```
---
output:
  pdf_document: default
  html_document: default
---

```{r lab3_setup, include=FALSE}

knitr::opts_chunk$set(error = TRUE, warning = TRUE, message = TRUE)
suppressPackageStartupMessages({
  if (requireNamespace("tidyverse", quietly = TRUE)) library(tidyverse)
  if (requireNamespace("RcppRoll", quietly = TRUE)) library(RcppRoll)
})

# Data helpers and canonical path to the CO₂ file in this lab
data_dir <- "data"
path_data <- function(...) file.path(data_dir, ...)
co2_path <- path_data("co2_mm_mlo.txt")
```


Lab: Data manipulation and visualization using the Tidyverse (Student Version)

Data files
The lab expects a folder named **data/** next to this Rmd. Files included:

```
data/499_GRN_ANT_mass_changes.csv
data/647_Global_Temperature_Data_File.txt
data/N_seaice_extent_daily_v3.0.csv
data/antarctica_mass_200204_202310.txt
data/co2_mm_mlo.txt
```

Lesson Overview

This lesson was written by Dr. Carl Boettiger, UC Berkeley. Part of his commitment to open science is to share his teaching materials (Thanks Carl!). Check out his [research group](https://www.carlboettiger.info/)

Part of this lesson is a bit of a scavenger hunt for trying to understand the publicly available data on climate monitoring. Some information on these datasets will be difficult to find, and hopefully inspire detailed meta-data documentation but everyone in this class in the future ;) In previous years, we have downloaded this data from online, but I worried that some data would be taken down so the data is included pre-downloaded.

Conservation/ecology Topics

- Become familiar with the primary data sources and evidence for global warming

Computational Topics

- Learn to discover and interpret essential metadata about how measurements are made
- Interpret Data provenance, "Raw" and "Derived" data
- Think about measurement uncertainty, resolution, and missing values in context of environmental science data
- Reading in data from the web into the R.
- Become familiar with variations in CSV / tabular data formats and how to handle them
- Encountering missing data
- Working with dates and date-time objects
- Plotting timeseries data
- Subsetting, reshaping data
- `apply` functions

Statistical Topics

- Interpret data visualizations
- Explore noise vs seasonality vs trends
- Understand the use of windowed averages

Demo: Evidence for Global Climate Change

In this module, we will explore several of the most significant data sources on global climate change. An introduction to these data sources can be found at NASA's [Climate Vital Signs website](http://climate.nasa.gov/vital-signs).

 We will begin by examining the carbon dioxide record from the Mauna Loa Observatory.

Why C02?

Carbon dioxide (CO2) is an important heat-trapping (greenhouse) gas, which is released through human activities such as deforestation and burning fossil fuels, as well as natural processes such as respiration and volcanic eruptions.

Parsing tabular data

One of the most common formats we will interact with is tabular data. Tabular data is often presented in *plain text*, which is not as simple as it sounds, (as we shall see in a moment). NASA points us to a raw data file maintained by [NOAA on one of it's FTP servers](ftp://aftp.cmdl.noaa.gov/products/trends/co2/co2_mm_mlo.txt).

So, where does this data come from? How does one measure atmospheric CO2 levels anyway?

Data Provenance

Knowing where our data come from and how values are measured is essential to proper interpretation of the results. Data scientists usually speak of *raw data* and *derived data*, but the intepretation of these terms is always relative. Typically *raw* simply means "the data I started with" and *derived* "the data I produced." Thus our "raw" data is almost always someone else's "derived" data, and understanding how they got to it can provide important insights to our analysis. One of the first questions we should ask of any data is "where does it come from?"
```

In particular, we usually want to make note of three things:

- 1. What is the uncertainty in the data?
- 2. What is the resolution of the data?
- 3. What do missing values mean?

1. What is the uncertainty in the data?

Almost all measurements come with some degree of uncertainty, or measurement error. Often we will not be able to know this precisely. Rather, we seek a a qualtiative understanding of the measurement process to give us some idea of the relative importance of errors in the measurement process influencing the value. We may later be able to infer a more precise description of measurement error from the data itself, but this will always require assumptions about both the data-generating process itself.

2. What is the resolution of the data?

Derived data often summarize raw data in some way. For instance, global climate data is frequently reported as monthly or even annual averages, even though the raw data may be collected day by day, or even minute by minute. Data may be averaged over space as well as time, such as weather measurements made in at separate stations. Weighted averages and more complex techniques are often used as well.

3. What do missing values mean?

Real world data almost always has missing values. Here, it is important we try to understand *why* values are missing so we know how to handle them appropriately. If there is a systematic reason behind why data are missing (say, days where snowfall or storms made the weather station inaccessible) they could bias our analysis (underestimating extreme cold days, say). If data are missing for an unrelated reason (the scientist is sick, or the instrument fails) then we may be more justified in simply ommitting the data. Often we cannot know the exact reason certain data are missing and this is just something we must keep in mind as a caveat to our infererence. Frequently our results will be independent of missing data, and sometimes missing data can be accurately inferred from the data that is available.

Measuring C02 levels

So how *are* atmospheric CO2 levels measured?

Researchers shine an infrared light source of a precise intensity through dry air in a container of precisely controlled volume & pressure, ensuring a consistent number of atoms in the chamber. CO2 absorbs some of this radiation as it passes through the chamber, and then a sensor on the opposite end measures the radiation it receives, allowing researchers to calculate the amount absorbed and infer the CO2 concentration. The data are reported in parts per million (ppm), a count of the number of CO2 molecules per million molecules of dry air. These calculations are calibrated by comparing against chambers that are prepared using known concentrations of CO2. For more information, see [NOAA documentation](http://www.esrl.noaa.gov/gmd/ccgg/about/co2_measurements.html).

Measurement uncertainty:

Importantly, the measurement error introduced here is rather small, roughly 0.2 ppm. As we shall see, many other factors, such as local weather and seasonal variation also influence the measurement, but the measurement process itself is reasonably precise. As we move to other sources of data these measurement errors can become much more significant.

Resolution:

What is the resolution of the CO2 data? Already we see our data are not the actual "raw" measurements the researchers at Mauna Loa read off their instruments each day, but have been reported as monthly averages.

Missing values:

The last column of the data set tells us for how many days that month researchers collected data. We see that they only started keeping track of this information in 1974, but have since been pretty diligent -- collecting data almost every day of the month (no breaks for weekends here! What do you think accounts for the gaps? How might you test your hypothesis? Would these introduce bias to the monthly averages? Would that bias influence your conclusion about rising CO2 levels?)

Spatially our Mauna Loa data has no aggregation -- the data is collected at only one location. How might the data differ if it were aggregated from stations all over the globe?

Importing Data

Goal. Load the NOAA Mauna Loa monthly CO_2 file that is bundled with this lab at `data/co2_mm_mlo.txt`.

Watch-outs (why parsing is tricky):
- Lines starting with `#` are **comments** that describe the columns.
- Some numeric columns use **sentinel codes** like `-9.99` to mean “missing” — we’ll convert those to real `NA`.
- The file includes a `decimal_date`, but we’ll also construct a proper **`Date`** from `year` + `month` for plotting.

We’ll first try a naive import (to see the problems), then a fix that **tells R about comments**, and finally a tidyverse-style read that
assigns column names, **converts sentinels to `NA`**, and **adds a `Date` column**.

Our first task is to read this data into our R environment. To this, we will use the `read.csv` function. Reading in a data file is called *parsing*, which sounds much more sophisticated. For good reason too -- parsing different data files and formats is a cornerstone of all pratical data science research, and can often be the hardest step.

So what do we need to know about this file in order to read it into R?

```{r, render=print}
### Let's try:
co2.test1 <- read.csv(co2_path)
head(co2.test1)
```

hmm... what a mess. Let's try defining the comment symbol:
```{r, render=print}
co2.test2 <- read.csv(co2_path,
                      comment = "#")
head(co2.test2)
```

Getting there, but not quite done. Our first row is being interpreted as column names. The documentation also notes that certain values are used to indicate missing data, which we would be better off converting to explicitly missing so we don't

get confused.

Seems like we need a more flexible way to load in the data to avoid further suffering. Let's try `readr::read_table` from `tidyverse`

```{r}

co2 <- read_table(co2_path, comment = "#",
                  col_names = c("year", "month", "decimal_date",
                                "average", "interpolated", "trend",
                                "stdev_days", "uncertainty_average"),
                  col_types = c("iidddidd"),
                  na = c("-1", "-99.99", "-9.99", "-0.99"))
co2
```

Success! We have read in the data. Celebrate!

Plotting Data with `ggplot`

Effective visualizations are an integral part of data science, poorly organized or poorly labelled figures can be as much a source of peril as understanding. Nevertheless, the ability to generate plots quickly with minimal tinkering is an essential skill. As standards for visualizations have increased, too often visualization is seen as an ends rather than a means of data analysis. See [Fox & Hendler (2011)](http://science.sciencemag.org/content/331/6018/705.short) for more discussion of this.

```{r}
ggplot(co2, aes(decimal_date, average)) + geom_line()
```

Alternative: Plotting Data with base R (no packages needed)
If you want to plot something quickly without loading any packages.

```{r}
plot(y=co2$average, x=co2$decimal_date, type="l",
     ylab="Average Co2", xlab="Decimal date")
```

But... it isn't quite as pretty as with ggplot. :/

Plotting multiple series

We often would like to plot several data values together for comparison,
for example the average, interpolated and trend CO_2 data. We can do
this in three steps:

1) subsetting the dataset to the columns desired for plotting

```{r}
co2_sub <- co2 %>%
    select(decimal_date, average, interpolated, trend)
co2_sub %>% head()
```


2) rearranging the data into a "long" data table where the data values
are stacked together in one column and there is a separate column that
keeps track of whether the data came from the average,
interpolated, or trend column. Notice by using the same name,
we overwrite the original co2_sub


```{r}
co2_sub <- co2_sub %>%
    pivot_longer(!decimal_date, names_to="series",
                 values_to="ppmv")
head(co2_sub)
```

3) Visualize the data using a line plot

```{r}
co2_sub %>%
 ggplot(aes(decimal_date, ppmv, col = series)) +
  geom_line()
```

Or we can take advantage of dplyr's nifty pipping abilities and
accomplish all of these steps in one block of code. Beyond being more
succinct, this has the added benefit of avoiding creating a new object
for the subsetted data.

```{r fig.height=3}
co2 %>%
  select(decimal_date, average, interpolated, trend) %>%
  gather(series, ppmv, -decimal_date) %>%
  ggplot(aes(decimal_date, ppmv, col = series)) +  geom_line()
```

What do we see?

Our "Figure 1" shows three broad patterns:

- A trend of steadily increasing CO2 concentration from 1950 to 2015
- Small, regular seasonal oscillation is visible in the data
- Increase appears to be accelerating (convex curve)

Understanding moving averages

Trend, cycle, or noise?

> "Climate is what you expect, weather is what you get"

Present-day climate data is often sampled at both finer temporal and spatial scales than we might be interested in when exploring long-term trends. More frequent sampling can reveal higher-frequency trends, such as the seasonal pattern we observe in the CO2 record. It can also reveal somewhat greater variability, picking up more random (stochastic) sources of variation such as weather patterns.

To reveal long term trends it is frequently valuable to average out this high-frequency variation. We could spend the whole course discussing ways such averaging or smoothing can be done, but instead we'll focus on the most common methods you will see already present in the climate data we examine. The monthly record data we analyze here already shows some averaging. How was this performed?

Moving averages

```{r}
co2 %>%
 mutate(annual = RcppRoll::roll_mean(average,
                                     n=12L,
                                     align = "left",
                                     fill = NA,
                                     na.rm=TRUE,
                                     normalize=FALSE)) ->
  co2
head(co2)
```

```{r}
co2 %>% ggplot(aes(decimal_date)) +
  geom_line(aes(y=average), col="blue") +
  geom_line(aes(y=annual), col="red")
```

Lab 2 Questions: Exploring and visualizing more data realted to global climate change.

Question 1:

Each of the last years has consecutively set new records on global climate. In this section we will analyze global mean temperature data. [Data from](http://climate.nasa.gov/vital-signs/global-temperature)

 -1a. Describe the data set. You can do additional searches to learn more about the data given the description on the data page is not detailed.
The data indicates the Land-Ocean Temperature Index which is a measure of how global average temperatures have changed over long periods of time. A temperature anomaly refers to the relative difference in surface temperture 1 year's data is, compared to a 30-year average.

- 1b. Describe what kind of column each data contains and what units it is measured in.
Column A is the year - numeric data,
Column B is the change in global average surface temperature compared to the Long-Term temperature average from 1951 to 1980 in degrees celcius,
Column C is the change in global average surface temperature , which is the value of column B, with five-year locally weighted smoothed scatterplot smotthing applied. This creates a smoother line through a time texperature series. This is also measured in degrees celcius.

Then address our three key questions in understanding this data:
- 1c. How are the measurements made? What is the associated measurement uncertainty?

year - this is a known fact, and thus there is no uncertainty
The surface temperature data in column B and then C was obtained by GISS (NASA's Goddard Institute for Space Studies) temperature analyses which incorporates surface temperature measurements from more than 20,000 weather stations, ship- and buoy-based observations of sea surface temperatures, and temperature measurements from Antarctic research stations. An algorythm is then used to consider the varied spacing of temperature stations around the globe and urban heat island effects on surface temperature.
Additionally, the measurement error could be increased due to local weatehr and season variations of temperatures, yet the long length of the data collection period, as well as the large number of GISS plots, streamlines these errors, and reduces uncertainty.

-1d. What is the resolution of the data?

Ultimately, the data is presented on the NASA website as the average temperature anomoly per year both with an without smoothing in degrees celcius.

-1e. Are their missing values? How should they be handled?
There doesn't appear to be any missing values in this data set.

Question 2:

- 2a: Construct the necessary R code to import and prepare for manipulation the [following data set] (http://climate.nasa.gov/system/internal_resources/details/original/647_Global_Temperature_Data_File.txt) You may wish to re-label the columns to something easier to read/understand.

```{r}
temp <-
read_table("https://data.giss.nasa.gov/gistemp/graphs/graph_data/Global_Mean_Estimates_based_on_Land_and_Ocean_Data/graph.txt",
                   col_types = "idd", skip = 5, na = "*", col_names = c("Year", "Annual_Mean", "5_year_Mean"))
temp
```

Question 3:

- 3a. Plot the trend in global mean temperatures over time.
- 3b. Describe what you see in the plot and how you interpret the patterns you observe.

```{r}
## your code here
```

```
ggplot(temp, aes(x = Year, y = Annual_Mean, color = "blue")) + geom_line() + geom_line(aes(x = Year, y = `5_year_Mean`, color = "red")) + labs(x = "Year", y = "Temperature Anomaly Celcius")

```
The plot illustrates how the smoothed, 5 year data folows the same overall trend in surface temperatures, however does not account for variations each year, and thus is less representative of smaller scale temperature anomalies. In turn, this 5_year_mean data would be less accurate in identifying natural disaters or deforestation or other events that may have promoted surface temperatures to decrease or increase. The Annual_mean is more effective for this purpose. However, the smoothed data does show a much simpler intepretation of the overall trend in the data, and could be helpful for understanding change over time, and analysing multi-year events such as wars or pandemics which may indirectly impact global surface temperatures.

Question 4: Evaluating the evidence for a "Pause" in warming?

The [2013 IPCC Report](https://www.ipcc.ch/pdf/assessment-report/ar5/wg1/WG1AR5_SummaryVolume_FINAL.pdf) included a tentative observation of a "much smaller increasing trend" in global mean temperatures since 1998 than was observed previously. This led to much discussion in the media about the existence of a "Pause" or "Hiatus" in global warming rates, as well as much research looking into where the extra heat could have gone. (Examples discussing this question include articles in [The Guardian](http://www.theguardian.com/environment/2015/jun/04/global-warming-hasnt-paused-study-finds), [BBC News](http://www.bbc.com/news/science-environment-28870988), and [Wikipedia](https://en.wikipedia.org/wiki/Global_warming_hiatus)).

- 4a. By examining the data here, what evidence do you find or not find for such a pause? Present an written/graphical analysis of this data (using the tools & methods we have covered so far) to argue your case.

```{r}
after.1998 <- temp %>%
  subset(Year >= 1998 & Year <= 2013) %>%
  mutate(YearBin = "after") %>% #specify if particular row was after 1998
  arrange(Year) %>% #naturally chronological so this orders it
  mutate(YearNum = 1:nrow(.)) # assigns number 1998 is year 1 etc

before.1998 <- temp %>%
  subset(Year < 1998 & Year >= 1983) %>%
  mutate(YearBin = "before") %>%
  arrange(Year) %>%
  mutate(YearNum = 1:nrow(.))

# join the data back together
new.temp <- rbind(before.1998, after.1998)

new.temp %>%
ggplot(aes(YearNum, col=YearBin)) + #This specifies the x axis/variable for all the geom_lines
  geom_line(aes(y= Annual_Mean, col=YearBin)) +
  geom_line(aes(y =`5_year_Mean`, col=YearBin)) +
  labs( x = "Year", y = "Surface Temp Annual Mean C")



```
The slope from after 1998 looks a little steeper than before 1998, however it is hard to tell without definitive slope line. This supports the data in the articles such as in the journal Science, which showed the rate of warming (0.116C per decade) over the past 15 years (after 1998) is almost the same, if not, higher, than the rate of warming before 1998 (0.113C per decade).

- 4b. What additional analyses or data sources would better help you refine your arguments?

Data on month by month surface temperatures and their averages may better inform the arguments but providing much more precise data points on a smaller time scale. To be able to see if there were missed recordings, or outliers in the surface temoertaure, these finer recorded data points would be much more beneficial.

Question 5: Rolling averages

-5a. What is the meaning of "5 year average" vs "annual average"?
The 5 year average takes the surface temperature from 5 consecutive years and creates an average, whereas the annual average takes the mean surface temperature for one singular year and is thus unique to that year alone.

-5b. Construct 5 year averages from the annual data. Construct 10 & 20-year averages. Plot the different averages and describe what differences you see and why.

```{r}
temp_10year <-temp %>%
   mutate(
   ten_year = floor(Year/10)*10
   ) %>%
   group_by(ten_year) %>%
   summarise(ten_yearMean = mean(Annual_Mean))

temp_20year <- temp %>%
  mutate(
   twenty_year = floor(Year/20)*20
   ) %>%
   group_by(twenty_year) %>%
   summarise(twenty_yearMean= mean(Annual_Mean))

temp_5year <- temp %>%
  mutate(
   five_year = floor(Year/5)*5
   ) %>%
   group_by(five_year) %>%
   summarise(five_yearMean= mean(Annual_Mean))

ggplot() +
  geom_line(data = temp_20year, aes(x = twenty_year, y= twenty_yearMean), color = "blue") +
  geom_line(data = temp_10year, aes(x = ten_year, y= ten_yearMean), color = "red") +
  geom_line(data = temp_5year, aes(x = five_year, y= five_yearMean), color = "green") +
  labs( x = "Year", y = "Surface Temp Mean C")
```

```

```{r}
### your code here
```

ANSWER: ...
The green line, which represents the 5 year average shows a much more detailed analysis of carbon dioxide fluctuations, which is expected due to the shorter time periods in which it is smoothed over. Hence, we can interpret this data with greater accuracy to predict why co2 levels peaked or dipped. Conversely, the most smoothed data represented by the blue line, especially when compared to the 5 year averages, is very inaccurate in representing smaller fluctations in the data, however, can give us a better visual of the overall increasing and decreasing trends over longer time scales.

Question 6: Melting Ice Sheets?

- [Data description](http://climate.nasa.gov/vital-signs/land-ice/)
- [Raw data file](http://climate.nasa.gov/system/internal_resources/details/original/499_GRN_ANT_mass_changes.csv)

6a. Describe the data set: what are the columns and units? Where do the numbers come from?

The data set includes the mass of both Antarctica and Greenland each year from 2002 to 2014, measured in gigatons (Gt). The numbers come from this article: JPL GRACE Mascon Ocean, Ice, and Hydrology Equivalent Water Height Release 06 Coastal Resolution Improvement (CRI) Filtered Version 1.0.

6b. What is the uncertainty in measurement? Resolution of the data? Interpretation of missing values?

- Uncertainty: The data is derived from glacial isostatic adjustment (GIA) estimates, which combines vertical and horizontal motion of the earth's surface, downward deflection of the goeid in previously glaciated regions and a linear contribution to the secular rate of geoid height change (A et al. and Ivins et al.). Uncertainty may arise from measurements due to seasonal and daily variations in temperature causing glacial melt, or accumulation. Additionally, differences in glacial thickness across the sheer distance between measurement locals could create uncertainty.

- Resolution:
The data employs a Coastal Resolution Improvement (CRI) filter that reduces leakage errors across coastlines. Additionally, durign the mathematical analysis, scientists considered known facts about the location of oceans, and the movement of water and ice masses. This process intends to reduce common types of errors, thus leaving the final data clean and as accurate as possible.

- Missing data: There are no missing values, however it is likely that by incoorperating real life geophysical information into the data analysis process, missing values were quantified and explained.

6c. Construct the necessary R code to import this data set as a tidy `Table` object.
```{r}
### check your working directory!

setwd("~/Library/CloudStorage/OneDrive-UniversityOfOregon/ds-environ-main/labs")

### your code

ice_table <- read.csv("/Users/sydneyjames/Library/CloudStorage/OneDrive-UniversityOfOregon/ds-environ-main/labs/data/499_GRN_ANT_mass_changes.csv")
```

- 6d: Plot the data and describe the trends you observe.

```{r}
### your code
colnames(ice_table) <- c("year", "Greenland", "Antarctica")
ggplot() +
  geom_line(data = ice_table, aes(x=year, y = Greenland), color = "green") +
  geom_line(data = ice_table, aes(x=year, y = Antarctica), color = "purple") +
  labs( x = "Year", y = "Ice Mass (Gigatons)")

```
Overall, the data shows a decreasing trend, indicating the ice mass in both Greenland and Antarctica are both declined between the years 2002 to 2014. The graph illustrates that the rate of ice mass decline in Antarctica is slower than that of Greenland, and it started with a smaller mass in gigatons.

Isn't this fun?!

#### Question 7: Rising Sea Levels?

- [Data description](http://climate.nasa.gov/vital-signs/sea-level/)
- [Raw data file](http://climate.nasa.gov/system/internal_resources/details/original/121_Global_Sea_Level_Data_File.txt)

- 7a. Describe the data set: what are the columns and units?
This data set contains the global averages of sea level change, or Global Mean Sea Level (GMSL) in cm over time from 1993 to 2025. The first column is the year, or point in time in relation to the year that the measurment was taken. The second column is the GMSL in cm, which was calculated as the area-weighted average over each mao in the time series of Simple Gridded Sea Surface Height, and computed over 10 and then 7 days, thus explaining the overlap in the data, and inconsistency between times. The third column is these 10/7 day observations smoothed over 60 days. Both of these observations were shifted to have zero mean over the calender 1993.

- 7b. Where do these data come from (try googling if the description isn't satisfactory)?

The data was based on observations conducted by satellite of sea surface height anomaly, measured by reference radar altimeter missions. The indicator values were calculated using NASA-HDR simple gridded sea surface height from standardizes reference missions.

- 7c. What is the uncertainty in measurement? Resolution of the data? Interpretation of missing values?

ANSWER:
- Uncertainty: As the data was obtained using satellite imagery, there is a high degree of uncertainty, as it relies soley on technological advancements in imaging a physical height. This method has a much higher degree of uncertainty than if the sea level was physically mesaured at a closer proximity using ml or mm, which has a greater degree of accuracy. Additionally,
```

there is uncertainty that may arise due to the seasonal and daily fluctuations of sea level due to ships, rain, flooding etc, coupled with the inconsistent fractions of the year that the data points relate to.
- Resolution: The data was smoothed over a 60 day average to reduce the level of uncertainty, especially in relation to timing. It was thus recorded in both 10 or 7 day intervals, and 60 day averages. The estimations were NOT however adjusted for Glacial Isostatic Adjustment, to account for the slight long-term depression of the sea floor which would lower the sea level.
-Missing data: There is no missing data as per this data set, however the inconstant time intervals could be representative of missing data points.

Question 8:

- 8a. Construct the necessary R code to import this data set as a tidy `Table` object.
```{r}
### your code

GMSL_data <- read.csv("/Users/sydneyjames/Library/CloudStorage/OneDrive-UniversityOfOregon/ds-environ-main/labs/data/GMSL_data.csv")
```

- 8b. Plot the data.

```{r}
### your code
ggplot() +
  geom_line(data = GMSL_data, aes(x=year_and_fraction, y = GMSL_cm), color = "blue") +
  geom_line(data = GMSL_data, aes(x=year_and_fraction, y = GMSL_60smooth), color = "red") +
  labs( x = "Year", y = "Global Mean Sea Level cm") +
  scale_x_continuous(
    breaks = seq(floor(min(GMSL_data$year_and_fraction)),
                 ceiling(max(GMSL_data$year_and_fraction)),
                 by = 5))
```

- 8c. Describe the trends you observe.

The global mean sea level as evidenced by this data, has increased since 1993 overall, however is a very consistent pattern of rising and falling of sea level in both the raw (blue), and the 60 day smoothed (red) data. I added more labels on the x axis to incriment the years by 5. By doing this, I can observe that there are 5 spikes in the data in every group of 5 years. Hence, each year, there is a clear minimum sea level, and maximum sea level. Over the entire data set, the fluctuation between the minimum and the maximum appears, based on pure observation of the data, to be similar in cm, however the total sea level increases.

Demo: Exploring Seasonal Oscillations

Circling back to our CO2 data from Mauna Loa, let's practice summarizing our data in different ways.

- Demonstrate that the periodic behavior truly is seasonal
- What month is the maximum? What is the minimum?
- What do you think could explain the seasonal cycle observed here?

```{r}
### Seasonal oscillation
# My version of visualising seasonality looks like this
co2 %>%
  group_by(year) %>%
  ### Note we need month[which.max] since not all years have all 12 months
  summarise(max_month = month[which.max(average)],
            min_month = month[which.min(average)]) %>%
  gather(id, value, -year) %>%
  mutate(month = month(value, label = TRUE)) %>% # needs lubridate package to work
  ggplot(aes(month, fill=id)) + geom_bar() + facet_wrap(~id) +labs(fill = "Oscillation")
```
From the data, we can see that May seems to have the highest Co2 each year, and September and October have the least recorded Co2. This could be explained by differences in carbon sequestration in different parts of the world, where warmer months have greater ecological biomass to absorb more carbon dioxide, and during colder months, there is less available photosynthesizing biomass to absorb it. However as this is global data, it is difficult to assume a relationship between temperature and co2 averages, as, for example, May in North America is warm, but cold in Australia.

Question 9 (Gradaute student or Extra credit) : Summarize the data of your choice.

Using any of the datasets from this demo/lab (co2, icesheets, temp etc.) come up with an interesting question to ask that necessitates you to summarize the data in some way, make a visualization, and answer your question.

- 9a. What dataset are you using? What is your question?

I will use the co2 data. My question is: How does the difference between co2 average in May and co2 average in December change from 1958 to 2025.

- 9b. Summarize your data to answer your question.

```{r}
co2_twelveSix <- co2 %>%
  filter(month == 6 | month == 12) %>%
  group_by(year) %>%
  select(year, month, average) %>%
  pivot_wider(names_from = month,
              values_from = average
              ) %>%
  mutate(deltaDecJun = `6` - `12`)

colnames(co2_twelveSix) <- c("year", "June", "Dec","deltaDecJun" )

co2_twelveSix

```

- 9c. Plot your data

```{r}
ggplot(co2_twelveSix, aes(x = year, y = deltaDecJun)) + geom_line()
```

- 9d. Interpret you plot. What is the answer to your question?

From the data, it is observed that there is a gradual decrease in the difference between December and May co2 values, each year, suggesting that while overall co2 ppm was increasing (as we know from previous questions), the seasonal fluctuations between summer and winter globally were also decreasing.