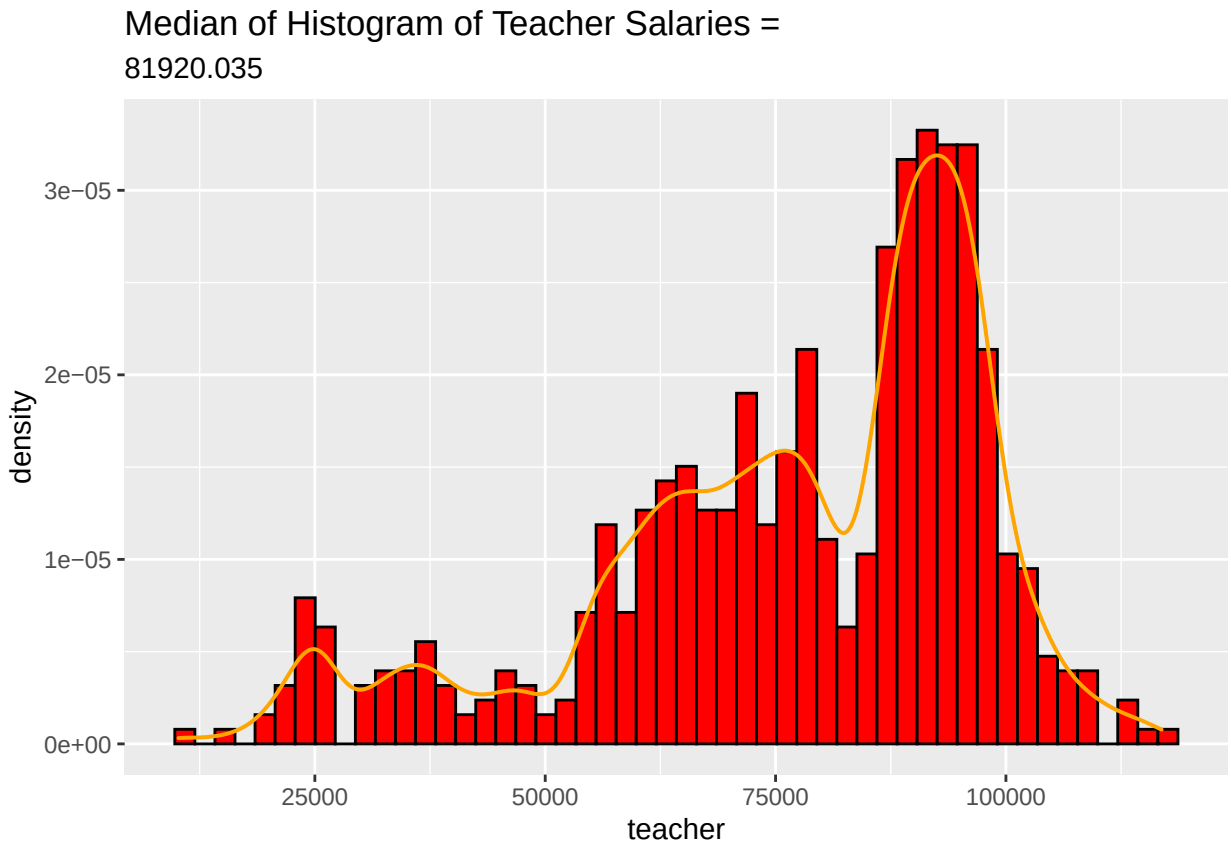

```
# log transformation of teacher salaires
teacher_hist_log = ggplot(data.frame(teacher), aes(log(teacher))) + geom_histogram(bins = 50,
    aes(y = ..density..), fill = "yellow", col = "black")

teacher_hist_log
```



Giving a log transformation to teacher salaries does not makes the distribution more symmetric at all, and does the opposite. The distribution becomes more skewed, with the tail on the left side becoming larger.

9. Using the "SJ" method to choose a bandwidth, draw an appropriate density estimate on this histogram. Also indicate the median teacher salary on the plot.

```
teacher_hist + geom_density(bw = "SJ", kernel = "gaussian", col = "orange", size = 0.7) +
    # indicate median of salaries on plot
ggtitle(paste("Median of Histogram of Teacher Salaries = "), median(teacher))
```



10. The California Department of Education website reports that the median salary for teachers in California is $83,059. Perform a nonparametric test of whether or not the median salary for teachers in Santa Barbara Unified School District is the same as for the whole state. What do you conclude?

```
wilcox.test(teacher, mu = 83059)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  teacher
## V = 66318, p-value = 8.996e-06
## alternative hypothesis: true location is not equal to 83059
```

```
n <- length(teacher)
less_teacher <- teacher[which(teacher < 83059)]
size_less <- length(less_teacher)
theta_hat <- size_less/n

z_score <- 2 * sqrt(n) * (0.5 - theta_hat)
p_value <- 2 * pnorm(z_score)
p_value
```

```
## [1] 0.8680854
```

H0: median salary for teachers in Santa Barbara Unified School District is the same as for the whole state Ha: median salary for teachers in Santa Barbara Unified School District is not the same as for the whole state

Using wilcox test of the median, because the pvalue of 8.996e-06 is less then alpha of 0.05, we reject the null hypothesis. There is suffieicent evidence to state that the median salary for teachers in Santa Barbara Unified School District is not the same as for the whole state and that it is stastically significant.

Using the test for median in lecture 2, the p value of 0.8680854 is greater than the alpha of 0.05, therefore we do not reject the nul hypothesis. There is not suffieicent evidence to state that the median salary for teachers in Santa Barbara Unified School District is not the same as for the whole state and that it is not stastically significant. It is possible to conlclude to accept the null hyptothesis, of which the teacher salary is the same across the whole state.