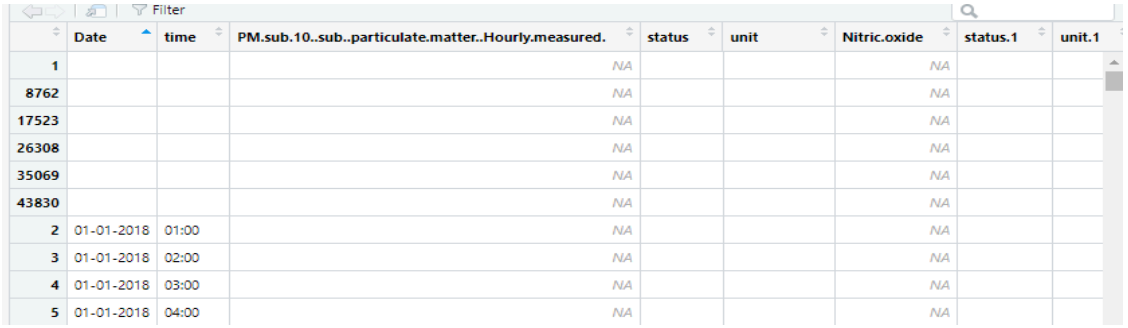
iDAV Coursework
Diwiz Budha Magar
Up2246287

1. Downloaded the csv files and viewed them. Saw that there were 4 spaces at the beginning that needed to be removed.

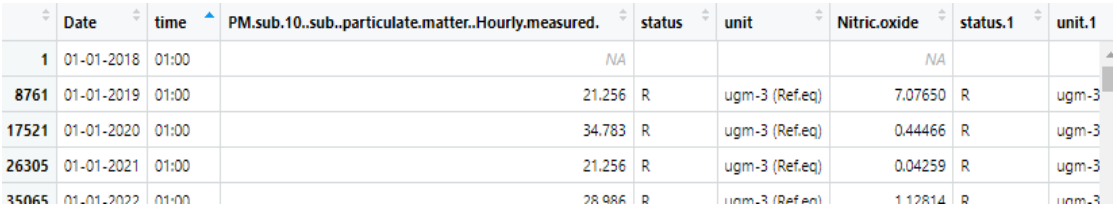
2. Used “df2018 <- read.csv("Data/POAR_2018.csv", skip = 4)” to load the csv and then put them in a data frame. Did the same for the rest of the 5 files.

3. Decided that I wanted to stack them one on top of the other as they were all in the same format, and wanted to work out from there.
df_total <- bind_rows(df2018, df2019, df2020, df2021, df2022, df2023)

4. After viewing the data frame I found out that there were certain rows with all empty/ null values so I decided to remove them. df_total <- bind_rows(df2018, df2019, df2020, df2021, df2022, df2023)

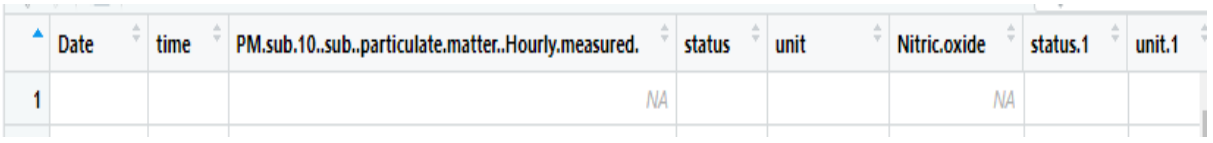


	Date	time	PM.sub.10..sub..particulate.matter..Hourly.measured.	status	unit	Nitric.oxide	status.1	unit.1
1			NA			NA		
8762			NA			NA		
17523			NA			NA		
26308			NA			NA		
35069			NA			NA		
43830			NA			NA		
2	01-01-2018	01:00	NA			NA		
3	01-01-2018	02:00	NA			NA		
4	01-01-2018	03:00	NA			NA		
5	01-01-2018	04:00	NA			NA		

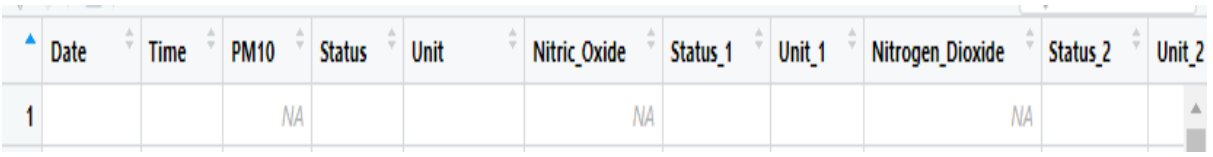


	Date	time	PM.sub.10..sub..particulate.matter..Hourly.measured.	status	unit	Nitric.oxide	status.1	unit.1
1	01-01-2018	01:00	NA			NA		
8761	01-01-2019	01:00	21.256	R	ugm-3 (Ref.eq)	7.07650	R	ugm-3
17521	01-01-2020	01:00	34.783	R	ugm-3 (Ref.eq)	0.44466	R	ugm-3
26305	01-01-2021	01:00	21.256	R	ugm-3 (Ref.eq)	0.04259	R	ugm-3
35065	01-01-2022	01:00	28.986	R	ugm-3 (Ref.eq)	1.12814	R	ugm-3

5. I then fixed the naming using rename() for better readability. All the column names were mutated.

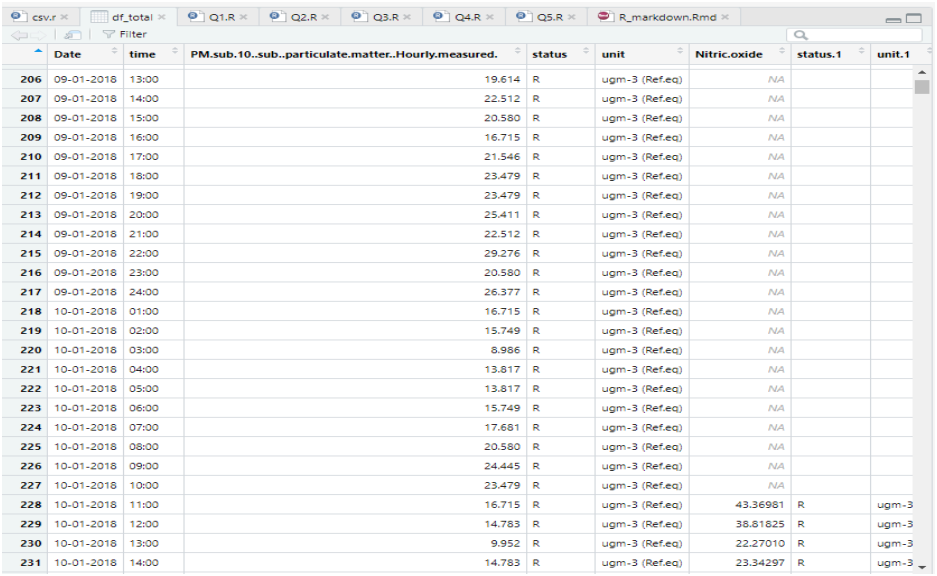


	Date	time	PM.sub.10..sub..particulate.matter..Hourly.measured.	status	unit	Nitric.oxide	status.1	unit.1
1			NA			NA		



	Date	Time	PM10	Status	Unit	Nitric_Oxide	Status_1	Unit_1	Nitrogen_Dioxide	Status_2	Unit_2
1			NA			NA			NA		

6. Used summary(df_total) to find out about the data frame to figure out there were null values and negative values. On closer inspection I found out that sometimes there were negative values and null values in only a few columns, not the entire row.

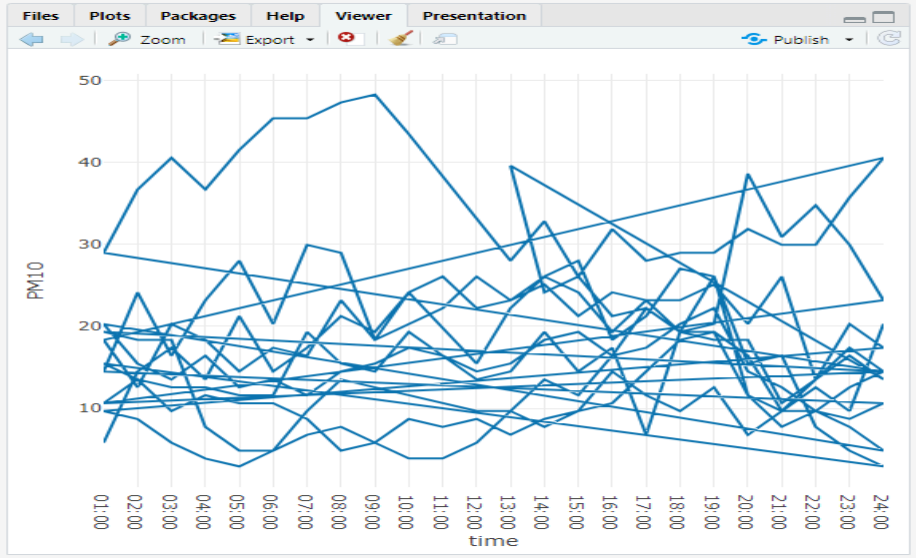


	Date	time	PM.sub.10..sub..particulate.matter..Hourly.measured.	status	unit	Nitric.oxide	status.1	unit.1
206	09-01-2018	13:00	19.614	R	ugm-3 (Ref.eq)	NA		
207	09-01-2018	14:00	22.512	R	ugm-3 (Ref.eq)	NA		
208	09-01-2018	15:00	20.580	R	ugm-3 (Ref.eq)	NA		
209	09-01-2018	16:00	16.715	R	ugm-3 (Ref.eq)	NA		
210	09-01-2018	17:00	21.546	R	ugm-3 (Ref.eq)	NA		
211	09-01-2018	18:00	23.479	R	ugm-3 (Ref.eq)	NA		
212	09-01-2018	19:00	23.479	R	ugm-3 (Ref.eq)	NA		
213	09-01-2018	20:00	25.411	R	ugm-3 (Ref.eq)	NA		
214	09-01-2018	21:00	22.512	R	ugm-3 (Ref.eq)	NA		
215	09-01-2018	22:00	29.276	R	ugm-3 (Ref.eq)	NA		
216	09-01-2018	23:00	20.580	R	ugm-3 (Ref.eq)	NA		
217	09-01-2018	24:00	26.377	R	ugm-3 (Ref.eq)	NA		
218	10-01-2018	01:00	16.715	R	ugm-3 (Ref.eq)	NA		
219	10-01-2018	02:00	15.749	R	ugm-3 (Ref.eq)	NA		
220	10-01-2018	03:00	8.986	R	ugm-3 (Ref.eq)	NA		
221	10-01-2018	04:00	13.817	R	ugm-3 (Ref.eq)	NA		
222	10-01-2018	05:00	13.817	R	ugm-3 (Ref.eq)	NA		
223	10-01-2018	06:00	15.749	R	ugm-3 (Ref.eq)	NA		
224	10-01-2018	07:00	17.681	R	ugm-3 (Ref.eq)	NA		
225	10-01-2018	08:00	20.580	R	ugm-3 (Ref.eq)	NA		
226	10-01-2018	09:00	24.445	R	ugm-3 (Ref.eq)	NA		
227	10-01-2018	10:00	23.479	R	ugm-3 (Ref.eq)	NA		
228	10-01-2018	11:00	16.715	R	ugm-3 (Ref.eq)	43.36981	R	ugm-3
229	10-01-2018	12:00	14.783	R	ugm-3 (Ref.eq)	38.81825	R	ugm-3
230	10-01-2018	13:00	9.952	R	ugm-3 (Ref.eq)	22.27010	R	ugm-3
231	10-01-2018	14:00	14.783	R	ugm-3 (Ref.eq)	23.34297	R	ugm-3

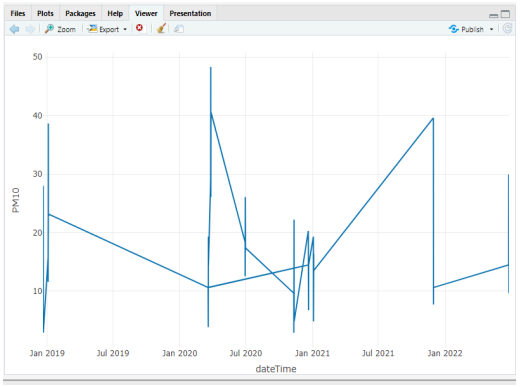
7. Hence, I decided to fix these issues individually instead of as a whole. Because if I used filter() then the entire row would be removed, which would lead to other values being affected.

8. Also saw that there were some units and status that were missing so, filled them up with similar values. This was unnecessary as I didn't use the column at all. The columns affected were all the unit and status.

9. I used str() to see that the date and time column were not in the correct data types so I tried to change them. This would lead to issues when plotting the graph.



10. I had even attempted to combine date time and work from there but the graph looked unfit for understanding.



11. I ended up deciding not to change the data types.

12. The next process was to combine the strings of date required and put them in a data frame called df_required. These affected all the columns which matched the given date.

	Date	Time	PM10	Status	Unit	Nitric_Oxide	Status_1	Unit_1	Nitrogen_Dioxide	Status_2	Unit_2
22	20-12-2018	22:00	7.730	R	ugm-3 (Ref.eq)	6.38996	R	ugm-3	33.53730	R	ugm
23	20-12-2018	23:00	4.831	R	ugm-3 (Ref.eq)	9.07864	R	ugm-3	23.45746	R	ugm
24	20-12-2018	24:00	2.899	R	ugm-3 (Ref.eq)	7.08832	R	ugm-3	20.81558	R	ugm
25	03-01-2019	01:00	15.459	R	ugm-3 (Ref.eq)	1.08658	R	ugm-3	28.12289	R	ugm
26	03-01-2019	02:00	13.527	R	ugm-3 (Ref.eq)	0.42061	R	ugm-3	14.43474	R	ugm
27	03-01-2019	03:00	12.561	R	ugm-3 (Ref.eq)	0.73607	R	ugm-3	16.12040	R	ugm
28	03-01-2019	04:00	12.561	R	ugm-3 (Ref.eq)	0.66597	R	ugm-3	13.21196	R	ugm
29	03-01-2019	05:00	11.594	R	ugm-3 (Ref.eq)	1.01648	R	ugm-3	12.25123	R	ugm
30	03-01-2019	06:00	11.594	R	ugm-3 (Ref.eq)	1.26183	R	ugm-3	17.53651	R	ugm

13. As my initial thought process was to split them and work on individual data frames so as to not affect the other data, I ended up creating separate data frames for each pollutant. Additionally for question 4 where I had to find the monthly average, I made a data frame with year as the filter. The affected columns were the Date, Time, the pollutant chosen (here PM10), the status and unit associated with.

csv.r | PM10 | df_required | Q1.R | Q2.R

Filter

	Date	Time	PM10	Status	Unit
1	20-12-2018	01:00	5.797	R	ugm-3 (Ref.eq)
2	20-12-2018	02:00	14.493	R	ugm-3 (Ref.eq)
3	20-12-2018	03:00	17.392	R	ugm-3 (Ref.eq)
4	20-12-2018	04:00	13.527	R	ugm-3 (Ref.eq)
5	20-12-2018	05:00	21.256	R	ugm-3 (Ref.eq)
6	20-12-2018	06:00	14.493	R	ugm-3 (Ref.eq)
7	20-12-2018	07:00	17.392	R	ugm-3 (Ref.eq)
8	20-12-2018	08:00	21.256	R	ugm-3 (Ref.eq)
9	20-12-2018	09:00	19.324	R	ugm-3 (Ref.eq)
10	20-12-2018	10:00	24.155	R	ugm-3 (Ref.eq)

14. I created a csv file for each of the pollutants and the years needed.

15. After the data was separated for question 1,2 ,and 3, worked on their individual csv file that was made. This led to the realization that there are sometimes null spaces of 1, 2, or more adjacent. Here I decided if the null values were 2 or less, the rows would be filled with averages and if there were more than 2 null values together they would be removed.

csv.r | Q1.R | Q1_PM10 | Q2.R | Q3.R

Filter

	Date	Time	PM10	Status	Unit
193	29-11-2021	01:00	NA	R	ugm-3 (Ref.eq)
194	29-11-2021	02:00	NA	R	ugm-3 (Ref.eq)
195	29-11-2021	03:00	NA	R	ugm-3 (Ref.eq)
196	29-11-2021	04:00	NA	R	ugm-3 (Ref.eq)
197	29-11-2021	05:00	NA	R	ugm-3 (Ref.eq)
198	29-11-2021	06:00	NA	R	ugm-3 (Ref.eq)
199	29-11-2021	07:00	NA	R	ugm-3 (Ref.eq)
200	29-11-2021	08:00	NA	R	ugm-3 (Ref.eq)
201	29-11-2021	09:00	NA	R	ugm-3 (Ref.eq)
202	29-11-2021	10:00	NA	R	ugm-3 (Ref.eq)
203	29-11-2021	11:00	NA	R	ugm-3 (Ref.eq)
204	29-11-2021	12:00	NA	R	ugm-3 (Ref.eq)
205	29-11-2021	13:00	39.614	R	ugm-3 (Ref.eq)
206	29-11-2021	14:00	24.155	R	ugm-3 (Ref.eq)

csv.r | Q1.R* | Q1_PM10_filtered | Q2.R | Q3.R

Filter

	Date	Time	PM10	Status	Unit
189	03-01-2021	21:00	9.662	R	ugm-3 (Ref.eq)
190	03-01-2021	22:00	13.527	R	ugm-3 (Ref.eq)
191	03-01-2021	23:00	16.425	R	ugm-3 (Ref.eq)
192	03-01-2021	24:00	13.527	R	ugm-3 (Ref.eq)
193	29-11-2021	13:00	39.614	R	ugm-3 (Ref.eq)
194	29-11-2021	14:00	24.155	R	ugm-3 (Ref.eq)
195	29-11-2021	15:00	26.087	R	ugm-3 (Ref.eq)
196	29-11-2021	16:00	21.256	R	ugm-3 (Ref.eq)
197	29-11-2021	17:00	22.223	R	ugm-3 (Ref.eq)
198	29-11-2021	18:00	19.324	R	ugm-3 (Ref.eq)
199	29-11-2021	19:00	26.087	R	ugm-3 (Ref.eq)
200	29-11-2021	20:00	11.594	R	ugm-3 (Ref.eq)

16. With regards to question 4 for the monthly averages,removing null values at the beginning would affect values of other columns which could skew the data. Hence, the data frame was split into individual columns based on pollutant, and again followed the steps where if there are less than 2 null values fill them with averages. The rest drop them. After that, take the averages of pollutants based on the months and make a table where values correspond to the month. Join the four pollutants and make the data columns pivot vertically to make it easier to plot.

Filter

	Date	Time	PM10	Status	Unit	Nitric_Oxide	Status_1	Unit_1	Nitrogen_Dioxide	Status_2	Unit_
1	01-01-2020	01:00	34.783	R	ugm-3 (Ref.eq)	0.44466	R	ugm-3	14.11784	R	ugm
2	01-01-2020	02:00	23.189	R	ugm-3 (Ref.eq)	1.90092	R	ugm-3	17.11645	R	ugm
3	01-01-2020	03:00	19.324	R	ugm-3 (Ref.eq)	1.60077	R	ugm-3	13.85954	R	ugm
4	01-01-2020	04:00	26.087	R	ugm-3 (Ref.eq)	2.26776	R	ugm-3	17.48324	R	ugm
5	01-01-2020	05:00	21.256	R	ugm-3 (Ref.eq)	1.86757	R	ugm-3	16.49643	R	ugm
6	01-01-2020	06:00	17.392	R	ugm-3 (Ref.eq)	0.50024	R	ugm-3	12.29452	R	ugm
7	01-01-2020	07:00	14.493	R	ugm-3 (Ref.eq)	0.36684	R	ugm-3	9.76285	R	ugm
8	01-01-2020	08:00	20.290	R	ugm-3 (Ref.eq)	1.66747	R	ugm-3	15.40932	R	ugm
9	01-01-2020	09:00	41.547	R	ugm-3 (Ref.eq)	2.83470	R	ugm-3	22.18962	R	ugm

csv.r × | Q1.R × | Q1_PM10_filtered × | Q2.R × | Q3.

Filter

	Date	Time	PM10	Status	Unit
1	01-01-2020	01:00	34.783	R	ugm-3 (Ref.eq)
2	01-01-2020	02:00	23.189	R	ugm-3 (Ref.eq)
3	01-01-2020	03:00	19.324	R	ugm-3 (Ref.eq)
4	01-01-2020	04:00	26.087	R	ugm-3 (Ref.eq)
5	01-01-2020	05:00	21.256	R	ugm-3 (Ref.eq)
6	01-01-2020	06:00	17.392	R	ugm-3 (Ref.eq)
7	01-01-2020	07:00	14.493	R	ugm-3 (Ref.eq)
8	01-01-2020	08:00	20.290	R	ugm-3 (Ref.eq)
9	01-01-2020	09:00	41.547	R	ugm-3 (Ref.eq)
10	01-01-2020	10:00	46.378	R	ugm-3 (Ref.eq)
11	01-01-2020	11:00	46.378	R	ugm-3 (Ref.eq)
12	01-01-2020	12:00	35.749	R	ugm-3 (Ref.eq)

	Month	PM10
1	January	20.68695
2	February	22.60013
3	March	20.36681
4	April	28.00645
5	May	20.30706
6	June	16.64347
7	July	12.24245
8	August	15.24615
9	September	15.33445
10	October	13.23414
11	November	19.06893
12	December	15.68792

	Month	PM10	NO	NO2	NOx_NO2
1	January	20.68695	18.050853	31.72067	59.39826
2	February	22.60013	11.162173	22.09751	39.21260
3	March	20.36681	8.173137	20.01943	32.55140
4	April	28.00645	2.984736	18.64371	23.22024
5	May	20.30706	3.018190	13.93417	18.56200
6	June	16.64347	4.739334	14.97264	22.23951
7	July	12.24245	7.609730	16.05933	27.72742
8	August	15.24615	7.936192	22.70388	34.87253
9	September	15.33445	10.022967	22.98490	38.35323
10	October	13.23414	11.567390	23.36762	41.10404
11	November	19.06893	9.165810	23.67544	37.72949
12	December	15.68792	11.449794	24.68824	42.24434

csv.r | Q1.R | Q1_PM10_filtered | Filter

	Month	Pollutants	Average
1	January	PM10	20.686947
2	January	NO	18.050853
3	January	NO2	31.720672
4	January	NOx_NO2	59.398260
5	February	PM10	22.600134
6	February	NO	11.162173
7	February	NO2	22.097507
8	February	NOx_NO2	39.212600
9	March	PM10	20.366815
10	March	NO	8.173137
11	March	NO2	20.019430
12	March	NOx_NO2	32.551398
13	April	PM10	28.006454
14	April	NO	2.984736
15	April	NO2	18.643714
16	April	NOx_NO2	23.220245
17	May	PM10	20.307063
18	May	NO	3.018190
19	May	NO2	13.934170
20	May	NOx_NO2	18.561996
21	June	PM10	16.643466
22	June	NO	4.739334
23	June	NO2	14.972636
24	June	NOx_NO2	22.239512
25	July	PM10	12.242454
26	July	NO	7.609730
27	July	NO2	16.059328

17. Followed the similar pattern for Q5, where I was calculating the yearly average of each pollutant.

Conclusion:-
After looking at the data and visualisation, it appears that the CAZ has made a positive impact on the air quality. There are always going to be outliers that might suggest otherwise, however looking at the graph, it's clear that the level of pollutants are less. The impact of implementing CAZ is reducing the level of pollutants in the area around Anglesea Road.

References

- Timeline for the year 2020 covid for assessing the graph.
  House of Commons Library. (2023, October 17). *Student loan statistics* (CBP-9068). UK Parliament. https://commonslibrary.parliament.uk/research-briefings/cbp-9068/

- Creating the bar graph
  Holtz, Y. (n.d.). *Grouped barplot with ggplot2*. The R Graph Gallery. https://r-graph-gallery.com/48-grouped-barplot-with-ggplot2.html

- Creating the line and scatter graph
  Holtz, Y. (n.d.). *The ggplot2 package*. The R Graph Gallery. https://r-graph-gallery.com/ggplot2-package.html

- Rmarkdown
  RStudio. (n.d.). *Layouts: Organizing dashboards using flexdashboard* [Article]. https://pkgs.rstudio.com/flexdashboard/articles/layouts.html#multiple-pages

- Summarise
  Wickham, H., François, R., Henry, L., & Müller, K. (2023). *dplyr: A Grammar of Data Manipulation* (Version 1.1.4) [R package]. https://CRAN.R-project.org/package=dplyr

- Creating the table with a column with month.
  Holtz, Y. (n.d.). *The ggplot2 package*. The R Graph Gallery. https://r-graph-gallery.com/ggplot2-package.html

- Filling in na values with up down.
  Wickham, H., Vaughan, D., & Girlich, M. (n.d.). *fill: Fill in missing values* [R package documentation]. https://tidyr.tidyverse.org/reference/fill.html

- Interpolate missing value
  Sanderson, S. (2024, November 28). *Post title* [Blog post]. Steve on Data. https://www.spsanderson.com/steveondata/posts/2024-11-28/