Math IA1 - Final

Formulate

Introduction

Keeping a business profitable requires, depending on the type of business, the owner to have great foresight into what the consumers want, and how much to sell it for depending on the demand. The difficult part of this is knowing how much demand there will be. This is why approximating the future is a valuable skill in the business world, this can be achieved relatively simply through mathematical means, and you can end up with an accurate mathematical representation of the future in regards to a businesses sales. In this Problem-Solving and modelling task, a method to create 2 mathematical models will be found. Both of which will attempt to forecast the profit for the client, given the clients data on the past year of sales. One model will approximate the **annual sales figures through to 2027**, and the other **monthly sales figures for 2024**.

Translation

Linear line of best fit:

This the is the simplest and most effective way to create forecasting models, it is found using linear regression. Following the equation: $y = a + mx$.

Residual

A residual is the vertical distance between a data point and the regression line, in the process of finding the line of best fit, the sum of squared **residuals** is minimized.

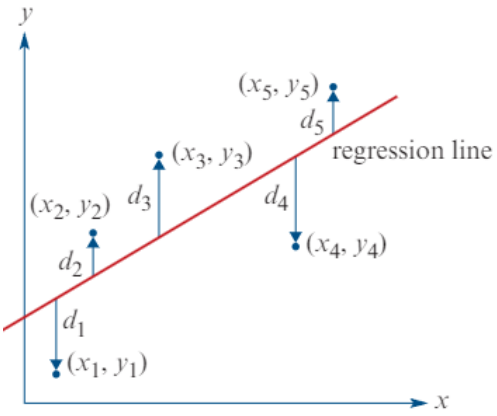


Figure 1, a graph showing a regression line, with the blue arrows showing residuals.

The Coefficient of Determination

(r^2), this represents the variance of a data set, $-1 < r^2 < 1$. The smaller r^2 is, the less the data follows a trend, the opposite for higher values.

Seasonalized data

Data is seasonalized if the data regularly undergoes predictable changes every year (for example, could be any time period).

Seasonal Indices

A score that shows how each annum differs from the average of all the seasons. It is found with:

$$SI = \frac{value\ for\ annum}{annum\ average}$$

Deseasonalization

Seasonalized data can be deseasonalised through the process of deseasonalisation. This process reveals the underlying trend of the data, and increases the value of the r^2 value. This in combination with the line of best fit, creates a great model to forecast with. In short, it is found with the following formula:

$$deseasonalised\ figure = \frac{actual\ figure}{SI}$$

Figure 2, a graph showing actual data vs deseasonalised data

Reseasonalisation

Once the forecasted data is found, the process of reseasonalisation reintroduces the seasonal effect. Completed by:

$$reseasonalized\ value = deseasonalized\ figure * SI$$

Excel Functions

Throughout the creation of the models, several excel functions will be used:

- `=SUM` , finds the sum of all selected rows.
- `=AVERAGE` , finds the average of all selected rows.
- `=LINEST(ys, [xs])` , given all the y, x values, will return the m, and a values of a line of best fit function, used in the equation $y = a + mx$.

Assumptions

- **Accurate Data**, it is assumed that the provided data is accurate, this directly affects the accuracy of the forecast.
- **Tax is included**, it is assumed that the sales data provided includes GST taxes. This assumption is important due to the fact that if it wasn't assumed, the forecast would be less than or greater than reality.
- **Tax variations**, it is assumed that tax remains the same, or at least on the same growth rate as included in the dataset. If not, the sales could decrease over time, but it is just tax going up.
- **Global Events**, it is assumed that no major global events will occur, that will directly affect the economy of the business location/nation, for example, COVID-19.
- **Australian Dollar**, it is assumed that the money used in the dataset is Australian Dollar ($), the Australian dollar is fairly stable, compared to other economies, this will ensure that sales don't vary.

Observations:

When plotting the provided data, the following observations were made:

- There is an outlier in the 19 month (July of 2020), this can affect the outcome of the forecast, as the least squares regression method changes it's output drastically if there is an outlier (due to the squaring of each value).
- It appears that the data is seasonalized, as seen by the constant fluctuations of the data in 1 year.
- Despite the seasonality, the overall trend appears to be upward.
- Finding the r^2 (correlation coefficient) of the dataset returns 0.21270961. This is a low correlation, this justifies the need for deseasonalization.

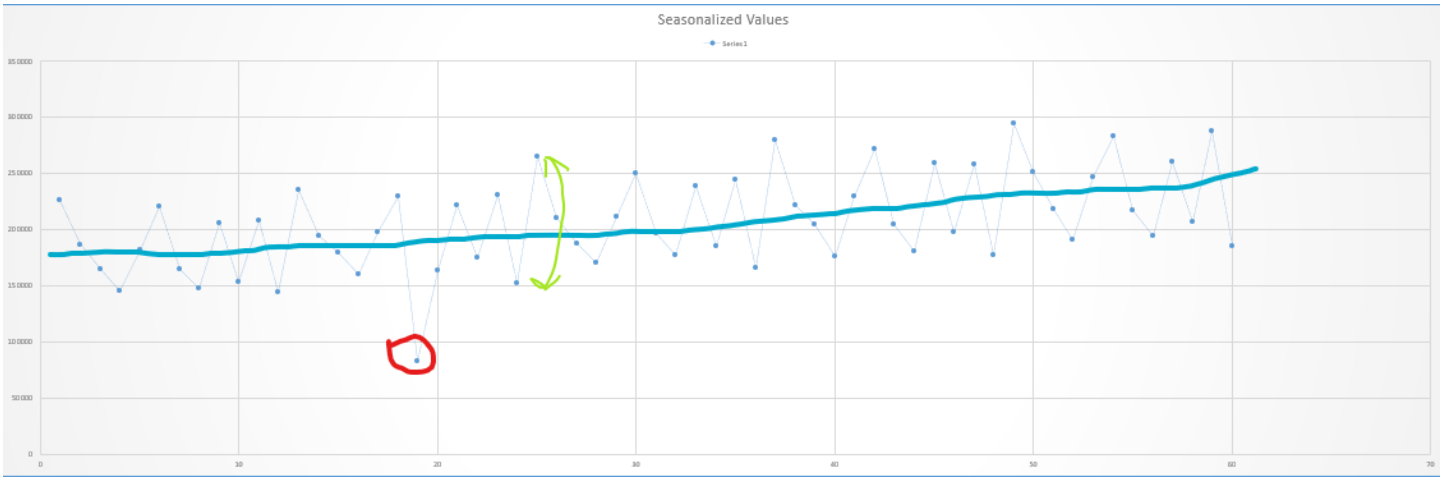


Figure 3, graph of the provided dataset

Solve

Dataset (2024.6)

	2019	2020	2021	2022	2023
Jan	$226,051.00	$235,132.00	$264,672.00	$279,628.00	$ 294,869.00
Feb	$186,945.00	$193,956.00	$210,287.00	$222,171.00	$ 251,780.00
Mar	$165,287.00	$179,832.00	$187,960.00	$205,072.00	$ 218,277.00
Apr	$145,315.00	$160,067.00	$170,631.00	$176,016.00	$ 191,340.00
May	$181,330.00	$198,298.00	$211,650.00	$229,360.00	$ 246,670.00
Jun	$220,043.00	$230,078.00	$249,996.00	$272,408.00	$ 282,702.00
Jul	$164,915.00	$ 83,033.00	$196,810.00	$205,079.00	$ 217,071.00
Aug	$148,049.00	$164,049.00	$176,987.00	$181,075.00	$ 194,564.00
Sep	$205,234.00	$221,707.00	$238,620.00	$258,855.00	$ 260,684.00
Oct	$153,741.00	$175,250.00	$185,033.00	$198,296.00	$ 206,363.00
Nov	$208,036.00	$230,961.00	$244,896.00	$258,120.00	$ 287,794.00
Dec	$144,515.00	$152,046.00	$165,743.00	$177,314.00	$ 185,535.00

Technology

Whilst the calculations discussed in Translation can be done by hand, **Excel** makes this process much easier for large datasets, such as provided dataset. The data from the provided Dataset can be copied directly into excel, and modified on a large scale, graphed and even find the line of best fit.

Working

Due to the seasonality of the provided dataset, finding a line of best fit for the data will not provide the optimal forecast. The data can be optimized in two ways:

1. Deseasonalisation - discussed more in Translation
2. Removing outliers - The provided dataset contains an outlier, which will affect the forecast heavily.

The deseasonalisation process:

1. Find the average value for each year, found by `=AVERAGE(B2:M2)`, where the `2` value is variable. (*Figure 4*)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Avg
2	2019	226051	186945	165287	145315	181330	220043	164915	148049	205234	153741	208036	144515	179122
3	2020	235132	193956	179832	160067	198298	230078		164049	221707	175250	230961	152046	194671
4	2021	264672	210287	187960	170631	211650	249996	196810	176987	238620	185033	244896	165743	208607
5	2022	279628	222171	205072	176016	229360	272408	205079	181075	258855	198296	258120	177314	221950
6	2023	294869	251780	218277	191340	246670	282702	217071	194564	260684	206363	287794	185535	236471

Figure 4, finding the average month value for each year

2. Now the SI's (Seasonal Index) for each month can be found by $\frac{actual\ value}{average\ month\ value}$, or in excel `=B2/$N2`. This is completed for all months in the dataset, with their respective average, and value. (*Figure 5*)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Avg
2	2019	226051	186945	165287	145315	181330	220043	164915	148049	205234	153741	208036	144515	179122
3	2020	235132	193956	179832	160067	198298	230078		164049	221707	175250	230961	152046	194671
4	2021	264672	210287	187960	170631	211650	249996	196810	176987	238620	185033	244896	165743	208607
5	2022	279628	222171	205072	176016	229360	272408	205079	181075	258855	198296	258120	177314	221950
6	2023	294869	251780	218277	191340	246670	282702	217071	194564	260684	206363	287794	185535	236471
7	SI's													
8	2019	=B2/$N2	1.04368	0.92276	0.81126	1.01233	1.22845	0.92069	0.82653	1.14578	0.8583	1.16142	0.8068	
9	2020	1.20785	0.99633	0.92378	0.82225	1.01863	1.18188		0.8427	1.13888	0.90024	1.18642	0.78104	
10	2021	1.26876	1.00805	0.90102	0.81795	1.01459	1.19841	0.94345	0.84842	1.14387	0.88699	1.17396	0.79452	
11	2022	1.25987	1.001	0.92396	0.79305	1.03339	1.22734	0.92399	0.81584	1.16628	0.89343	1.16297	0.79889	
12	2023	1.24696	1.06474	0.92306	0.80915	1.04313	1.19551	0.91796	0.82278	1.10239	0.87268	1.21704	0.7846	

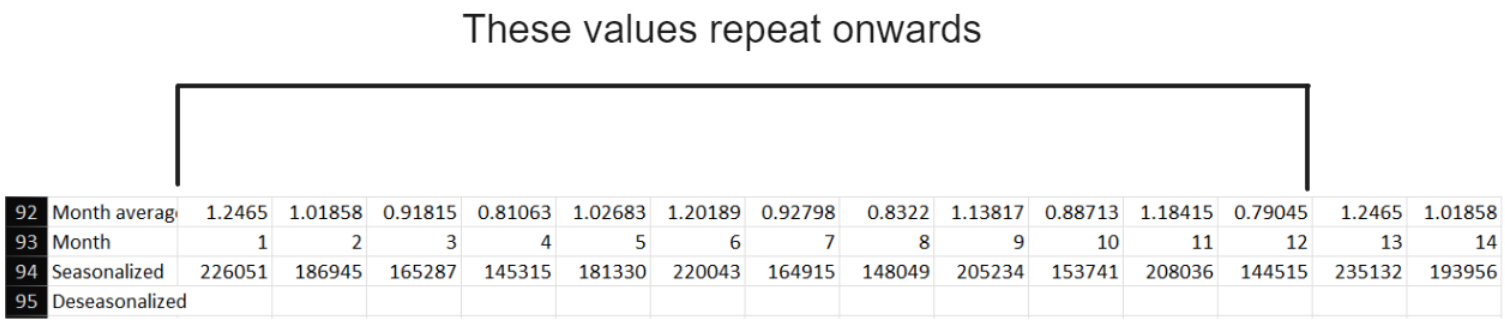
Figure 5, finding the SI's for each month

3. The average for each months' SI can be found with `=AVERAGE(B8:B12)`. (*Figure 6*)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Avg
2	2019	226051	186945	165287	145315	181330	220043	164915	148049	205234	153741	208036	144515	179122
3	2020	235132	193956	179832	160067	198298	230078		164049	221707	175250	230961	152046	194671
4	2021	264672	210287	187960	170631	211650	249996	196810	176987	238620	185033	244896	165743	208607
5	2022	279628	222171	205072	176016	229360	272408	205079	181075	258855	198296	258120	177314	221950
6	2023	294869	251780	218277	191340	246670	282702	217071	194564	260684	206363	287794	185535	236471
7	SI's													
8	2019	1.262	1.04368	0.92276	0.81126	1.01233	1.22845	0.92069	0.82653	1.14578	0.8583	1.16142	0.8068	
9	2020	1.20785	0.99633	0.92378	0.82225	1.01863	1.18188		0.8427	1.13888	0.90024	1.18642	0.78104	
10	2021	1.26876	1.00805	0.90102	0.81795	1.01459	1.19841	0.94345	0.84842	1.14387	0.88699	1.17396	0.79452	
11	2022	1.25987	1.001	0.92396	0.79305	1.03339	1.22734	0.92399	0.81584	1.16628	0.89343	1.16297	0.79889	
12	2023	1.24696	1.06474	0.92306	0.80915	1.04313	1.19551	0.91796	0.82278	1.10239	0.87268	1.21704	0.7846	
13	Average	=AVERAGE(B8:B12)	0.91892	0.81073	1.02441	1.20632	0.92652	0.83125	1.13944	0.88233	1.18036	0.79317		

Figure 6, finding the average SI for each month

4. Unfortunately, when graphing in excel and excel sees multiple lines, it assumes it is multiple separate datasets. So for our data we have to spread it across one line, and number each month, not naming. Additionally, there should be 3 rows, Seasonalized, Deseasonalised, and a repeating 12 values of all the month averages over the dataset. (*Figure 7*)



92	Month averag	1.2465	1.01858	0.91815	0.81063	1.02683	1.20189	0.92798	0.8322	1.13817	0.88713	1.18415	0.79045	1.2465	1.01858
93	Month	1	2	3	4	5	6	7	8	9	10	11	12	13	14
94	Seasonalized	226051	186945	165287	145315	181330	220043	164915	148049	205234	153741	208036	144515	235132	193956
95	Deseasonalized														

Figure 7, spread out version of dataset, (The data continues further than image)

5. Each deseasonalised value can be found using the formula $\frac{actual\space value}{monthly\space average}$, or in excel `B16/B14`, where `B` varies along columns.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Avg	
2	2019	226051	186945	165287	145315	181330	220043	164915	148049	205234	153741	208036	144515	179122	
3	2020	235132	193956	179832	160067	198298	230078		164049	221707	175250	230961	152046	194671	
4	2021	264672	210287	187960	170631	211650	249996	196810	176987	238620	185033	244896	165743	208607	
5	2022	279628	222171	205072	176016	229360	272408	205079	181075	258855	198296	258120	177314	221950	
6	2023	294869	251780	218277	191340	246670	282702	217071	194564	260684	206363	287794	185535	236471	
7	SI's														
8	2019	1.262	1.04368	0.92276	0.81126	1.01233	1.22845	0.92069	0.82653	1.14578	0.8583	1.16142	0.8068		
9	2020	1.20785	0.99633	0.92378	0.82225	1.01863	1.18188		0.8427	1.13888	0.90024	1.18642	0.78104		
10	2021	1.26876	1.00805	0.90102	0.81795	1.01459	1.19841	0.94345	0.84842	1.14387	0.88699	1.17396	0.79452		
11	2022	1.25987	1.001	0.92396	0.79305	1.03339	1.22734	0.92399	0.81584	1.16628	0.89343	1.16297	0.79889		
12	2023	1.24696	1.06474	0.92306	0.80915	1.04313	1.19551	0.91796	0.82278	1.10239	0.87268	1.21704	0.7846		
13	Average	1.24909	1.02276	0.91892	0.81073	1.02441	1.20632	0.92652	0.83125	1.13944	0.88233	1.18036	0.79317		
14		1.2465	1.01858	0.91815	0.81063	1.02683	1.20189	0.92798	0.8322	1.13817	0.88713	1.18415	0.79045	1.2465	1.01858
15	Month	1	2	3	4	5	6	7	8	9	10	11	12	13	14
16	Seasonalized	226051	186945	165287	145315	181330	220043	164915	148049	205234	153741	208036	144515	235132	193956
17	Deseasonalize	=B16/B14		180022	179263	176592	183081	177714	177901	180319	173301	175684	182827	188633	190419

Figure 8, finding the deseasonalised data points, (data continues past image)

6. Additionally, we can calculate the r^2 value of the deseasonalised data using the `=RSQ()` function in excel. This returns 0.830007922, this proves that deseasonalisation was necessary, as it increased the correlation by a product of ≈ 4.

With the data optimized, the actual forecasting can begin:

1. Plot the Seasonalized data & Deseasonalised data.

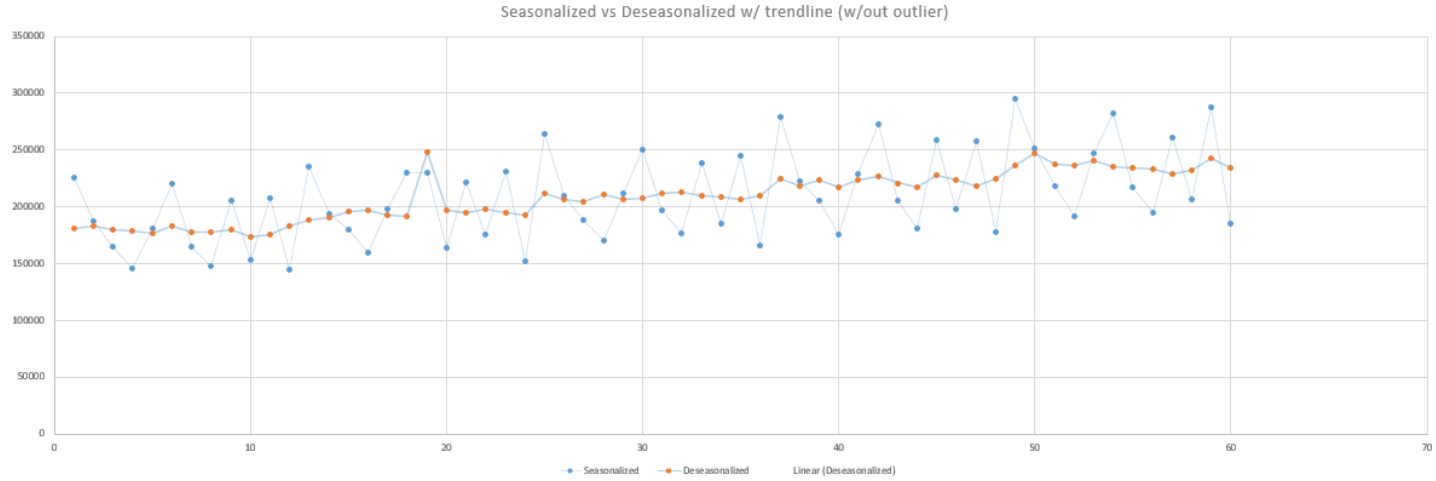


Figure 9, a plot with the seasonalized and deseasonalised data

2. Add a trendline to the plot to the deseasonalised data

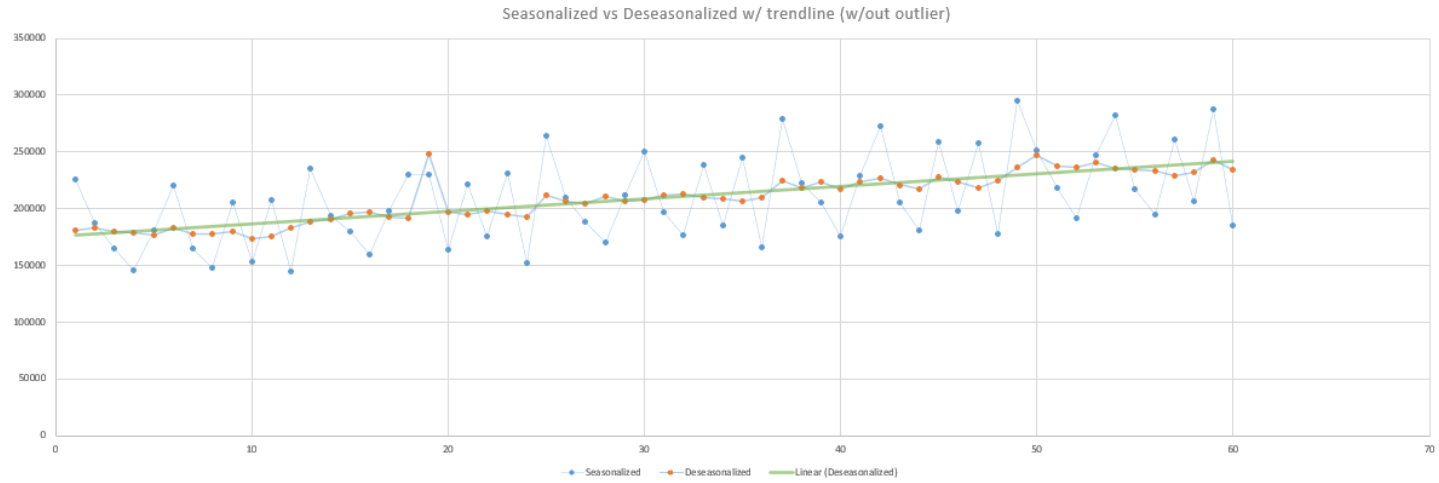


Figure 10, line of best fit for deseasonalised data

3. When adding a trendline to the graph in excel, excel rounds the data, thus the trendline is inaccurate. To find the correct values for a and b, we can use the `=LINEST(known_ys, [known_xs])` function, and enter the Deseasonalised data as the `known_ys`, and Month data for the `[known_xs]`. (`=LINEST(B17:BI17,B15:BI15)`), as seen in *Figure 11*

Month	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Seasonalized	226051	186945	165287	145315	181330	220043	164915	148049	205234	153741	208036	144515	235132	193956	179832	160067	198298	230078	230078	164049	221707
Deseasonalize	181348	183536	180022	179263	176592	183081	177714	177901	180319	173301	175684	182827	188633	190419	195864	197461	193117	191430	247934	197127	194792

m	b
=LINEST(B17:BI17,B15:BI15)	

LINEST(**known_ys**, [known_xs], [const], [stats])

Figure 11, LINEST function

This returns the a and b values ($y = ax + b$). Therefore, the equation is

$$monthly\ prediction = 1105.687097 * month + 175372.9$$

a	b
1105.6871	175373

4. Now, in the month row, instead of going to 60, go to 120 (which is December of 2028).
5. For the deseasonalised row (all month values > 60), we can use the line of best fit formula to forecast, by substituting the $month$ for the month. (*Figure 12*)

1.26121	1.03276	0.92819	0.81898	1.03464	1.21818	0.8308	0.83971	1.15087	0.89136	1.19227	0.80101	1.26121	1.03276	0.92819	0.81898	1.03464	1.21818	0.8308	0.83971	1.15087
1.2465	1.01858	0.91815	0.81063	1.02683	1.20189	0.92798	0.8322	1.13817	0.88713	1.18415	0.79045	1.2465	1.01858	0.91815	0.81063	1.02683	1.20189	0.92798	0.8322	1.13817
61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81
306247	251917	227435	201582	255807	302533	207247	210398	289635	225311	302690	204244	322981	265619	239751	212449	269534	318696	218271	221540	304905
=A2D*BJ15+B20			246137	247243	248348	249454	250560	251665	252771	253877	254982	256088	257194	258299	259405	260511	261616	262722	263828	264934	

Figure 12, extrapolating/forecasting the data to 2028 (data continues past image)

Where `$A$20` is a, and `$B$20` is b.

6. To reseasonalise the data, we can use the formula $reseasonalised\ value = deseasonalised\ value * SI$

0.790446	1.246504	1.018576	0.918147	0.810625	1.026831	1.201891	0.92798	0.8322	1.138174	0.887134	1.184149	0.790446	1.24
60	61	62	63	64	65	66	67	68	69	70	71	72	
185535	=BJ17*BJ14		224974.7	199524.7	253876.3	298487.6	231488.3	208515.7	286439	224241.6	300627.8	201549.8	3192
234721.9	242819.8	243925.5	245031.2	246136.9	247242.6	248348.3	249453.9	250559.6	251665.3	252771	253876.7	254982.4	2560

7. Finally, we can graph the whole dataset with the predictions, up to 2028:

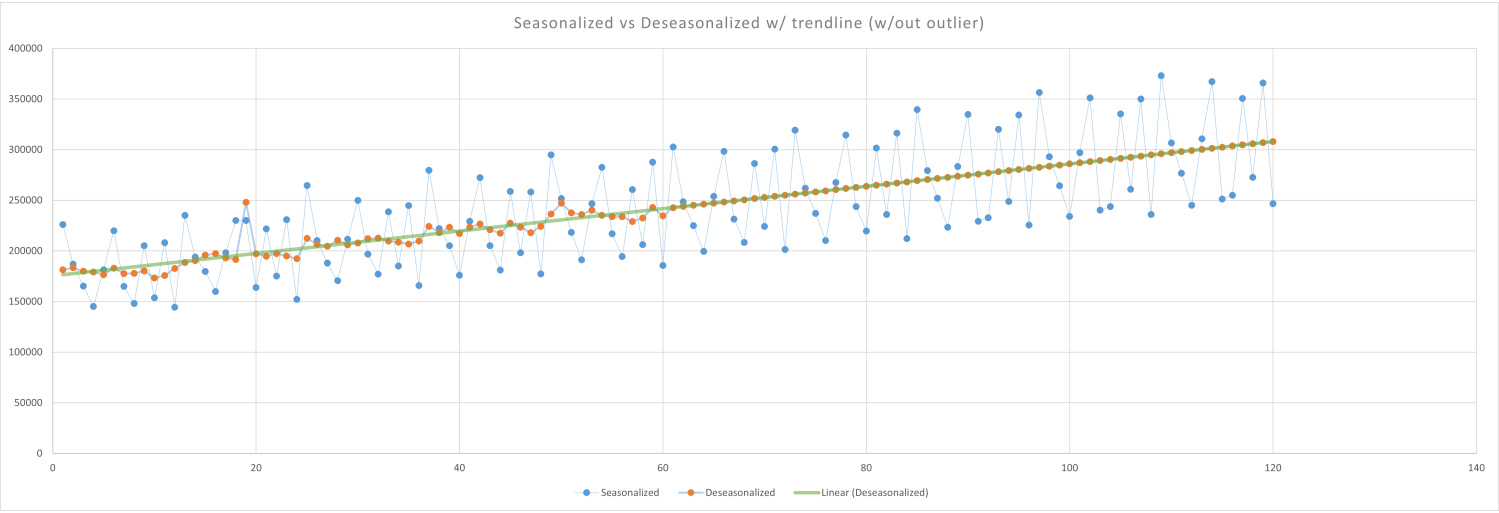


Figure 13, a showing the seasonalised, deseasonalised, and reseasonalised data with forecasting up to 2028 (2028 being month 120)

Solution

Monthly forecast of 2024.

Data	Deseasonalised (AUD)	Seasonalised (AUD)
January	$302,675.90	$242,819.80
February	$248,456.70	$243,925.50
March	$224,974.70	$245,031.20
April	$199,524.70	$246,136.90
May	$253,876.30	$247,242.60
June	$298,487.60	$248,348.30
July	$231,488.30	$249,453.90
August	$208,515.70	$250,559.60
September	$286,439.00	$251,665.30
October	$224,241.60	$252,771.00
November	$300,627.80	$253,876.70
December	$201,549.80	$254,982.40

Figure 14, a table containing the monthly predictions of 2024

Yearly forecast to 2028 (2024 - 2028)

The same process discussed in Working can be performed on the sum of each years' sales.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Avg	Total
2	2019	226051	186945	165287	145315	181330	220043	164915	148049	205234	153741	208036	144515	179121.75	=SUM(B2:M2)
3	2020	235132	193956	179832	160067	198298	230078	83033	164049	221707	175250	230961	152046	185367.4167	SUM(number
4	2021	264672	210287	187960	170631	211650	249996	196810	176987	238620	185033	244896	165743	208607.0833	2503285
5	2022	279628	222171	205072	176016	229360	272408	205079	181075	258855	198296	258120	177314	221949.5	2663394
6	2023	294869	251780	218277	191340	246670	282702	217071	194564	260684	206363	287794	185535	236470.75	2837649

Figure 15, the sum of each years' sales

1	2149461		m	b
2	2224409		181536.1	1931031.3
3	2503285			
4	2663394			
5	2837649	2024		
6	3020247.9	2025		
7	3201784	2026		
8	3383320.1	2027		
9	3564856.2	2028		

2024	2025	2026	2027	2028
$ 2,837,649.00	$ 3,020,247.90	$ 3,201,784.00	$ 3,383,320.10	$ 3,564,856.20

Figure 16, short summary of forecasting the sum of sales.

- Sum of each years' sales, found in *Figure 15*.
- The line of best fit equation found using the `LINEST` function.
- Forecasted data by extrapolation, using the line of best fit equation.
- The forecasted data in monetary form.

2024	2025	2026	2027	2028
$ 2,837,649.00	$ 3,020,247.90	$ 3,201,784.00	$ 3,383,320.10	$ 3,564,856.20

Figure 17, a table containing the yearly predictions to 2028

Evaluate & Verify

Justifications

During the process of creating these models, several decisions were made.
It was decided that, in order to get the best forecasting models, the provided data would have to be optimized, to get the highest r^2 value. This was completed by removing outliers, and deseasonalizing the data. These optimizations changed the r^2 value from ≈ 0.2 to ≈ 0.8.
The Sum of Squared Residuals was the chosen method of creating a function to fit the data, this method is easy to complete and explain. However, it comes with a downside to not being as accurate. Because of its simplicity, and expandability, it was decided to use Excel to do all of the calculations and graphing.

Strengths

The created models are effective in predicting sales several years into the future, given that specific economical future events won't occur (Assumptions). This was achieved by optimizing the data - removing outliers, and deseasonalizing, to create a strong r^2 value. Additionally, the model reseasonalises the forecasted data, ensuring that each months' prediction will be based on past years data.

Limitations

The future cannot be modelled not nearly perfectly, especially with a simple linear model, taking into account only 1 factor (sales). The businesses sales could start to dip due to a global economic event, or their sales could increase at a higher rate, it is impossible to know from the data provided.

Conclusion

The aim of this Problem-Solving and modelling task was to provide 2 forecasting models for a business, given 5 years of monthly sale history. The first model was to forecast the monthly sales of 2024, and the second the yearly sales to 2028. This was completed for the first model by optimizing the provided data to increase the r^2 value, and then using The Least Squares Regression Method to find a line of best fit for the data. A similar process was completed for the second model, but with the sum of sales of each year as the data to create a line of best fit on. Using these models, the data was extrapolated to the required dates.
Overall, this modelling task demonstrated the application of relevant knowledge and skills to solve a real-world problem using a systematic approach.

Bibliography

Linear Regression. (2024). Yale.edu. http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm

---

Subtract ≈ 166 words due to tables.