21MAP500 Project

Last updated: 02, 06, 2022

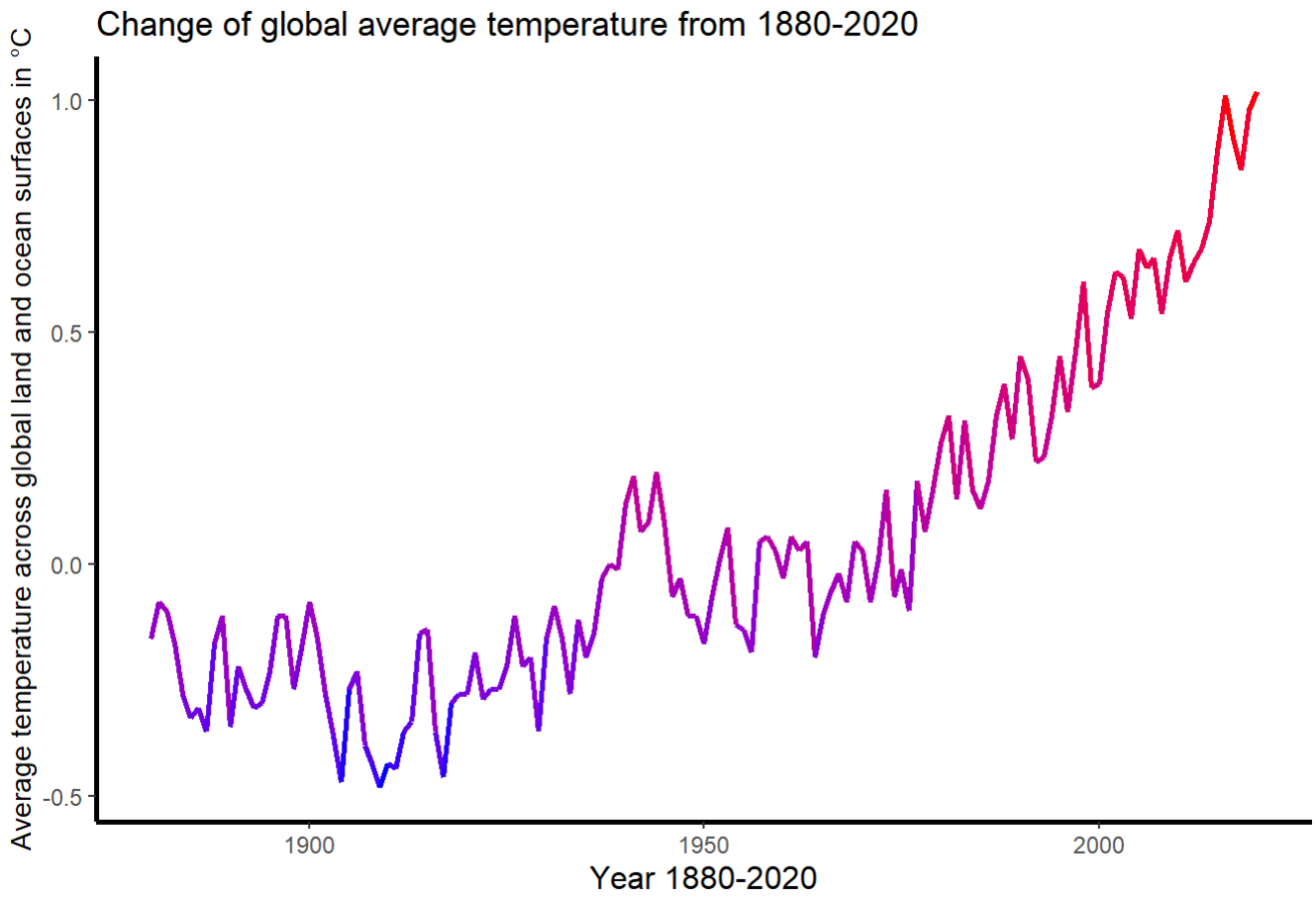
1 Question

1.1

```
read_lines(file = here("data", "nasa_global_temperature.txt"), n_max = 5L) #read the file first
 to see what kind of data we are using
```

```
## [1] "Land-Ocean Temperature Index (C)" "-------------------------------"
## [3] ""                                 "Year No_Smoothing  Lowess(5)"
## [5] "----------------------------"
```

```
read_table(here("data", "nasa_global_temperature.txt"), skip = 3) -> nasa_temp #read the file in
to a tibble dataframe and save into a new  variable

nasa_temp %>%
  slice(2:142) %>% #get rid of the first row
  select(Year, No_Smoothing) %>% #select the columns we need
  mutate(Year = parse_date(Year, format = "%Y")) %>%  #change type of "Year" to date type
  rename(date = Year, temp = No_Smoothing) -> nasa_temp  #rename "Year" to "date", "No_Smoothin
g" to "temp" and save to variable nasa_temp

nasa_temp %>%
  ggplot(mapping = aes(x = date, y = temp, colour = temp)) +
  geom_line(size = 1) +  #use "date" as x axis, "temp" as y axis to plot a line graph, change li
ne size
  labs(    #change titles
    title = "Change of global average temperature from 1880-2020",
    x = "Year 1880-2020",
    y = "Average temperature across global land and ocean surfaces in °C",
    caption = "Global average temperature has became higher than before."
  ) +
  scale_color_continuous(low = "blue", high = "red") +  #change colour to emphasis temperature r
ise
  guides(colour = FALSE) +  #remove redundant legend
  theme_classic()+  #remove background
  theme(
        axis.line.x = element_line(size = 1),      #adjust x axis size
        axis.line.y = element_line(size = 1),      #adjust y axis size
        axis.title.x = element_text(size = 12),    #adjust x axis title size
        axis.title.y = element_text(size = 10.5)   #adjust y axis title size
        )
```



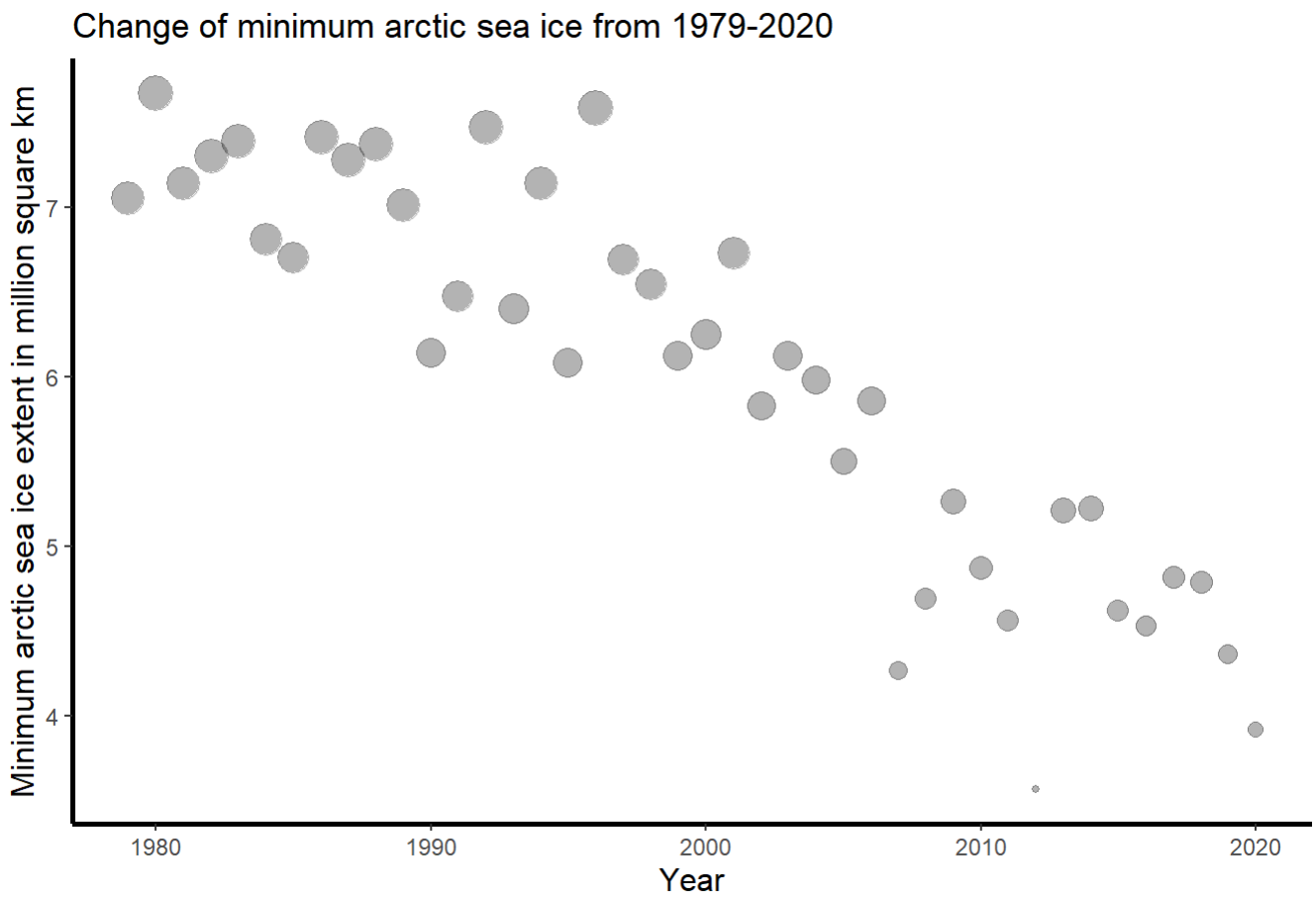
Global average temperature has became higher than before.

1.2

```
read_lines(file = here("data", "nasa_arctic_sea_ice.csv"), n_max = 10)  #read the file first to
 see what kind of data we are using
```

```
##  [1] "year; mo;    data-type; region; extent;   area"
##  [2] "1979;  9;      Goddard;      N;   7,05;   4,58"
##  [3] "1980;  9;      Goddard;      N;   7,67;   4,87"
##  [4] "1981;  9;      Goddard;      N;   7,14;   4,44"
##  [5] "1982;  9;      Goddard;      N;   7,30;   4,43"
##  [6] "1983;  9;      Goddard;      N;   7,39;   4,70"
##  [7] "1984;  9;      Goddard;      N;   6,81;   4,11"
##  [8] "1985;  9;      Goddard;      N;   6,70;   4,23"
##  [9] "1986;  9;      Goddard;      N;   7,41;   4,72"
## [10] "1987;  9;      Goddard;      N;   7,28;   5,64"
```

```
nasa_ice <- read_csv2(file = here("data", "nasa_arctic_sea_ice.csv"),
          col_select=c(year, extent)) #read the file into a tibble data frame and save into a ne
w variable



nasa_ice %>%
  mutate(across(year, as.character)) %>%  #change "year" into character type
  mutate(year = parse_date(year, format = "%Y")) %>%  #then to date type
  rename(date = year, ice = extent) -> nasa_ice  #rename columns and save the result to variable
nasa_ice

nasa_ice %>%
  ggplot(mapping = aes(x = date, y = ice, size = ice)) +
  geom_point(alpha = 0.3) -> p #use "date" as x axis, "ice" as y axis to plot a scatter plot

p +
  labs(    #change titles
    title = "Change of minimum arctic sea ice from 1979-2020",
    x = "Year",
    y = "Minimum arctic sea ice extent in million square km",
    caption = "The Arctic ice started decline around 1995"
  ) +
  theme_classic() + #remove background
  theme(
        axis.line.x = element_line(size = 1),     #adjust x axis size
        axis.line.y = element_line(size = 1),     #adjust y axis size
        axis.title.x = element_text(size = 12),   #adjust x axis title size
        axis.title.y = element_text(size = 12)    #adjust y axis title size
        ) +
  guides(size = FALSE)  #remove legend
```



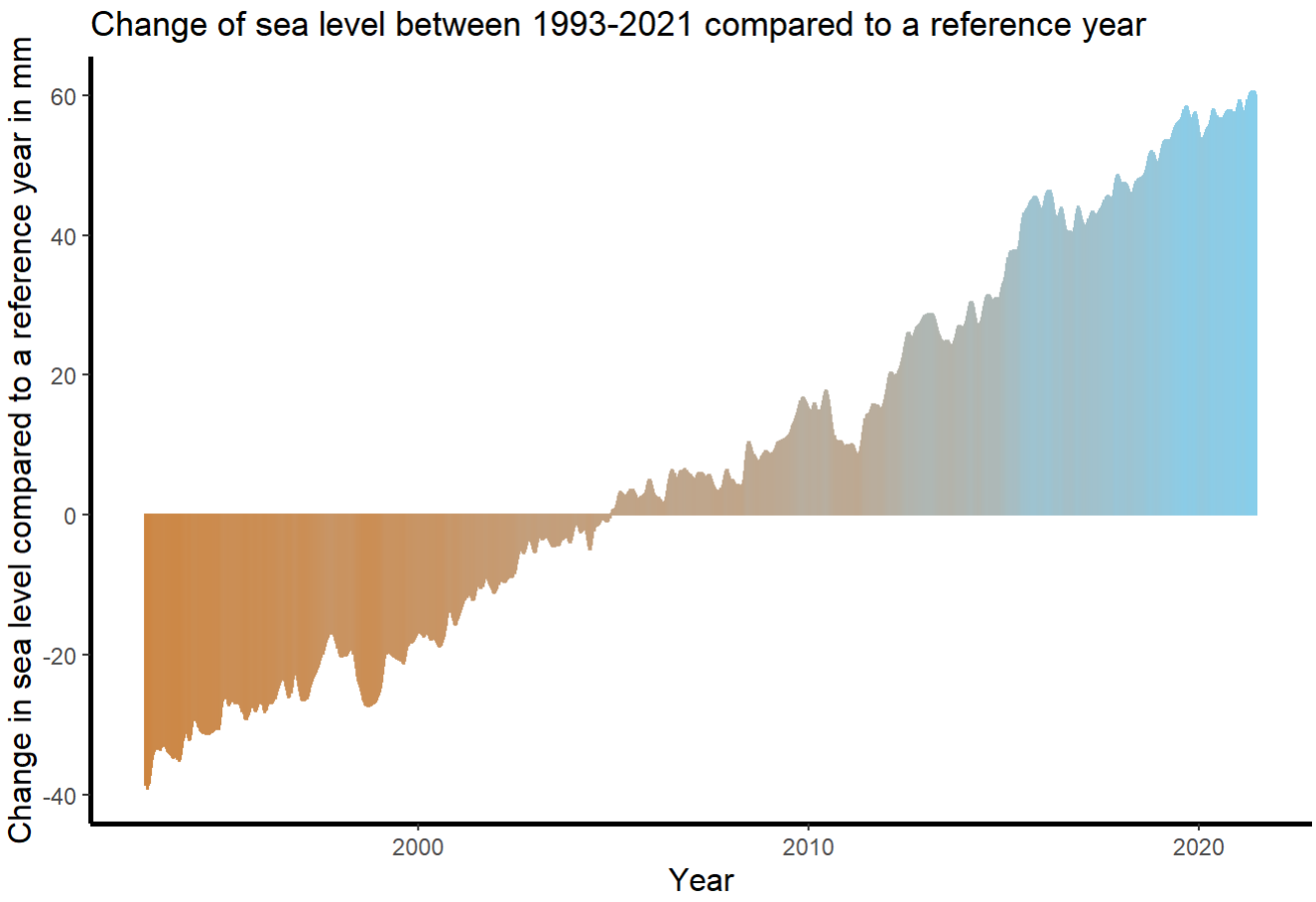
The Arctic ice started decline around 1995

1.3

```
read_lines(file = here("data", "nasa_sea_level.csv"), n_max = 10)  #read the file first to see w
hat kind of data we are using
```

```
##  [1] "HDR Global Mean Sea Level Data"
##  [2] "HDR"
##  [3] "HDR This file contains Global Mean Sea Level (GMSL) variations computed at the NASA God
dard Space Flight Center under the "
##  [4] "HDR auspices of the NASA Sea Level Change program. The GMSL was generated using the Int
egrated Multi-Mission Ocean Altimeter Data for "
##  [5] "HDR Climate Research (http://podaac.jpl.nasa.gov/dataset/MERGED_TP_J1_OSTM_OST_ALL_V5).
It combines Sea Surface Heights from "
##  [6] "HDR TOPEX/Poseidon, Jason-1, OSTM/Jason-2, and Jason-3 to a common terrestrial referenc
e frame with all inter-mission biases, range and "
##  [7] "HDR geophysical corrections applied and placed onto a georeferenced orbit. This creates
a consistent data record throughout "
##  [8] "HDR time, regardless of the instrument used."
##  [9] "HDR"
## [10] "HDR The data can be found below. A separate figure file, Global_Sea_Level_Graph, was ge
nerated using the GMSL data (listed in "
```

```
nasa_sea <- read_table(file = here("data", "nasa_sea_level.csv"), skip = 48, col_names = FALSE)
#read the file into a tibble skip  the first 48 lines we don't need and save into a new variable


nasa_sea %>%
  select(X3, X12) %>%  #select the column we need
  rename(date = X3, sea = X12) %>% #rename the column to "date" and "sea"
  mutate(date = date_decimal(date, tz = "UTC")) -> nasa_sea #change the type to date type and sa
ve to variable nasa_sea

nasa_sea %>%
  ggplot(mapping = aes(x = date, y = sea, colour = sea)) +
  geom_col() -> p #use "date" as x axis, "sea" as y axis to plot a histogram


p +
  labs(    #change titles
    title = "Change of sea level between 1993-2021 compared to a reference year",
    x = "Year",
    y = "Change in sea level compared to a reference year in mm",
    caption = "Around 2005, sea level started to exceed the reference year sea level."
  ) +
  scale_color_continuous(low = "tan3", high = "skyblue") + #change colour to emphasis sea level
 risen
  theme_classic() + #remove background
  theme(
        axis.line.x = element_line(size = 1),     #adjust x axis size
        axis.line.y = element_line(size = 1),     #adjust y axis size
        axis.title.x = element_text(size = 12),   #adjust x axis title size
        axis.title.y = element_text(size = 12),   #adjust y axis title size
        panel.grid.major.x = element_blank(),     # remove thick vertical grid lines
        panel.grid.minor.x = element_blank()      # remove thin vertical grid lines
        ) +
  guides(colour = FALSE)  #remove legend
```



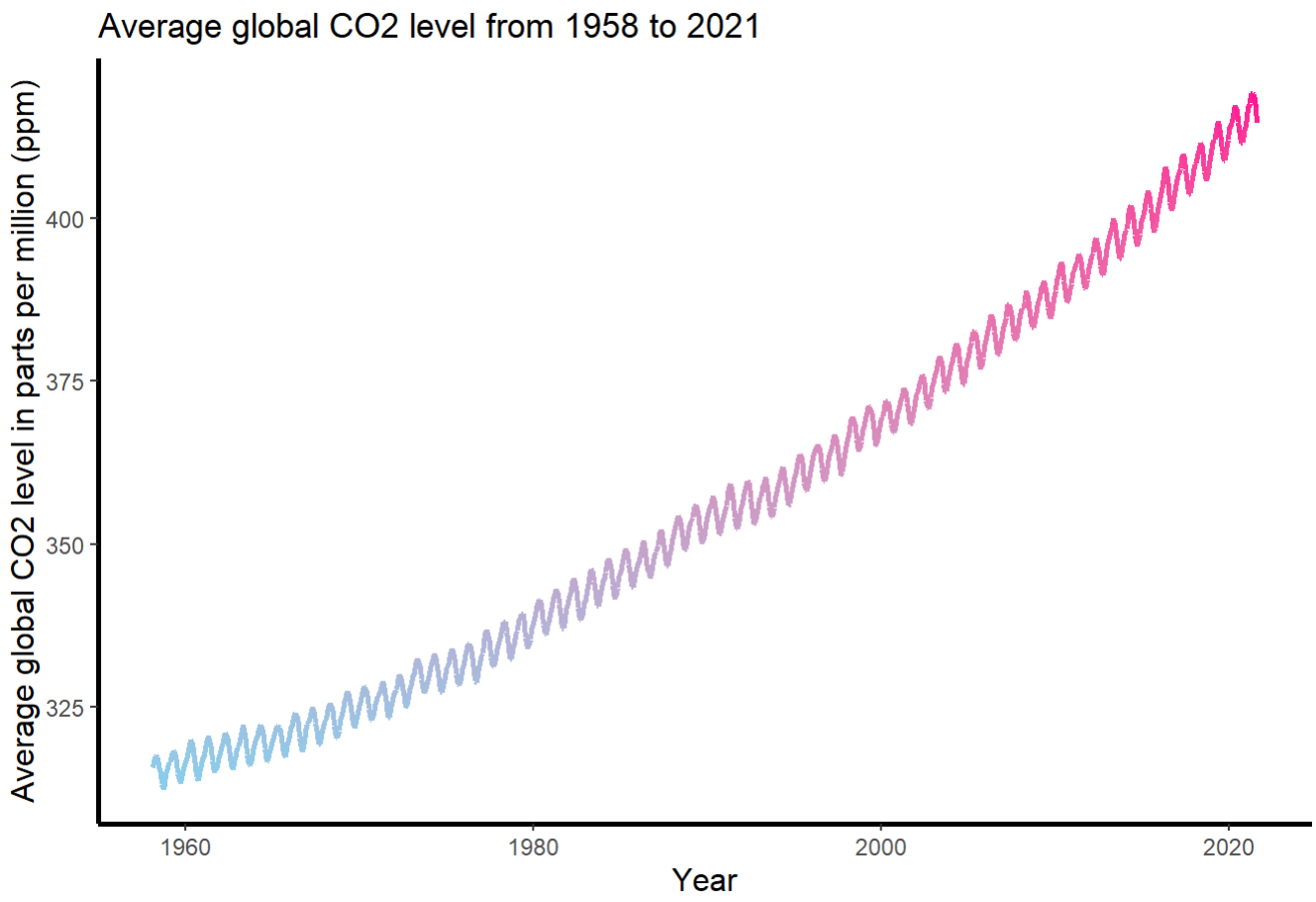
Around 2005, sea level started to exceed the reference year sea level.

1.4

```
read_lines(file = here("data", "nasa_carbon_dioxide.txt"), n_max = 10) #read the file first to s
ee what kind of data we are using
```

```
##  [1] "# ----------------------------------------------------------------------"
##  [2] "# USE OF NOAA GML DATA"
##  [3] "# "
##  [4] "# These data are made freely available to the public and the"
##  [5] "# scientific community in the belief that their wide dissemination"
##  [6] "# will lead to greater understanding and new scientific insights."
##  [7] "# The availability of these data does not constitute publication"
##  [8] "# of the data.  NOAA relies on the ethics and integrity of the user to"
##  [9] "# ensure that GML receives fair credit for their work.  If the data "
## [10] "# are obtained for potential use in a publication or presentation, "
```

```
nasa_co2 <- read_table(file = here("data", "nasa_carbon_dioxide.txt"),
                       skip = 52,) #read the file into a tibble and save into a new variable
nasa_co2 %>%
  select(average, alized) %>%  #select the columns we need
  rename(date = average, co2 = alized) %>%  #rename the columns to "date" and "co2"
  mutate(date = date_decimal(date)) %>%   #use decimal_date to calculate the fraction year
  mutate(date = as_date(date)) %>% #change to date type
  mutate(date = floor_date(date, unit = "month")) %>%  #round date to the first day of the month
  filter(co2 > 0) -> nasa_co2 # filter out outliers


nasa_co2 %>%
  ggplot(mapping = aes(x = date, y = co2, colour = co2)) +
  geom_line(size = 1) -> p #visualize data with a scatter plot, "date" as x axis, "co2" as y axi
s



p +
  labs(    #change titles
    title = "Average global CO2 level from 1958 to 2021",
    x = "Year",
    y = "Average global CO2 level in parts per million (ppm)",
    caption = "The global co2 level has increased dramatcally since 1958."
  ) +
  scale_color_continuous(low = "skyblue", high = "deeppink") + #change colour to emphasis co2 ri
sen
  theme_classic() + #remove background
  theme(
        axis.line.x = element_line(size = 1),     #adjust x axis size
        axis.line.y = element_line(size = 1),     #adjust y axis size
        axis.title.x = element_text(size = 12),   #adjust x axis title size
        axis.title.y = element_text(size = 12),   #adjust y axis title size
        panel.grid.major.x = element_blank(),     #remove thick vertical grid lines
        panel.grid.minor.x = element_blank()      #remove thin vertical grid lines
        ) +
  guides(colour = FALSE)  #remove legend
```



The global co2 level has increased dramatcally since 1958.

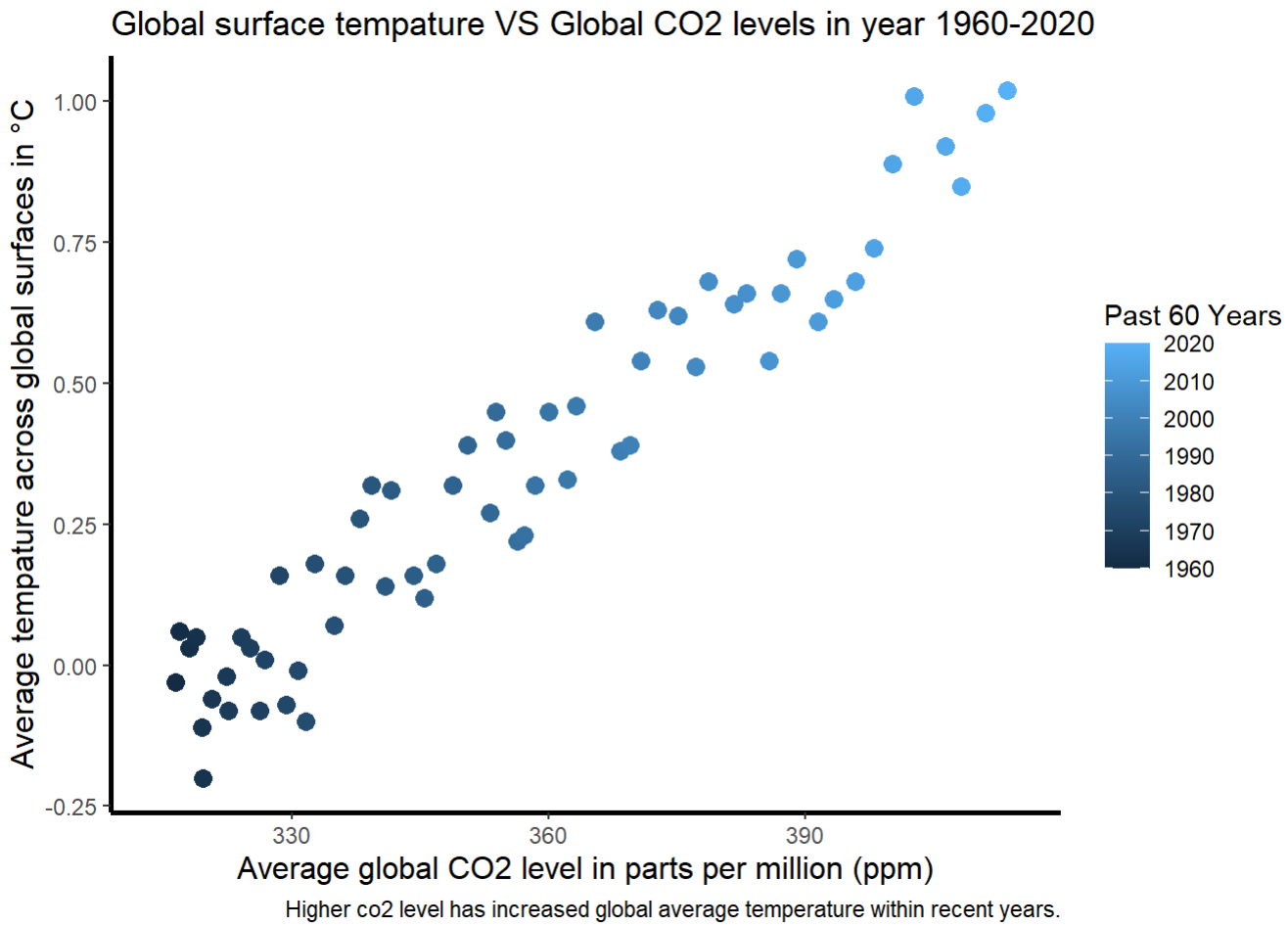
1.5

```
nasa_sea %>%
  mutate(date = as_date(date)) ->nasa_sea #convert nasa_sea$date into date type instead of datet
ime type to be join with other later


nasa_temp %>%
  full_join(nasa_ice, by = c("date" = "date"), keep = FALSE) %>%
  full_join(nasa_sea, by = c("date" = "date"), keep = FALSE) %>%
  full_join(nasa_co2, by = c("date" = "date"), keep = FALSE) -> nasa #join all the tibble togeth
er and save in to "nasa"


nasa
```

```
## # A tibble: 1,855 x 5
##    date        temp   ice   sea   co2
##    <date>     <dbl> <dbl> <dbl> <dbl>
##  1 1880-01-01 -0.16    NA    NA    NA
##  2 1881-01-01 -0.08    NA    NA    NA
##  3 1882-01-01 -0.1     NA    NA    NA
##  4 1883-01-01 -0.17    NA    NA    NA
##  5 1884-01-01 -0.28    NA    NA    NA
##  6 1885-01-01 -0.33    NA    NA    NA
##  7 1886-01-01 -0.31    NA    NA    NA
##  8 1887-01-01 -0.36    NA    NA    NA
##  9 1888-01-01 -0.17    NA    NA    NA
## 10 1889-01-01 -0.11    NA    NA    NA
## # ... with 1,845 more rows
```

1.6

```
nasa %>%
  select(date,co2, temp) %>%  #select only the columns we need
  filter(date >= "1960-01-01" & date <= "2020-01-01") %>% # filter out the time frame we want
  ggplot(mapping = aes(x = co2, y = temp, colour = date)) +
  geom_point(size = 3)  + #use "co2" as x axis, "temp" as y axis, "date for colour" to visualize
the scatterplot, and use use size to change the point size
  labs(title = "Global surface tempature VS Global CO2 levels in year 1960-2020",
        x = "Average global CO2 level in parts per million (ppm)",
        y = "Average tempature across global surfaces in °C",
        caption = "Higher co2 level has increased global average temperature within recent year
s.",
        colour =  "Past 60 Years"
  ) + # use labs() to change the titles and axis name
  theme_classic() +  #remove background
  theme(
        axis.line.x = element_line(size = 1),     #adjust x axis size
        axis.line.y = element_line(size = 1),     #adjust y axis size
        axis.title.x = element_text(size = 12),   #adjust x axis title size
        axis.title.y = element_text(size = 12)
      )
```



2 Question

2.1

```
read_table(here("data", "luthi_carbon_dioxide.txt"), skip = 773) %>%  #read in the file and skip the first 773 lines
  rename(CO2 = `CO2(ppmv)`) %>% #rename column name
  rename(yrbp = `Age(yrBP)`) -> historic_co2  # rename column name and save to new variable

historic_co2
```

```
## # A tibble: 1,096 x 2
##     yrbp   CO2
##    <dbl> <dbl>
##  1   137  280.
##  2   268  275.
##  3   279  278.
##  4   395  279.
##  5   404  282.
##  6   485  278.
##  7   559  281.
##  8   672  282.
##  9   754  280.
## 10   877  278.
## # ... with 1,086 more rows
```

2.2

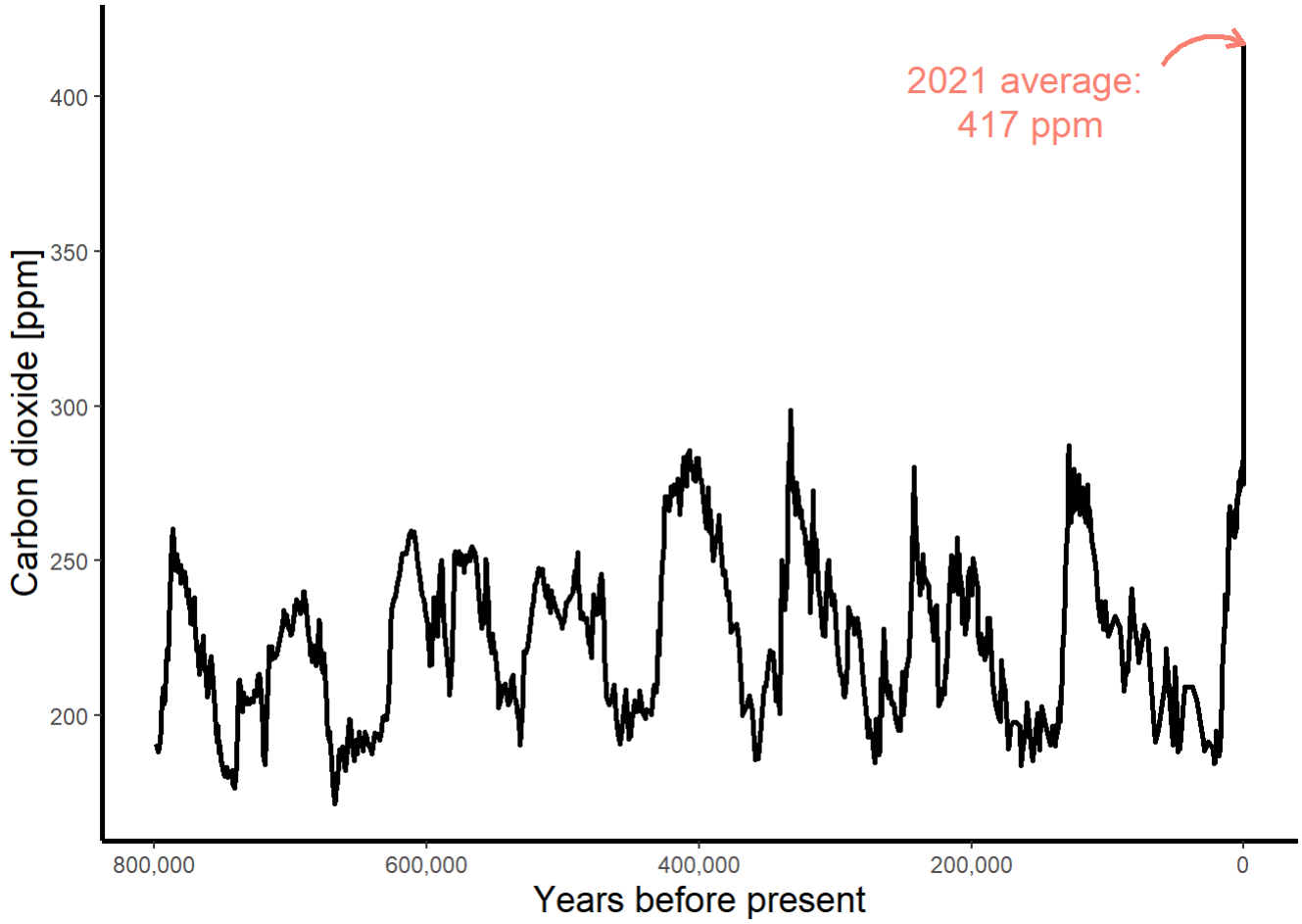
```
historic_co2 %>%
  mutate(yrbp = yrbp + 13) -> historic_co2_modified  #mutate the yrbp reference year to 2021


nasa_co2 %>%
  mutate(date = year(date)) %>% # retrieve the year of the date
  mutate(yrbp = 2021 - date) %>%  # mutate the yrbp reference year to 2021
  select(yrbp, co2) %>%  #select the columns we need
  group_by(yrbp) %>%  #group yrbp for calculation
  summarise(co2 = mean(co2, na.rm = TRUE)) %>%  #average the "co2" column
  full_join(historic_co2_modified, by = c("yrbp" = "yrbp", "co2" = "CO2")) -> combined_co2  #joi
n nasa_co2 tibble and historic_co2_modified tibble together as requested and save in to new vari
able combined_co2

combined_co2
```

```
## # A tibble: 1,160 x 2
##     yrbp   co2
##    <dbl> <dbl>
##  1     0  417.
##  2     1  414.
##  3     2  412.
##  4     3  409.
##  5     4  407.
##  6     5  404.
##  7     6  401.
##  8     7  399.
##  9     8  397.
## 10     9  394.
## # ... with 1,150 more rows
```

2.3

```
combined_co2 %>%
  ggplot(mapping = aes(x = yrbp, y = co2)) +  #set "yrbp as x axis and "co2" as y axis as the qu
estion
  geom_line(size = 1) +  #choose geom_line to create line graph and set line size as 1 to make i
t thicker
  scale_x_reverse(breaks = c(800000, 600000, 400000, 200000, 0),  #reverse the x axis value and
 setting the x axis breaks from 800,000 to 0
                  labels = c("800,000", "600,000", "400,000", "200,000", "0")  #change the scien
tific notation to the label we want
                  ) +
  labs(x = "Years before present",  #change x axis title
       y = "Carbon dioxide [ppm]",  #change y axis title
       ) +
  theme_classic() + #change the theme to classic to remove the background grid
  theme(axis.title.x = element_text(size = 14), #adjust x axis title size
        axis.title.y = element_text(size = 14), #adjust y axis title size
        axis.line.x = element_line(size = 1),   #adjust x axis line size
        axis.line.y = element_line(size = 1)    #adjust y axis line size
        ) -> p # save to variable p



point_coords <- c(combined_co2$yrbp[1], combined_co2$co2[1])  #set the start point coordinate
label_coords <- point_coords + c(60000, -7)   #set the label coordinate



#curve annotation
p_curve <-  #save the curve annotation to p_curve variable
  annotate(
    geom = "curve",  #create the curve line
    x = label_coords[1],  #set the start point coordinate for x
    y = label_coords[2],  #set the start point coordinate for y
    xend = point_coords[1],   #set the end point coordinate for x
    yend = point_coords[2],   #set the end point coordinate for y
    curvature = -0.4,   #set the curvature
    arrow = arrow(length = unit(2.5, "mm")),   #add arrow on the start point set size
    colour = "salmon",   #set colour to "salmon"
    size = 1    #set line size
  )



# Plot curve and point annotation with explanatory text
p +
  p_curve +  #combine the plot and curve annotation and annotation text together
  annotate(
    geom = "text",  #set annotation text
    x = label_coords[1] + 100000,  #set text coordinate
    y = label_coords[2],  #set y text coordinate
    label = c("2021 average:\n 417 ppm"),  #set the label content
    hjust = 0.5,  #adjust the horizontal position to the middle
```

```
    vjust = 1,    #adjust the vertical position
    lineheight = 1,  #adjust the line height in between lines
    colour = "salmon",  #set colour
    size = (5)  #set font size
    )
```



3 Question

3.1

```
#install.packages("readxl")

read_lines(file = here("data", "nsidc_sea_ice_daily_extent.xlsx"), n_max = 10)  #read the file f
irst to understand what data type it is
```

```
## [1] "<?xml version=\"1.0\" encoding=\"UTF-8\"?>"
## [2] "<Relationships xmlns=\"http://schemas.openxmlformats.org/package/2006/relationships\"><R
elationship Id=\"rId1\" Type=\"http://schemas.openxmlformats.org/officeDocument/2006/relationshi
ps/officeDocument\" Target=\"xl/workbook.xml\"/><Relationship Id=\"rId2\" Type=\"http://schemas.
openxmlformats.org/package/2006/relationships/metadata/core-properties\" Target=\"docProps/core.
xml\"/><Relationship Id=\"rId3\" Type=\"http://schemas.openxmlformats.org/officeDocument/2006/re
lationships/extended-properties\" Target=\"docProps/app.xml\"/><Relationship Id=\"rId4\" Type=
\"http://schemas.openxmlformats.org/officeDocument/2006/relationships/custom-properties\" Target
=\"docProps/custom.xml\"/>"
```

```
read_xlsx(here("data", "nsidc_sea_ice_daily_extent.xlsx")) -> sea_ice_1   #read in the file as t
ibble and save to new variable

sea_ice_1 %>%
  select(!47:49)%>%  #remove the column we do not need
  rename(month = ...1, day = ...2) %>%   #rename the first 2 columns
  fill(month) %>%   #fill up the missing value by default direction will go down
  pivot_longer(cols = 3:46, names_to = c("year"), values_to = "extent") %>%   #pivot the tibble
 longer by making the years columns in to one column, name "year" and the values to "extent"
  mutate(year = parse_date(year, format = "%Y")) %>%   #change the "year" column to date type
  mutate(year = year(year)) %>%          #retrieve only the year part of the date
  mutate(year = as.integer(year)) %>%    #change in to integer type as requested
  mutate(day = as.integer(day)) %>%      #change in to integer type as requested
  #mutate(month = as_date(month, format = "%B"))   #does not work
  mutate(month = case_when(      #manually changing the month to integer
    month == "January" ~ 1L,
    month == "February" ~ 2L,
    month == "March" ~ 3L,
    month == "April" ~ 4L,
    month == "May" ~ 5L,
    month == "June" ~ 6L,
    month == "July" ~ 7L,
    month == "August" ~ 8L,
    month == "September" ~ 9L,
    month == "October" ~ 10L,
    month == "November" ~ 11L,
    month == "December" ~ 12L,
  )) -> sea_ice_1


sea_ice_1
```

```
## # A tibble: 16,104 x 4
##    month   day  year extent
##    <int> <int> <int>  <dbl>
##  1     1     1  1978   NA
##  2     1     1  1979   NA
##  3     1     1  1980   14.2
##  4     1     1  1981   14.3
##  5     1     1  1982   NA
##  6     1     1  1983   14.3
##  7     1     1  1984   NA
##  8     1     1  1985   NA
##  9     1     1  1986   14.0
## 10     1     1  1987   NA
## # ... with 16,094 more rows
```

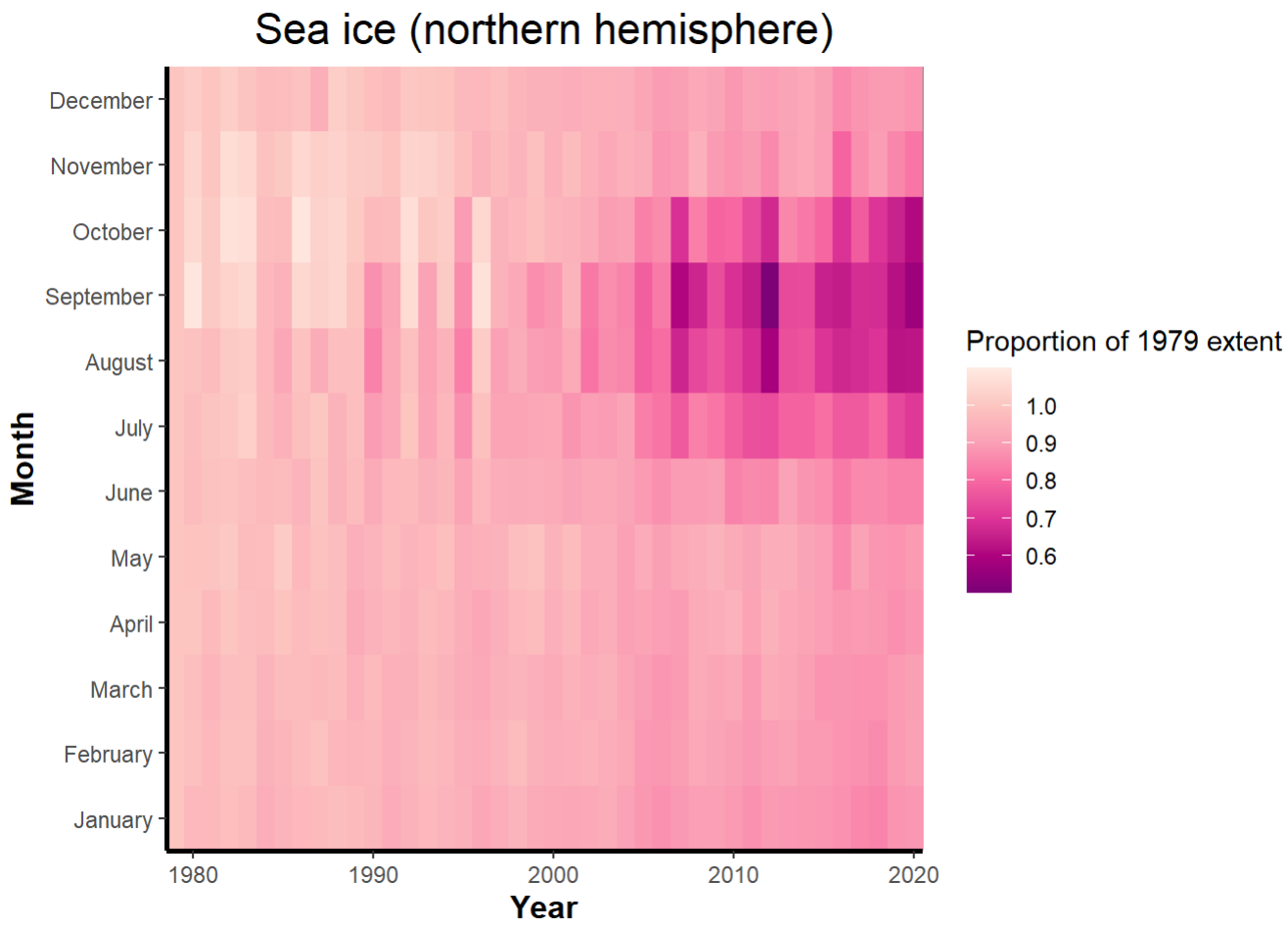
3.2

```
sea_ice_1 %>%
  group_by(month, year) %>%  #group by month and year
  summarise(avg_month_extent = mean(extent, na.rm = TRUE)) -> monthly_avg_extent   #calculate th
e average value of extent column and save in to new variable

sea_ice_1 %>%
  group_by(month, year) %>%   #group by month and year
  filter(year == 1979) %>%    #filter to get value in 1979
  summarise(month_specific_baseline_extent = mean(extent, na.rm = TRUE)) %>%   #calculate the av
erage of "extent" in 1979
  select(!year) -> baseline  #select all the columns except for "year"


monthly_avg_extent %>%
  left_join(baseline, by = c("month" = "month")) %>%    #join two tibble together
  mutate(proportion_baseline_extent = avg_month_extent / month_specific_baseline_extent) %>%   #
calculate the "proportion baseline extent" as instructed
  select(year, month, proportion_baseline_extent) -> sea_ice_2  #select the columns we need and
 save it to new variable

sea_ice_2
```

```
## # A tibble: 528 x 3
## # Groups:   month [12]
##     year month proportion_baseline_extent
##    <int> <int>                      <dbl>
##  1  1978     1                     NaN
##  2  1979     1                       1
##  3  1980     1                       0.964
##  4  1981     1                       0.967
##  5  1982     1                       0.985
##  6  1983     1                       0.969
##  7  1984     1                       0.939
##  8  1985     1                       0.955
##  9  1986     1                       0.966
## 10  1987     1                       0.971
## # ... with 518 more rows
```

3.3

```
sea_ice_2 %>%
  ggplot(mapping = aes(x = year, y = month, fill = proportion_baseline_extent)) +  #use "year" a
s x axis, "month" as y axis and  "proportion_baseline_extent" as fill value
  geom_tile() -> p   #to create a tile graph and save to p

p + scale_x_continuous(expand = expansion(add = c(-1,-1))) +     #remove padding
  scale_y_continuous(expand = expansion(add = c(0,0)),           #remove padding
                     breaks = c(1:12), labels = c("January",     #set the x axis breaks and labe
ls
                                                  "February",
                                                  "March",
                                                  "April",
                                                  "May",
                                                  "June",
                                                  "July",
                                                  "August",
                                                  "September",
                                                  "October",
                                                  "November",
                                                  "December"
                                                  )) +

  labs(title = "Sea ice (northern hemisphere)",   #create title
    x = "Year",                                   #create x axis title
    y = "Month",                                  #create y axis title
    fill = "Proportion of 1979 extent"            #create legend axis title
    ) +
  scale_fill_distiller(palette = "RdPu",          #change the fill value color
                       breaks = c(0.6, 0.7, 0.8, 0.9, 1.0),   #set legend break
                       limits = c(0.5, 1.1)                   #set legend limit
                       ) +
  theme_classic() +   #set the theme same as question
  theme(plot.title = element_text(hjust = 0.5,        #adjust title position
                                  size = rel(1.5)),
        axis.line.x = element_line(size = 1),         #adjust x axis size
        axis.line.y = element_line(size = 1),         #adjust y axis size
        axis.title.x = element_text(size = 12, face = "bold"),   #adjust x axis title size and f
ont
        axis.title.y = element_text(size = 12, face = "bold")    #adjust y axis title size and f
ont
        )
```



4 Question

4.1

```
read_lines(file = here("data", "stop_and_search.csv"), n_max = 10)   #read the file first to und
erstand what data it is
```

```
##  [1] "\"Measure\",\"Time\",\"Time_type\",\"Ethnicity\",\"Ethnicity_type\",\"Legislation_type
\",\"Geography\",\"Geography_type\",\"Number of stop and searches\",\"Total number of stop and s
earch carried out in this year in this area (excluding cases where the ethnicity was unreported)
\",\"Population by ethnicity\",\"Rate per 1,000 population by ethnicity\",\"Proportion of total
stop and searches of this ethnicity in the financial year (excludes unreported)\",\"\""
##  [2] "\"Number of stops and searches carried out (excluding vehicle only searches)\",\"2006/0
7\",\"Financial year\",\"All\",\"All\",\"All\",\"All - excluding BTP\",\"Police Force Area
\",\"1,037,735\",\"932,065\",\"56,075,912\",\"18.51\",\"N/A\",\"\""
##  [3] "\"Number of stops and searches carried out (excluding vehicle only searches)\",\"2007/0
8\",\"Financial year\",\"All\",\"All\",\"All\",\"All - excluding BTP\",\"Police Force Area
\",\"1,214,693\",\"1,125,647\",\"56,075,912\",\"21.66\",\"N/A\",\"\""
##  [4] "\"Number of stops and searches carried out (excluding vehicle only searches)\",\"2008/0
9\",\"Financial year\",\"All\",\"All\",\"All\",\"All - excluding BTP\",\"Police Force Area
\",\"1,495,379\",\"1,409,802\",\"56,075,912\",\"26.67\",\"N/A\",\"\""
##  [5] "\"Number of stops and searches carried out (excluding vehicle only searches)\",\"2009/1
0\",\"Financial year\",\"All\",\"All\",\"All\",\"All - excluding BTP\",\"Police Force Area
\",\"1,345,334\",\"1,284,197\",\"56,075,912\",\"23.99\",\"N/A\",\"\""
##  [6] "\"Number of stops and searches carried out (excluding vehicle only searches)\",\"2009/1
0\",\"Financial year\",\"All\",\"All\",\"Section 1\",\"All - excluding BTP\",\"Police Force Area
\",\"1,169,012\",\"1,284,197\",\"56,075,912\",\"20.85\",\"N/A\",\"\""
##  [7] "\"Number of stops and searches carried out (excluding vehicle only searches)\",\"2009/1
0\",\"Financial year\",\"All\",\"All\",\"Section 44/47a\",\"All - excluding BTP\",\"Police Force
Area\",\"85,310\",\"1,284,197\",\"56,075,912\",\"1.83\",\"N/A\",\"\""
##  [8] "\"Number of stops and searches carried out (excluding vehicle only searches)\",\"2009/1
0\",\"Financial year\",\"All\",\"All\",\"Section 60\",\"All - excluding BTP\",\"Police Force Are
a\",\"119,639\",\"1,284,197\",\"56,075,912\",\"2.13\",\"N/A\",\"\""
##  [9] "\"Number of stops and searches carried out (excluding vehicle only searches)\",\"2010/1
1\",\"Financial year\",\"All\",\"All\",\"All\",\"All - excluding BTP\",\"Police Force Area
\",\"1,272,799\",\"1,220,198\",\"56,075,912\",\"22.7\",\"N/A\",\"\""
## [10] "\"Number of stops and searches carried out (excluding vehicle only searches)\",\"2010/1
1\",\"Financial year\",\"All\",\"All\",\"Section 1\",\"All - excluding BTP\",\"Police Force Area
\",\"1,220,248\",\"1,220,198\",\"56,075,912\",\"21.76\",\"N/A\",\"\""
```

```
stop_search_1_raw <- read_csv(here("data", "stop_and_search.csv"))  #read in the file as tibble
 and save to new variable


stop_search_1_raw %>%
  clean_names() %>%   #change column names to snake case
  rename(stops = number_of_stop_and_searches,   #rename columns
         population = population_by_ethnicity,
         rate = rate_per_1_000_population_by_ethnicity
          ) %>%
  mutate(stops = parse_number(stops, na = c("", "NA")),   #change the "stops" type to integer
         population = parse_number(population, na = c("", "NA")),   #change the "population" typ
e to integer
         rate = parse_number(rate, na = c("", "NA"))    #change the "rate" type to integer
          ) %>%
  filter(ethnicity == "All" |        #filter out the ethnicity we need
          ethnicity == "Asian" |
          ethnicity == "Black" |
          ethnicity == "White" |
          ethnicity == "Other"
           ) %>%
  rename(year = time) %>%  #rename "time" to "year"
  select(year, ethnicity, legislation_type, geography, stops, population, rate) -> stop_search_1
#select the columns we need and save to new variable

stop_search_1
```

```
## # A tibble: 2,926 x 7
##    year    ethnicity legislation_type geography           stops population  rate
##    <chr>   <chr>     <chr>            <chr>               <dbl>      <dbl> <dbl>
##  1 2006/07 All       All              All - excluding B~ 1.04e6   56075912 18.5
##  2 2007/08 All       All              All - excluding B~ 1.21e6   56075912 21.7
##  3 2008/09 All       All              All - excluding B~ 1.50e6   56075912 26.7
##  4 2009/10 All       All              All - excluding B~ 1.35e6   56075912 24.0
##  5 2009/10 All       Section 1        All - excluding B~ 1.17e6   56075912 20.8
##  6 2009/10 All       Section 44/47a   All - excluding B~ 8.53e4   56075912  1.83
##  7 2009/10 All       Section 60       All - excluding B~ 1.20e5   56075912  2.13
##  8 2010/11 All       All              All - excluding B~ 1.27e6   56075912 22.7
##  9 2010/11 All       Section 1        All - excluding B~ 1.22e6   56075912 21.8
## 10 2010/11 All       Section 44/47a   All - excluding B~ 8.93e3   56075912  0.17
## # ... with 2,916 more rows
```

4.2

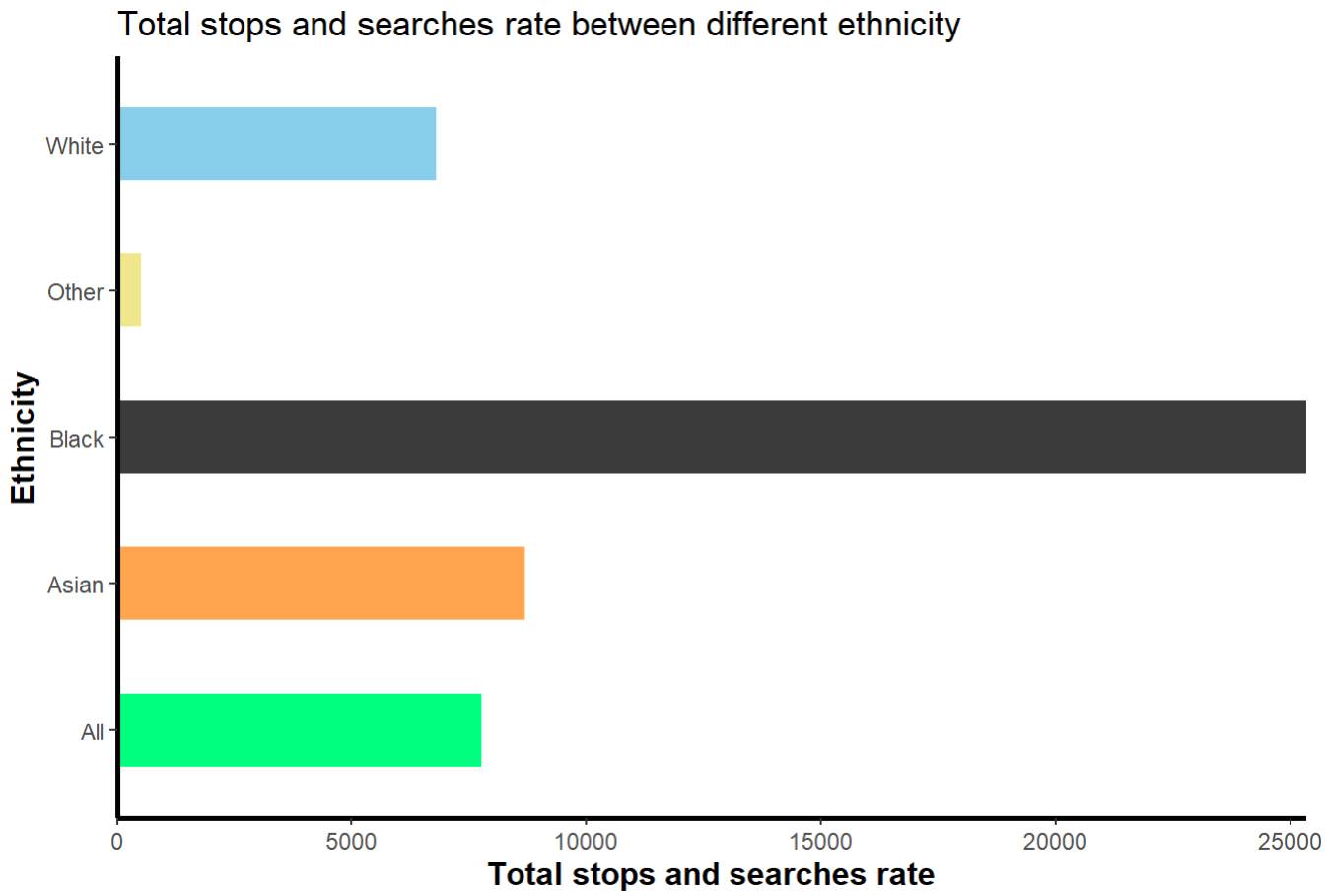
```
stop_search_1 %>%
  group_by(year, ethnicity, legislation_type, geography) %>%   #group by the columns value to cr
eate new tibble that will be join later
  summarise(rate_white = mean(rate, na.rm = TRUE)) %>%   #calculate the average rate
  filter(ethnicity == "White") -> avg_rate_white    #filter out only the white ethnicity and sav
e it to new variable


stop_search_1 %>%
  left_join(avg_rate_white, by = c("year" = "year", "legislation_type" = "legislation_type", "ge
ography" =  "geography")) %>%  #join the "avg_rate_white" to "stop_search_1"
  rename(ethnicity = ethnicity.x) %>%   #rename column
  mutate(relative_disparity = rate / rate_white) %>%   #calculate "relative_disparity" as instru
cted and create new column to save value
  select(!ethnicity.y) -> stop_search_2   #select only the columns we need and save to new varia
ble


stop_search_2
```

```
## # A tibble: 2,926 x 9
##    year    ethnicity legislation_type geography  stops population  rate rate_white
##    <chr>   <chr>     <chr>            <chr>      <dbl>      <dbl> <dbl>      <dbl>
##  1 2006/07 All       All              All - ex~ 1.04e6   56075912 18.5       14.4
##  2 2007/08 All       All              All - ex~ 1.21e6   56075912 21.7       16.8
##  3 2008/09 All       All              All - ex~ 1.50e6   56075912 26.7       19.7
##  4 2009/10 All       All              All - ex~ 1.35e6   56075912 24.0       18.0
##  5 2009/10 All       Section 1        All - ex~ 1.17e6   56075912 20.8       16.3
##  6 2009/10 All       Section 44/47a   All - ex~ 8.53e4   56075912  1.83       1.26
##  7 2009/10 All       Section 60       All - ex~ 1.20e5   56075912  2.13       0.99
##  8 2010/11 All       All              All - ex~ 1.27e6   56075912 22.7       17
##  9 2010/11 All       Section 1        All - ex~ 1.22e6   56075912 21.8       16.7
## 10 2010/11 All       Section 44/47a   All - ex~ 8.93e3   56075912  0.17       0.12
## # ... with 2,916 more rows, and 1 more variable: relative_disparity <dbl>
```

4.3

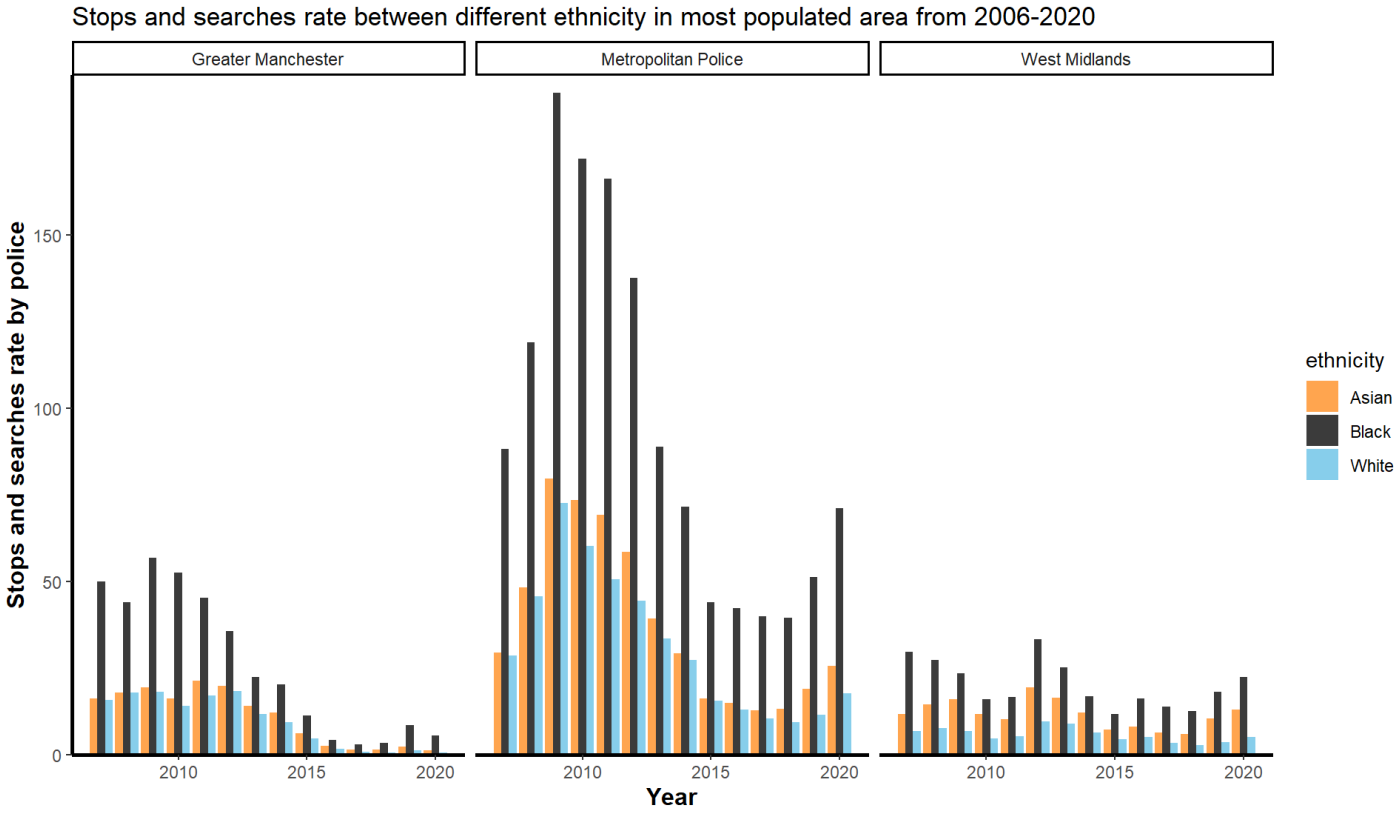
4.3.1 Question: Does black ethnicity has higher chance to be stop and search by police in general?

```
stop_search_2 %>%
  group_by(ethnicity) %>%   #grouped by same ethnicity
  mutate(ethnicity_rate = mean(rate, na.rm = TRUE)) %>%   #average the rate between different et
hnicity
  select(ethnicity, ethnicity_rate) %>%    #select the column we need
  ggplot(mapping = aes(x = ethnicity_rate, y = ethnicity, fill = ethnicity)) +    #set x axis an
d y axis and fill with ethnicity
  geom_col(width = 0.5) -> p   #plot the graph and save to variable p

p +
  theme_classic() +
  theme(
    axis.line.x = element_line(size = 1),  #adjust x axis line size
    axis.line.y = element_line(size = 1),  #adjust y axis line size
    axis.title.x = element_text(size = 12, face = "bold"),   #adjust x axis title size and font
    axis.title.y = element_text(size = 12, face = "bold")    #adjust y axis title size and font
        ) +
  labs(
       title = "Total stops and searches rate between different ethnicity",   #change titles, ax
es labels and caption
       x = "Total stops and searches rate",
       y = "Ethnicity",
       caption = "Black ethnicity has much higher chance to be stop and search by police."
      ) +
  scale_x_continuous(
    expand = expansion(add = c(0, 5))) +  #remove padding
  scale_fill_manual(values = c("springgreen", "tan1", "gray23", "khaki", "skyblue")) +   #change
the colour of ethnicity in legend
  guides(fill = FALSE)  #remove legend
```



4.3.2 Question: What are the stops and searches rate difference between Black, Asian and White ethnicity in three most populated area?

```
stop_search_2 %>%
  mutate(year = as_date(year, format = "%Y/%y")) %>%   #change the character type to date type
  mutate(year = year(year)) %>%  #retrieve only the year
  filter(geography != "All - excluding BTP" &   #excluding area in "all"
         geography !="All - excluding BTP and Greater Manchester" &
         geography != "All - including BTP" &
         geography != "All - including BTP and excluding Greater Manchester") %>%
  arrange(desc(population)) %>%   #reordered the population to find out most populated area
  filter(geography == "Metropolitan Police" |    #filter out the area we need
          geography == "West Midlands" |
          geography == "Greater Manchester"
         ) %>%
  select(year, ethnicity, stops, rate, geography) -> top_3_populated_df   #select the columns we
need and save to new variable

top_3_populated_df %>%
  filter(ethnicity %in% c("Black", "Asian", "White")) %>%   #filter out the ethnicity we want to
observe
  ggplot(mapping = aes(x = year,  y = rate, fill = ethnicity)) +  #set the axes to plot graph
  geom_col(position = "dodge") +    #set position "dodge" to avoid stacking
  facet_wrap(facets = vars(geography)) + #use facet function to plot different graph in differen
t area
  labs(
       title = "Stops and searches rate between different ethnicity in most populated area from
 2006-2020",   #change titles and caption
       x = "Year",
       y = "Stops and searches rate by police",
       caption = "In all three area black ethnicity has the highest chance to be stop and search
by police, where white ethnicity has the lowest,"
      ) +
  scale_fill_manual(values = c("tan1", "gray23", "skyblue")) +    #change the colour of ethnicit
y in legend
  theme_classic() +    #remove background
  theme(
    axis.line.x = element_line(size = 1),  #adjust x axis line size
    axis.line.y = element_line(size = 1),  #adjust y axis line size
    axis.title.x = element_text(size = 12, face = "bold"),   #adjust x axis title size and font
    axis.title.y = element_text(size = 12, face = "bold")    #adjust y axis title size and font
        ) +
  scale_y_continuous(
    expand = expansion(add = c(0, 5)))   #remove padding
```

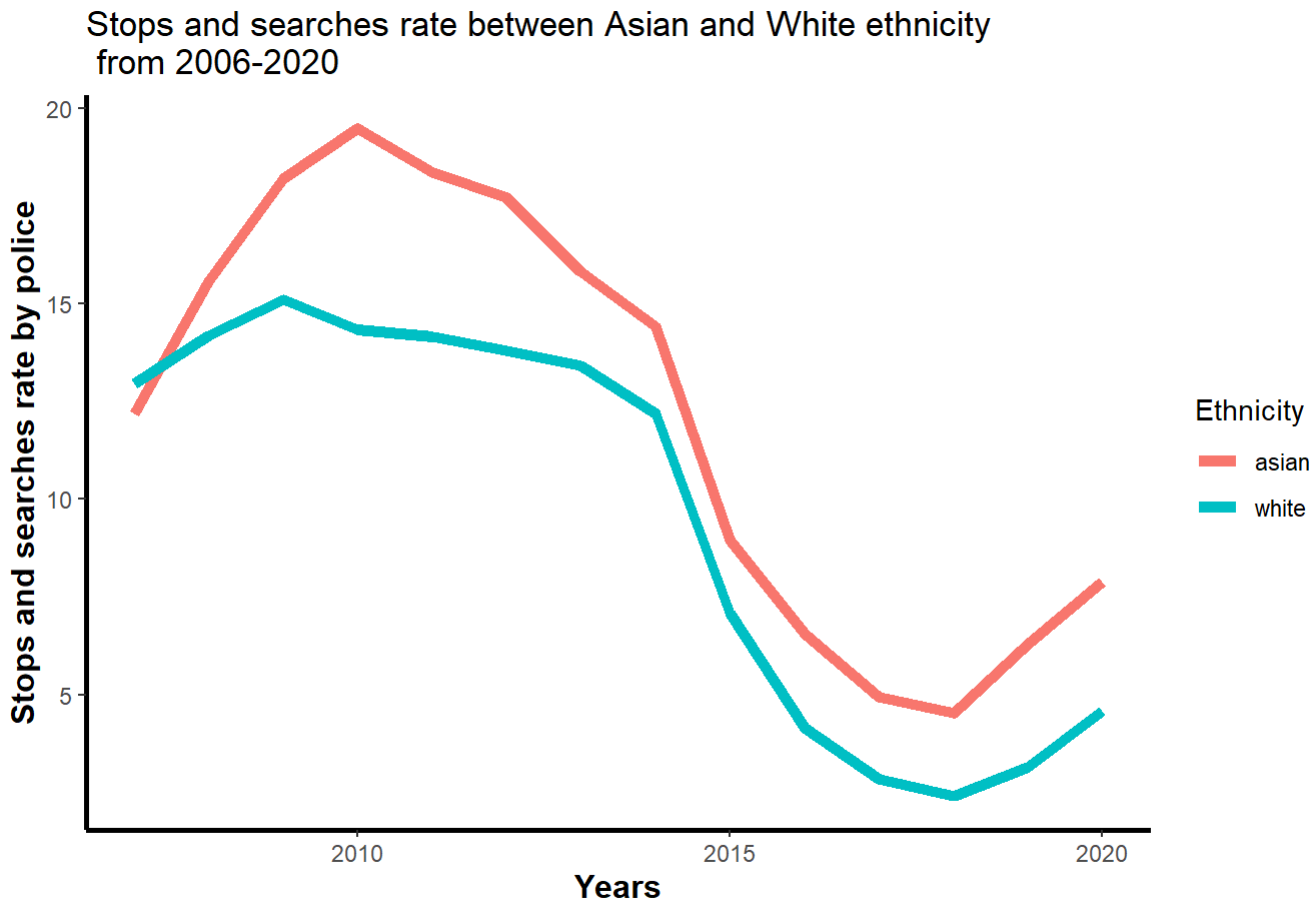


In all three area black ethnicity has the highest chance to be stop and search by police, where white ethnicity has the lowest,

My thought was that more populated area should have more multi-culture experiences and the police will be less bias than other area, but the result shows that other ethnicity still have much higher chance to be stop and search than white ethnicity.

4.3.3 Question: What is the rate for Asian ethnicity to be stop and search compare to white ethnicity?

```
stop_search_2 %>%
  mutate(year = as_date(year, format = "%Y/%y")) %>%   #change the character type to date type
  mutate(year = year(year)) %>%   #retrieve only the year
  filter(ethnicity == "Asian") %>%   #filter out Asian
  group_by(year) %>%   #grouped by year for calculation
  mutate(asian = mean(rate, na.rm = TRUE)) %>%   #calculate average rate for Asian
  mutate(white = mean(rate_white, na.rm = TRUE)) %>%  #calculate average rate for White
  select(year, asian, white) -> df  #select columns we need save to new variable

df %>%
  pivot_longer(cols = 2:3, names_to = "ethnicity", values_to = "rate") %>%   #transform tibble t
o make ethnicity to one column
  ggplot(mapping = aes(x = year, y = rate, colour = ethnicity)) +  #create line graph include As
ian and white stop and search rate
  geom_line(size = 2) -> p   #make line thicker


p +
  labs(
       title = "Stops and searches rate between Asian and White ethnicity \n from 2006-2020",
#change titles, axes labels and caption
       x = "Years",
       y = "Stops and searches rate by police",
       caption = "Asian ethnicity has a higher rate to be stop and search by the police than Whi
te ethnicity.",
       colour = "Ethnicity"
      ) +
  theme_classic() +    #remove background
  theme(
    axis.line.x = element_line(size = 1),  #adjust x axis line size
    axis.line.y = element_line(size = 1),  #adjust y axis line size
    axis.title.x = element_text(size = 12, face = "bold"),   #adjust x axis title size and font
    axis.title.y = element_text(size = 12, face = "bold")    #adjust y axis title size and font
        )
```



Asian ethnicity has a higher rate to be stop and search by the police than White ethnicity.