As part of term project, I wanted to analyze the US educational data set available at the below link. This data set contains data per financial school districts of each state and for every year from 1993 till 2016

Source Link: <https://www.kaggle.com/noriuk/us-education-datasets-unification-project/version/4>

Data set: finance\_districts.csv

Variables in the data set:

The data set consists of financials of each school district in each state for different years. It has the following variables:

STATE - State of Financial School district

ENROLL - The U.S. Census Bureau's count for students in the state. Should be comparable to GRADES\_ALL

NAME - Name of the school district

YRDATA - Year that the record pertains to

TOTALREV: The total amount of revenue for the state.

TFEDREV - Federal Revenue

TSTREV - State Revenue

TLOCREV - Local Revenue

TOTALEXP: The total expenditure for the state.

TCURINST - Instruction Expenditure

TCURSSVC - Supportive Services Expenditure

TCURONON - Other Expenditure

TCAPOUT - Capital Outlay Expenditure

**Statistical/Hypothetical Question:**

By exploring this data set regarding financial school districts and their enrollment numbers, I want to find out whether there is a any statistical correlation between total revenues of the state and number of enrollment numbers for the school districts. If there is correlation, how much have an impact Total Revenues have on enrollment.

**Outcome of your EDA:**

The assumption that I had before exploring this data set was that the school districts that are in higher revenue states will have more chance of higher enrollments in the school. After performing EDA,

I did find statistical correlation between Total Revenues and Enrollment of the school districts. So my assumption was correct. One more observation that I made is not all states will respond similarly to the total revenue numbers. For example, take state like LOUISIANA, even though Total Revenues increase for this state, Enrollment numbers will not raise proportionally when compared to other states.

**What was missing during the analysis**

Having demographic information of each school districts like population, family size, number of school going kids, family income etc. would have added more sense to the analysis. Also, one of these or couple of these could be the confounding factors that I highlighted above with STATE – LOUISIANA.

**Variables that could have helped in the analysis**

As states above, demographic information could have helped more in the analysis in finding the actual enrollment prediction for the school districts.

**Assumptions made correct or incorrect**

No. The assumptions that I made that enrollments in school districts are based on Total Revenues of the state was correct, backed by the higher correlation factor and the linear models

**Challenges faced**

In the selected data, there were some other aggregated data sets available, which I wanted to explore and compare with the financial school district data set that I selected. But some of the variables that I wanted to explore like GRADES\_ALL\_G etc. are not available through those sheets. I feel enrollment numbers depend a lot on population or number of families living in that school district. Having a demographic data for these school district would have been an interesting analysis, I would like to try. But that information was not available readily, so I couldn’t venture into that analysis. One more challenge is that some of the states like California, New York are too big to be compared with other smaller states like Vermont, but before realizing this when I performed EDA overall financial school district data set, numbers were highly skewed – and didn’t find an ideal number of bins to represent these variables into proper histogram. Plotting PDF and CDF without using the thinkstats2, thinkplot modules has been a challenge, at some places I gave up trying to figure out other means and ended up using these modules. I would like to explore these in free time.

# Modules used:

To do this analysis, I have used the following libraries:

Pandas – to import, hold and transform the data as required

NumPy – for function like sorting, random number generation and permutation of array etc.

SciPy – for using the inbuilt stats modules for calculating descriptive statistics like – Mean, Median, Std, Var, PDF, CDF etc.

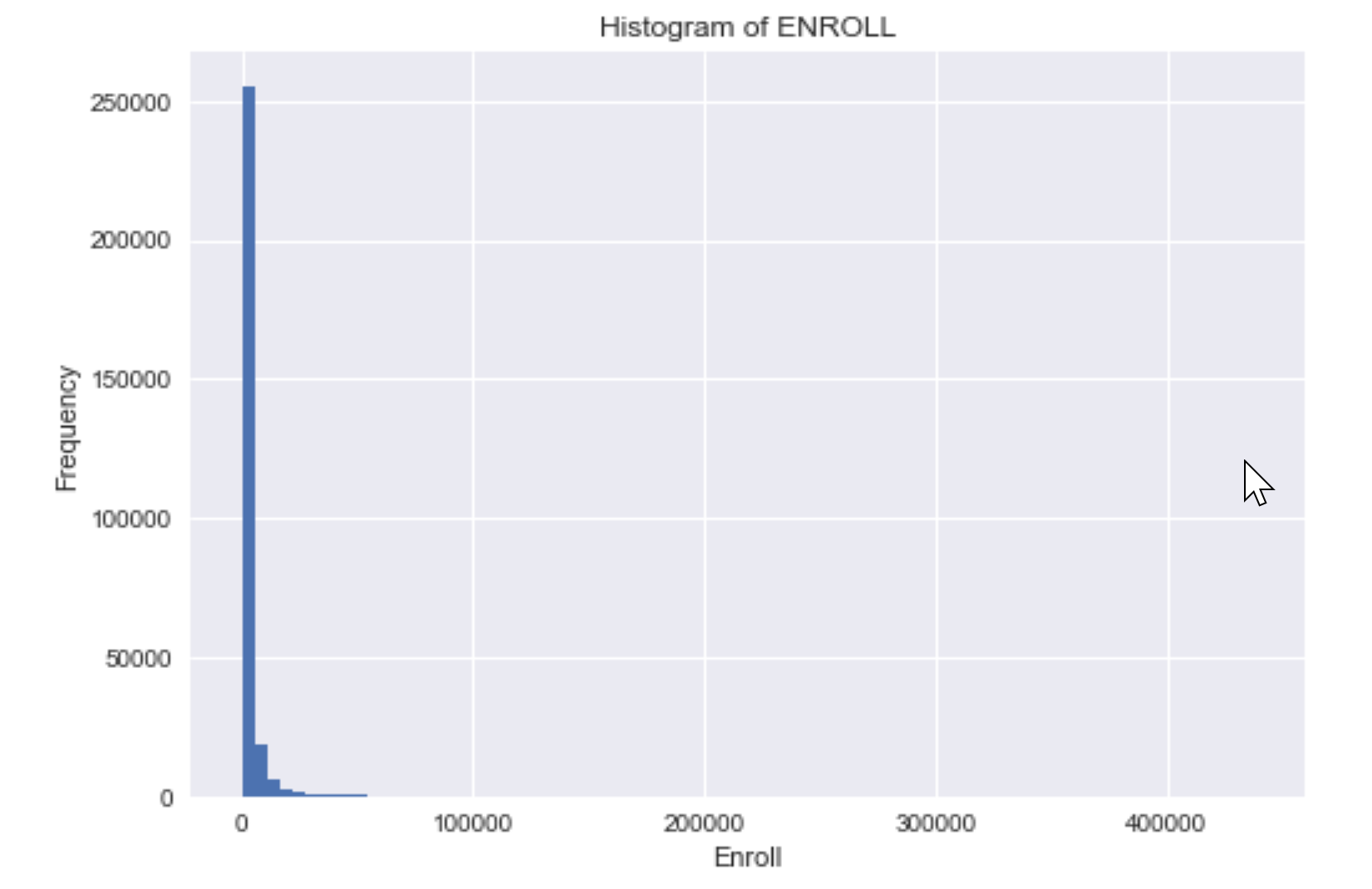
Matplotlib, Seaborn – for visualizing the data in terms of charts.

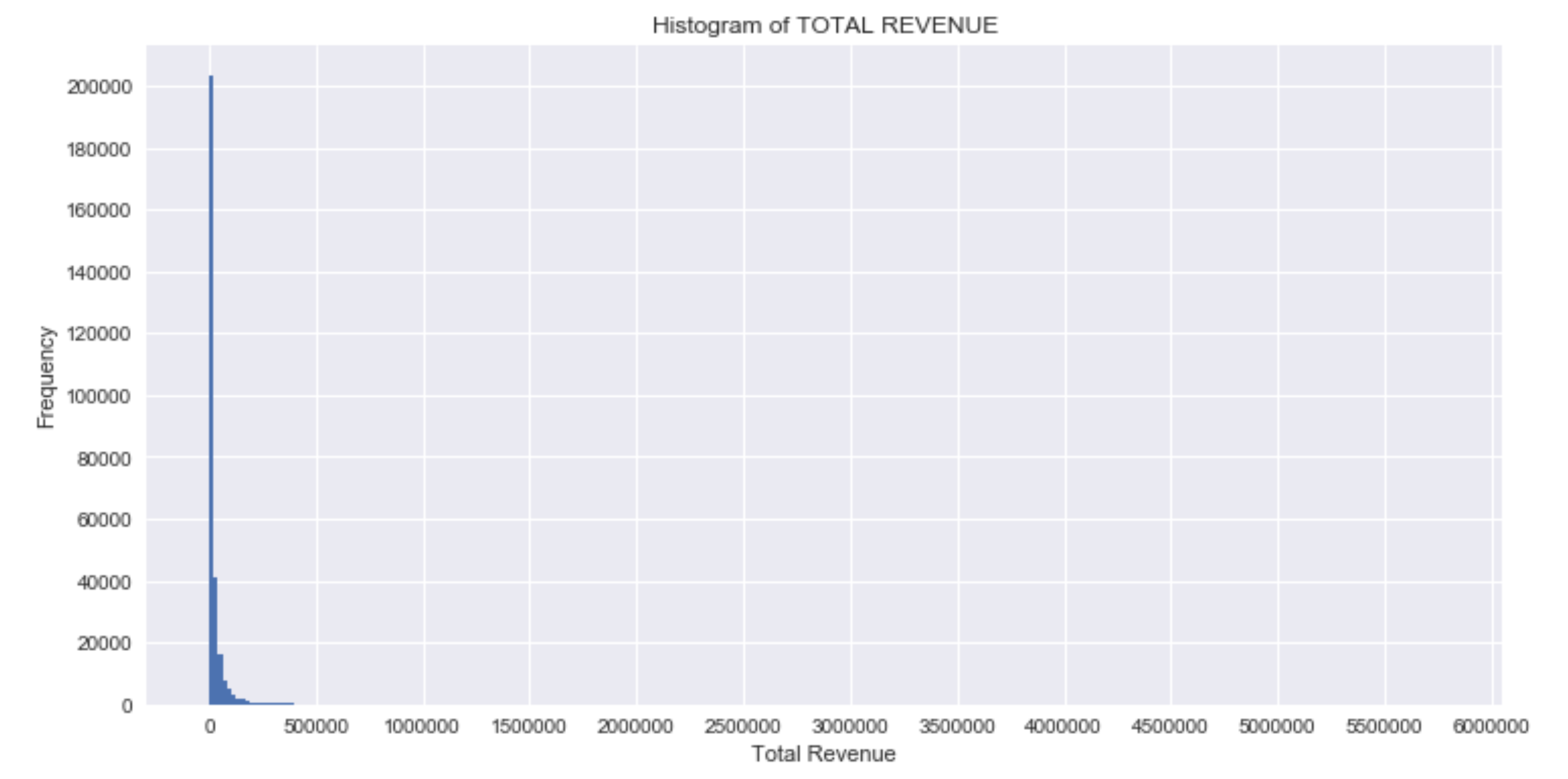
Thinkstats2, thinkplot – for reusing some of the wrappers for plotting PMF, CDF etc.

# Descriptive statistics

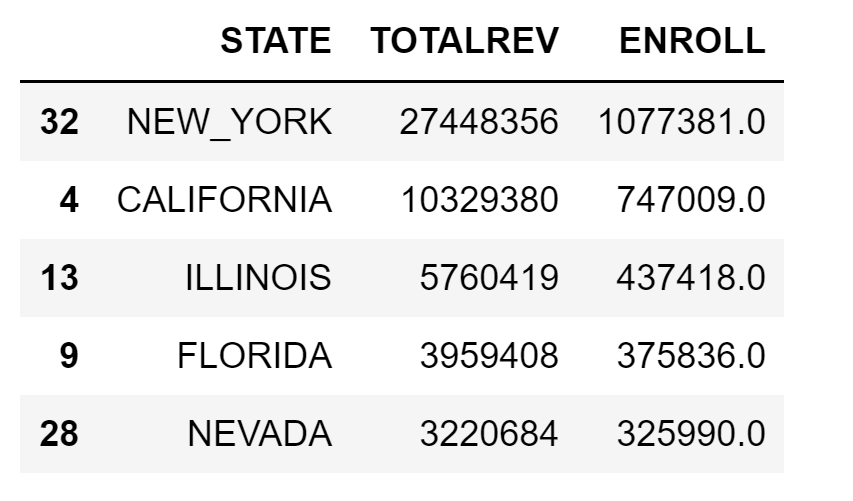
Initially cleaned the dataset to remove the null values, especially in the variables that are required for analysis – Total Revenue, Enroll.

Histograms of ENROLL and TOTALREV

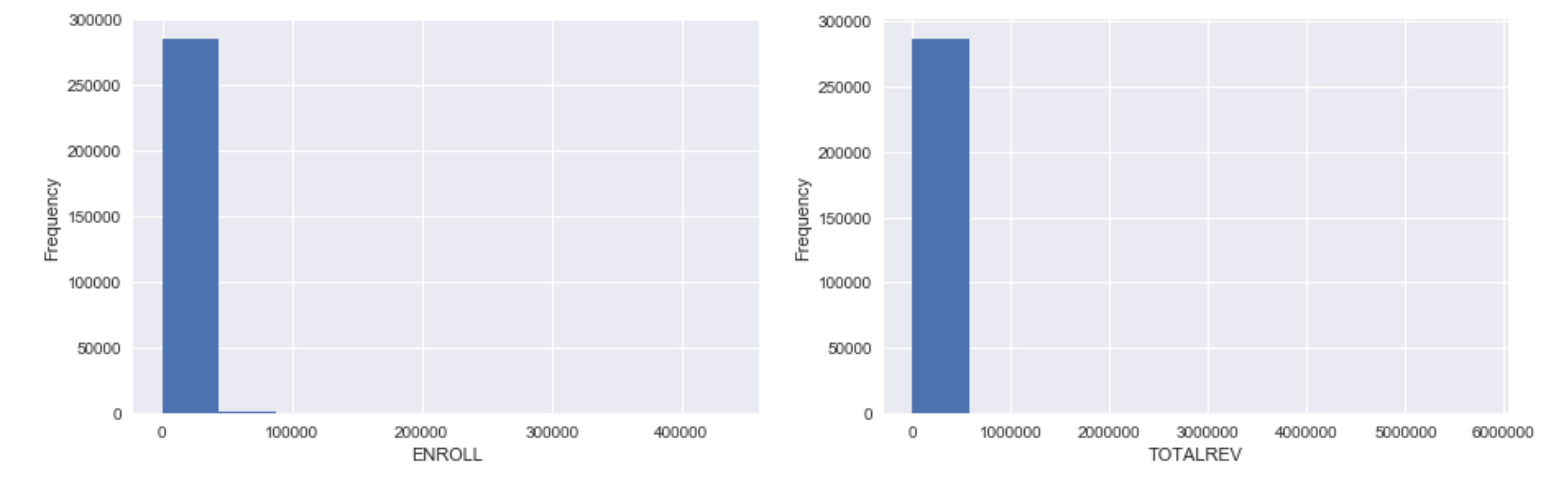
****

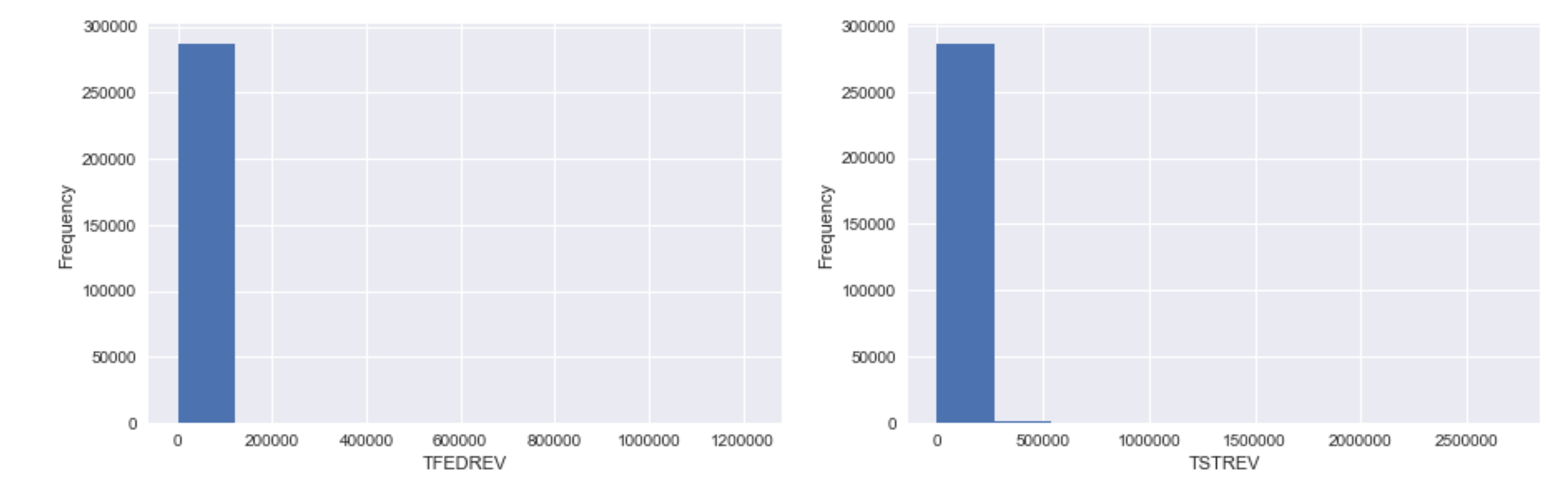
****

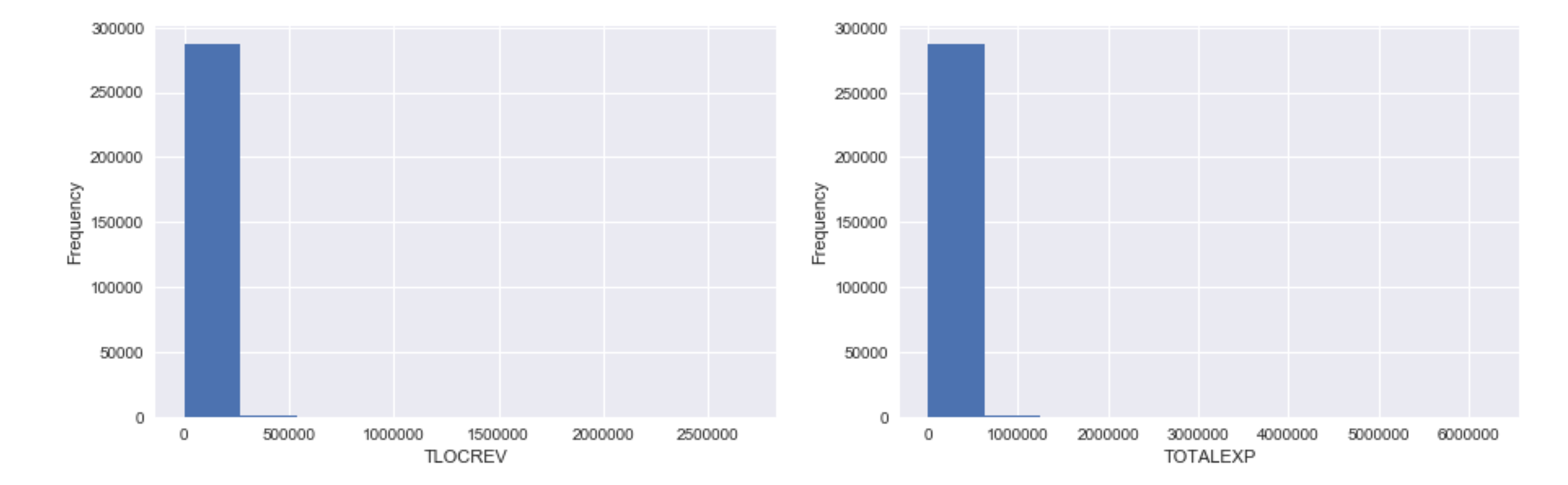
As can be seen from the histogram plots, the numbers are highly ‘right’ skewed with long tail towards higher end of the scale. After some analysis found that the TOTAL REVENUE for California and New York are almost triple and double of the next state in terms of total revenue. This could result in highly skewed analysis, so wanted to treat them separately and removed from the original data for further analysis.



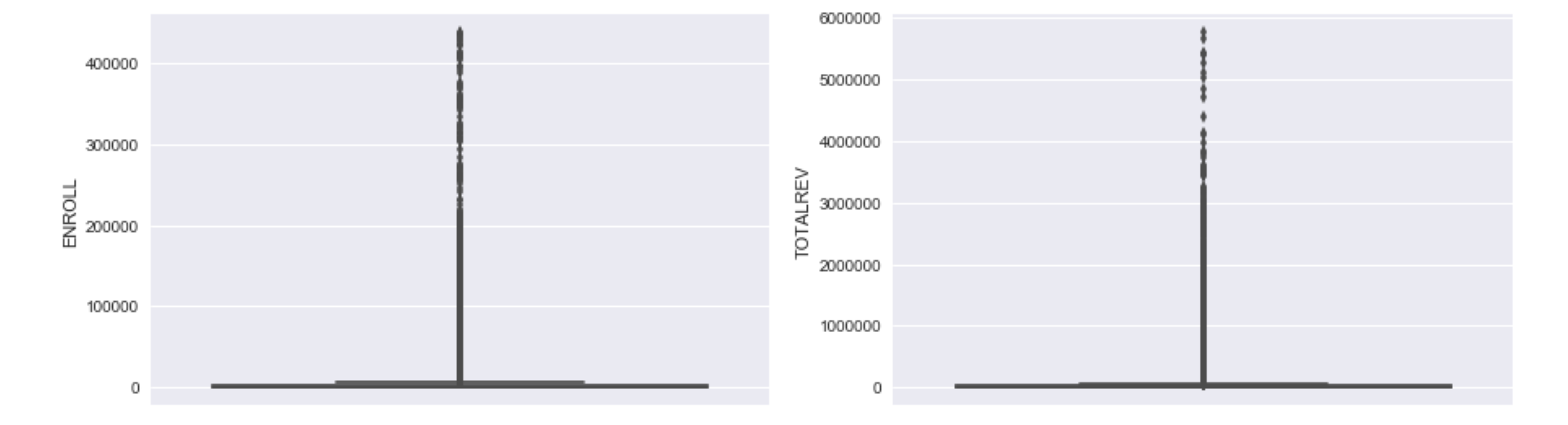
Histograms after removing those 2 states:

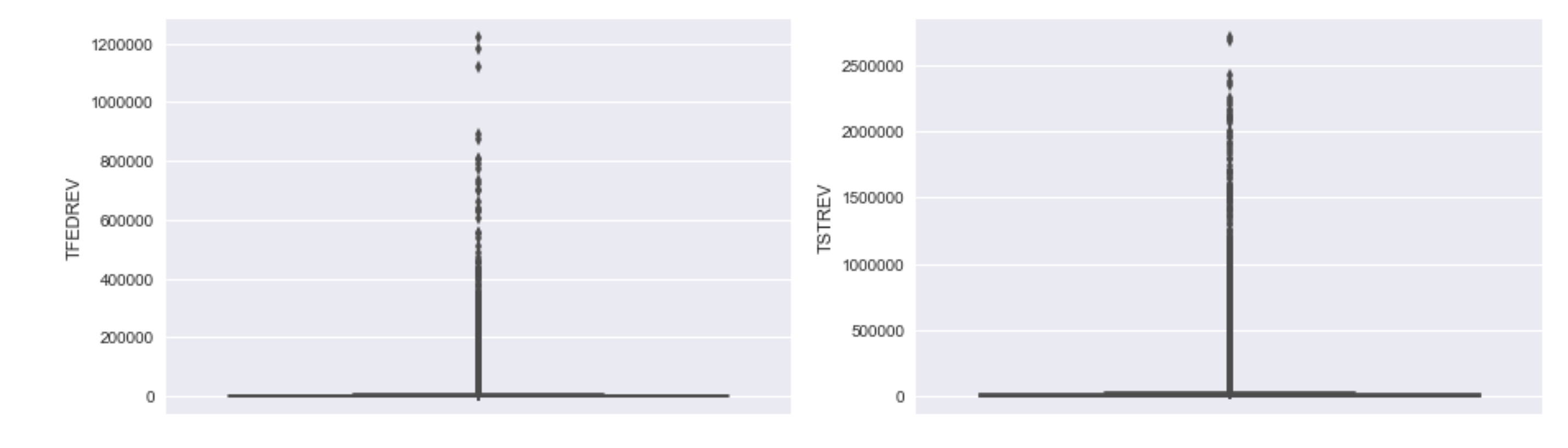


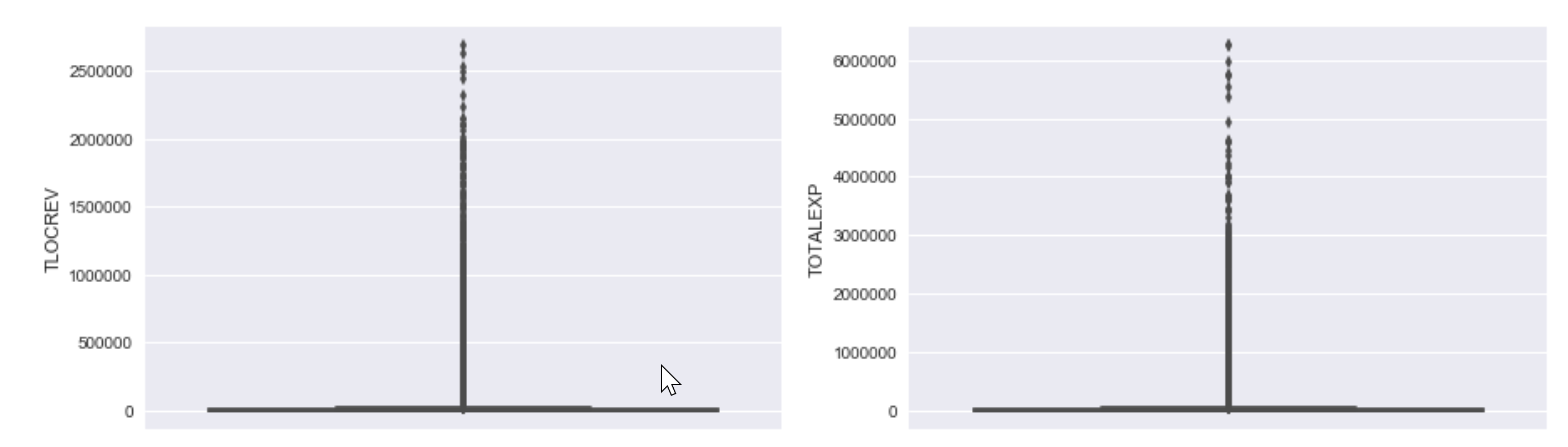




**Outliers in data set:**

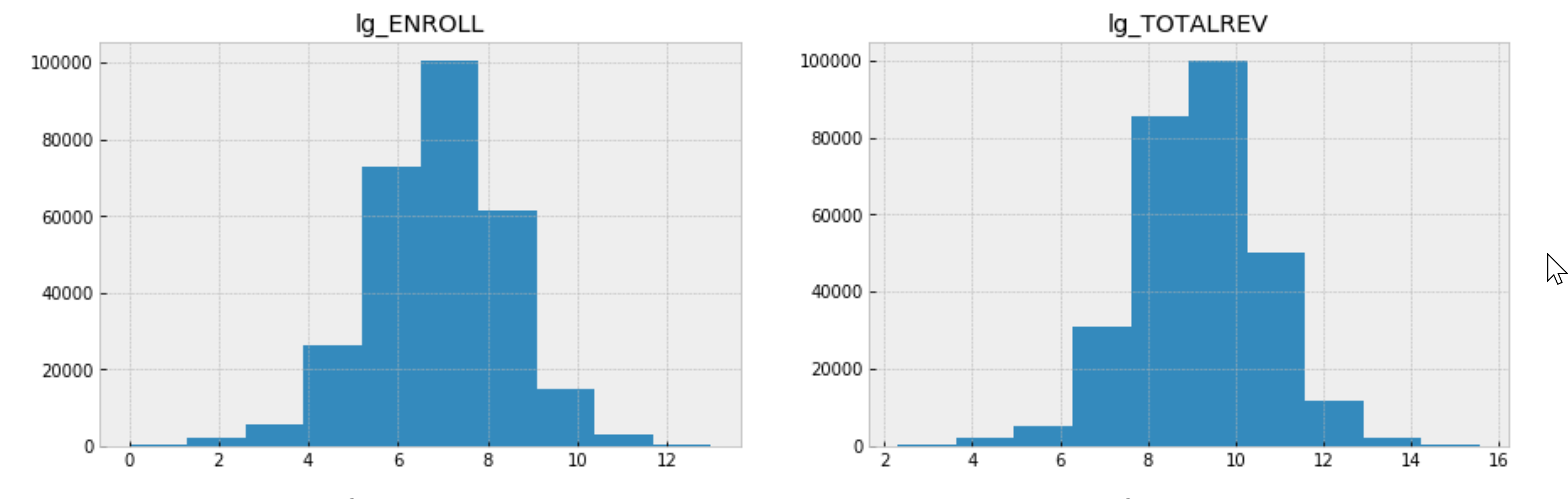
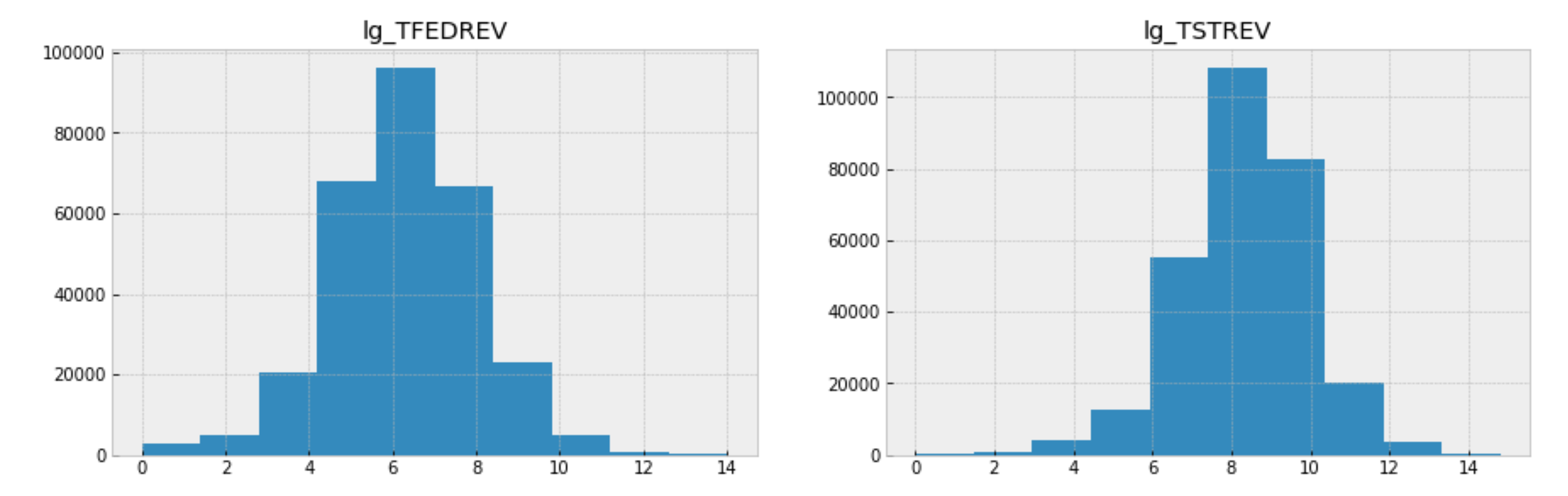


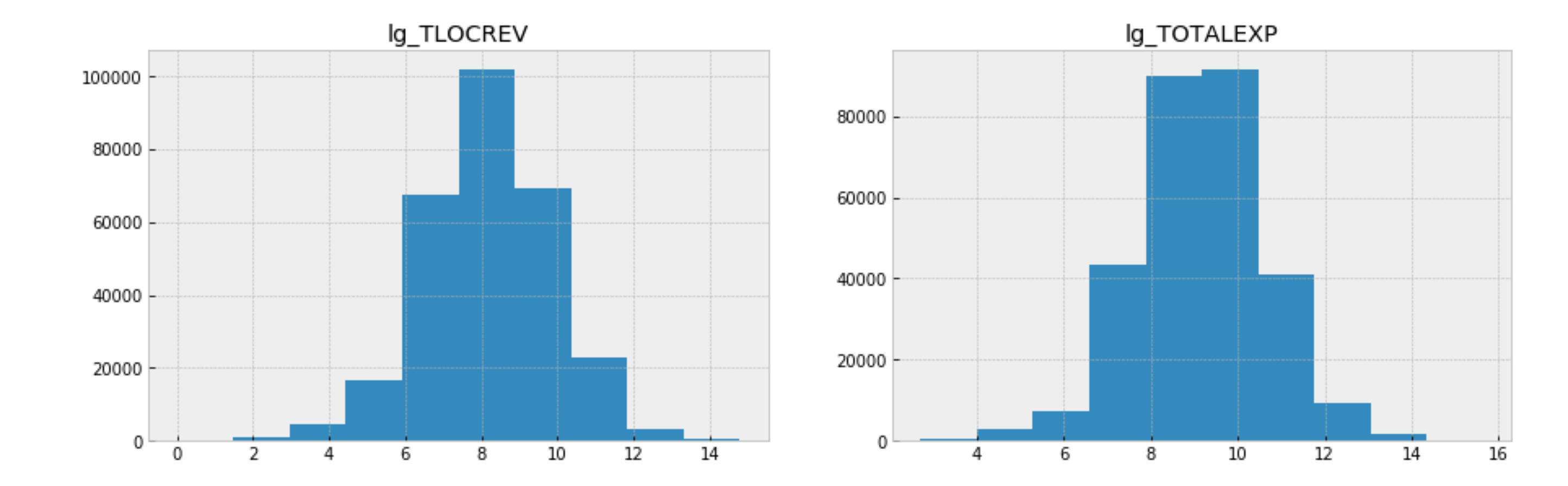




After separating those states, I observed that there are still some outliers in the data set, so transformed the numbers by applying log function across all the numerical variables.

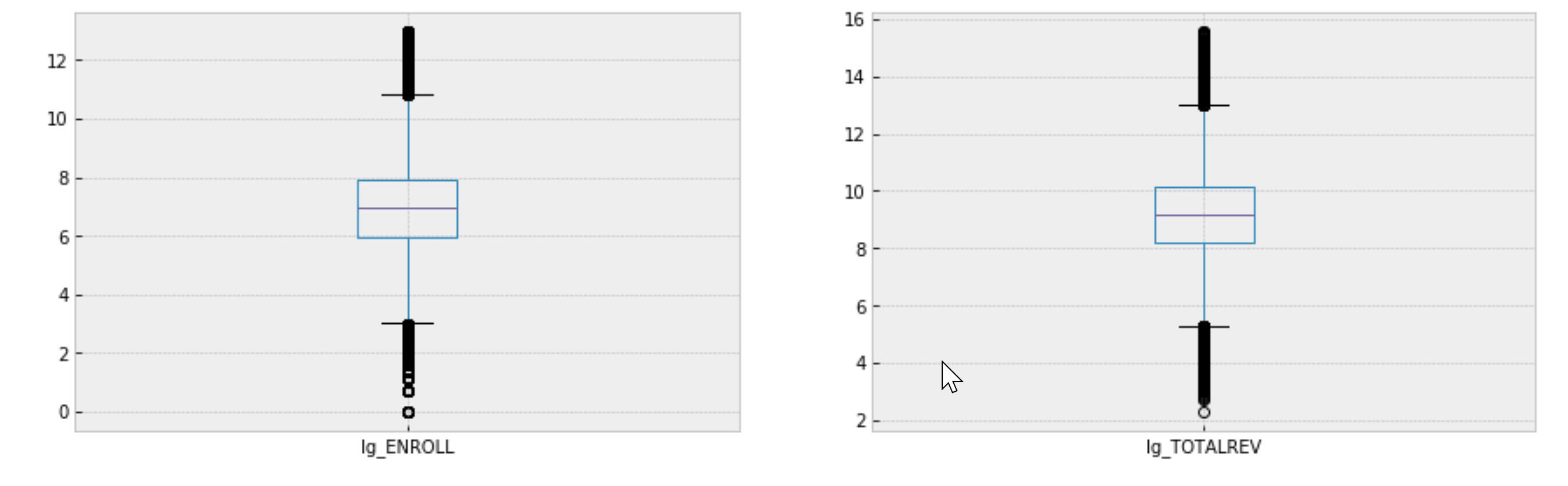
**Histograms of Log Transformed variables**

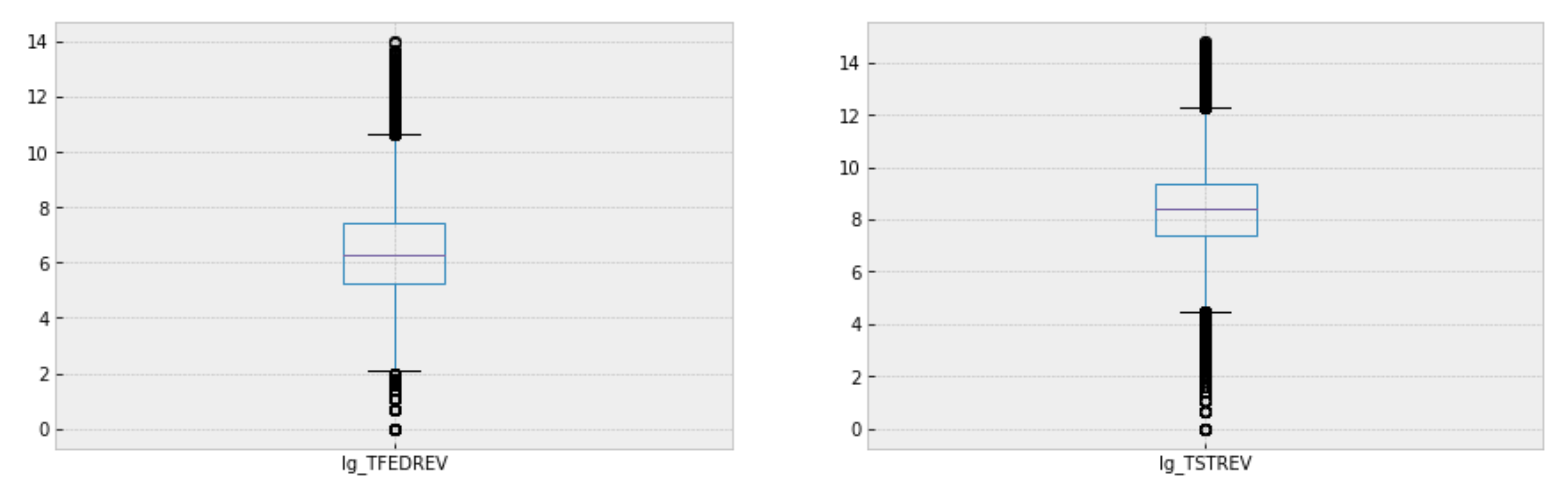
**** ****

****

Histograms of log transformed variables doesn’t appear as much skewed as they were earlier without log transformations. All variables except for lg\_TOTALEXP appear in unimodal distribution where as lg\_TOTALEXP is in bimodal distribution.

**Outliers after log transforming variables:**

****

****

****

Log transformation gives a better representation of these variables with lesser outliers.

# Descriptive Characteristics of the variables:

**Descriptive Characteristics for ENROLL**

Mean, Median, Mode of ENROLL, 3170.830384 1073.000000 180.000000

Spread - Variance, Standard deviation of ENROLL, 108893770.472233 10435.217797

Skew of ENROLL, 16.793895

Interquartile range of ENROLL, 388.000000 1073.000000 2720.000000

**Descriptive Characteristics for TOTALREV**

Mean, Median, Mode of TOTALREV, 30255.458977 9715.000000 2276.000000

Spread - Variance, Standard deviation of TOTALREV, 10677274625.544109 103330.898697

Skew of TOTALREV, 18.316904

Interquartile range of TOTALREV, 3653.000000 9715.000000 25200.250000

**Descriptive Characteristics for TFEDREV**

Mean, Median, Mode of TFEDREV, 2472.746674 539.000000 1.000000

Spread - Variance, Standard deviation of TFEDREV, 150278906.593238 12258.829740

Skew of TFEDREV, 32.430302

Interquartile range of TFEDREV, 187.000000 539.000000 1649.000000

**Descriptive Characteristics for TSTREV**

Mean, Median, Mode of TSTREV, 13981.876206 4608.000000 40.000000

Spread - Variance, Standard deviation of TSTREV, 2323006348.923632 48197.576173

Skew of TSTREV, 19.203716

Interquartile range of TSTREV, 1623.000000 4608.000000 11501.250000

**Descriptive Characteristics for TLOCREV**

Mean, Median, Mode of TLOCREV, 13800.836066 3524.000000 25.000000

Spread - Variance, Standard deviation of TLOCREV, 2613455677.474381 51121.968638

Skew of TLOCREV, 19.163061

Interquartile range of TLOCREV, 1224.000000 3524.000000 10725.250000

**Descriptive Characteristics for TOTALEXP**

Mean, Median, Mode of TOTALEXP, 30560.999228 9660.000000 1638.000000

Spread - Variance, Standard deviation of TOTALEXP, 11182506231.429939 105747.369856

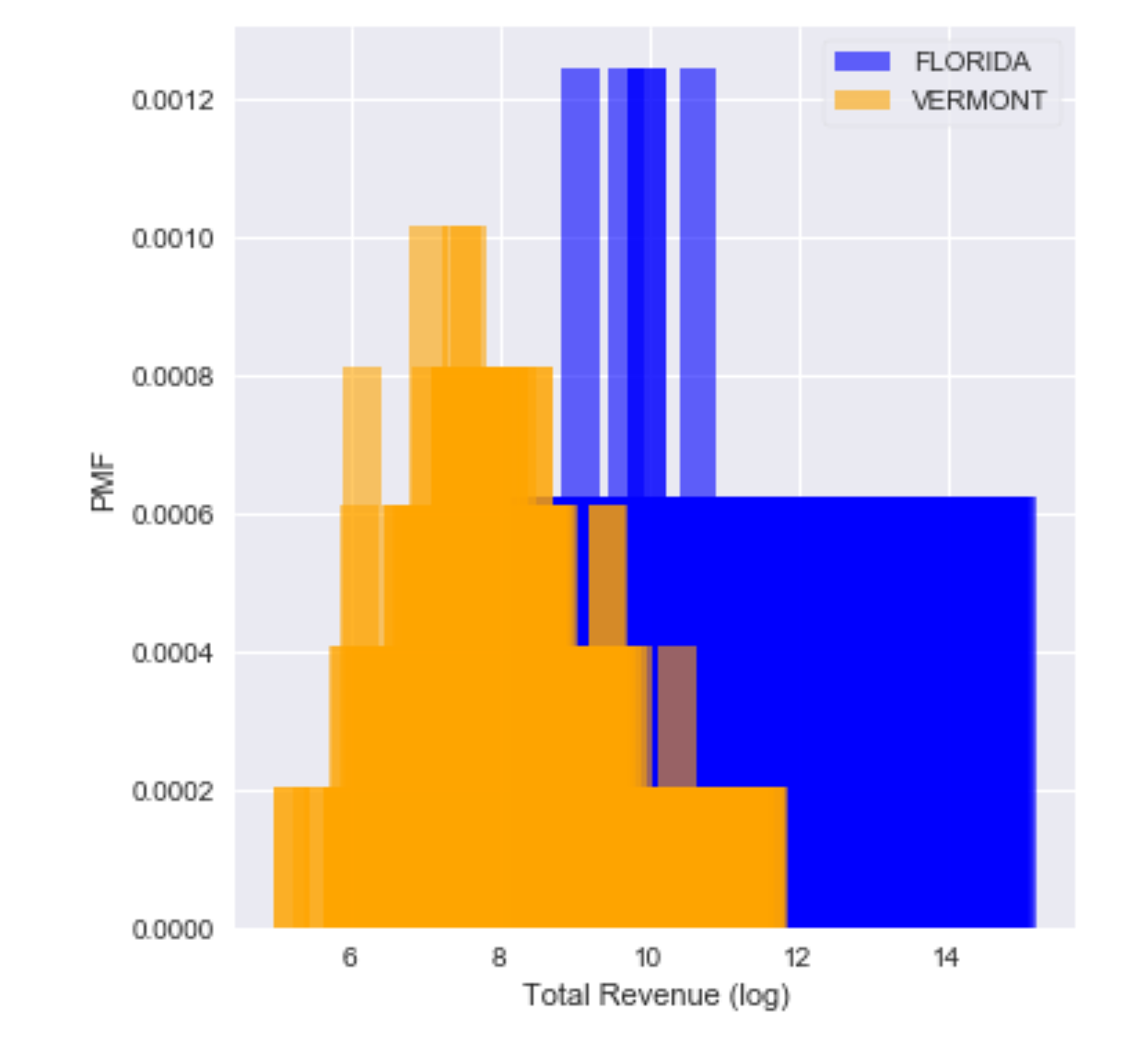
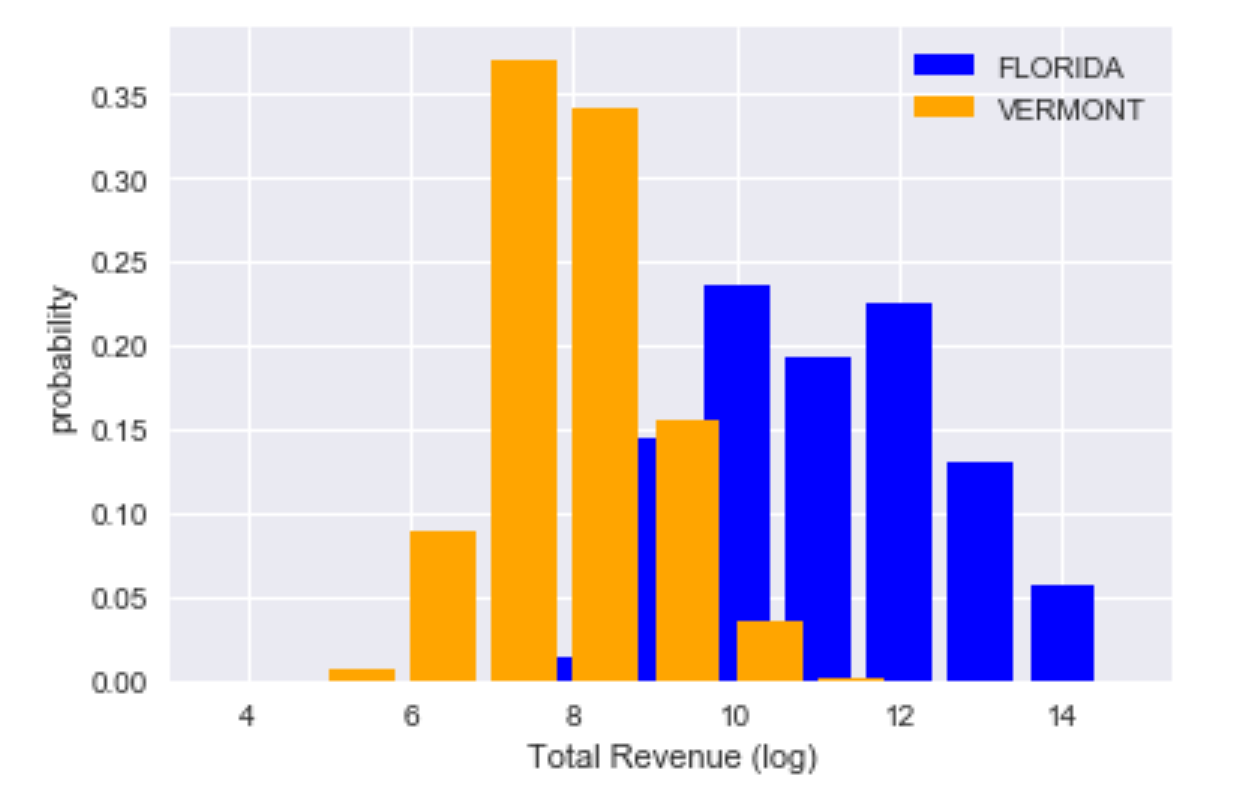
Skew of TOTALEXP, 19.065186

Interquartile range of TOTALEXP, 3592.000000 9660.000000 25343.000000

Skew is far greater than 1, highlighting that the numbers for every column are skewed heavily towards right with long tail towards higher scale.

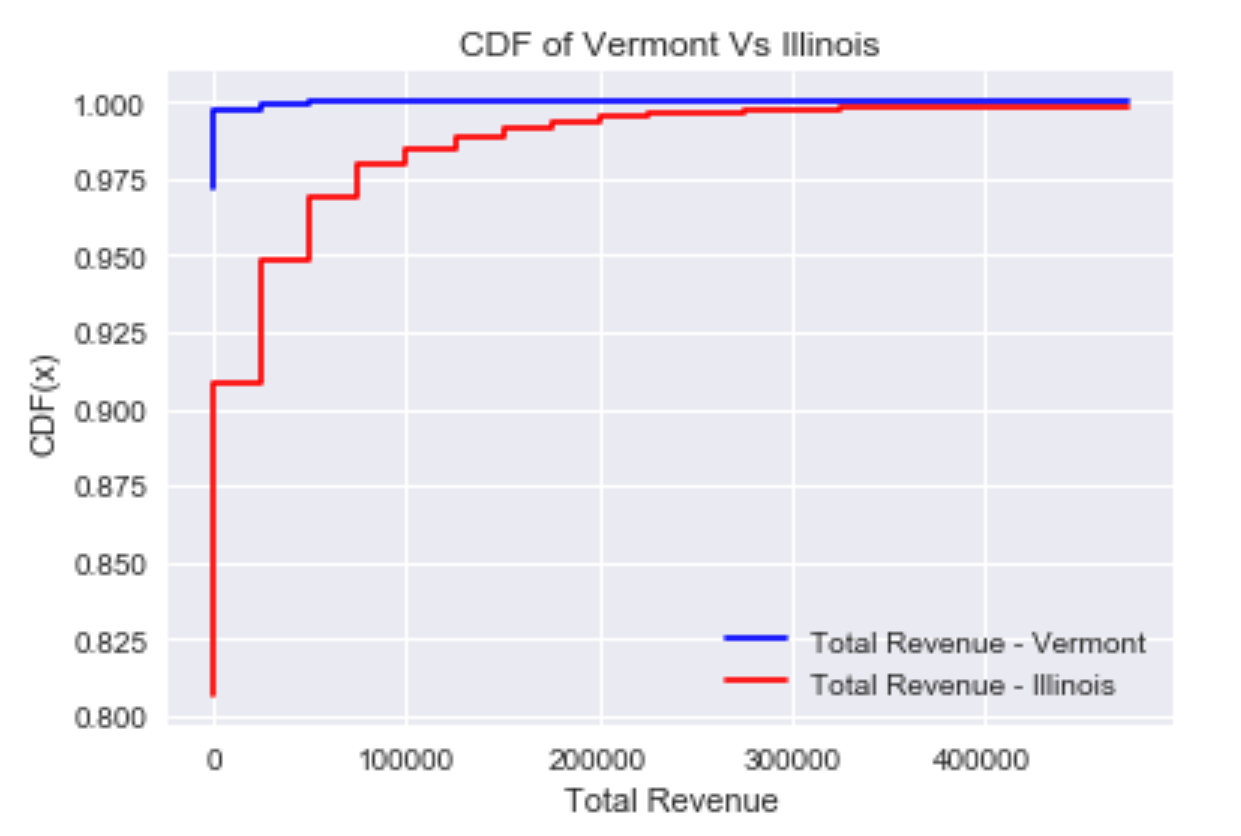
**Plotting PMFS**

For plotting PMS and probability plots, I have selected Vermont and Florida

Based on the comparisons of PMF's Vermont - school districts are more likely to have lesser total revenues than Illinois school districts

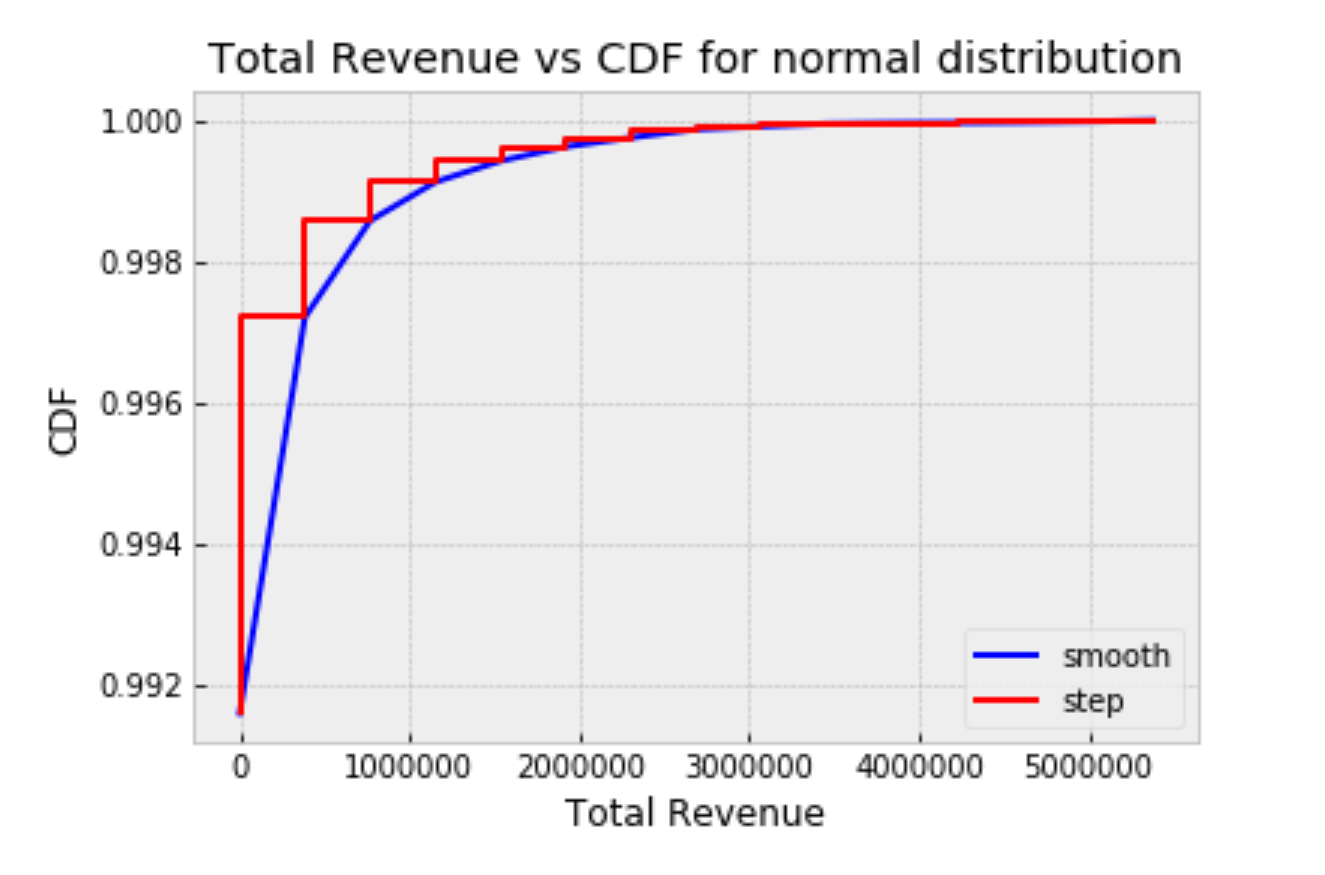
# Calculating CDF

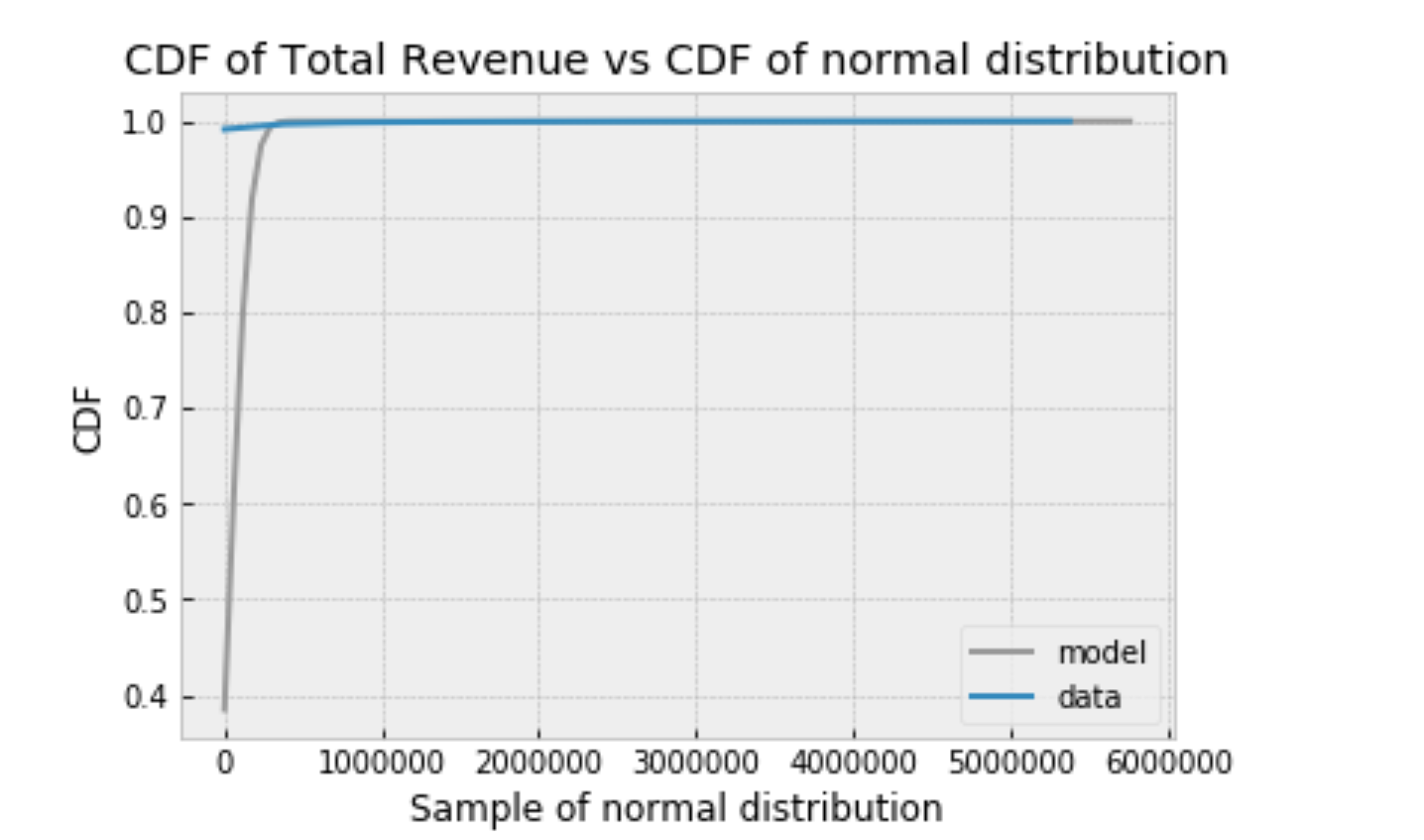


Overall school districts in Illinois have higher total revenue than Vermont and 98% of the total revenues for all school districts in Illinois is less than 100,000. Whereas for Vermont, almost 97% of the total revenues for school districts are below 50000$.

Put it in another way, Illinois school districts have higher chance of having more Total Revenue.

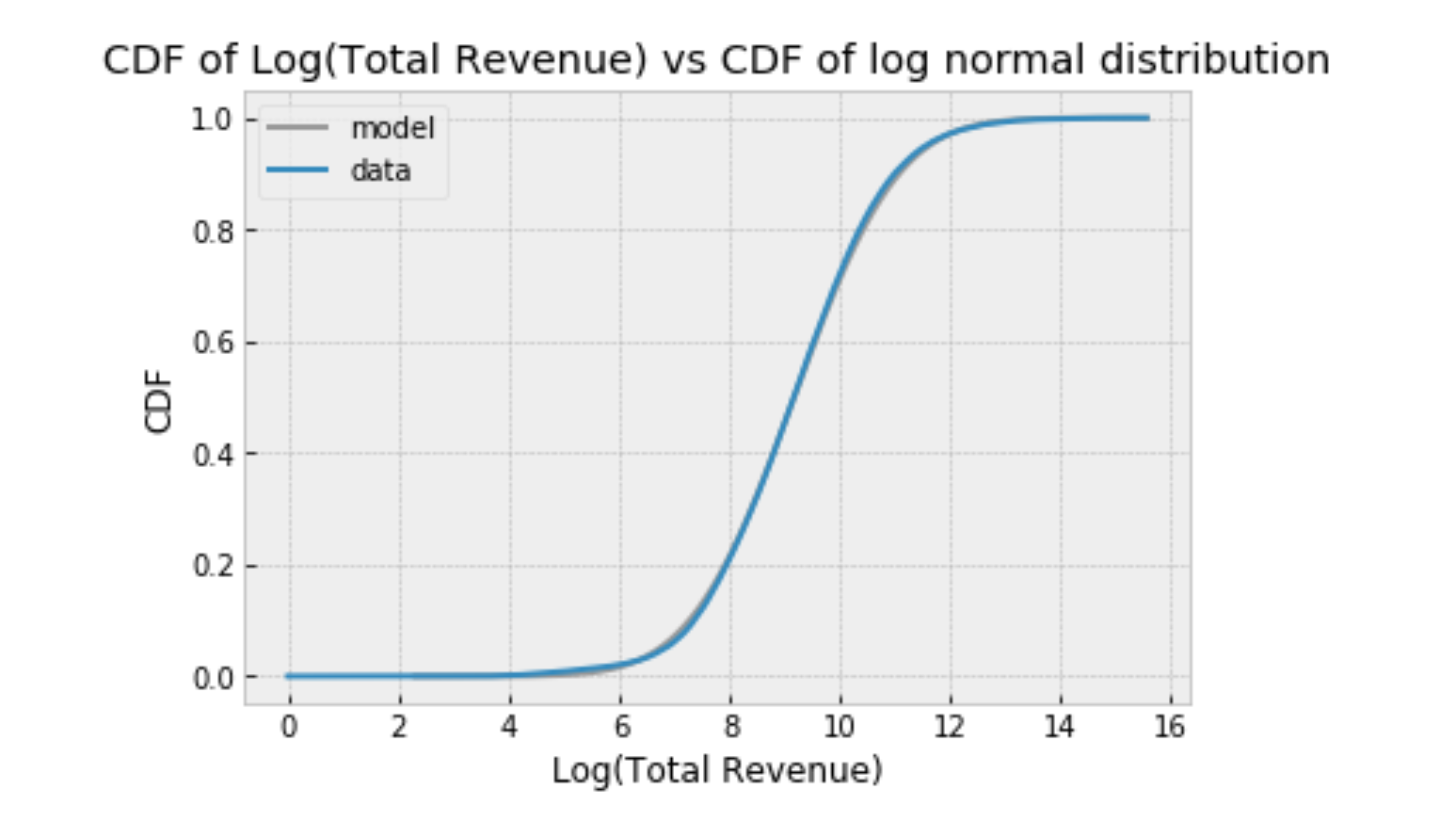
# Plotting analytical distributions





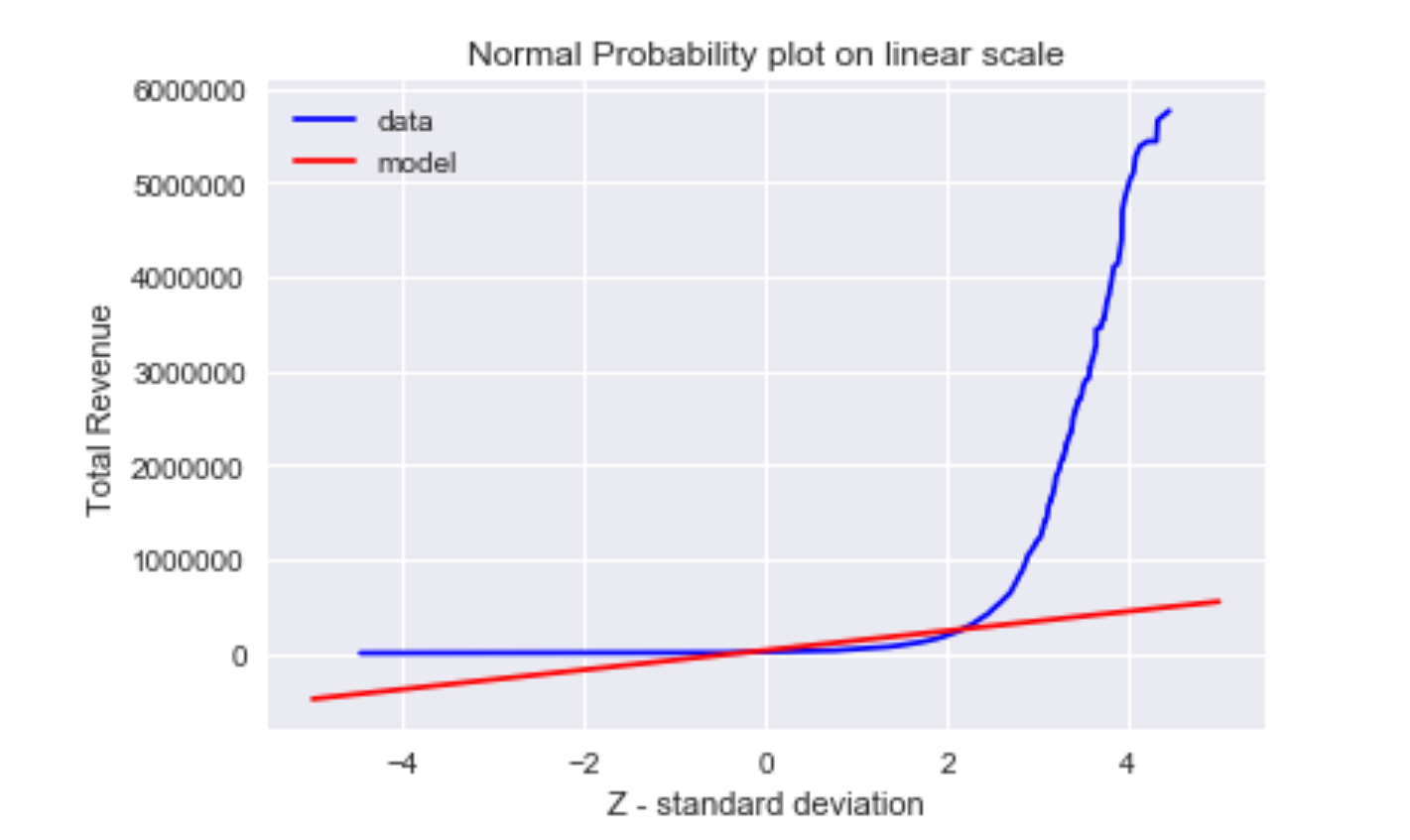
From the above overlapping plot of normal distribution vs total revenue, we can infer that Normal distribution doesn’t represent the total revenue. Let us try to see if log normal distribution is applicable.

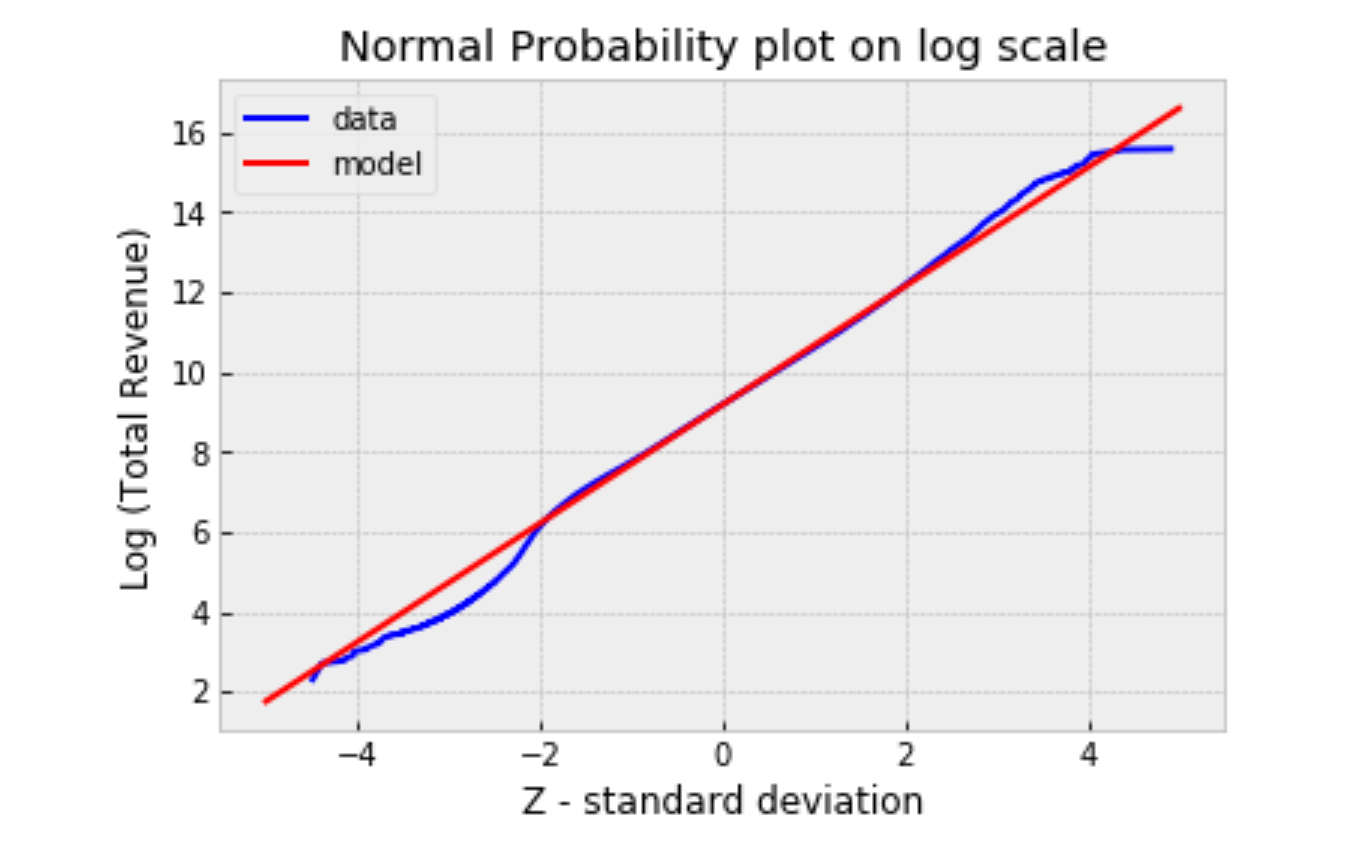
To find out I will use the log transformed total revenue column - findistdf.lg\_TOTALREV



From the above overlapping plot of log normal distribution vs log(total revenue), we can infer that log normal distribution perfectly fits for the variable total revenue

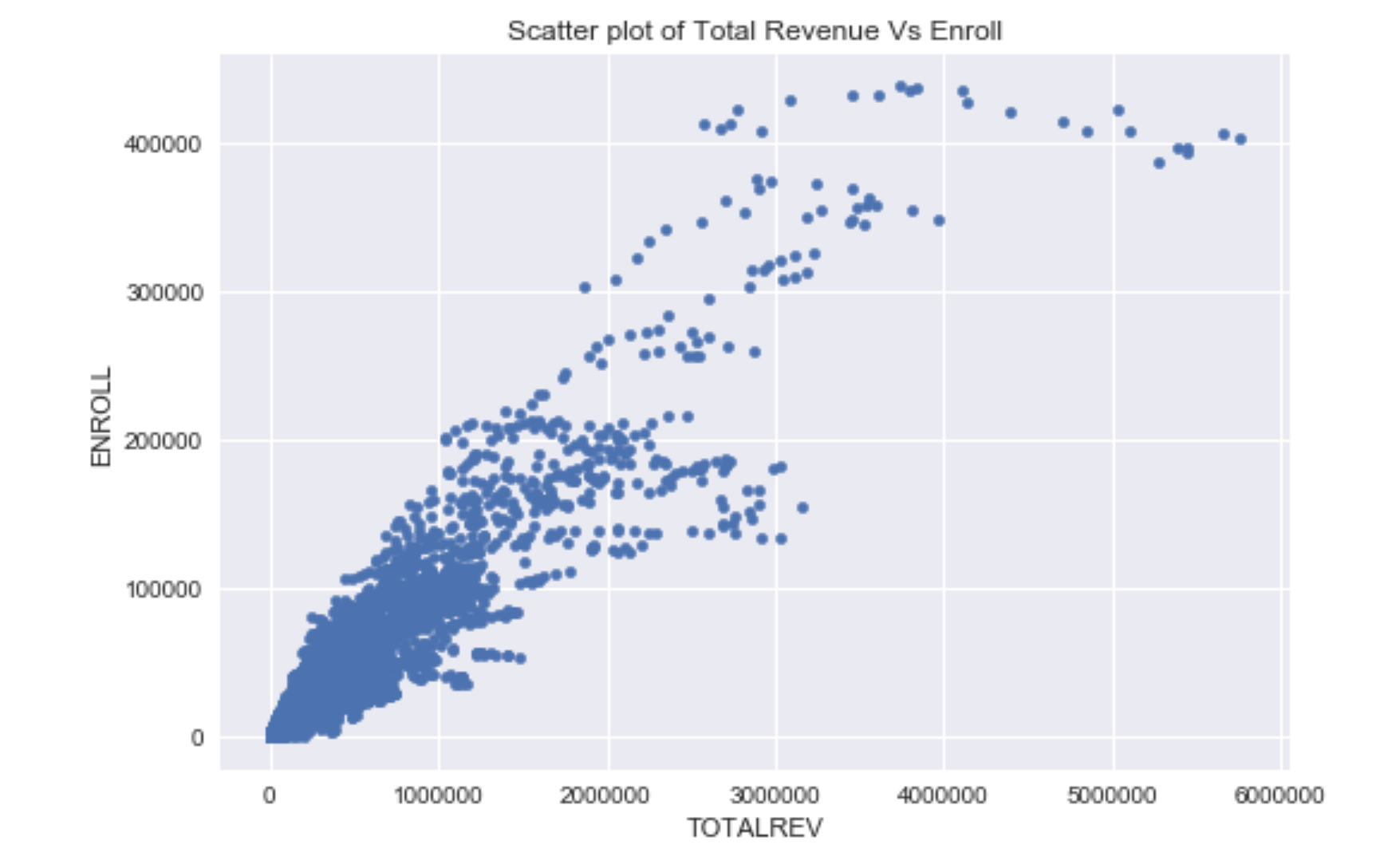
# Probability plots for total revenue and Log (total revenue)





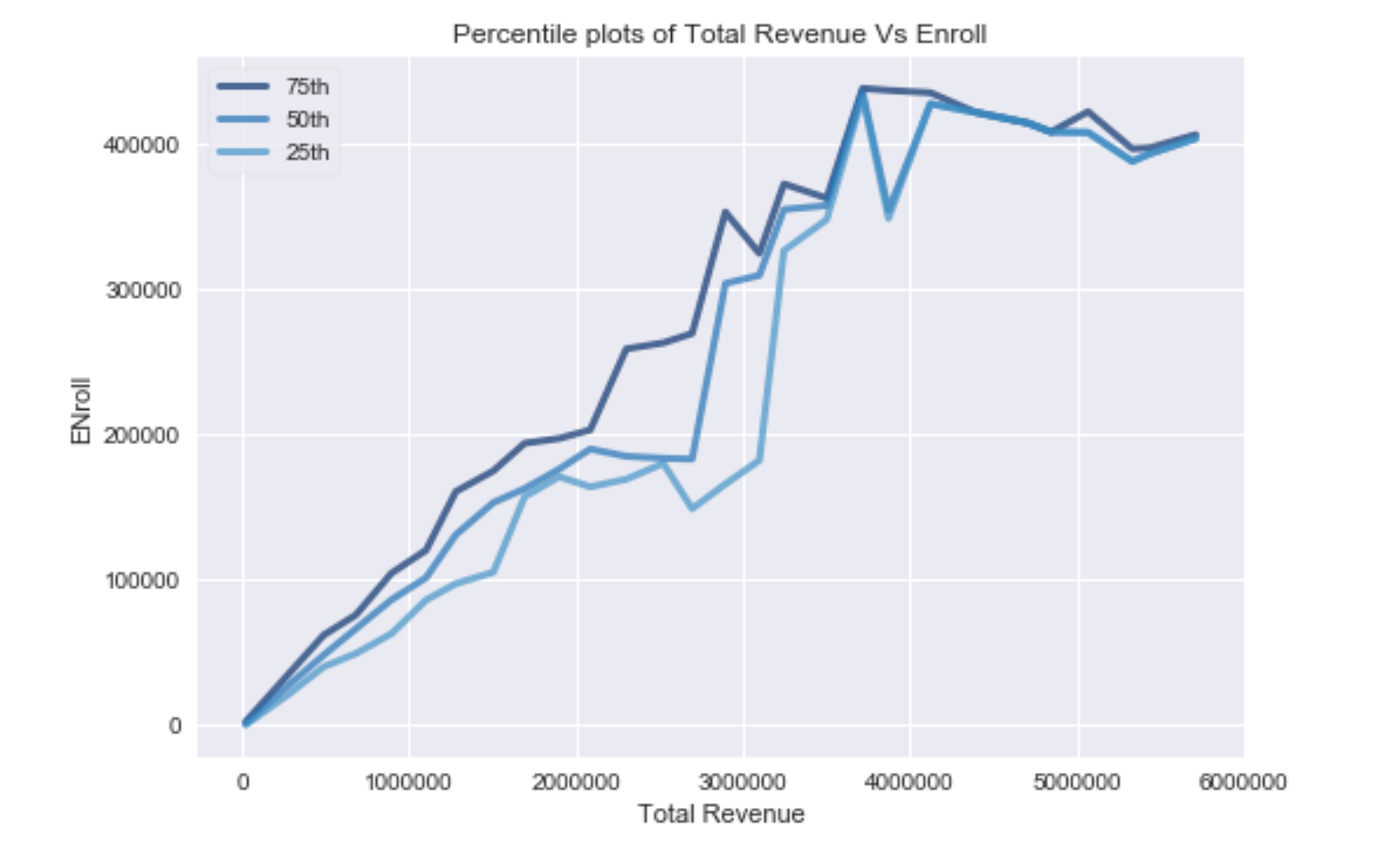
From the above two normal probability plots, we can infer that data deviates substantially from normal model whereas log normal model fits perfectly to the data within 2 standard deviations (between -2 to 2) but deviates from the log normal model significantly for the school districts with lower and higher end of the (log) total revenue scale.

# Scatter plots and Correlation analysis



The above chart shows that school districts with higher total revenue has better enrollment than the school districts with lower total revenue, agrees with one of our assumptions.

# Characterizing Relationships



Above percentiles plot of Total Revenue Vs Enroll, relationship is linear up to 3500000$, after that

relationship is going in the wrong direction.

# Covariance and Correlation

Covariance between Total Revenue and Enroll is - 1024052236.613642

Correlation between Total Revenue and Enroll is - 0.9497119340376089

Correlation value of 0.95 indicates that total revenue and enroll variables are strongly and positively correlated; and it implies that school districts with higher total revenue tend to have higher enrollments in

those schools.

But our distributions are highly skewed and are not normal distributions, so let’s find out the Spearman's Rank correlation as well.

# Hypothesis Testing

Defining Null Hypothesis: My earlier assumption is that school districts with higher total revenue will have higher enrollments in the school. Based on that, my Null hypothesis is that there is no relationship between Total revenue and school enrollments for school districts. The p-value for this correlation testing is to find out the probability of having such a high observed correlation of 0.95 by pure chance should be significant (p-value > 0.05).

Let us find out with Hypothesis testing.

Defined the function ‘samplepermute’ that create a sample data from existing data by randomly permutating the total revenue’s array and then calculates the correlation between Enroll and Total Revenue numbers. Iterates this process for 100 times to find out the count of samples that have higher correlation than observed correlation of 0.95.

It turns out that the probability of having such a high (observed - 0.95) correlation between Total Revenue & Enroll by chance is 0. Hence null hypothesis that there is no correlation between Total Revenue and Enroll is false.

# Regression Analysis - Linear Least Square Model

# 

# Above plot confirms the linear relationship between Total Revenue and Enroll.

# Goodness of Linear Least Square fit

# Goodness of linear least square fit can be found by comparing the Root mean square error between with model and without model.

# Without any model, RMSE of predicted Enroll numbers is represented by its standard deviation – which here in this case is 10435.

# With Linear Least Square fit model, RMSE of predicted Enroll numbers from known Total Revenues is calculated by finding the residuals from prediction (Observed Enroll – Predicted Enroll) and finding the standard deviation from the residual. In this case it is 3267.

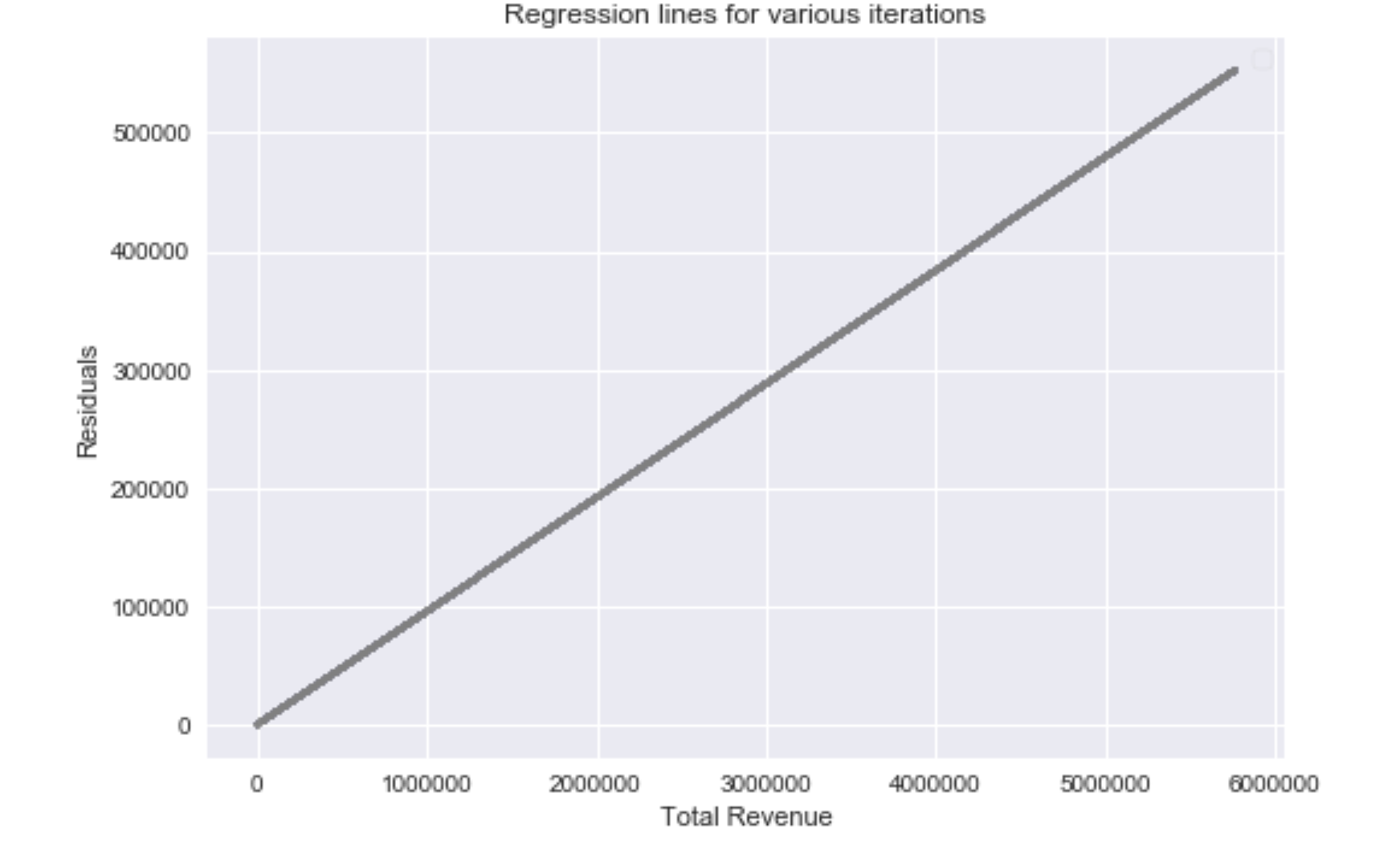
# As predicting Enrollment numbers with Linear Least Square model results in lesser standard deviation, in this case knowing the total revenue and predicting enrollment numbers from it has significantly helped for better prediction and reducing the error.

The Coefficient of Determination is 0.90 indicates that total revenue helps predict almost 90% of the variance in the enrollment numbers for school districts.

# Testing Linear Model

Probability that the slope in the sampling distribution falls below 0 (p-value) is 0; as it is less than

0.001 indicating that the relation between Total Revenue and Enroll is statistically significant and not by chance.



After repeated sampling the regression line roughly stayed in the same place, so it is a low variance model.

# Regression Analysis - Ordinary Least Square Model

Model 1 – I have considered Total Revenue as the independent variables and variable Enroll dependent.

*formula = 'findistdf1.ENROLL ~ findistdf1.TOTALREV'*



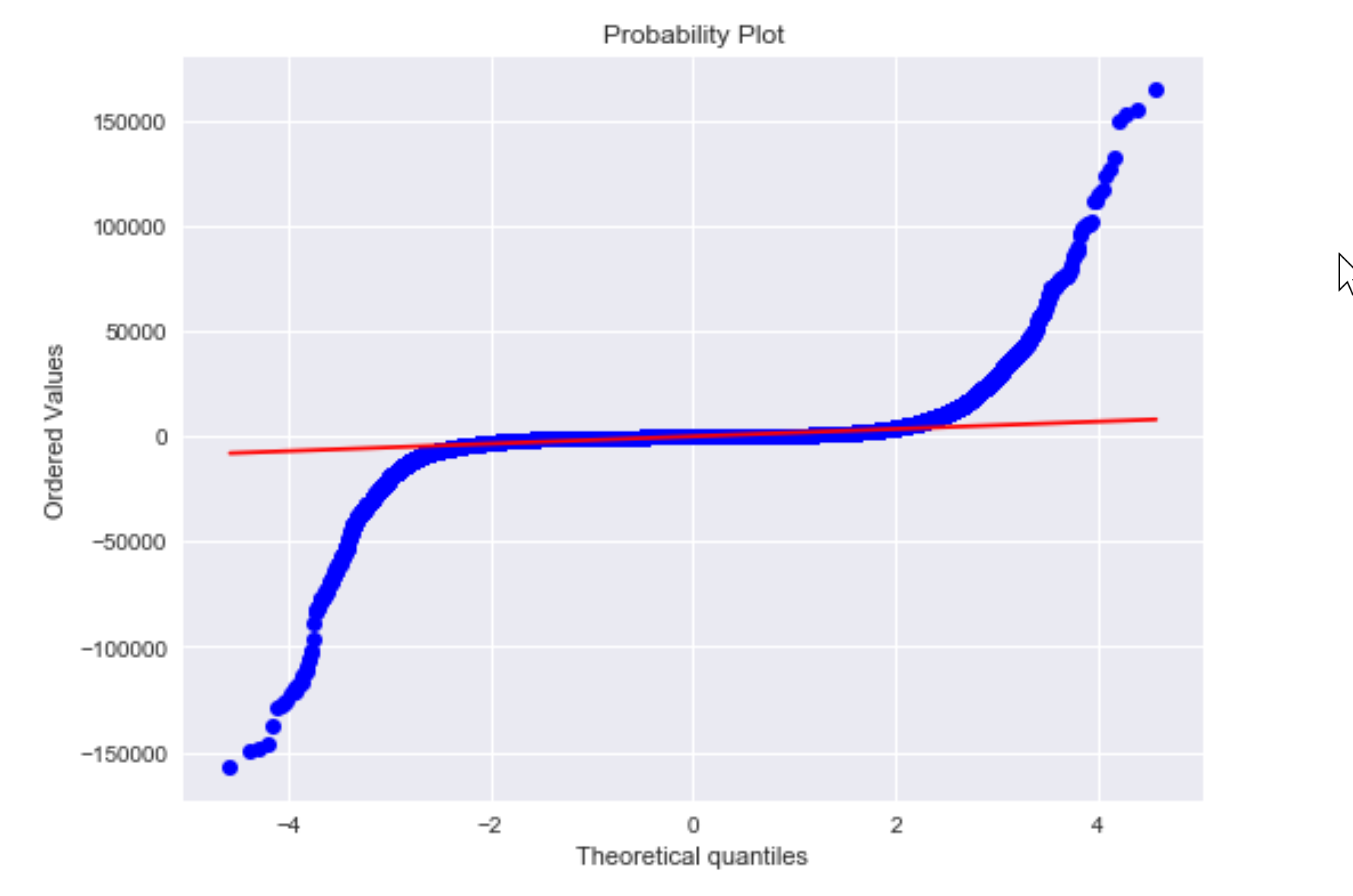
inter, slope = (269.0338863686441, 0.09590984885567201)

Interpreting the coefficients: slope value of 0.095 infers that unit increase in total revenue is associated with 0.095 unit increase in enroll numbers for the school districts.

And P-value is less than 0.001, the estimated slope is significant.

# Coefficient of determination

R- Square value of 0.90 shows that variation in enrollment can be explained by variation in Total Revenue up to 90%. As more variance is being explained by the model, once again proves the fit of the model.



Above normality plot indicates that this model is good fit only between quartile -2 to +2. After that it

significantly deviates from linear model.

# p-values for the model coefficients

results.pvalues

Intercept 0.0

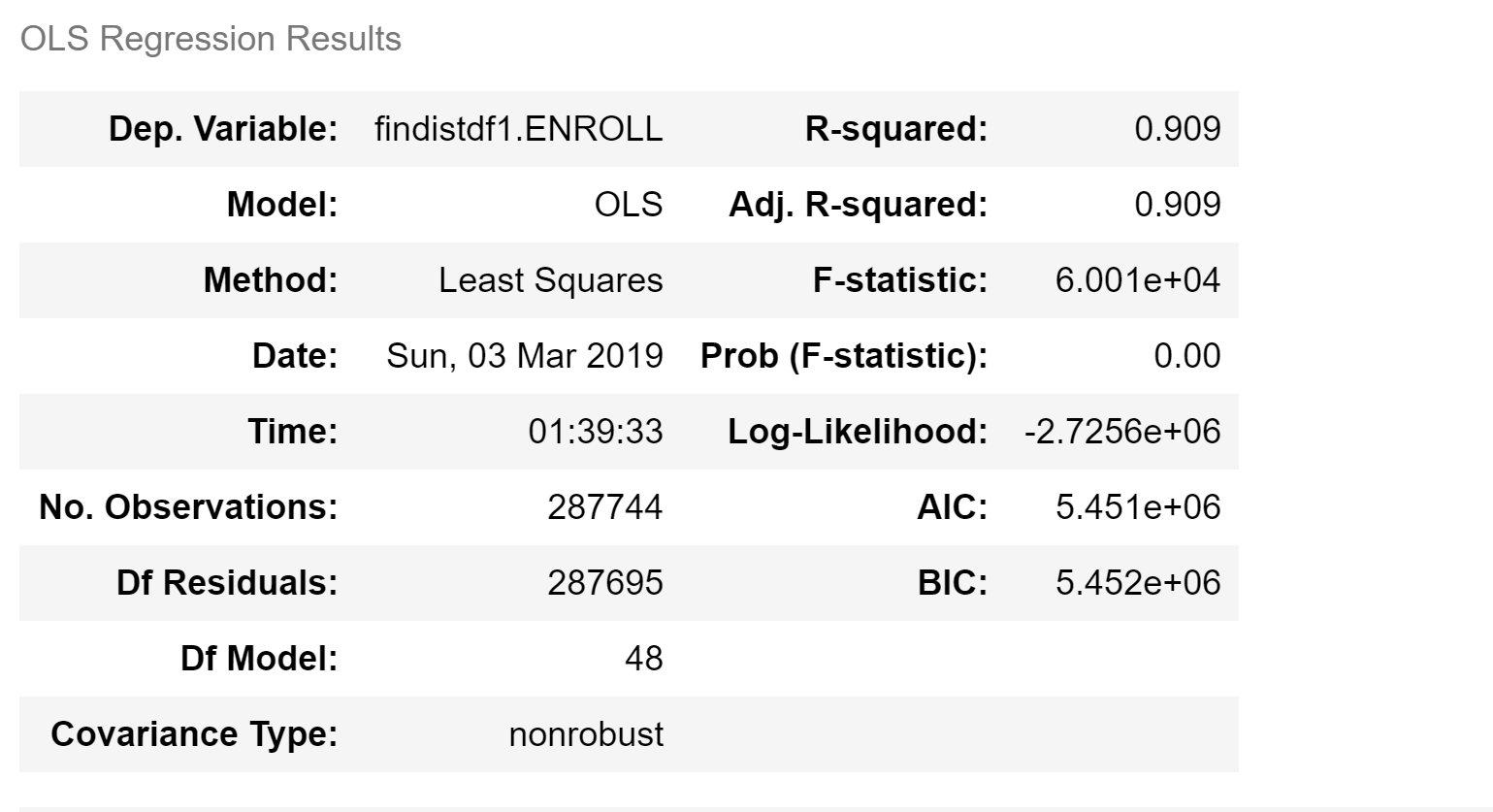
findistdf1.TOTALREV 0.0

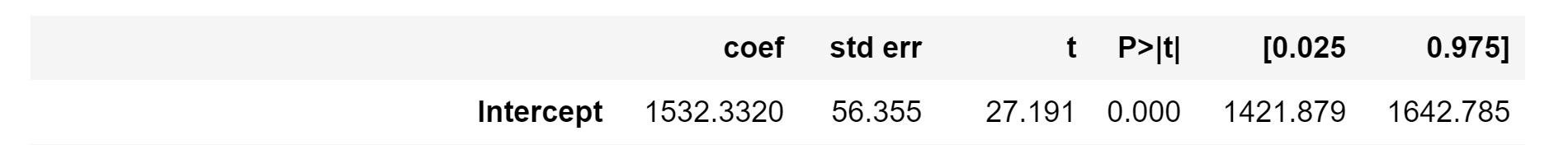
Again p-values are way less than 0.05, indicating that the relation between dependent and independent variables is genuine.

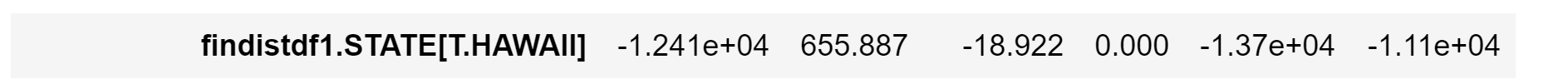
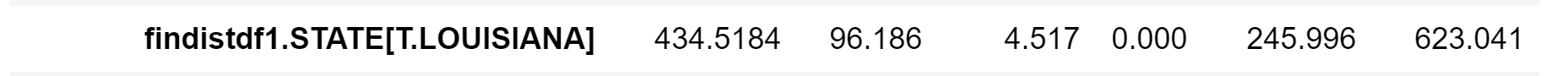
# Multiple Linear Regression

**Model 2** – I have considered Total Revenue and State as the independent variable and variable Enroll as dependent.

*formula2 = 'findistdf1.ENROLL ~ findistdf1.TOTALREV + findistdf1.STATE'*







# Comparing model 1 vs model2:

1) Adjusted R-square of model2 is .909, better than Adjusted R-square of model1 (0.902) - indicating that model2 can explain slightly more variation in dependent variable compared to model1.

2) AIC of model1 is 5.473e+06, which is slightly higher than AIC of model2 5.451e+06, indicating that model2 (enrollment as a function of total revenue and state) is slightly better model among the 2 models.

3) HAWAII has the highest absolute coefficient - indicating that ENROLL'ment numbers change hugely with a single unit of variation in TOTALREV for that state. In other words, we can probably see more enrollment numbers for every same number of units increase in total revenue compared to all other states (all remaining things being constant).

4) LOUISIANA has the lowest absolute coefficient value - indicating that ENROLL numbers will change at a slower rate compared to all other states for the same unit of increase in the total revenue (all remaining things being constant).

5) Overall adding STATE to the ordinary least square model, improved the model very slightly but not significantly. But interesting aspect of adding STATE to the equation is it gives us insights into how each is the relationship between Total Revenue and Enroll for each STATE.