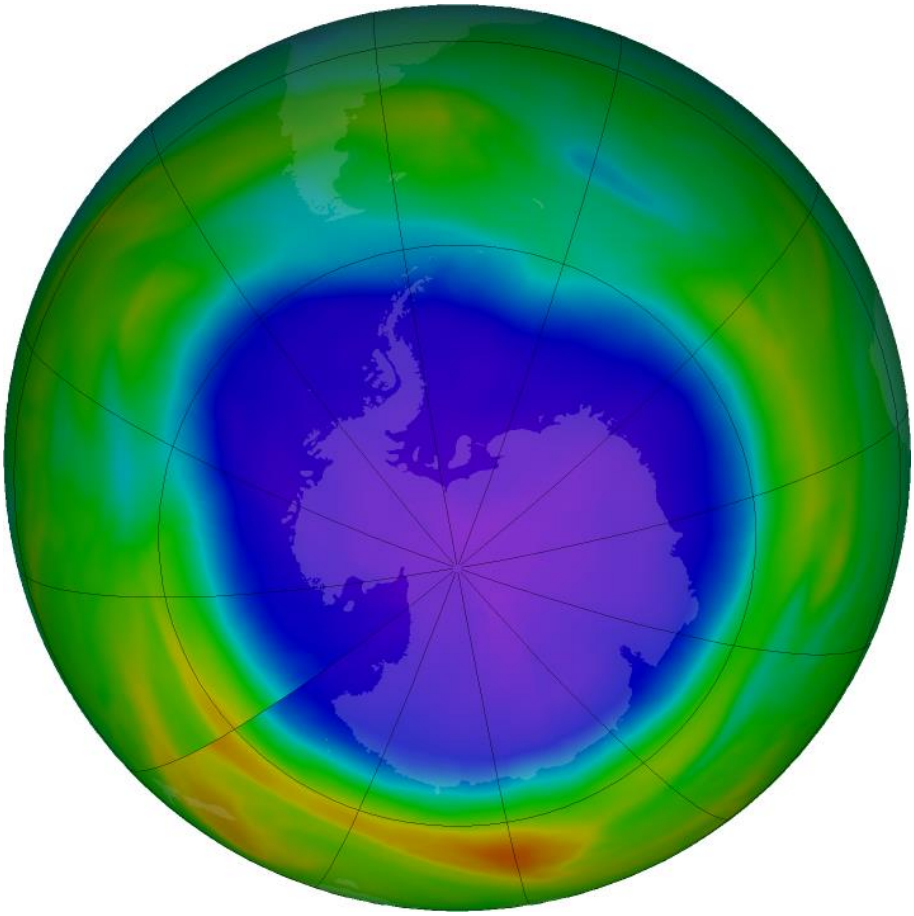
GDV – Ozone depletion

Plot the fluctuation of ozone depletion area over the years (1979-2022) for each day and show significant global events.



Thomas Mandelz
06.01.2022
Link: Github Repository for attachments and python code

Contents

Idea and history 3
Reading datasource and python setup 3
LO1: Visualization basics, chart types 4
Line chart 4
Scatter plot 4
Map Projections from NASA 5
Review 6
LO2: Visual Perception 7
Size variable 7
Color variable 7
Orientation visual variable 8
Gestalt Theory 8
LO3: Design Principles vs. Data 10
Data Quality (validity, accuracy, completeness, consistency, uniformity) 10
Data Cleaning 11
LO4: Grammar of Graphics Tools 12
Data 12
Aesthetics 12
Scale 12
Geometric Objects 12
Statistics 13
Facets 13
Coordinate System 13
LO5: Evaluation 14
User Testing outline decisions 14
Facilitator Role 14
Participants 14
Test script 14
Reflection after user testing 15
Bibliography 16
Appendix 19

Idea and history

As I listened to an episode of the Startalk podcast (deGrasse Tyson, 2021) I was reminded of the ozone depletion which was a threat in my infant and teen years. Since then, I have not heard of any developments in recent years. So, I decided to investigate this topic and get some experience in data visualisation as well.

After some initial research I found several data sources with different complexities by NASA.

My first source (Ozonewatch, 2021d) is an aggregated source which gives you the values of square kilometres of ozone depletion in antarctica per year, month or day. The second source (Ozonewatch, 2021c) is from NASA as well but this includes a lot of raw data from different satellites and longitudes as well as latitudes.

I chose the first source because my second source is too complex to visualize efficiently and in a timely manner.

Reading datasource and python setup

Using the pandas and matplotlib packages in python needs some setup. I choose the "ggplot" style for the plots to ensure a comprehensive look and feel on all plots. To enlarge the plots, I chose a bigger figure size as defined in the default settings. Later in the report I started to use the plotly library as my main plotting library.

LO1: Visualization basics, chart types

I then started to read about the different types of visualisations like bar plots, line plots and other basic visualisation plots. To learn more about the different plot types I went through my data source and looked for a use case for each plot in my data.

Line chart

The line bar can be used to show a continuously positive value on the x axis, for instance time (Yi & Sapountzis, 2021), and show the respective y points in a coherent line. I used this to show that the ozone depletion area increased dramatically over the years (1979 - 2000). A line bar will also show missing values (or outliers) like in the year 1995.

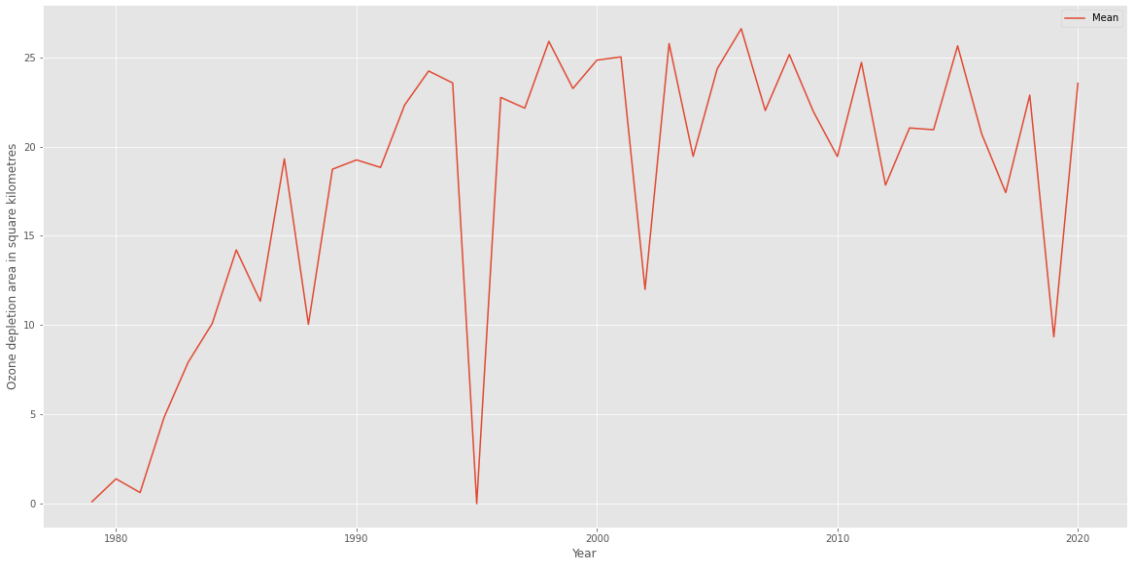


Figure 1 Line Plot Mean ozone depletion over years 1979-2020

Scatter plot

A scatter plot is best used to show a relation between two variables (Castañón, 2019). They can also be extended to show the relation between n to 1 variable (mean, max, min to the year variable). It will also visualize outliers quite easily like the value in year 1995 and some of the following years.

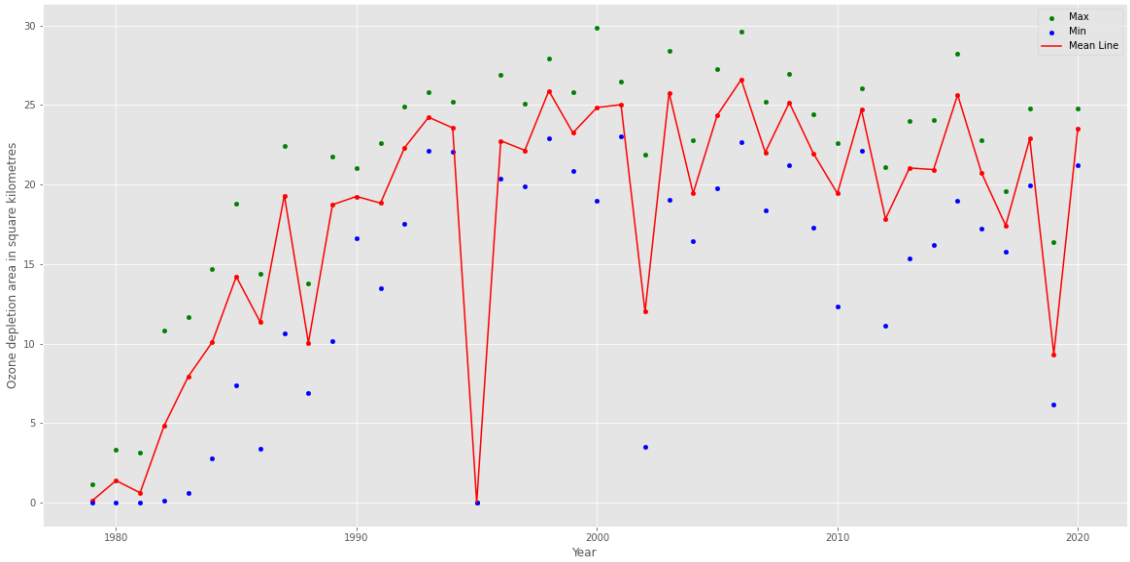


Figure 2 Scatter & Line Plot Mean, Max, Min ozone depletion over years 1979-2020

Map Projections from NASA

NASA made some interesting map projections (Ozonewatch, 2021b) (Ozonewatch, 2021a) which are much more intuitive for a broad audience.

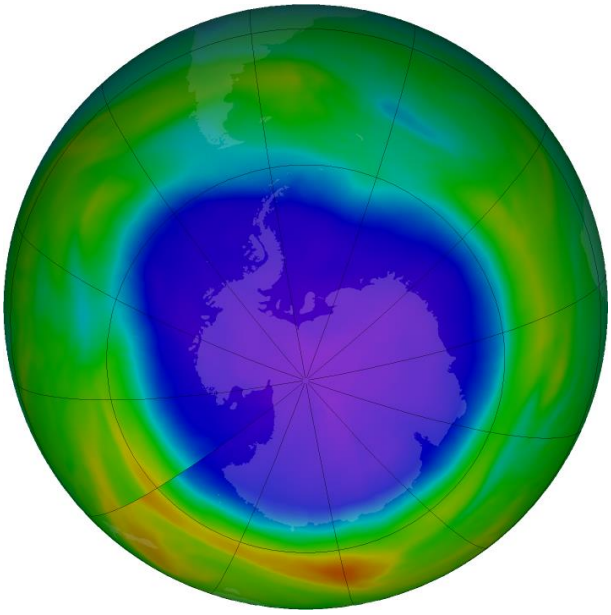


Figure 3 Ozone Map Projection 26.09.2021

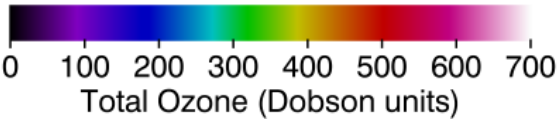


Figure 4 Ozone Dobson units legend

Review

I tried many different plots for my dataset. While there are many possibilities and plot types, the best for my dataset is a combination of a scatterplot with a line plot. They show multiple variables (max, min, mean) and highlight the mean value over the years with a line.

LO2: Visual Perception

In the following pages I will show some of the differences a visual variable produces when changed on a plot.

Size variable

The size variable can make a difference in lines really easy to spot (Wiki, 2017) but in my plot the lines overlap each other which makes it hard to distinguish among them. Also, values are not readable anymore. This makes the size variable not a good visual variable for my dataset.

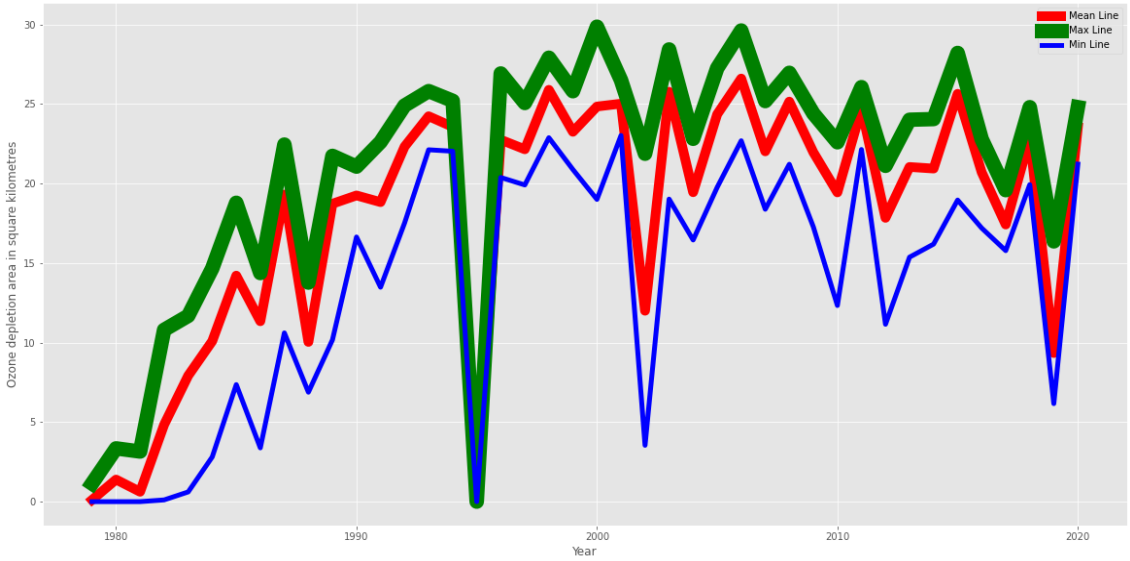


Figure 5 Line Plot size variable differences

Color variable

The color variable can be used to represent data values like the mean, max and min values. With different colors in one plot the audience can easily distinguish (Wilke, 2019) between the different values. Always consider using colors which are easily distinguishable from another. Make sure to be as inclusive for colorblind people as possible. Ggplot and other libraries have color palettes which complement each other well.

Figure 6 Line Plot color variable differences

Orientation visual variable

The orientation variable is used to show data in another way (Wiki, 2017). I changed the axis of the plot data. Now the data is upside down and looks different. For a continuous time variable, the X axis is more visually pleasing.

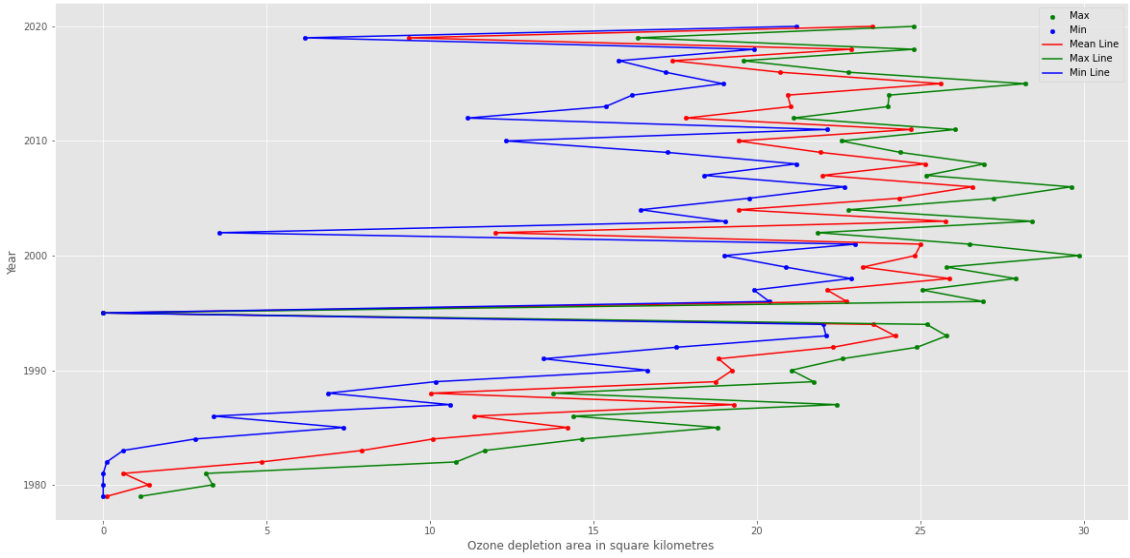


Figure 7 Line Plot orientation variable differences

Gestalt Theory

While reading Gestalt Theory resources (Guberman, 2017) I found some general rules for Gestalt Theory which are the following:

- Gestalt means 'seeing the whole picture all at once'.

- Something that is made of many parts and yet is somehow more than or different from the combination of its parts.
- When you put the parts together, you get the whole - in other words, the Gestalt.
- Humans naturally follow lines or curves
- The mind will attempt to fill in detail that isn't actually there

LO3: Design Principles vs. Data

After I have learned about the basics of visualization and perception, I was ready to make a more complex time series. For now, I only had data for each year. Nasa provides daily data over the last years. I wanted to show events on specific days, so it makes sense to have the scale of the time axis on a daily scope.

Data Quality (validity, accuracy, completeness, consistency, uniformity)

With the previous data sets I was already positive that the daily data is of high quality as well. I proceeded to check the following metrics which are described on (omnisci, 2021).

Accuracy

NASA is a very reliable source. On their website (Ozonewatch, 2021f) they describe their measuring process accurately as well as the concepts of ozone mapping as well as units like "Dobson Units". They also provide the raw data of their satellites which are measurements of the real atmosphere and not aggregated.

Completeness

Because NASA is constraint by the physical world and, they are a public agency of the American government there are some incomplete datasets and missing values.

I asked the curator of the data repository, why there is a big gap between 1994 and 1996 and what might cause the other missing values, and he answered with the following response: "There were no satellites available during that period that measured global total column ozone. The last TOMS instrument was onboard the Meteor-3 satellite that ended in 1994. The Earth Probe TOMS satellite wasn't launched until 1996. Other missing data here and there throughout the data record are due to technical issues with either the instrument or the satellite." (E. R. Nash, personal communication, December 14, 2021)

By "that period" he means the big gap in data between September 1994 and July 1996.

Consistency

The Data of NASA is very consistent and has the same format overall. It also is accessible by the same link. The only change in the URL they make is the year which I used to my advantage which you can see in the "LO4 - Data" chapter.

Validity

NASA collects data with their applied scientific method which falls in the right format and range of the real world.

Timeliness

The data is updated daily which is a very good time expectation for my timeseries. "Timeliness refers to the time expectation for accessibility and availability of information" (Loshin, 2009)

Data Cleaning

After checking the data quality I started the cleaning process and I used the following guide (Elgabry, 2019) to go through cleaning.

Inspection

I inspected the dataset after merging all years together it is very clean. The "date" column is formatted as a string, but no changes need to be made to convert it into a datetime value. In the "data" column there are float values in a string which are easily mutated to a float value. There are several values missing which had the value of "-9999". They need to be converted to a NAN value.

Cleaning

For the dataset not much cleaning is needed. I replaced the "-9999" values with NAN values so the plotting library removes these values and doesn't use them with the outlier "-9999" value. Then I transformed the "date" column into a datetype value and the "data" column to a float. I don’t want to assign a substitutional value to the missing values because they are valid as there were no measurements in these timeframes.

There were no duplicates or irrelevant data to remove.

Verifying

After the cleaning process I reinspected the data and tried a first plot without annotations or labels. The plot with the raw data looked good and no constraints were broken, or unforeseen outliers were detected.

Reporting

The data of NASA has a high quality from the source on. Working with it is very easy and didn't need much cleaning.

LO4: Grammar of Graphics Tools

For this LE 4 I followed the following article (Sarkar, 2018) and its suggestions what to consider when designing a visualisation. Following there are the 7 design principles which I explain with some examples of my ozone event time series.

Data

Previously I used only the yearly aggregated observation from NASA. I looked for the data sources which are less aggregated already. The daily data points are available on a NASA site, but they are separated in a file for each year with its own link. So, I was looking for an automated download for each year. NASA does not provide a single file or download for all the daily data. So, I started thinking how I would achieve this. I recognised that the URL "https://ozonewatch.gsfc.nasa.gov/meteorology/figures/ozone/to3caps_{Year}_toms+omi+omps.txt" includes a year. I then checked if this is the same for the other downloads. Indeed, this is the case for all the downloads. From experience I reckoned that a loop through these URL's should be enough to download all the files and merge them together. I found Urllib (Github contributors, 2011/2021) a python library which can download txt files from a website and save a file from it. By then removing the first 4 lines of the text file and setting the 5th row as the column names I prepared a big chunk of the data. Missing observations are set to the value -9999. I removed these and put the NAN value instead for easier handling with the python packages. I looped through each year and added them to a full dataset. Now I have a full time series with daily values from 1979-Today.

Aesthetics

My X axis will be the continuous date from the data points. For a time series it's best to use a time variable as the X axis to ensure continuity. No encodings are needed as there are not multiple data dimensions plotted. The time series highlights are encoded by color and an annotation sticker.

Scale

Ozone is measured in Dobson Units. Dobson Units are a measurement of concentration of gases in a specific volume. One Dobson unit is equivalent to 0.4462 mmol/m2 (Wikipedia contributors, 2021) which could be used to scale if for a broader scientific audience. NASA has a much more intuitive scaling method. They explain that 100 Dobson Units when compressed to their fullest, is about 1 mm high (Ozonewatch, 2018). This shows how 100 Dobson Units is a very small volume.

Geometric Objects

For a time series it makes the most sense to use a continuous plot like line or scatter. Because I want to add annotations on important events, I identified the scatter plot as the most useful. With the scatter plot I can highlight a specific point in the series.

Statistics

I considered showing the spread of the daily measurements, but I did not find hourly data, so I have only one data point per day. Because some specific dates should be highlighted, I don't want to make monthly aggregations.

Facets

NASA does many different measurements of the atmosphere. So there could be a visualisation which shows different gases (Oxygen, Hydrogen, Ozone, Methane, etc.) beside each other in a time series (NEO, 2021). I choose to only show ozone measurements for time and volume constraints of this project.

Coordinate System

For my time series I will use a cartesian coordinate system with the X and Y axis for plotting. The aggregated data of NASA does not have coordinates or regional data available so I can only use 2 axes and not a map projection. NASA of course has further datasets with coordinates of the measured data and also made some map (Ozonewatch, 2021e).

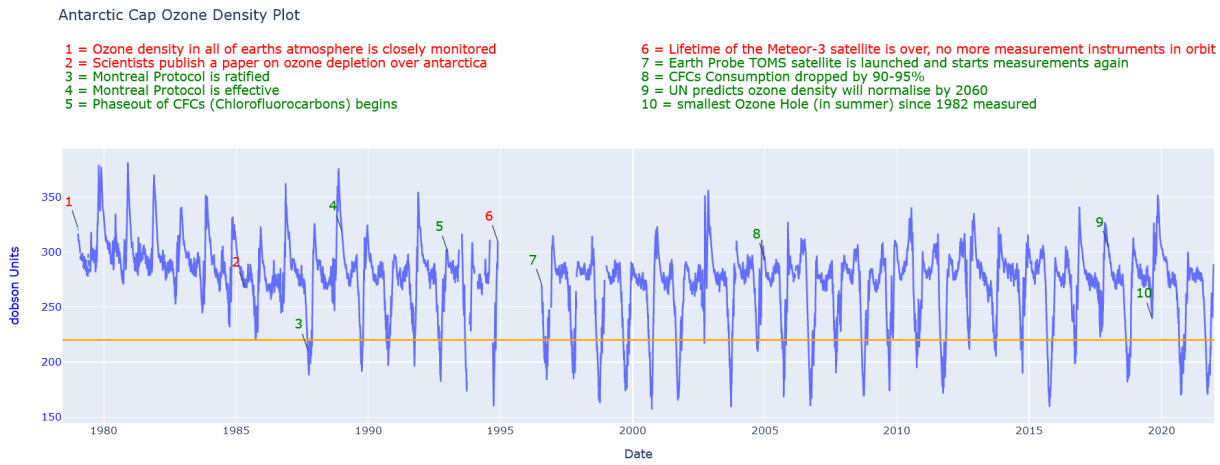


Figure 8 Time Series Plot https://tmandelz.github.io/GDV/ozoneplot.html

LO5: Evaluation

For the last LO I wanted an external evaluation by users for my time series plot. For this I started reading into the source material. Soon I chose to make a user study and defined the outline of it.

User Testing outline decisions

First, I had the choice between a qualitative or a quantitative approach to the user study. Because the time series does not have a fixed task to be measured by a metric (Budiu, 2017), as its purely informational, I have chosen a qualitative approach. I expected to gather either additions or improvements for the understanding of the plot or it's storytelling of the topic of ozone depletion by the participating users.

The second choice I needed to make is whether the form of user testing is Remote or In-Person. Because my participants are not meeting up regularly at the same physical location and it would be hard to meet each one bilaterally, I have chosen the remote form of user testing. I also did not moderate their sessions because I made a test form which they can use for clarification but I was available for questions (Schade, 2013) which was needed and helped a lot in gathering insights of the uncertainties the users where having.

Facilitator Role

I have chosen myself as the test facilitator because I want to learn how the evaluate my own visualisations by going through the whole process. To create a script for testing is my main part in this role. The moderation part is not included here as I chose an unmoderated form of a user study.

Participants

The following Users have participated in my user testing:

- User 1 - Economics Student - male
- User 2 - Migration and ethnic studies Student - female
- User 3 - Agricultural Science Consultant - female
- User 4 - Political Scientist - male
- User 5 - Speech Therapist - female

Test script

The following Test script was given to the participants. They should start by looking at the plot and answer some questions to get into the features of the plot as well as the dimensions. Afterwards they shall answer some questions regarding their problems and additions they would like.

Description

Please have a look at the provided plot and make yourself familiar with it. Answer the questions in the tasks section beneath by inspecting the plot in the .html file which you find

in the attachment. Afterwards reflect on the process by answering the questions in the improvements and additions section.

Tasks

- How many Dobson Units were measured on the 22nd of September of 2000?
- In what year was the first measurement of a depleted area of the ozone layer?
- Where is the longest gap in Data and what is the most likely reason for it?

Improvements

What are the improvements you would make to the plot? Where did you have problems answering?

Additions

What are additions to the plot you would make? (Legend, Descriptions, etc.)

Reflection after user testing

The users had issues with the following topics:

- zooming feature needs further explanation
- some formatting and language changes they would make to the event descriptions
- adding a trend line to make the trend visual
- restructuring and repositioning the legends of the events

With these user inputs I started a remake of the plot and added multiple features such as rolling means and better descriptions of the functions.

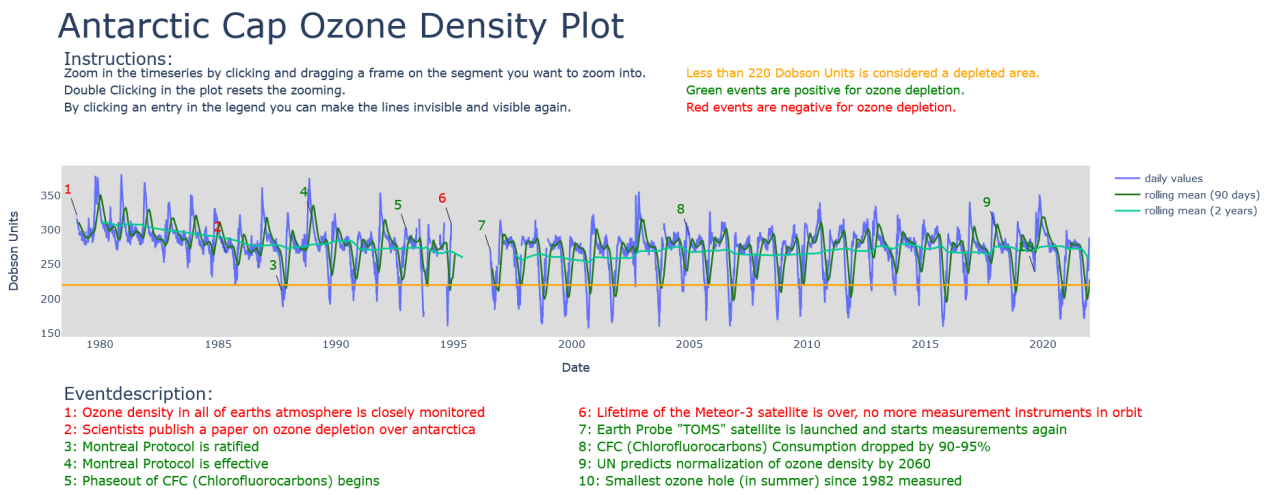


Figure 9 Time Series Plot with additions and improvements [(https://tmandelz.github.io/GDV/ozoneplotRedone.html)](https://tmandelz.github.io/GDV/ozoneplotRedone.html)

Bibliography

Budiu, R. (2017, October 1). *Quantitative vs. Qualitative Usability Testing* [Article]. Nielsen Norman Group. https://www.nngroup.com/articles/quant-vs-qual/

Castañón, J. (2019, November 8). *10 Visualizations Every Data Scientist Should Know*. Medium. https://towardsdatascience.com/10-viz-every-ds-should-know-4e4118f26fc3

deGrasse Tyson, N. (2021, September 14). *Cosmic Queries in the O-zone: Saving the World with Susan Solomon & Stephen Andersen - StarTalk Radio* (No. 36). https://www.startalkradio.net/show/cosmic-queries-in-the-o-zone-saving-the-world-with-susan-solomon-stephen-andersen/

Elgabry, O. (2019, March 2). *The Ultimate Guide to Data Cleaning* [Blogpost]. Medium. https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4

Github contributors. (2021). *Urllib3/urllib3* (1.26.7) [Python]. https://github.com/urllib3/urllib3 (Original work published 2011)

Guberman, S. (2017). Gestalt Theory Rearranged: Back to Wertheimer. *Frontiers in Psychology*, *8*, 1782. https://doi.org/10.3389/fpsyg.2017.01782

Loshin, D. (2009). Master data management. In *Master data management* (1st edition). Elsevier/Morgan Kaufmann.

NEO, N. (2021, November 29). *NASA Earth Observations (NEO)* [Article]. NASA Earth Observations (NEO); NASA Earth Observations (NEO). https://neo.gsfc.nasa.gov/

omnisci. (2021, December 13). *What is Data Quality? Definition and FAQs | OmniSci* [Article]. Data Quality Definition. https://www.omnisci.com/technical-glossary/data-quality

Ozonewatch. (2021a). *Ozone Legend*. https://ozonewatch.gsfc.nasa.gov/images/ozone_legend_full.png

Ozonewatch. (2021b). *Ozone Map Projection 26.09.2021*. https://ozonewatch.gsfc.nasa.gov/ozone_maps/images/Y2021/M09/OZONE_D2021-09-26_G%5E716X716.IOMPS_PNPP_V21_MGEOS5FP_LSH.PNG

Ozonewatch, N. (2018, October 18). *Nasa Ozone Watch: Dobson Unit facts* [Article]. Dobson Unit Facts. https://ozonewatch.gsfc.nasa.gov/facts/dobson.html

Ozonewatch, N. (2021c, September 26). *NASA | Ozone and Air Quality*. NASA | Ozone and Air Quality. https://ozoneaq.gsfc.nasa.gov/data/ozone/#

Ozonewatch, N. (2021d, September 26). *NASA Ozone Watch: Latest status of ozone*. Latest Status of Ozone. https://ozonewatch.gsfc.nasa.gov/

Ozonewatch, N. (2021e, November 1). *Ozone Hole Watch: Southern ozone map for November 2021*. Southern Ozone Map for November 2021. https://ozonewatch.gsfc.nasa.gov/Scripts/big_image.php?date=2021-11&hem=S&section=MAPS

Ozonewatch, N. (2021f, December 1). *NASA Ozone Watch: Antarctic ozone maps for 2021*. NASA Ozone Watch. https://ozonewatch.gsfc.nasa.gov/monthly/SH.html

R. Nash, E. (2021, December 14). *Ozone Depletion Article—Data Questions* [Personal communication].

Sarkar, D. (DJ). (2018, September 13). *A Comprehensive Guide to the Grammar of Graphics for Effective Visualization of Multi-dimensional…*. Medium. https://towardsdatascience.com/a-comprehensive-guide-to-the-grammar-of-graphics-for-effective-visualization-of-multi-dimensional-1f92b4ed4149

Schade, A. (2013, October 12). *Remote Usability Tests: Moderated and Unmoderated* [Article]. Nielsen Norman Group. https://www.nngroup.com/articles/remote-usability-tests/

Todorovic, D. (2008). Gestalt principles Proximity_principle. *Scholarpedia*, *3*(12), 5345. https://doi.org/10.4249/scholarpedia.5345

Wiki, G. (2017, November 21). *Visual variable—GIS Wiki | The GIS Encyclopedia* [Article]. Visual Variable. http://wiki.gis.com/wiki/index.php/Visual_variable

Wikipedia contributors. (2021). Dobson unit. In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Dobson_unit&oldid=1054036929

Wilke, C. O. (2019). *Fundamentals of Data Visualization*. O'Reilly Media, Inc. https://clauswilke.com/dataviz/aesthetic-mapping.html

Yi, M., & Sapountzis, M. (2021, September 30). *Essential Chart Types for Data Visualization* [Article]. Chartio. https://chartio.com/learn/charts/essential-chart-types-for-data-visualization/

Appendix

Figure 1 Line Plot Mean ozone depletion over years 1979-2020 4

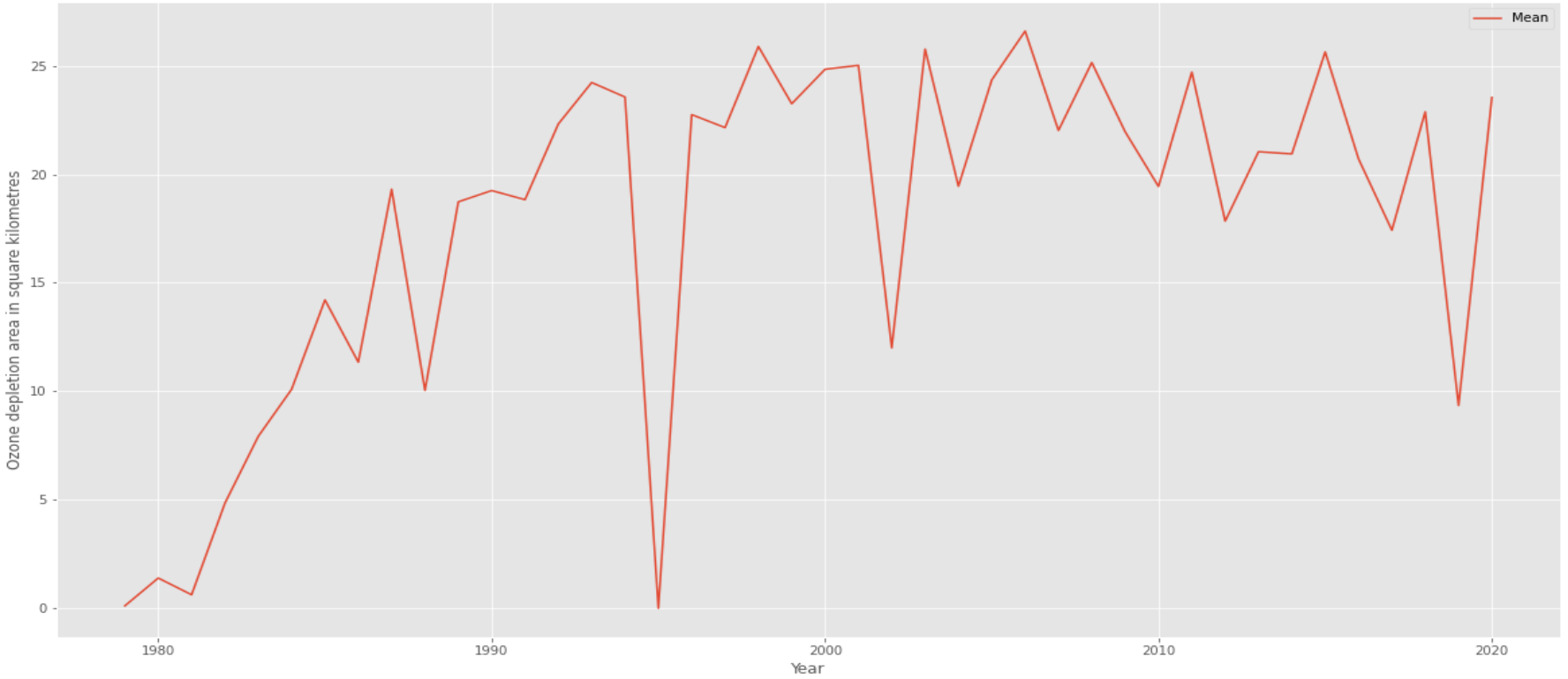


Figure 2 Scatter & Line Plot Mean, Max, Min ozone depletion over years 1979-2020 5

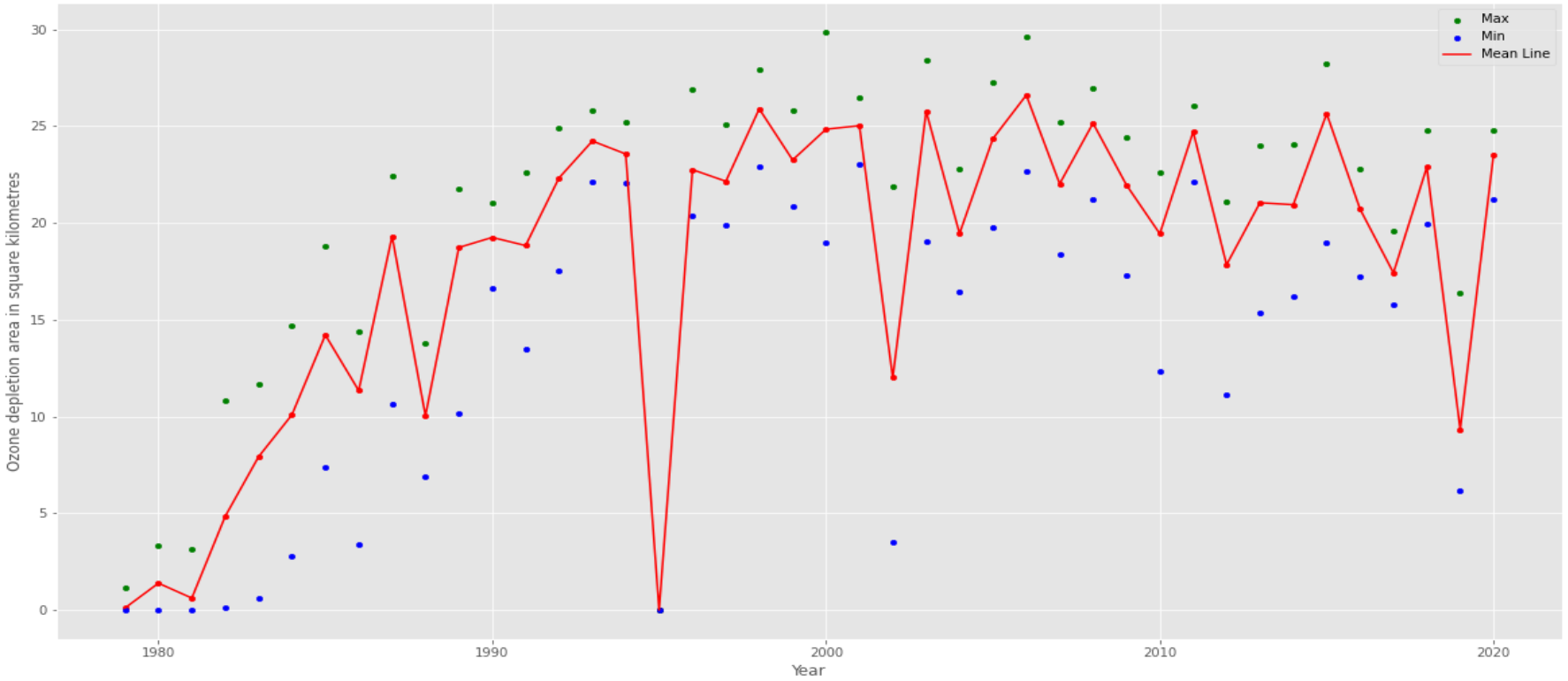


Figure 3 Ozone Map Projection 26.09.2021 5

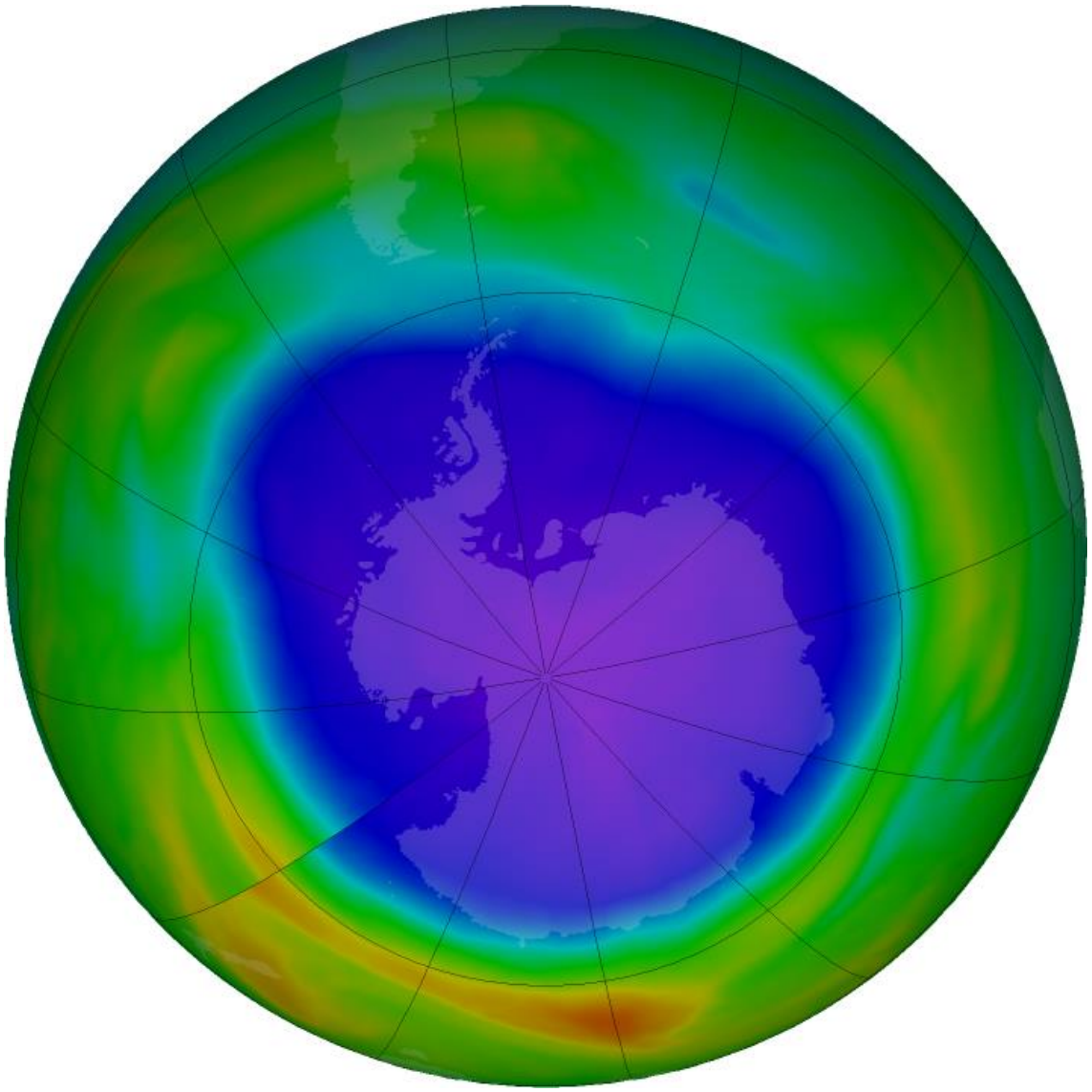


Figure 4 Ozone Dobson units legend..... 5

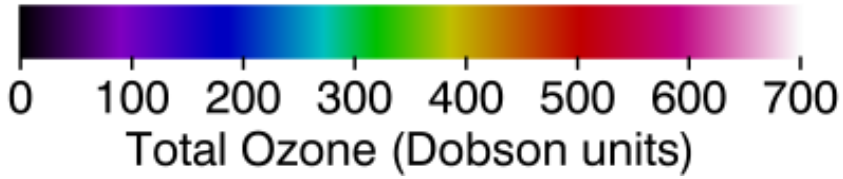


Figure 5 Line Plot size variable differences 7

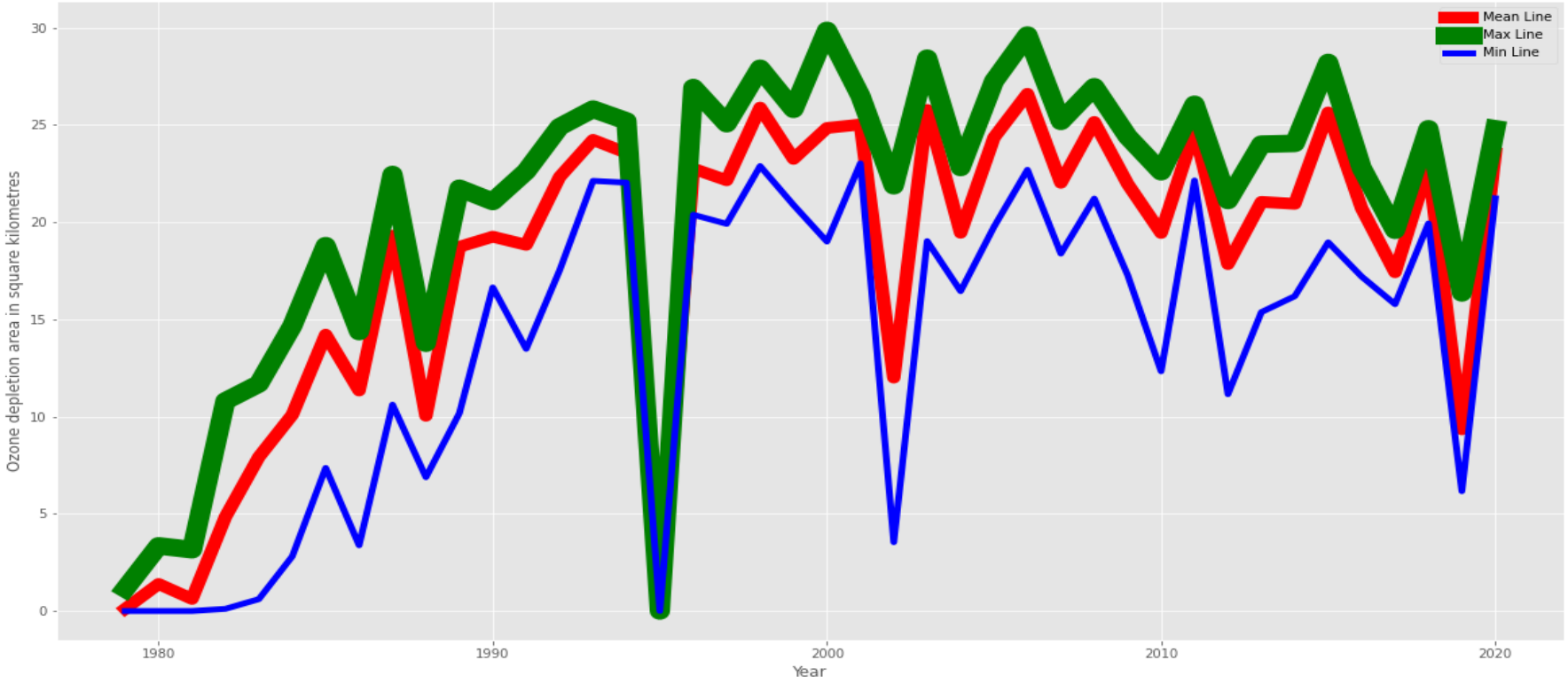


Figure 6 Line Plot color variable differences 8

Figure 7 Line Plot orientation variable differences 8

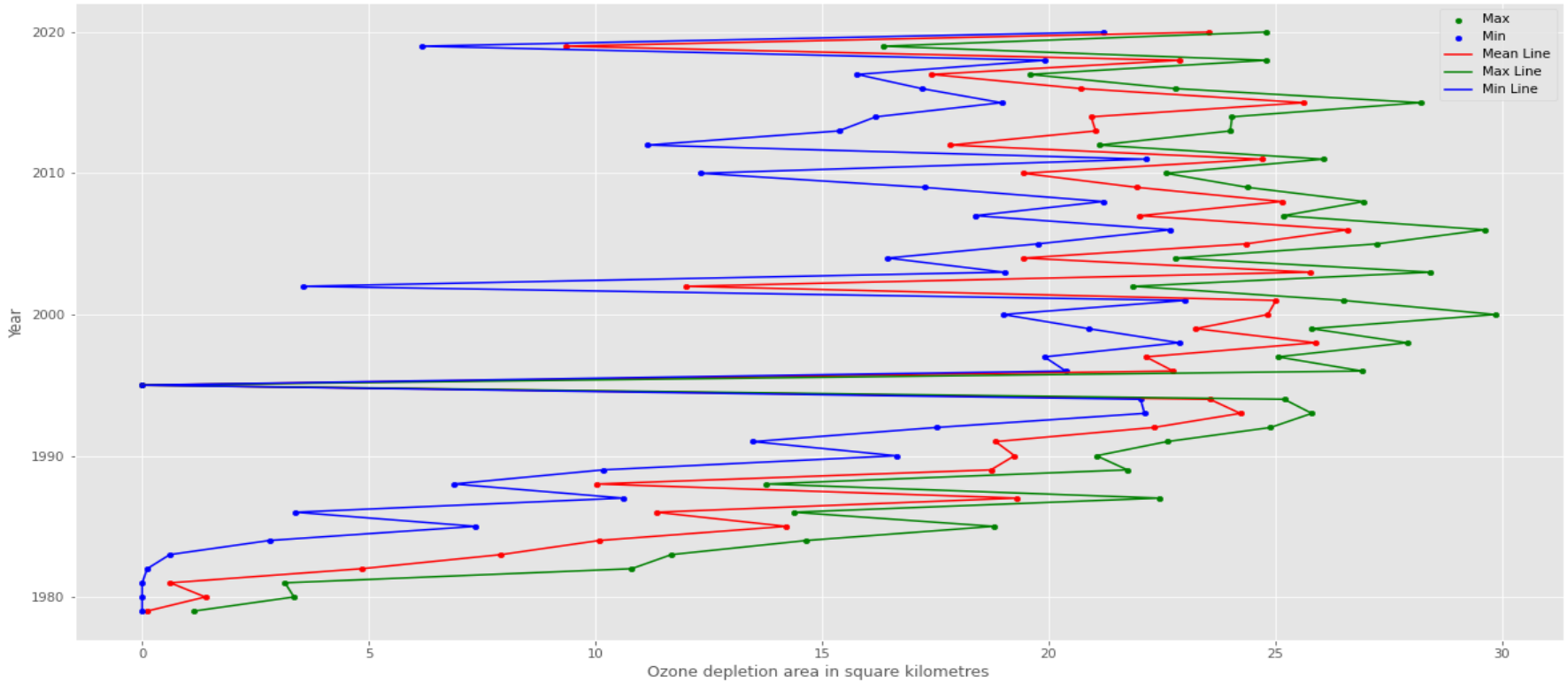


Figure 8 Time Series Plot (https://tmandelz.github.io/GDV/ozoneplot.html)..... **Error! Bookmark not defined.**

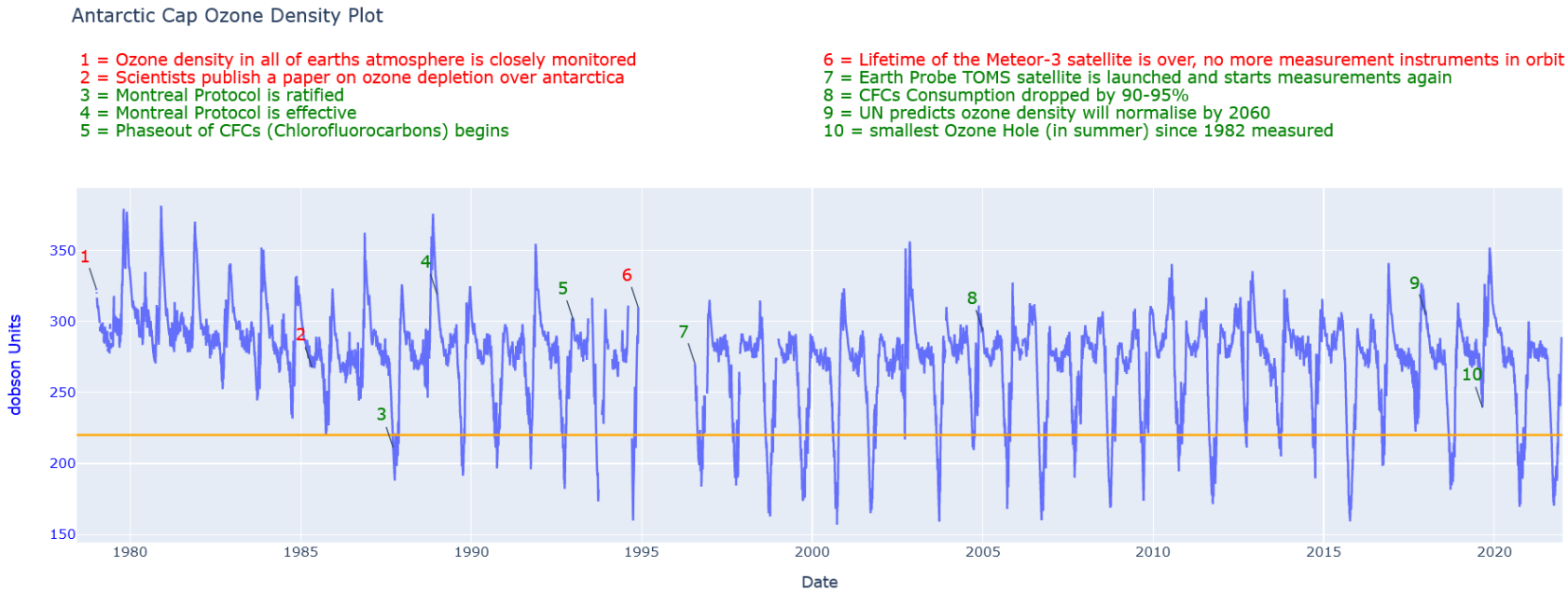


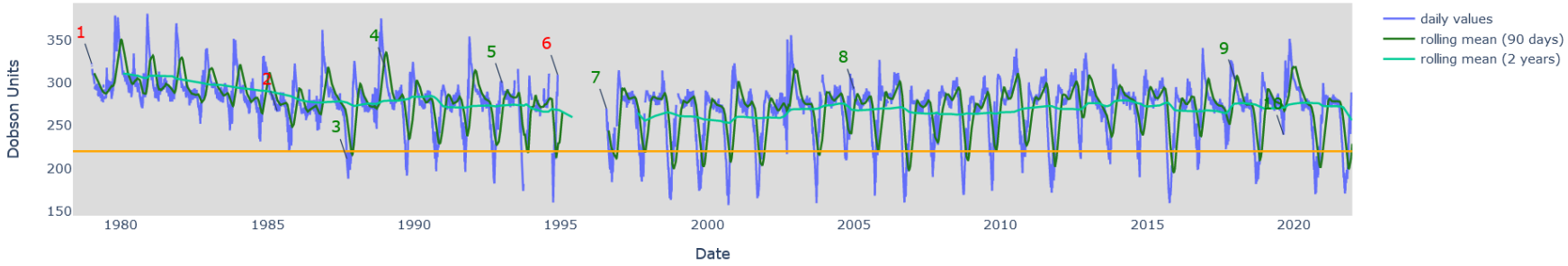
Figure 9 Time Series Plot with additions and improvements (https://tmandelz.github.io/GDV/ozoneplotRedone.html) ..... **Error! Bookmark not defined.**

Antarctic Cap Ozone Density Plot

Instructions:

Zoom in the timeseries by clicking and dragging a frame on the segment you want to zoom into.
Double Clicking in the plot resets the zooming.
By clicking an entry in the legend you can make the lines invisible and visible again.

Less than 220 Dobson Units is considered a depleted area.
Green events are positive for ozone depletion.
Red events are negative for ozone depletion.



Eventdescription:

1: Ozone density in all of earths atmosphere is closely monitored
2: Scientists publish a paper on ozone depletion over antarctica
3: Montreal Protocol is ratified
4: Montreal Protocol is effective
5: Phaseout of CFC (Chlorofluorocarbons) begins
6: Lifetime of the Meteor-3 satellite is over, no more measurement instruments in orbit
7: Earth Probe "TOMS" satellite is launched and starts measurements again
8: CFC (Chlorofluorocarbons) Consumption dropped by 90-95%
9: UN predicts normalization of ozone density by 2060
10: Smallest ozone hole (in summer) since 1982 measured