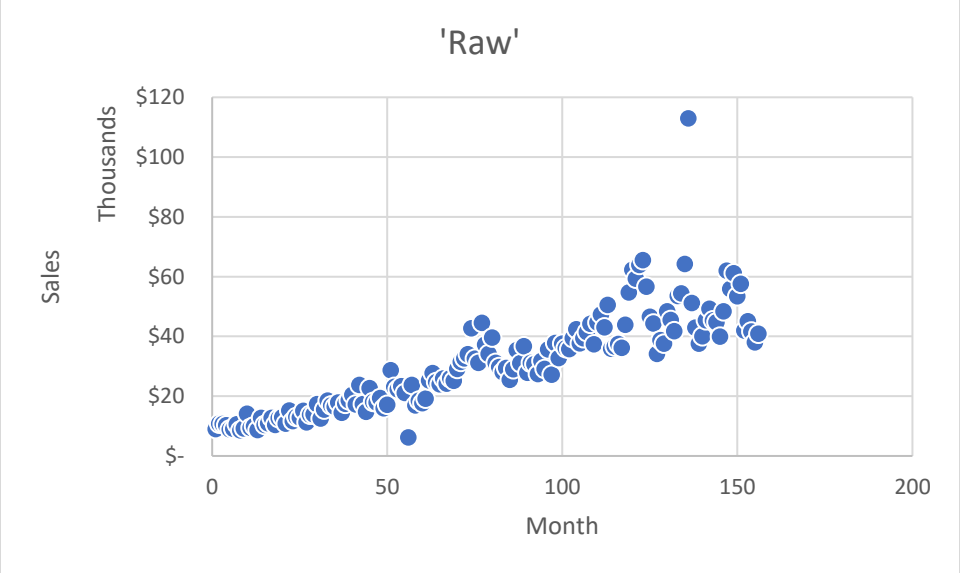
These data represent monthly sales of a product from a business. The data is numeric and only includes dates and dollar values. There are 157 months represented.

The initial plan for data exploration is to look for gaps in data (missing months or missing sales data,) duplicate values, and outliers.

Row Labels	Sum of Raw
Jan-09	$ 9,000
Feb-09	$ 10,800
Mar-09	$ 10,700
Apr-09	$ 10,300
May-09	$ 8,900
Jun-09	$ 9,100
Jul-09	$ 10,600
Aug-09	$ 8,600
Sep-09	$ 9,200
Oct-09	$ 14,100
Nov-09	$ 9,400
Dec-09	$ 10,000
Jan-10	$ 8,700
Feb-10	$ 12,700
Mar-10	$ 10,400

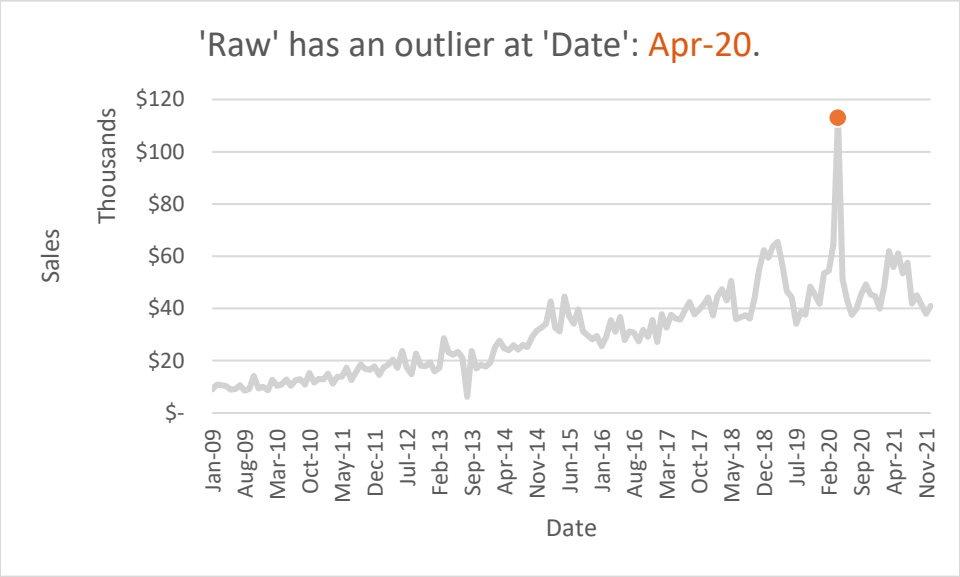


No gaps or duplicate values were found.

The 157th data record (January 2022) was removed so that only complete years (January to December) would remain. Date records were split into two columns, month and year, so that sorting and grouping by certain months across multiple years would be easier.

The data has outliers that are explainable with domain knowledge. The largest value was an exceptionally large one-time order. The smallest value was due to a system shutdown. The outliers are not representative of normal sales. The

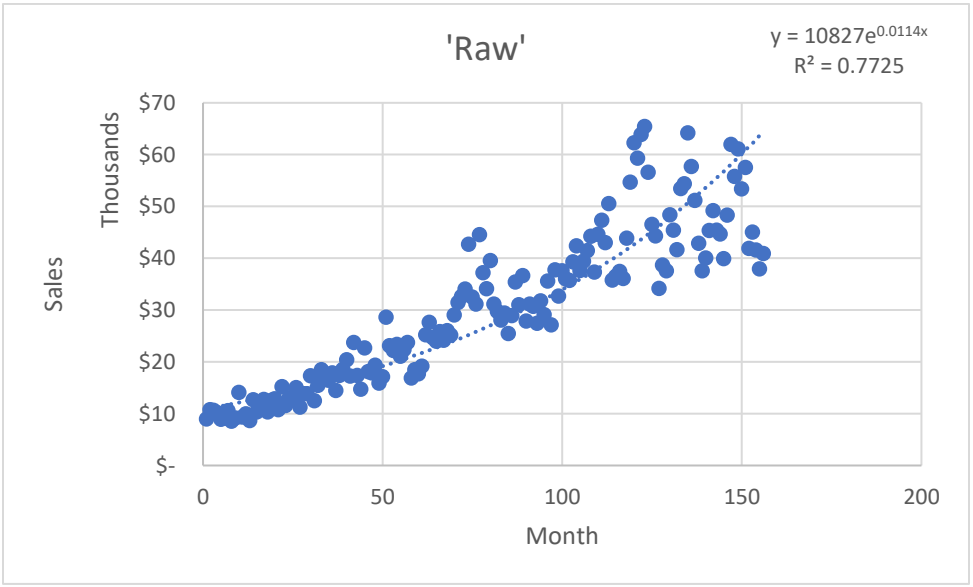
outlier values were replaced by averaging the sales of the month preceding and the month immediately following the outlier month.

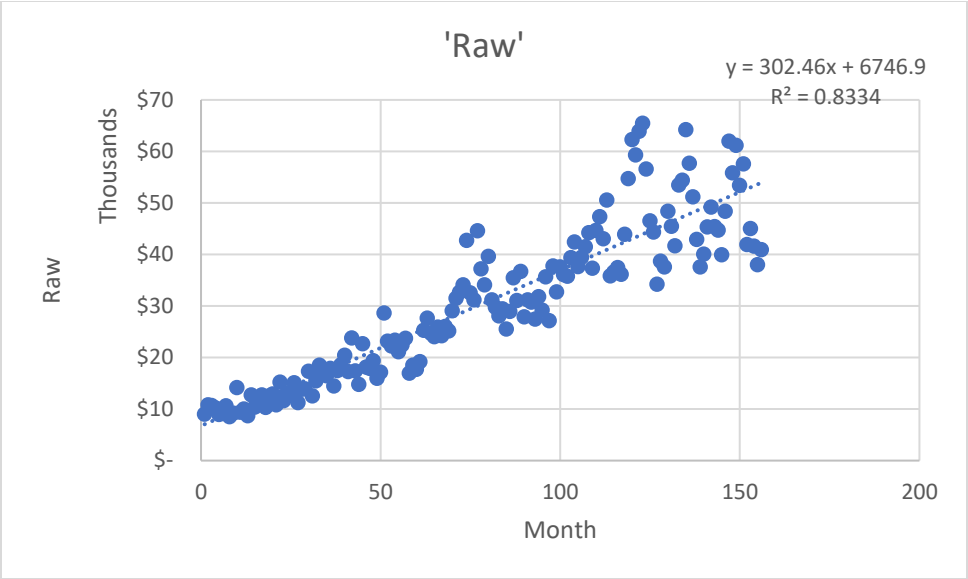


There are three hypotheses to test:

- Is variance in the last three years larger than in preceding years. The graph suggests this is so, but is it true?
- January is believed to be a typically slow sales month. Do January sales average less than the rest of the year?
- Are any of the months of the year reliably strong or weak in terms of sales?

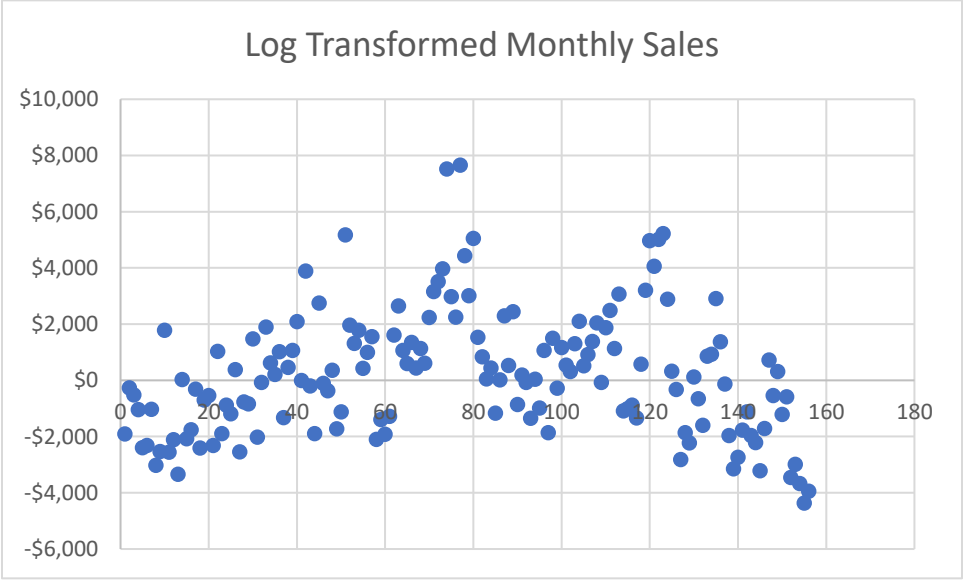
Trendlines were fit to see if the data was better represented by an exponential growth curve or by linear regression. The thought is that the data should grow exponentially because of the long time series (over 10 years) if for no other reason than price inflation.



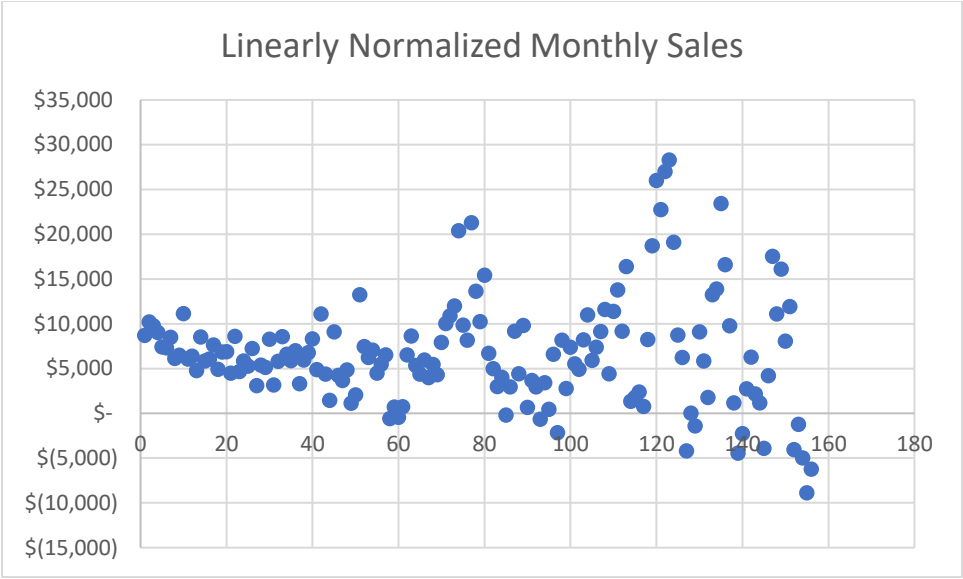


The R-squared value for linear regression is slightly better than for exponential regression.

The data was normalized for growth over time. A log (ln) transformation was used when considering the data as exponential. When treating the data as linear, the growth factor ($302.46 per month sales increase) was subtracted out to normalize sales to "flat."

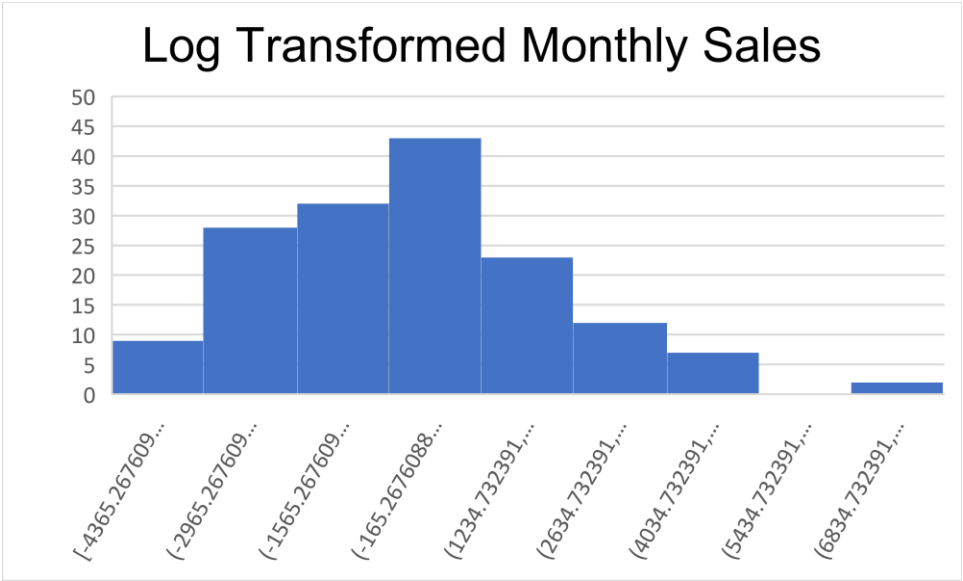


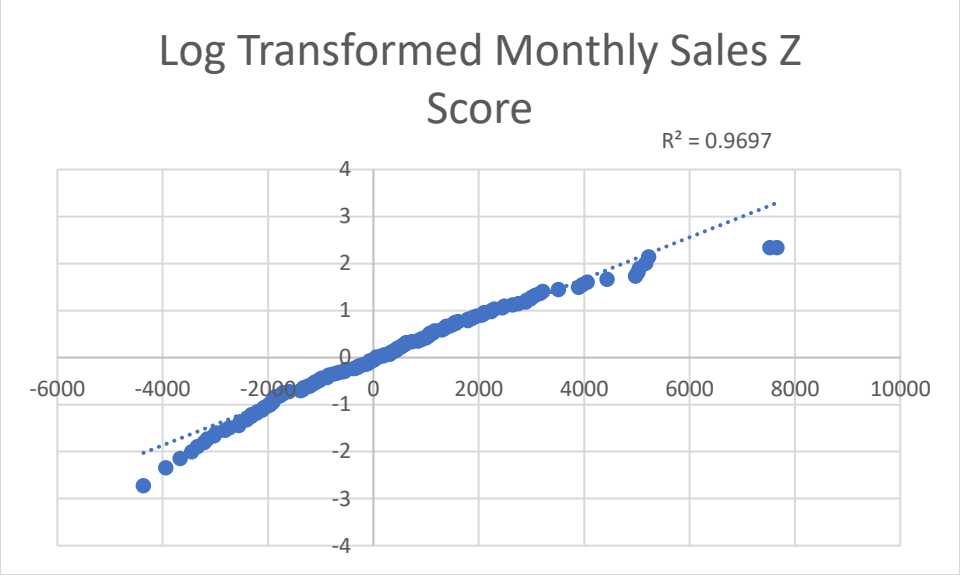
This shows what the sales would have looked like had there been no growth over time, assuming the growth was exponential.

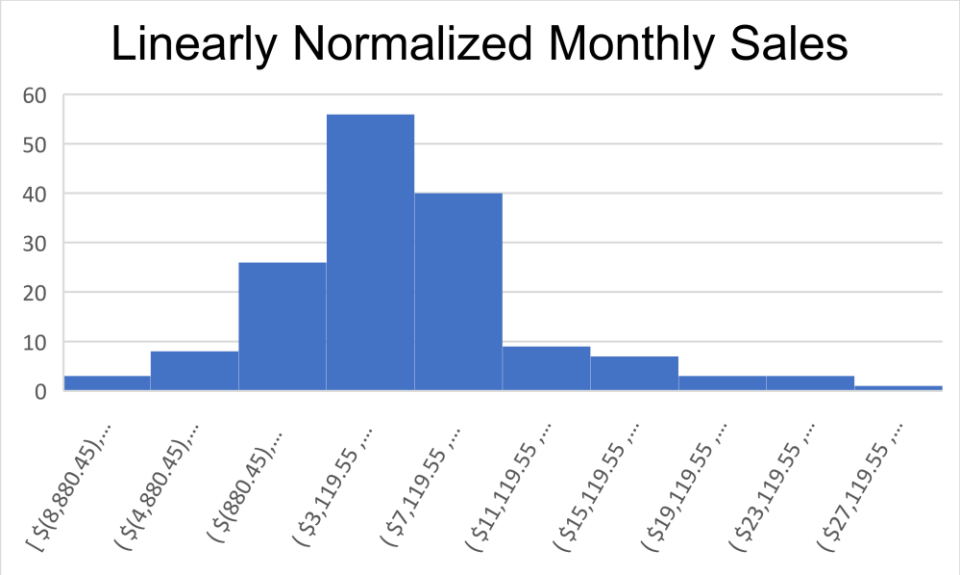


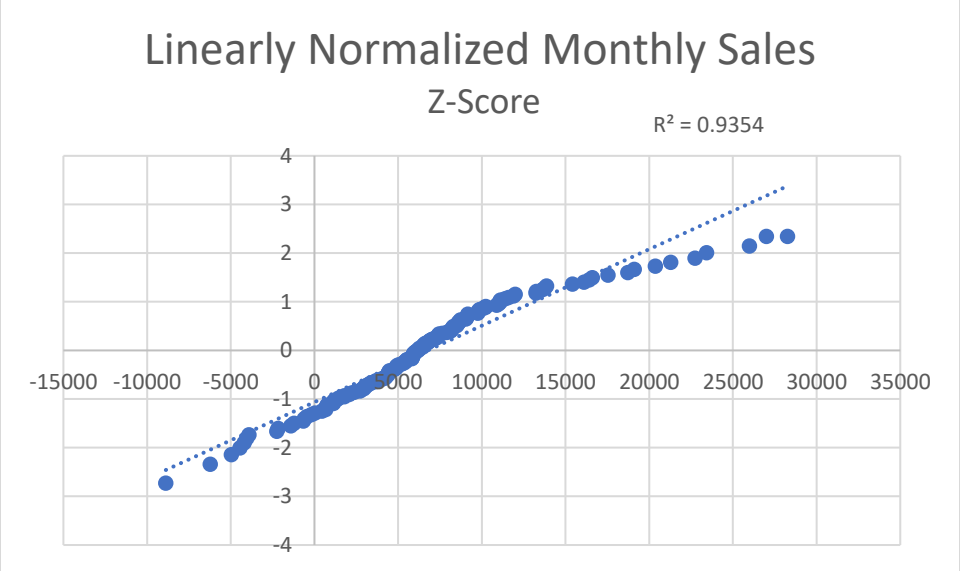
This shows what the sales would have looked like had there been no growth over time, assuming the growth was linear.

A histogram and a QQ plot show that the normalized data is in fact, normally distributed (though slightly right skewed) in both cases (treating the raw data as showing exponential growth and treating it as though it shows linear growth.) The log transformed data is slightly more normal, as shown by the slightly larger R-squared value on the QQ plot.




Log Transformed Monthly Sales Z Score
R² = 0.9697
4
3
2
1
0
-1
-2
-3
-4
-6000
-4000
-2000
0
2000
4000
6000
8000
10000


Linearly Normalized Monthly Sales
60
50
40
30
20
10
0
[$(8,880.45),...
($(4,880.45),...
($(880.45),...
($3,119.55 ,...
($7,119.55 ,...
($11,119.55 ,...
($15,119.55 ,...
($19,119.55 ,...
($23,119.55 ,...
($27,119.55 ,...


Linearly Normalized Monthly Sales
Z-Score
R² = 0.9354
4
3
2
1
0
-1
-2
-3
-4
-15000
-10000
-5000
0
5000
10000
15000
20000
25000
30000
35000

From this point onward, we will only consider exponential growth and the log transformed data. Exponential and Linear growth models seem to fit well, but considering the nature of business and inflation, over a longer time period, exponential growth is more likely.

Hypothesis 1: The normalized monthly sales in the last three years has more variance than the preceding years.

The sales data seems to "fan out" as time progresses. Is there really more variance in the last three years, or is it an artifact of growth? A 50% swing in sales that have grown to a million dollars a month is more visible than a 50% swing when sales are just one hundred dollars a month. We will test for a difference in variance with normalized data to isolate sales swings from growth.

We use a Two-Sample F-Test to see if the variances are equal when comparing the first ten years of sales to the last three years.

The null hypothesis is that the variances are equal. We test at a 95% level (alpha = .05)

F-Test Two-Sample for Variances		
	last 3 years	*first 10 years*
Mean	-706.3969796	487.7979913
Variance	6002072.496	4265959.154
Observations	36	120
df	35	119
F	1.40696905	
P(F<=f) one-tail	0.090458846	
F Critical one-tail	1.522087455	

Since F Critical one-tail is greater than F, we cannot reject the null hypothesis, so we assume that the variances in the *normalized data* ARE in fact equal and that what looks to be increasingly larger spreads between strong and weak sales months is accounted for by sales growth.

Hypothesis 2: January is typically a slow sales month.

Normalized data must be used to compare January to the other 11 months (combined.) If raw data is used, even a "slow" month in 2020 will dwarf a "strong" month in 2010 because of sales growth.

A T-Test will be used to compare the means of all the Januarys against all the other months.

An F-Test is first used to see if the variances in the two populations are equal. The null hypothesis is that the variances are equal.

F-Test Two-Sample for Variances		
	Januarys	*Other Months*
Mean	-629.0287245	288.6911966
Variance	5456799.451	4800674.607
Observations	13	143
df	12	142
F	1.136673467	
P(F<=f) one-tail	0.335442126	
F Critical one-tail	1.820927215	

Since F Critical one-tail > F, we cannot reject the null hypothesis. We assume the variances are equal. Alpha = .05

t-Test: Two-Sample Assuming Equal Variances		
	Januarys	*Other Months*
Mean	-629.0287245	288.6911966
Variance	5456799.451	4800674.607
Observations	13	143
Pooled Variance	4851801.219	
Hypothesized Mean Difference	0	
df	154	
t Stat	-1.438255539	
P(T<=t) one-tail	0.076195329	
t Critical one-tail	1.654808385	
P(T<=t) two-tail	0.152390659	
t Critical two-tail	1.975488058	

Since -t Critical two-tail < t Stat < t Critical two -tail (-1.975 < -1.438 < 1.975) we are unable to reject the null hypothesis at the 95% confidence level. We conclude that the idea that January is typically a slow sales month is not true.

Just for fun, the T-test was rerun with different values for Alpha. We can conclude that January IS in fact a slow sales month, but only at the 80% confidence level.

Hypothesis 3: There are strong and weak sales months

Similar to the method used to examine January, the means of each individual month could be tested against all other months to see if there are any months that are reliably strong or weak in terms of sales. This could hint at a cyclic nature to the sales.

Summary of Data Quality:

This data was high quality, likely because it was so simple. It was from one source, represented sales of a single product, and covered a relatively short time frame (13 years.) There were no missing or duplicate records. The outliers were real data, not anomalies. The outliers were removed in order to get a more accurate sense of the sales performance of the product, simply because the events that were tied to the outliers were one-time events and not expected to repeat.

Future Study:

There is quite a lot of variance in the sales of this product. Since the data is of good quality and is reliable, the variance is real. A next step would be to get sales data by customer to see if the wide swings in sales are the result of varying orders from a single customer (or a few customers) that get added to steady sales of the product to a core customer base. If by-customer data does not show any significant variability, we should investigate to see if it's cyclic in any way. We could try grouping the current data by quarters or by years to see if there are repeating cycles. We could try correlating the sales data to public data sets on stock market price and economic activity to see if these product sales are simply driven by macroeconomic forces.