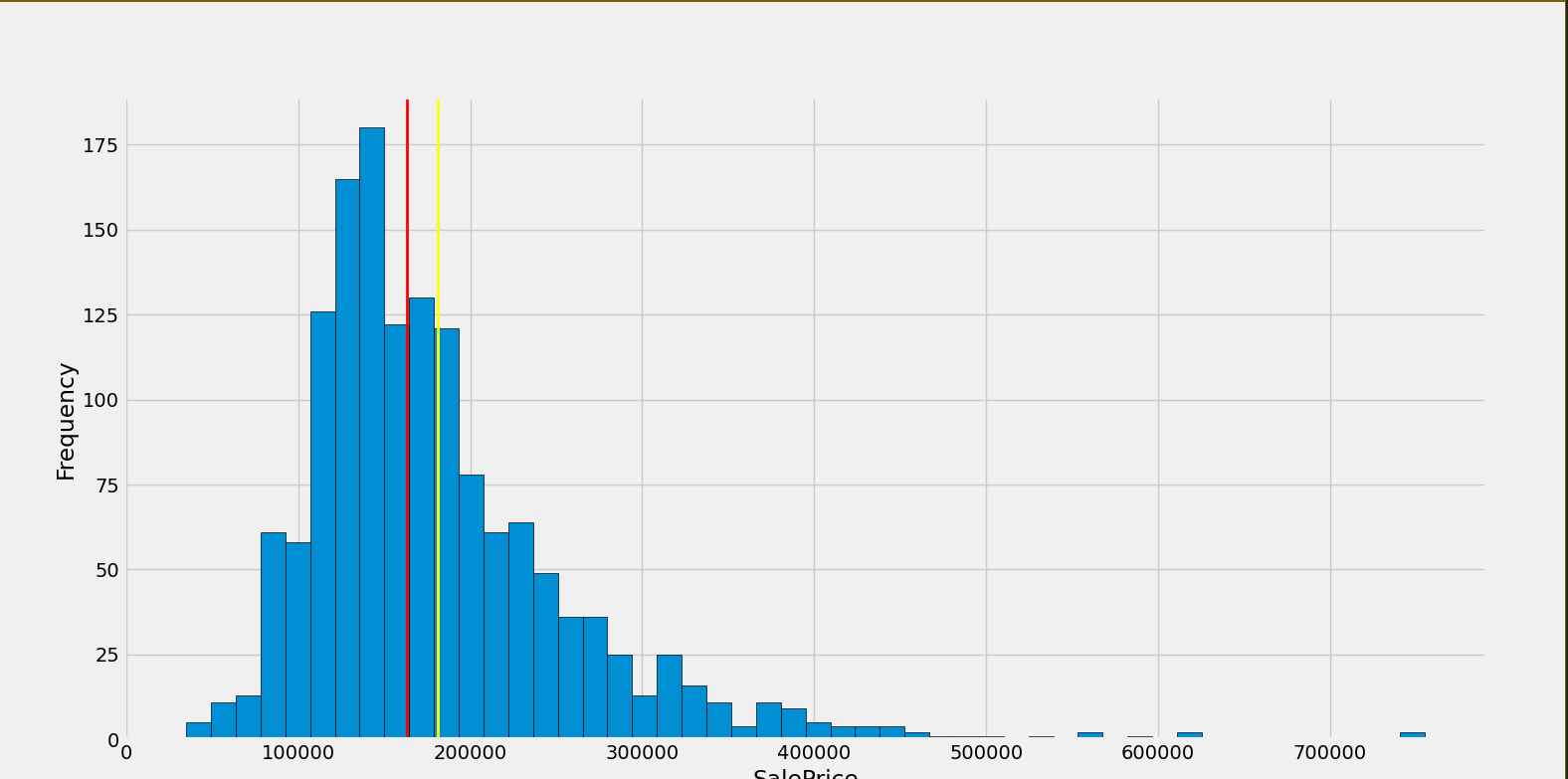
**SalePrice Description**

****

count 1460.000000

mean 180921.195890

std 79442.502883

min 34900.000000

25% 129975.000000

50% 163000.000000

75% 214000.000000

max 755000.000000

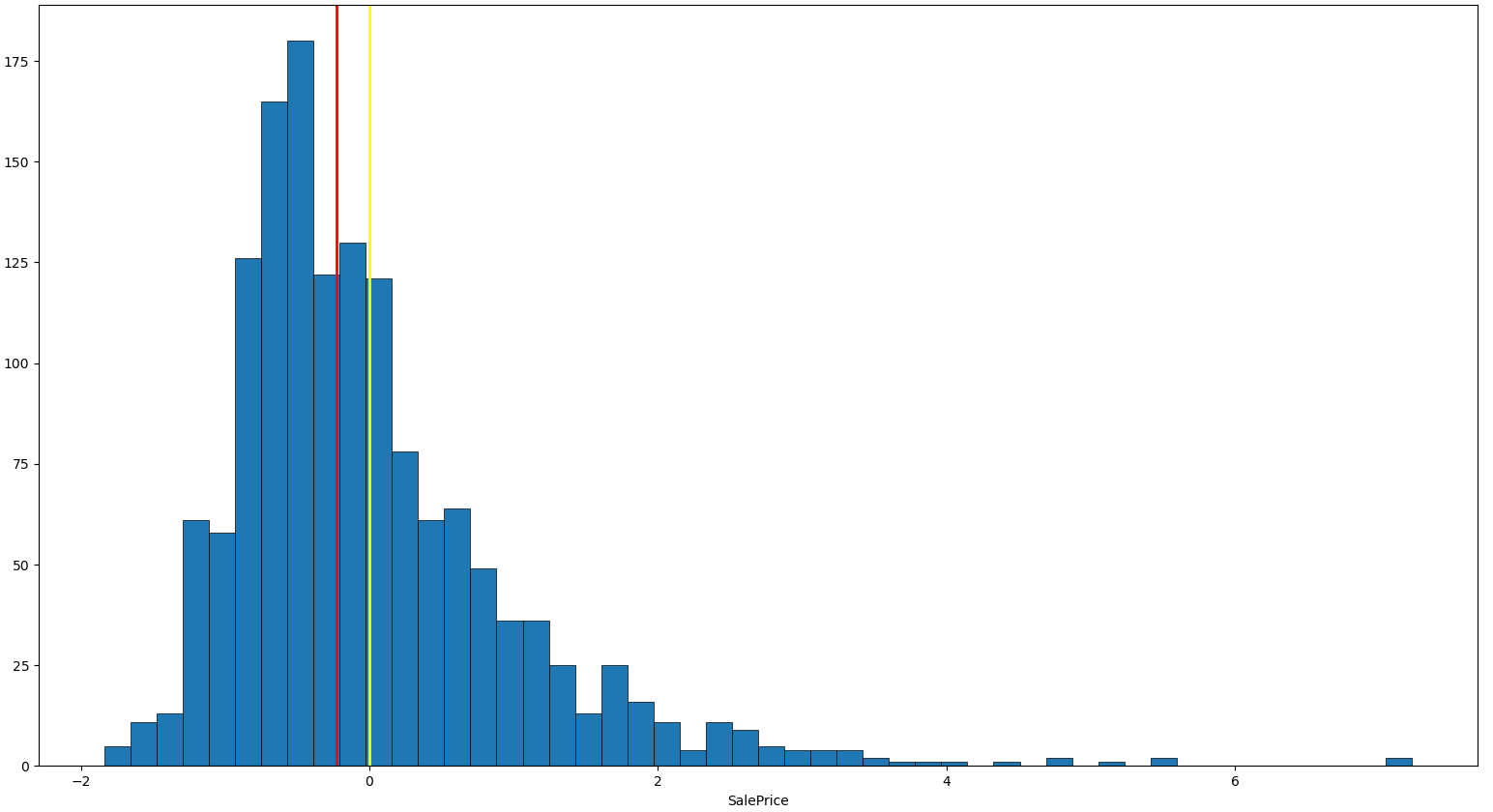
Skewness 1.8828757597682129

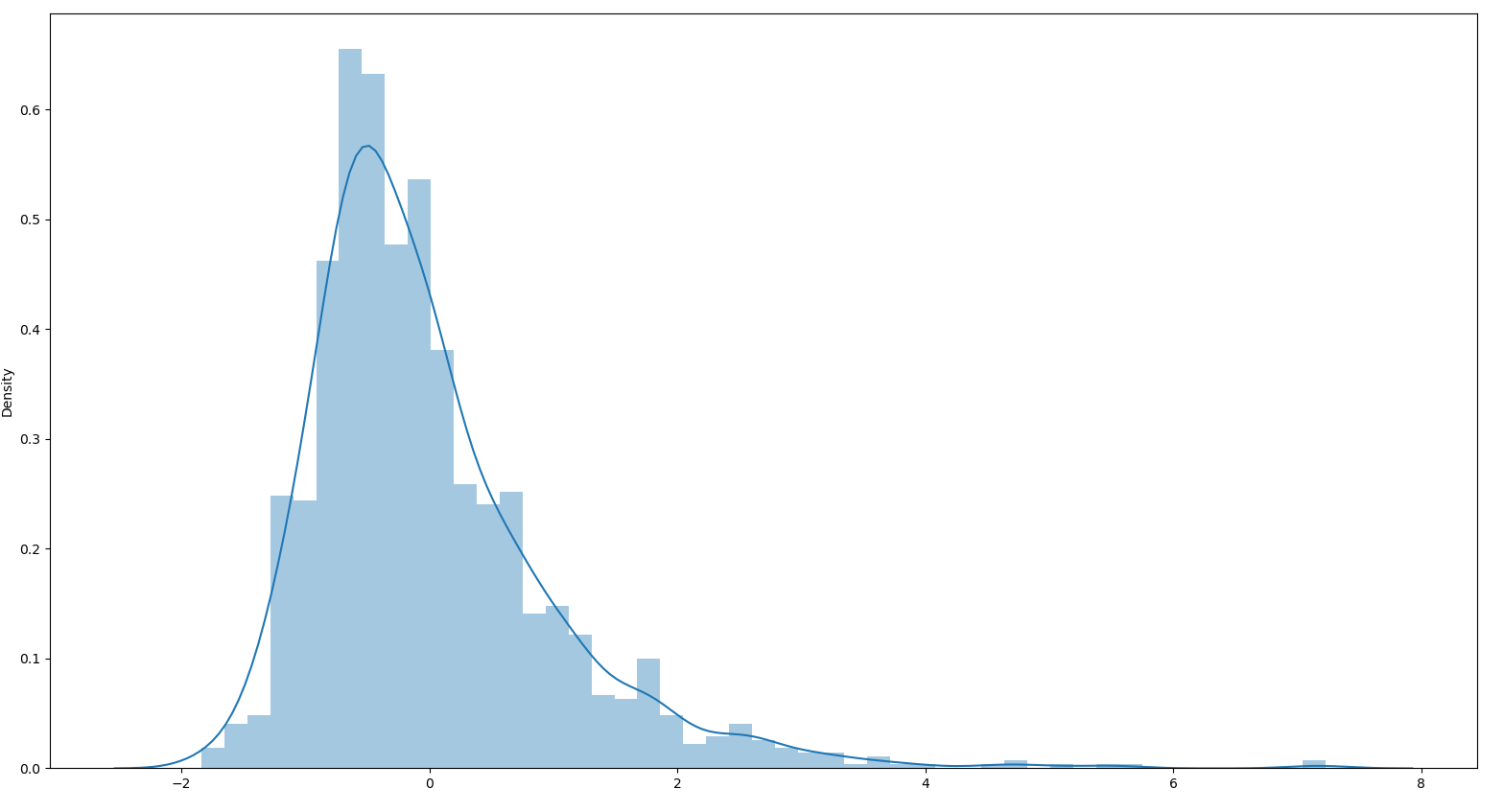
Kurtosis: 6.536281860064529

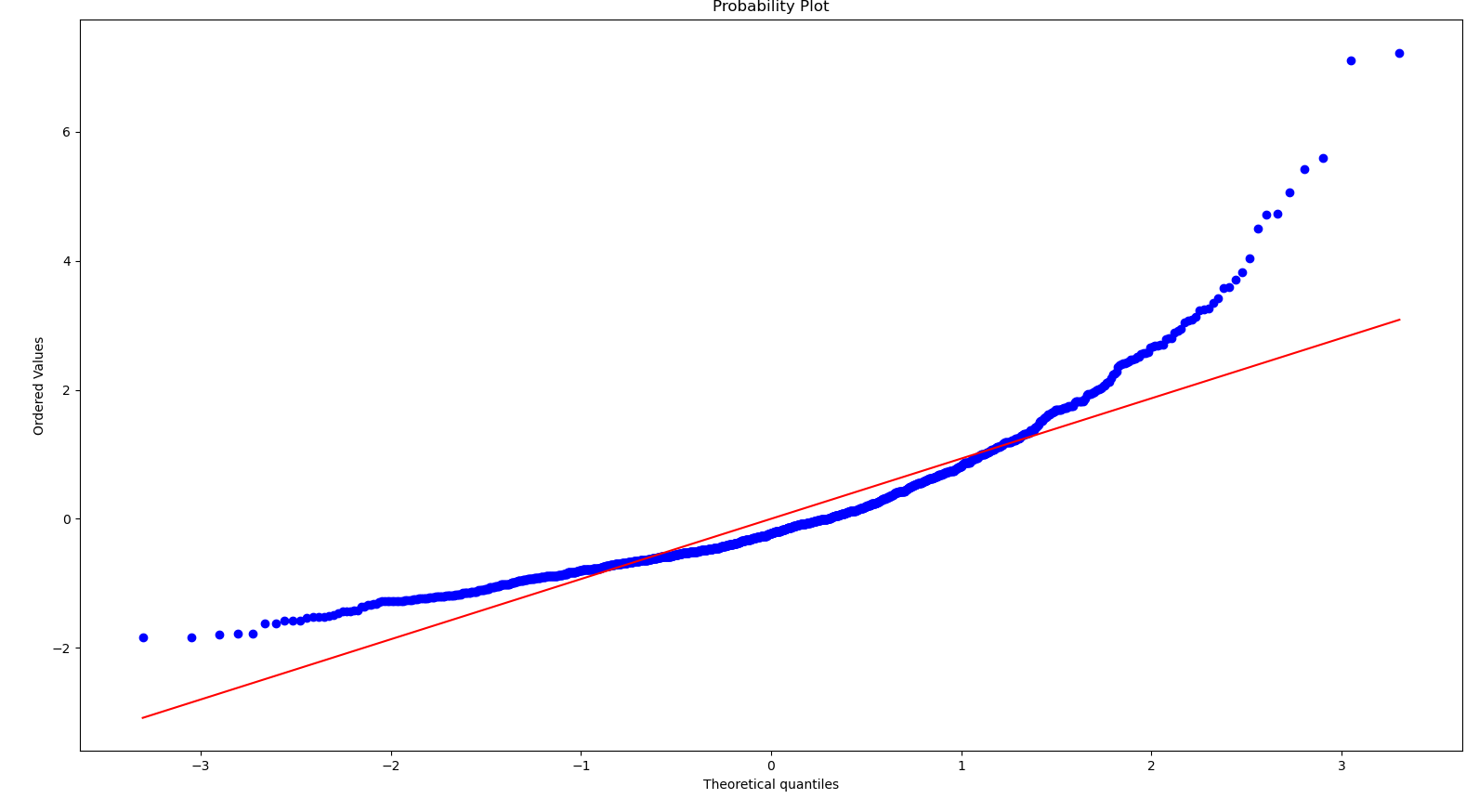
Conclusion:

1. Positive skewness
2. Show peakedness
3. Deviate from normal distribution

After Standard Scaling:







Description:

count 1.460000e+03

mean 1.362685e-16

std 1.000343e+00

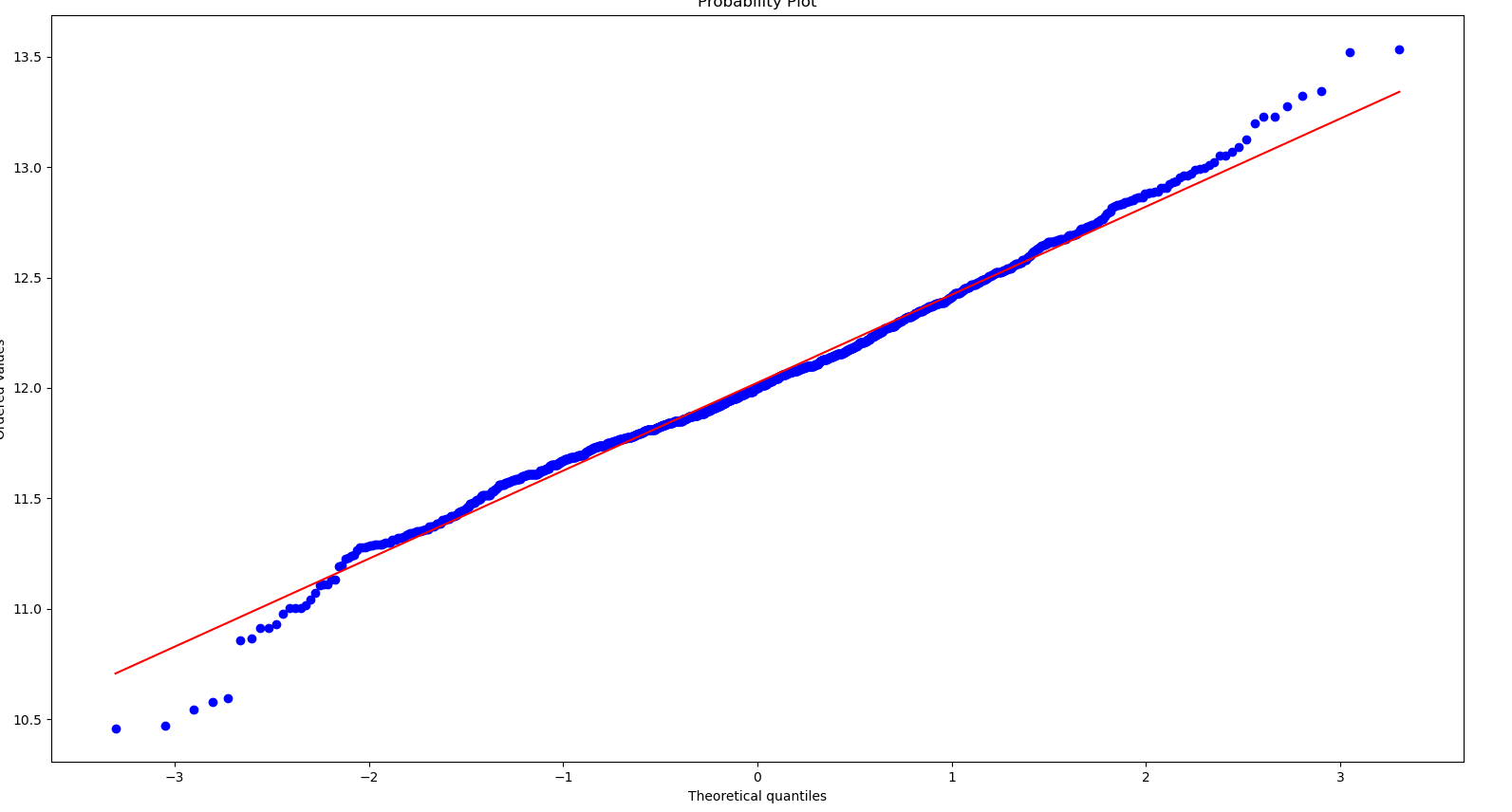
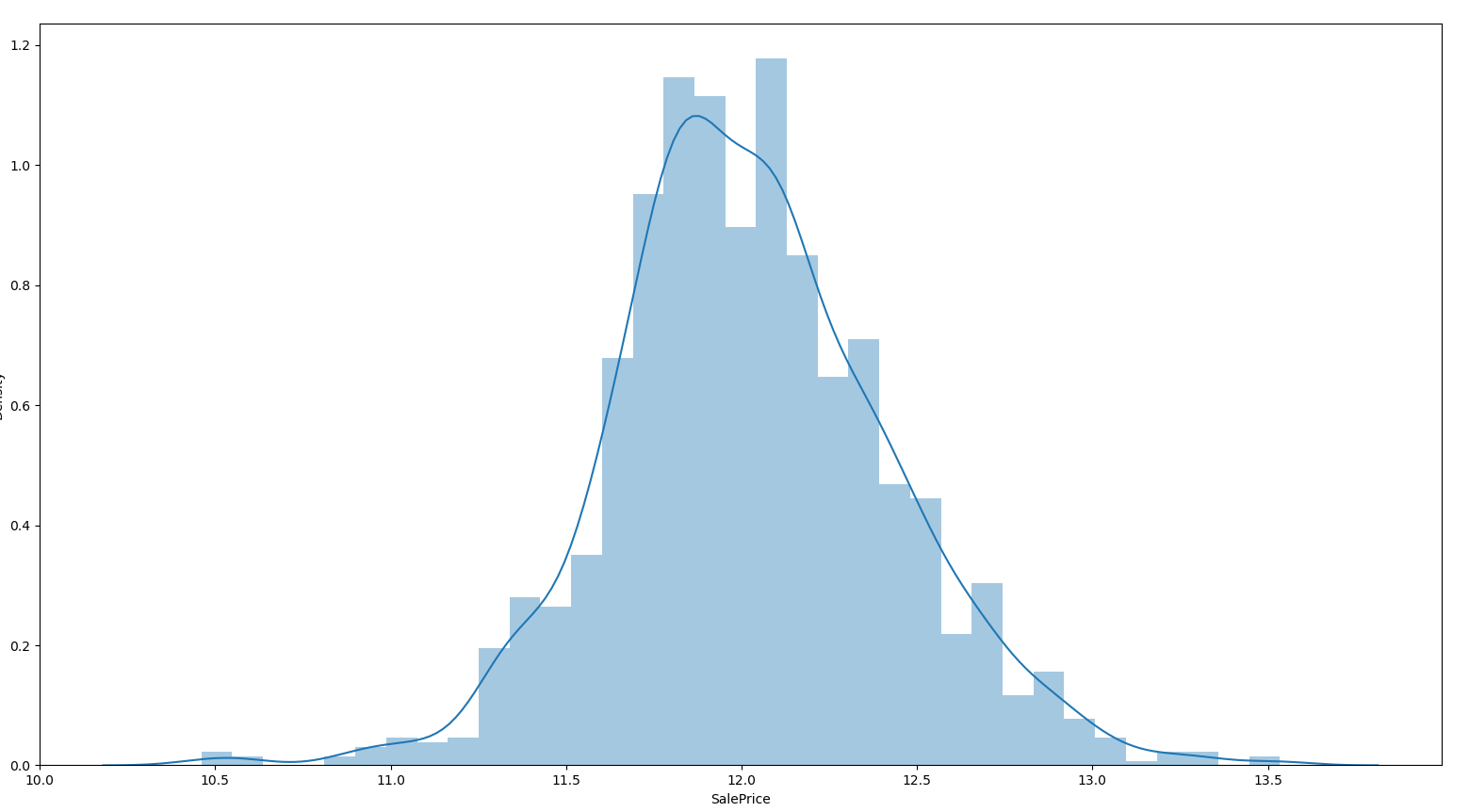
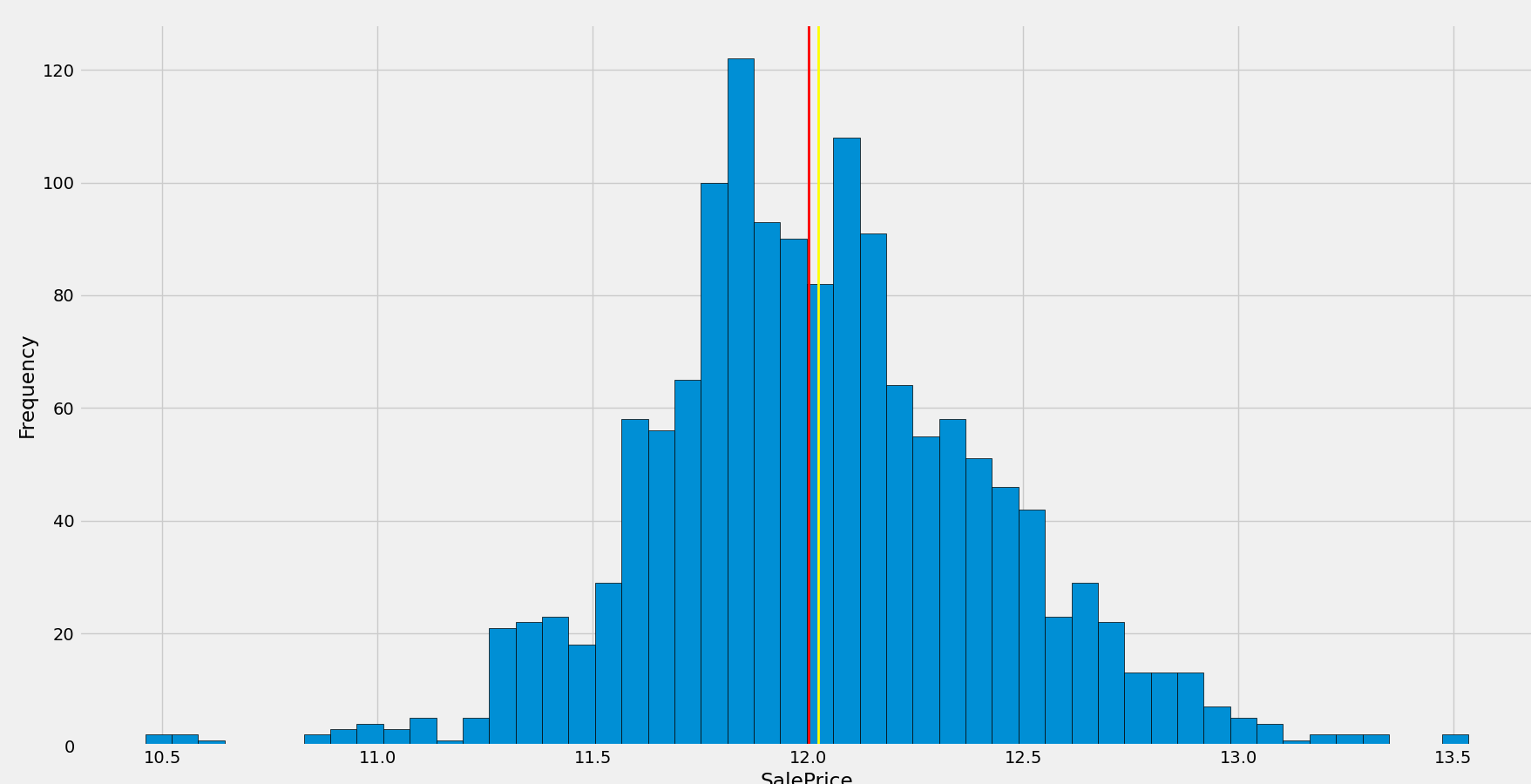
min -1.838704e+00

25% -6.415162e-01

50% -2.256643e-01

75% 4.165294e-01

max 7.228819e+00

After Only Log Transform

**After Log Transform:**

count 1460.000000

mean 12.024051

std 0.399452

min 10.460242

25% 11.775097

50% 12.001505

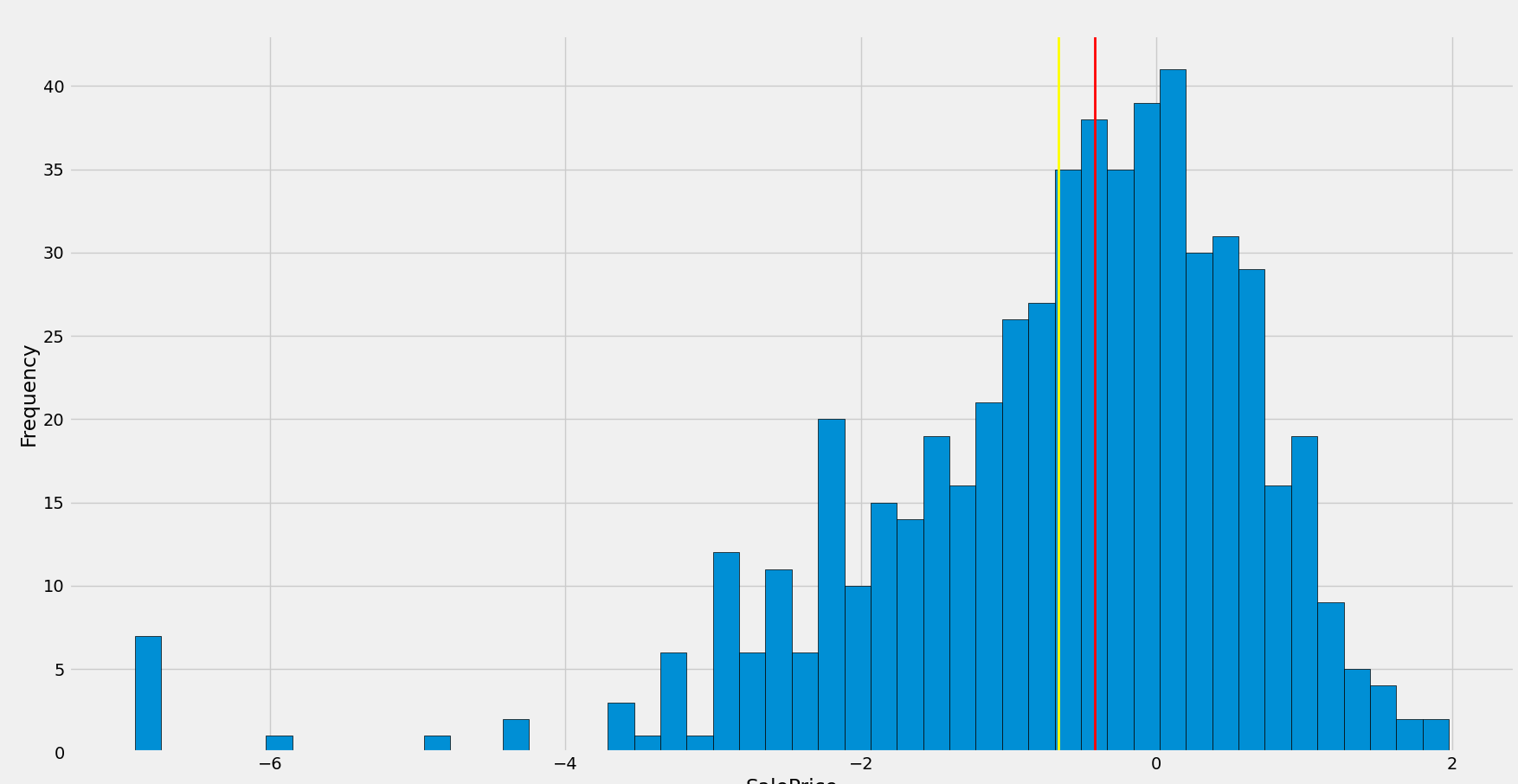
75% 12.273731

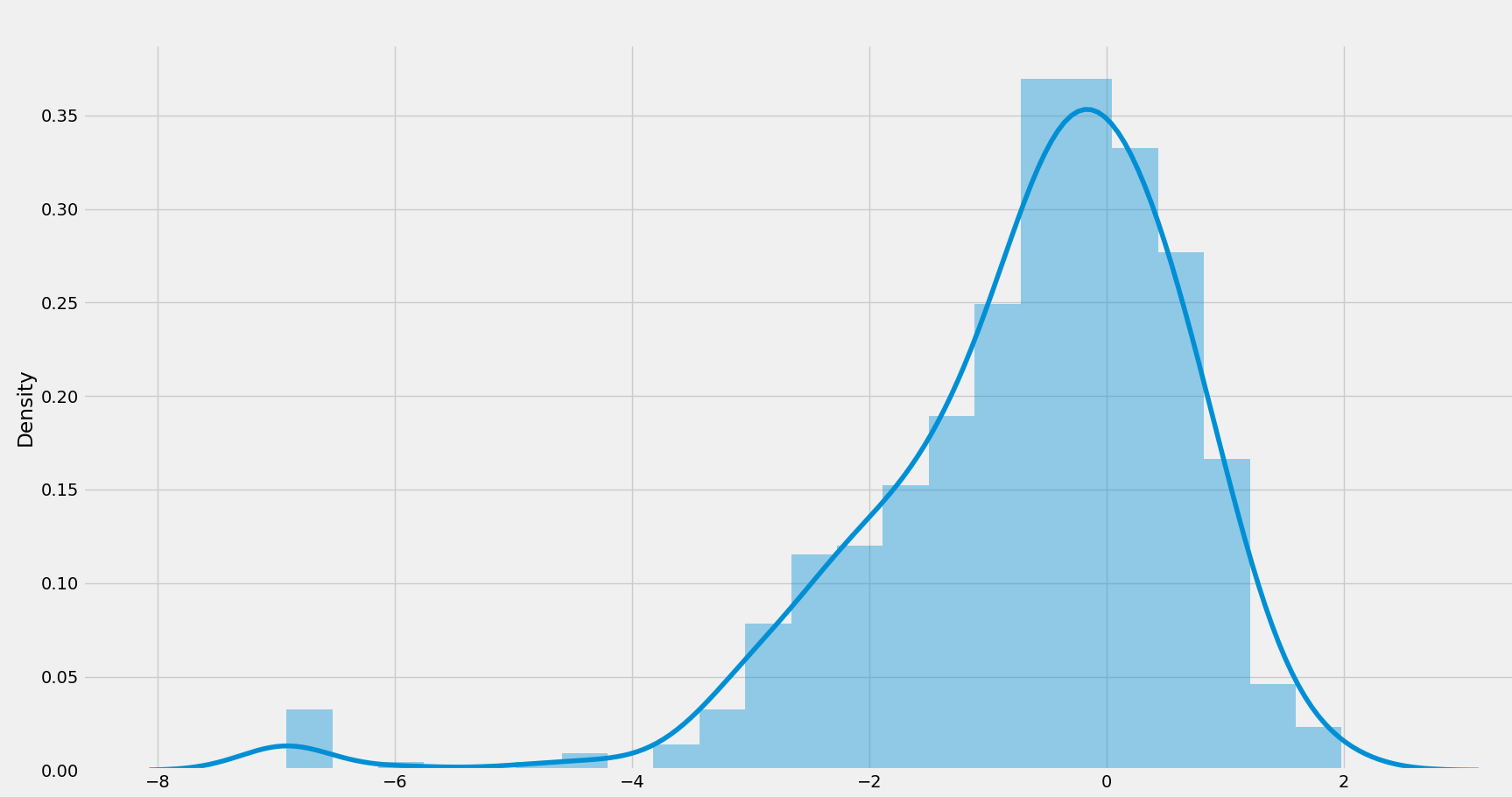
max 13.534473

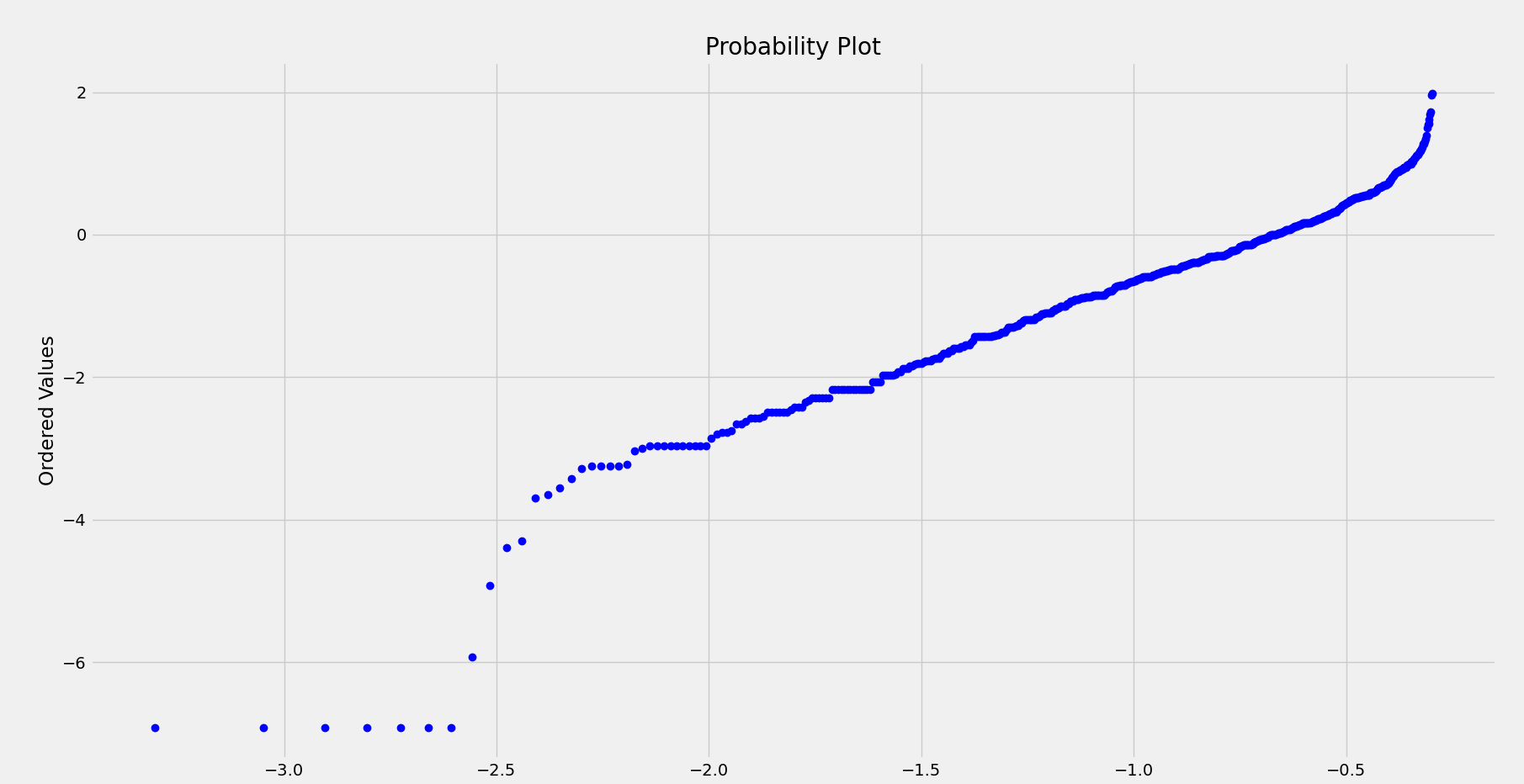
0.12133506220520406

0.8095319958036296

After Scaling and Log







**Description:**

count 560.000000

mean -0.664414

std 1.367256

min -6.915481

25% -1.375026

50% -0.418100

75% 0.223735

max 1.978076

-1.5412563509016897

4.444195398147645

26

This is called [the re-transformation problem](https://web.archive.org/web/20170712024346/https:/www.herc.research.va.gov/include/page.asp?id=cost-regression). I'm going to make your model a little simpler to talk about it:

lnY=β0+β1X1+β2X2+β3X22+ϵln⁡Y=β0+β1X1+β2X2+β3X22+ϵ

Now, that model does not make predictions for YY, it makes predictions for lnYln⁡Y. It is tempting to make predictions for YY by just taking predictions for lnYln⁡Y and exponentiating them like Y^=exp(lnYˆ)Y^=exp⁡(ln⁡Y^). This is wrong (i.e. biased), though:

lnYYE{Y|X}E^{Y|X}=β0+β1X1+β2X2+β3X22+ϵ=exp(β0+β1X1+β2X2+β3X22)exp(ϵ)=exp(β0+β1X1+β2X2+β3X22)E{exp(ϵ)|X}=exp(lnYˆ)E{exp(ϵ)|X}ln⁡Y=β0+β1X1+β2X2+β3X22+ϵY=exp⁡(β0+β1X1+β2X2+β3X22)exp⁡(ϵ)E{Y|X}=exp⁡(β0+β1X1+β2X2+β3X22)E{exp⁡(ϵ)|X}E^{Y|X}=exp⁡(ln⁡Y^)E{exp⁡(ϵ)|X}

The best predictor of YY is its expectation. If we could conclude that E{exp(ϵ)|X}=1E{exp⁡(ϵ)|X}=1, then we could just exponentiate like you are suggesting above. But Jensen's inequality says that since E{ϵ|X}=0E{ϵ|X}=0, it must be that E{exp(ϵ)|X}>1E{exp⁡(ϵ)|X}>1. So, we have to use some kind of adjustment. The adjustment is called Duan's Smearing Estimator. It is just the sample mean of the exponentiated prediction errors (residuals) from the original model, (1/N)∑exp(ei)(1/N)∑exp⁡(ei). So the right way to re-transform from the log model back to predictions on Y is:

Y^j=exp(lnYˆj)⋅1N∑i=1Nexp(ei)Y^j=exp⁡(ln⁡Y^j)⋅1N∑i=1Nexp⁡(ei)

To your questions. On the parameters, whether you need to re-transform depends on what you are trying to measure. The parameter β2β2 measures the amount that YY goes up (in percents) for a one unit increase in X1X1. So, if β2=0.04β2=0.04, that says that YY goes up 4% for each one unit X1X1 goes up. Similarly, for each unit X2X2 goes up, YY goes up β2+2β3X2β2+2β3X2 percent.

If you want to measure the amount that YY goes up *in units* when X2X2 goes up by one unit, then you need to re-transform:

Y∂Y∂X1E^{∂Y∂X1}=exp(β0+β1X1+β2X2+β3X22)exp(ϵ)=exp(β0+β1X1+β2X2+β3X22)exp(ϵ)β1=exp(lnYˆ)⋅1N∑exp(ei)⋅β1Y=exp⁡(β0+β1X1+β2X2+β3X22)exp⁡(ϵ)∂Y∂X1=exp⁡(β0+β1X1+β2X2+β3X22)exp⁡(ϵ)β1E^{∂Y∂X1}=exp⁡(ln⁡Y^)⋅1N∑exp⁡(ei)⋅β1

Notice that the answer depends on Y^Y^, totally unlike "regular" regression. You should expect this, though. The model is non-linear, so the derivative depends on the point of evaluation. For your more complicated model, you have to be careful to apply the chain rule properly---that is, where I have β1β1, you will have a complicated expression with ββs and powers of your various XXs and such.

For the confidence intervals, again, the question is what you are trying to measure. If you are happy with knowing how many percents YY goes up when X1X1 goes up by one, then the "regular" confidence intervals you get from the usual regression output are fine. If you want to measure the number of units that YY goes up when X1X1 goes up by one, then it's more complicated. Actually, it's very complicated in that case---you should use bootstrapping to do it. You can use something called the delta method, but it is a pain.

Root mean squared error of the prediction is easy to calculate, once you have re-transformed back to predicted YY:

RMSEP=1N−1∑(Y^i−Yi)2−−−−−−−−−−−−−−−−−√RMSEP=1N−1∑(Y^i−Yi)2

where Y^iY^i comes from the formula above.

5

You didn't give any details about why you think the outputs are wildly unlikely, but my guess is that your errors are not normally distributed. That "smearing adjustment" (bias correction) you're using is only valid if the errors are normal.

There is a more general smearing adjustment you can use, which is easy to implement. If I recall correctly, and I think I do, the steps are:

1. Compute exp(Xβ^)exp⁡(Xβ^), i.e. the retransformed but unadjusted prediction.
2. Regress YY against exp(Xβ^)exp⁡(Xβ^) without an intercept. Call the resulting regression coefficient γγ.
3. Compute the adjusted retransformed prediction as γexp(Xβ^)γexp⁡(Xβ^).

It's about the most intuitive thing you can do--forget the theory based on the normal distribution and just estimate the multiplier that gets the job done.