```r
setwd("/Users/adazhong/Library/Mobile Documents/com~apple~CloudDocs/DSO 545/project/submission")

library(data.table)
library(fuzzyjoin)
library(dplyr)
library(ggplot2)
```

```r
# read icecore data from NOAA
df1 = fread("https://www.ncei.noaa.gov/pub/data/paleo/icecore/antarctica/epica_domec/edc3deuttemp2007.t
            skip = "last 1000 years)")
str(df1)
```

```
## Classes 'data.table' and 'data.frame':   5800 obs. of  5 variables:
##  $ Bag        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ ztop       : num  0 0.55 1.1 1.65 2.2 2.75 3.3 3.85 4.4 4.95 ...
##  $ Age        : num  -50 -43.5 -37.4 -31.6 -24.5 ...
##  $ Deuterium  : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Temperature: num  NA NA NA NA NA NA NA NA NA NA ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```r
df2 = fread("https://www.ncei.noaa.gov/pub/data/paleo/icecore/antarctica/epica_domec/edc-co2-2008.txt",
            skip = "Timescale EDC3_gas_a") %>%
  rename(Age = "Age(yrBP)", CO2 = "CO2(ppmv)")
str(df2)
```

```
## Classes 'data.table' and 'data.frame':   1096 obs. of  2 variables:
##  $ Age: int  137 268 279 395 404 485 559 672 754 877 ...
##  $ CO2: num  280 275 278 279 282 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```r
appx = function(x, y) {
  ifelse(abs(x-y) < 100, T, F)
}

# join co2 level and temperature by year
df3 = fuzzy_inner_join(df1, df2, by = c("Age"),
                       match_fun = list(`appx`))
str(df3)
```

```
## 'data.frame':    2733 obs. of  7 variables:
##  $ Bag        : int  13 14 15 16 17 18 19 20 21 22 ...
##  $ ztop       : num  6.6 7.15 7.7 8.25 8.8 ...
##  $ Age.x      : num  38.4 46.8 55.1 64.4 73.2 ...
##  $ Deuterium  : num  -391 -385 -378 -394 -399 ...
##  $ Temperature: num  0.88 1.84 3.04 0.35 -0.42 0.05 0.05 -0.52 0.79 -0.55 ...
##  $ Age.y      : int  137 137 137 137 137 137 137 137 137 137 ...
##  $ CO2        : num  280 280 280 280 280 ...
```

```r
# linear regression - co2(ppmv)~temperature
lin = lm(df3$Temperature~df3$CO2)
x = list(df3$CO2)
```

```r
pred = predict(lin, x)
df3$Pred.Temperature = pred

# visualize model accuracy
# statistically significant correlation between co2 and temperature
cols = c("Pred.Temperature" = "red", "Temperature" = "blue")
ggplot(df3, aes(Age.x))+
  geom_line(aes(y = Pred.Temperature, color = "Pred.Temperature"),alpha = 0.7) +
  geom_line(aes(y = Temperature, color = "Temperature"), alpha = 0.7) +
  scale_x_reverse() +
  scale_color_manual(name = "", values = cols) +
  ylab("Temperature in Celsius ") +
  xlab("Number of Years Before Present") +
  ggtitle("Assessing the Quality of Linear Reg. Model")
```