Data Science Project 1

Audrey Holloman, Nicole Kadosh, Leah Newman, Hannah Godfrey

Updated: Monday, December 17, 2018 @ 07:43:56 PM

DATA FILES

- Mauna Loa Monthly CO2 dataset
- Mauna Loa Weekly CO2 dataset
- Global Monthly CO2 dataset
- Global Daily CO2 dataset
- Global Growth dataset

Refer to the Mauna Loa data source files and global data source files for more information on how the data were collected and descriptions of the variables included in each dataset.

```
library(dplyr)
library(gapminder)
library(ggplot2)
library(plotly)
library(gridExtra)

Mauna <- read.csv(file = url("https://raw.githubusercontent.com/STAT-JET-ASU/DataScience1/master/Projec
Weekly <- read.csv(file = url("https://raw.githubusercontent.com/STAT-JET-ASU/DataScience1/master/Proje
Monthly <- read.csv(file = url("https://raw.githubusercontent.com/STAT-JET-ASU/DataScience1/master/Proj
Daily <- read.csv(file = url("https://raw.githubusercontent.com/STAT-JET-ASU/DataScience1/master/Projec
Growth <- read.csv(file = url("https://raw.githubusercontent.com/STAT-JET-ASU/DataScience1/master/Proje
```

DATA EXPLORATIONS

- Write a brief description of each data set's origin and variables, similar to here and here.
- Run `glimpse()` and `summary()` on each dataset to examine its structure and contents.

Mauna Loa Monthly

Description

This data set describes the amount of Carbon Dioxide in parts per million (ppm) collected at the Mauna Loa observatory. The global, daily, monthly and growth are based off of this data set. Because of this observatory's location (11,300 feet above sea level and in the middle of the pacific ocean), it is able to collect very clear readings that are not interrupted by city or industrial areas. This model shows both the raw data (red line) readings each month and the interpolated data that averages the raw data (black line). It is ploted over time broken down in months and year.

Variables

- Year: calendar year
- Month: calendar month
- Decimaldate: combining year and month; the fraction of that year
- Average: monthly mean CO$_2$ mole fraction determined from daily averages
- Interpolated: average values from the preceding column and interpolated values where data is missing

- Trend: the ppm of CO_2 each day
- Numdays: number of days

```
glimpse(Mauna)
```

```
## Observations: 727
## Variables: 7
## $ year         <int> 1958, 1958, 1958, 1958, 1958, 1958, 1958, 1958, 1...
## $ month        <int> 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 1, 2, 3, 4, 5, 6...
## $ decimaldate  <dbl> 1958.208, 1958.292, 1958.375, 1958.458, 1958.542,...
## $ average      <dbl> 315.71, 317.45, 317.50, -99.99, 315.86, 314.93, 3...
## $ interpolated <dbl> 315.71, 317.45, 317.50, 317.10, 315.86, 314.93, 3...
## $ trend        <dbl> 314.62, 315.29, 314.71, 314.85, 314.98, 315.94, 3...
## $ numdays      <int> -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -...
```

```
summary(Mauna)
```

```
##       year          month         decimaldate      average
##  Min.   :1958   Min.   : 1.000   Min.   :1958   Min.   :-99.99
##  1st Qu.:1973   1st Qu.: 4.000   1st Qu.:1973   1st Qu.:328.43
##  Median :1988   Median : 6.000   Median :1988   Median :351.31
##  Mean   :1988   Mean   : 6.495   Mean   :1988   Mean   :349.56
##  3rd Qu.:2003   3rd Qu.: 9.000   3rd Qu.:2004   3rd Qu.:375.70
##  Max.   :2018   Max.   :12.000   Max.   :2019   Max.   :411.24
##   interpolated      trend          numdays
##  Min.   :312.7   Min.   :314.6   Min.   :-1.00
##  1st Qu.:328.6   1st Qu.:329.3   1st Qu.:-1.00
##  Median :351.3   Median :351.4   Median :24.00
##  Mean   :353.6   Mean   :353.6   Mean   :18.34
##  3rd Qu.:375.7   3rd Qu.:376.1   3rd Qu.:28.00
##  Max.   :411.2   Max.   :409.0   Max.   :31.00
```

Mauna Lao Weekly

Description

This data depicts the changes in the amount of carbon dioxide (CO_2) in parts per million (ppm) in weekly increments. Instead of showing the amount of CO_2 in parts per million (ppm), this data has been compiled to show the *change* from 1 year ago, 10 years ago, and from 2000. This data is meaningful because it gives different and more immediately effective information. This data is better able to reflect what the state of the atmospheres is like in the recent past. Therefore, it reflects trends of development and change more immediate to us.

Variables

- Start Year: the beginning year
- Start Month: the beginning month
- Start Day: the beginning day
- Decimal: fraction of time in each month
- CO_2 ppm: the amount of CO_2 measured in parts per million
- Numdays: number of days
- X1yr_ago: the change in CO_2 starting 1 year ago
- X10yr_ago: the change in CO_2 starting 10 years ago

- Since2000: the change in CO$_2$ from 2000

```
glimpse(Weekly)
```

```
## Observations: 2,316
## Variables: 9
## $ startyear  <int> 1974, 1974, 1974, 1974, 1974, 1974, 1974, 1974, 197...
## $ startmonth <int> 5, 5, 6, 6, 6, 6, 6, 7, 7, 7, 7, 8, 8, 8, 8, 9, 9, ...
## $ startday   <int> 19, 26, 2, 9, 16, 23, 30, 7, 14, 21, 28, 4, 11, 18,...
## $ decimal    <dbl> 1974.380, 1974.399, 1974.418, 1974.437, 1974.456, 1...
## $ CO2ppm     <dbl> 333.34, 332.95, 332.32, 332.18, 332.37, 331.59, 331...
## $ numdays    <int> 6, 6, 5, 7, 7, 6, 6, 6, 5, 7, 4, 5, 6, 6, 7, 5, 4, ...
## $ X1yr_ago   <dbl> -999.99, -999.99, -999.99, -999.99, -999.99, -999.9...
## $ X10yr_ago  <dbl> -999.99, -999.99, -999.99, -999.99, -999.99, -999.9...
## $ since1800  <dbl> 50.36, 50.06, 49.57, 49.63, 50.07, 49.60, 50.04, 50...
```

```
summary(Weekly)
```

```
##    startyear      startmonth        startday        decimal
##  Min.   :1974   Min.   : 1.000   Min.   : 1.00   Min.   :1974
##  1st Qu.:1985   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:1985
##  Median :1996   Median : 7.000   Median :16.00   Median :1997
##  Mean   :1996   Mean   : 6.528   Mean   :15.72   Mean   :1997
##  3rd Qu.:2007   3rd Qu.: 9.000   3rd Qu.:23.00   3rd Qu.:2008
##  Max.   :2018   Max.   :12.000   Max.   :31.00   Max.   :2019
##      CO2ppm           numdays         X1yr_ago          X10yr_ago
##  Min.   :-1000.0   Min.   :0.000   Min.   :-1000.0   Min.   : -999.99
##  1st Qu.:  345.9   1st Qu.:5.000   1st Qu.:  344.5   1st Qu.:  329.45
##  Median :  362.5   Median :6.000   Median :  360.6   Median :  347.55
##  Mean   :  353.7   Mean   :5.843   Mean   :  323.8   Mean   :   39.33
##  3rd Qu.:  384.0   3rd Qu.:7.000   3rd Qu.:  382.1   3rd Qu.:  364.59
##  Max.   :  411.9   Max.   :7.000   Max.   :  410.2   Max.   :  388.88
##    since1800
##  Min.   : -999.99
##  1st Qu.:   65.90
##  Median :   82.69
##  Mean   :   76.09
##  3rd Qu.:  104.11
##  Max.   :  129.39
```

```
Weekly$since2000 <- Weekly$since1800
```

Global Monthly

Description

This data shows the change in Carbon Dioxide (CO$_2$) in parts per million per month over the complete past four years and the data collected from this year. The data is collected in the Mauna Loa observatory in Hawaii. This data is very accurate depictions of the about of CO$_2$ in the environment because of the observatory's remote location in the middle of the pacific and how high it is above sea level (11,300 feet). Therefore, this data is not as impacted by industrial activity. This data can be used to make models that can lead to new insights on the amount of green house gases in our atmosphere and help make predictions about global warming across the globe.

Variables

- Year: calender year

- Month: calender month
- Decimal: fraction of time in each month
- Average: average of the readings
- Trend: the ppm of CO_2 each day

```
glimpse(Monthly)
```

```
## Observations: 463
## Variables: 5
## $ year    <int> 1980, 1980, 1980, 1980, 1980, 1980, 1980, 1980, 1980, ...
## $ month   <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 1, 2, 3, 4, 5, ...
## $ decimal <dbl> 1980.042, 1980.125, 1980.208, 1980.292, 1980.375, 1980...
## $ average <dbl> 338.45, 339.15, 339.48, 339.87, 340.30, 339.86, 338.34...
## $ trend   <dbl> 337.83, 338.10, 338.13, 338.25, 338.78, 339.08, 339.19...
```

```
summary(Monthly)
```

```
##       year          month           decimal        average     
##  Min.   :1980   Min.   : 1.000   Min.   :1980   Min.   :336.9  
##  1st Qu.:1989   1st Qu.: 3.000   1st Qu.:1990   1st Qu.:352.9  
##  Median :1999   Median : 6.000   Median :1999   Median :366.6  
##  Mean   :1999   Mean   : 6.462   Mean   :1999   Mean   :368.9  
##  3rd Qu.:2008   3rd Qu.: 9.000   3rd Qu.:2009   3rd Qu.:385.4  
##  Max.   :2018   Max.   :12.000   Max.   :2019   Max.   :408.9  
##      trend      
##  Min.   :337.8  
##  1st Qu.:352.9  
##  Median :367.5  
##  Mean   :368.9  
##  3rd Qu.:385.4  
##  Max.   :407.7  
```

Global Daily

Description

The estimated daily global trend value for CO_2 is determined from the daily averaged CO2 data from the four NOAA/ESRL/GMD Baseline observatories. A trend curve is determined for each observatory record at daily intervals. Then, an estimated global trend is computed by averaging four individual trend curves at each daily interval. This data is subject to change, but the changes are usually minor.

Variables

- Year: calendar year
- Month: calendar month
- Day: calendar day
- Trend: the ppm of CO_2 each day

```
glimpse(Daily)
```

```
## Observations: 3,936
## Variables: 4
## $ year  <int> 2008, 2008, 2008, 2008, 2008, 2008, 2008, 2008, 2008, 20...
## $ month <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
## $ day   <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1...
## $ trend <dbl> 384.05, 384.06, 384.06, 384.07, 384.07, 384.08, 384.08, ...
```

```
summary(Daily)
```

```
##       year          month             day           trend      
##  Min.   :2008   Min.   : 1.000   Min.   : 1.0   Min.   :384.1  
##  1st Qu.:2010   1st Qu.: 3.000   1st Qu.: 8.0   1st Qu.:389.1  
##  Median :2013   Median : 6.000   Median :16.0   Median :395.1  
##  Mean   :2013   Mean   : 6.427   Mean   :15.7   Mean   :395.3  
##  3rd Qu.:2016   3rd Qu.: 9.000   3rd Qu.:23.0   3rd Qu.:401.4  
##  Max.   :2018   Max.   :12.000   Max.   :31.0   Max.   :408.3  
```

Global Growth

Description

Data from March 1958 through April 1974 have been obtained by C. David Keeling of the Scripps Institution of Oceanography (SIO) and were obtained from the Scripps website (scrippsco2.ucsd.edu). The annual mean rate of growth of carbon dioxide in a year is the difference in concentration between the end of December and the start of January. This represents the amount of carbon dioxide added and removed to the atmosphere. The final estimate for the annual mean growth rate of the previous year in March by using the average of the most recent November-February months as the trend value for January 1. The uncertainty in the Mauna Loa annual mean growth rate is estimated from the standard deviation of the differences between monthly mean values determined independently by the Scripps Institution of Oceanography and by NOAA/ESRL.

This dataset includes the year with the annual increase (growth rate) and its corresponding uncertainty (error).

Variables

- year: calender year
- anninc: annual increase
- unc: uncertainty

```
glimpse(Growth)
```

```
## Observations: 59
## Variables: 3
## $ year   <int> 1959, 1960, 1961, 1962, 1963, 1964, 1965, 1966, 1967, 1...
## $ anninc <dbl> 0.96, 0.71, 0.78, 0.56, 0.57, 0.49, 1.10, 1.10, 0.61, 0...
## $ unc    <dbl> 0.31, 0.27, 0.27, 0.27, 0.28, 0.27, 0.26, 0.28, 0.34, 0...
```

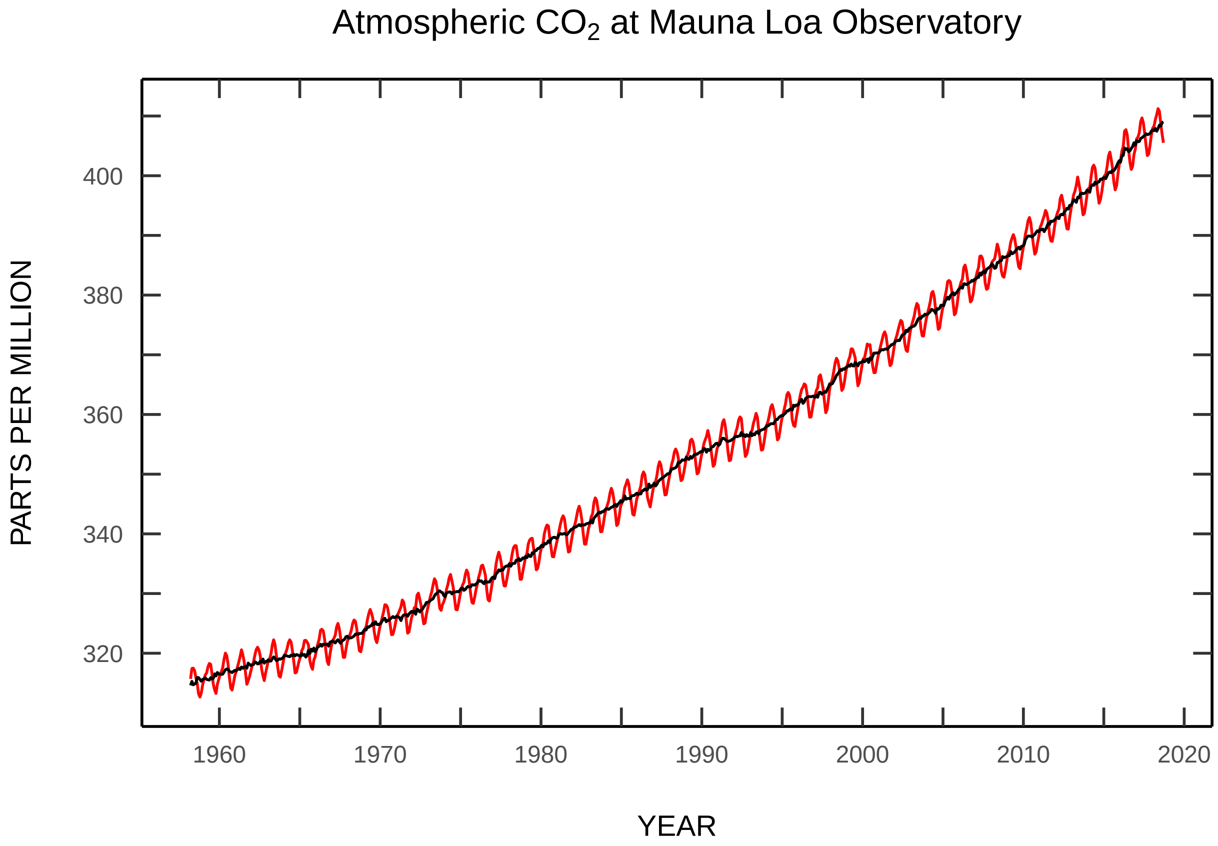
```
summary(Growth)
```

```
##       year          anninc           unc        
##  Min.   :1959   Min.   :0.490   Min.   :0.0500  
##  1st Qu.:1974   1st Qu.:1.040   1st Qu.:0.0800  
##  Median :1988   Median :1.460   Median :0.1000  
##  Mean   :1988   Mean   :1.536   Mean   :0.1575  
##  3rd Qu.:2002   3rd Qu.:2.005   3rd Qu.:0.2700  
##  Max.   :2017   Max.   :2.940   Max.   :0.3400  
```

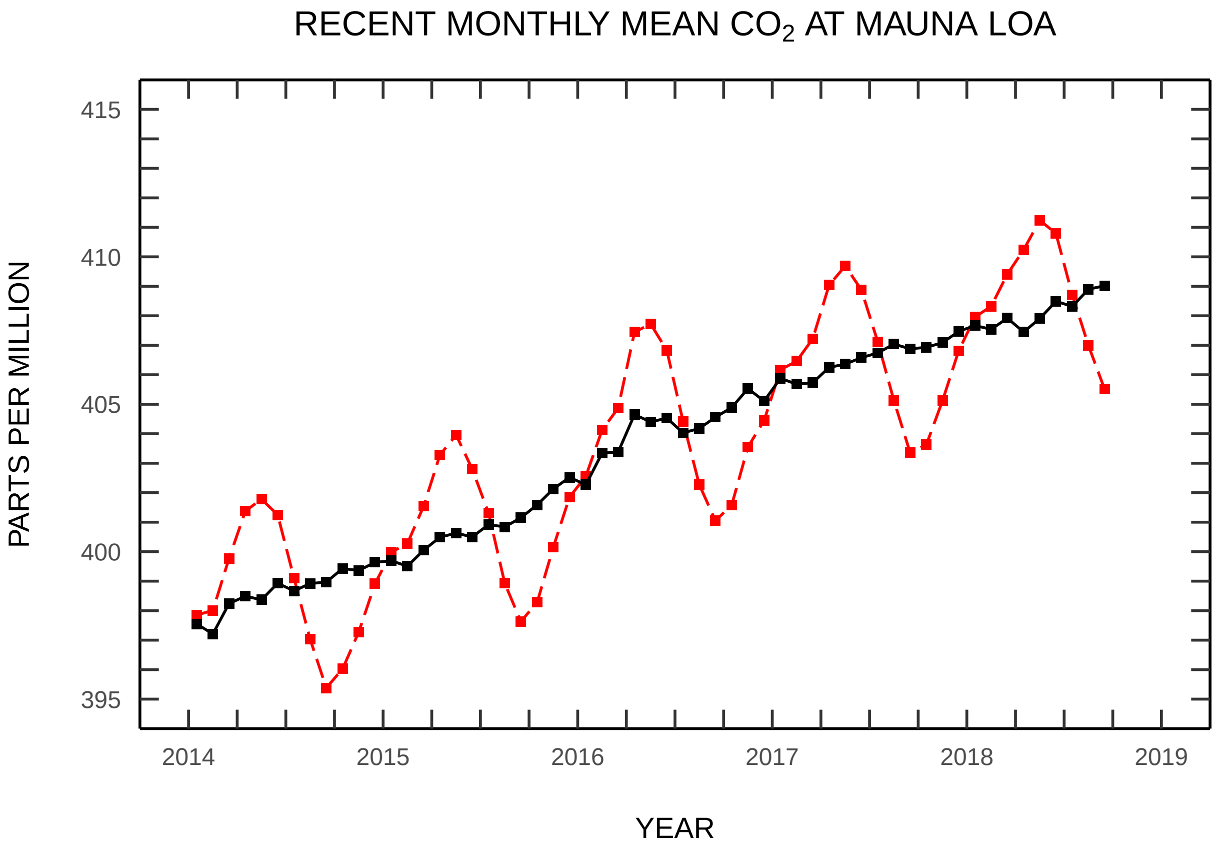
DATA VISUALIZATIONS

- Replicate the plots shown on this web page and this web page. You do not need to include the NOAA / Scripps logos or labels. Your results should look similar to Dr. Thomley's replications here.

```r
attach(Mauna)
theme_update(plot.title = element_text(hjust = 0.5))
ggplot(Mauna) + theme_classic() + geom_line(aes(x = decimaldate, y = interpolated), colour = "red") + g

element_text(margin=unit(c(0.5,0.5,.5,.5), "cm")), axis.text.y =

element_text(margin=unit(c(0.5,0.5,.5,.5), "cm")), axis.text.x.top =

element_text(margin = unit(c(.5,.5,.5,.5), "cm"))
) + scale_x_continuous(sec.axis = dup_axis(labels = NULL, name = ""), labels = c("1960", "", "1970", ""
```



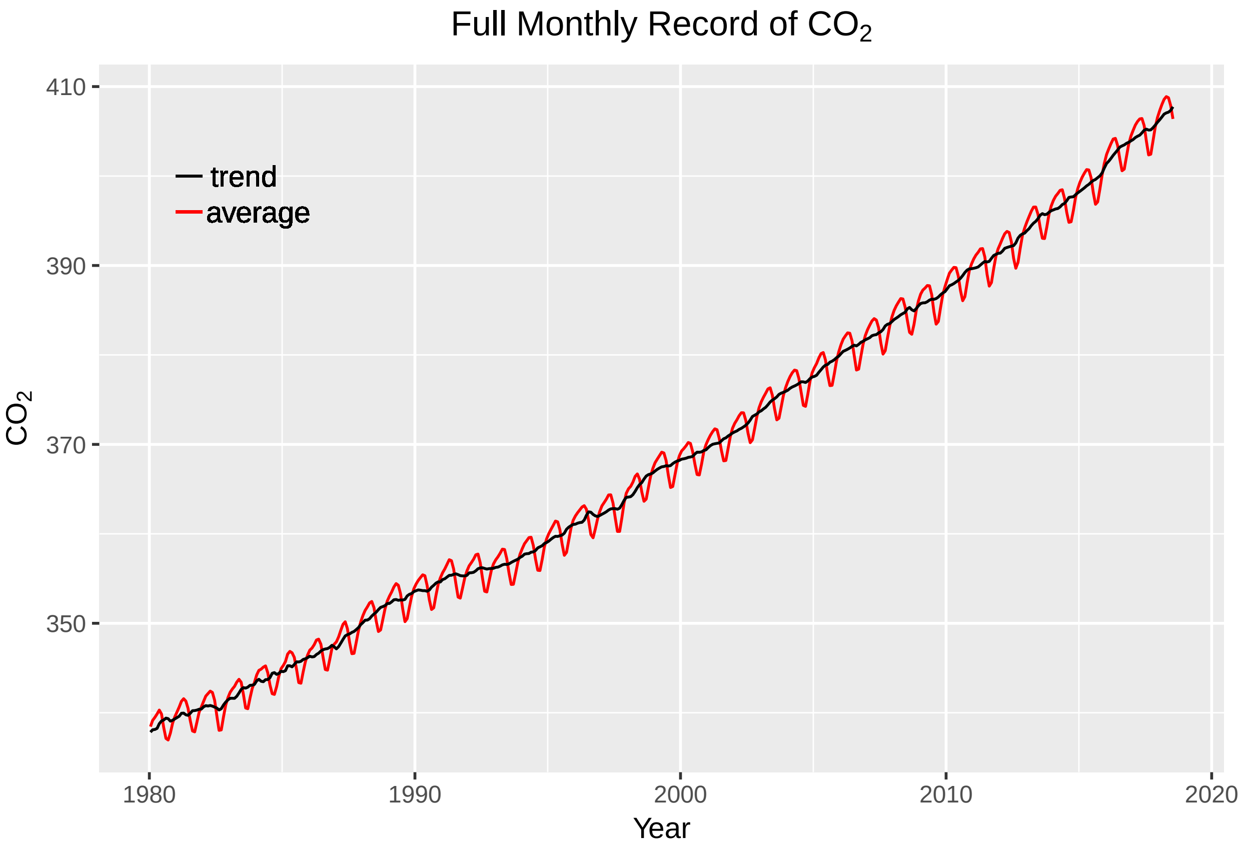
```r
MaunaNew <- Mauna %>% filter(year > 2013)
ggplot(MaunaNew) + theme_classic() + geom_line(aes(x = decimaldate, y = interpolated), linetype = "long

element_text(margin=unit(c(0.5,0.5,.5,.5), "cm")), axis.text.y =

element_text(margin=unit(c(0.5,0.5,.5,.5), "cm")), axis.text.x.top =

element_text(margin = unit(c(.5,.5,.5,.5), "cm"))
) + scale_x_continuous(sec.axis = dup_axis(labels = NULL, name = ""), labels = c("2014", rep("", 3), "2
```



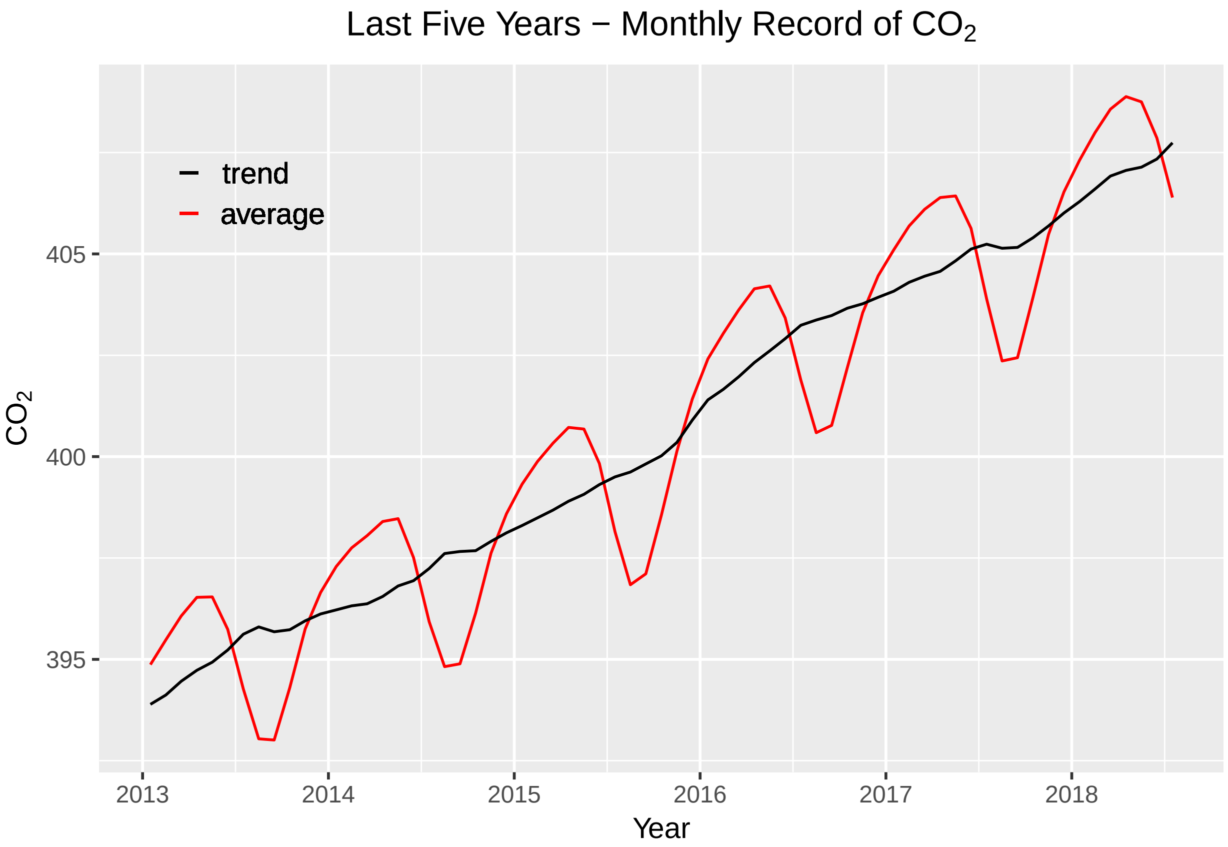
- Create time series plots to show the “full record” and “last five years” for the global monthly CO$_2$ data. Make your own choices with regard to axis formatting, line and point styles, colors, labels, etc.

Monthly

```
attach(Monthly)
FullMonthlyRecord <-
  ggplot(Monthly) +
  geom_line(aes(x = decimal, y = average), colour = "red") +
  geom_line(aes(x = decimal, y = trend)) +
  geom_segment(aes(x = 1981, y = 400, xend = 1982, yend = 400)) +
  geom_text(aes(x = 1982.3, y = 400, label = "trend", hjust = "left")) +
  geom_segment(aes(x = 1981, y = 396, xend = 1982, yend = 396), colour = "red") +
  geom_text(aes(x = 1984.1, y = 396, label = "average", hjsut = "left")) +
  xlab("Year") +
  ylab(expression("CO"[2]*"")) +
  ggtitle(expression("Full Monthly Record of CO"[2]*""))
print(FullMonthlyRecord)
```

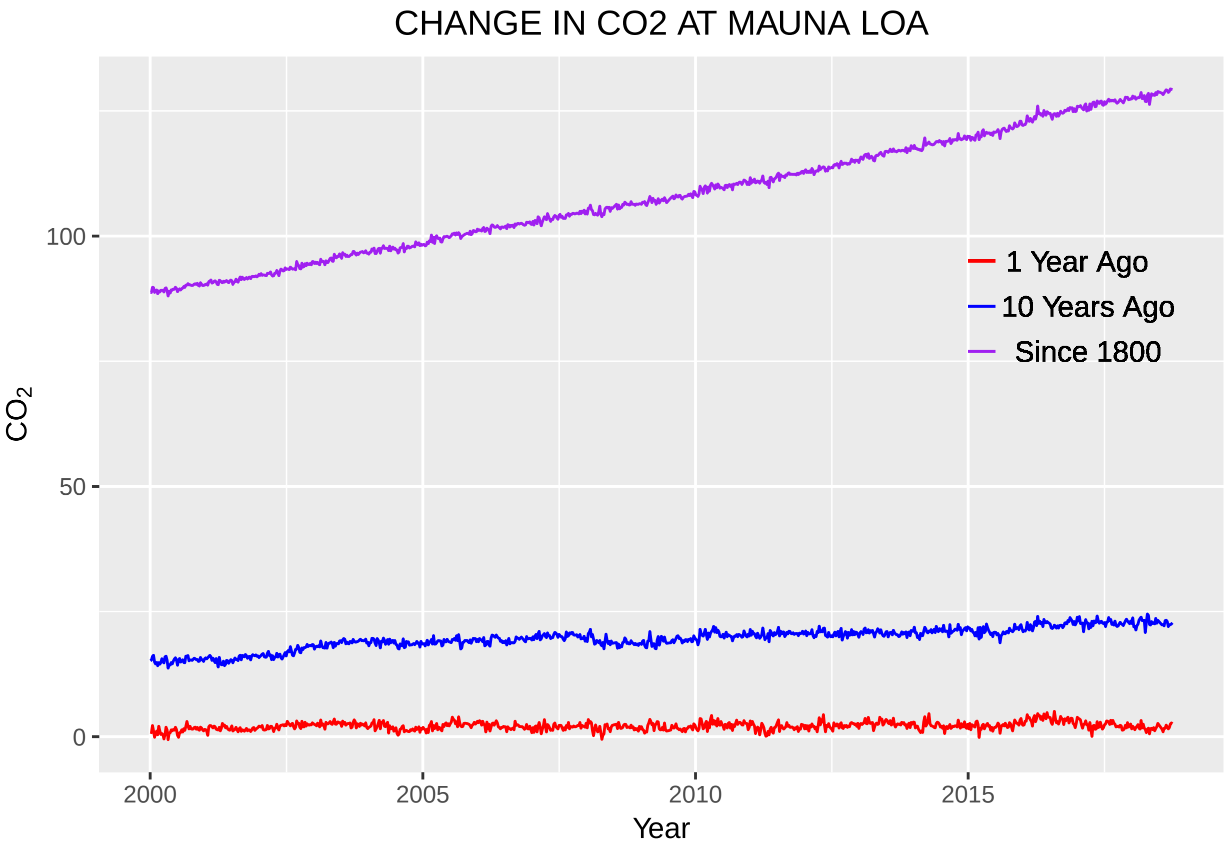


```
MonthlyNew <- Monthly %>% filter(year > 2012)
LastFiveYearsMonthlyRecord <-
  ggplot(MonthlyNew) +
  geom_line(aes(x = decimal, y = average), colour = "red") +
  geom_line(aes(x = decimal, y = trend)) +
  geom_segment(aes(x = 2013.2, y = 407, xend = 2013.3, yend = 407)) +
  geom_text(aes(x = 2013.43, y = 407, label= "trend", hjust = "left")) +
  geom_segment(aes(x = 2013.2, y = 406, xend = 2013.3, yend = 406), colour = "red") +
  geom_text(aes(x = 2013.7, y = 406, label = "average", hjsut = "left")) +
  xlab("Year") +
  ylab(expression("CO"[2]*"")) +
  ggtitle(expression("Last Five Years - Monthly Record of CO" [2]*""))
print(LastFiveYearsMonthlyRecord)
```



- Use the weekly Mauna Loa data to create a plot showing the change in CO2 for one year, 10 years, and since 2000. You will need to create new variables for the 1-year and 10-year change in CO$_2$.

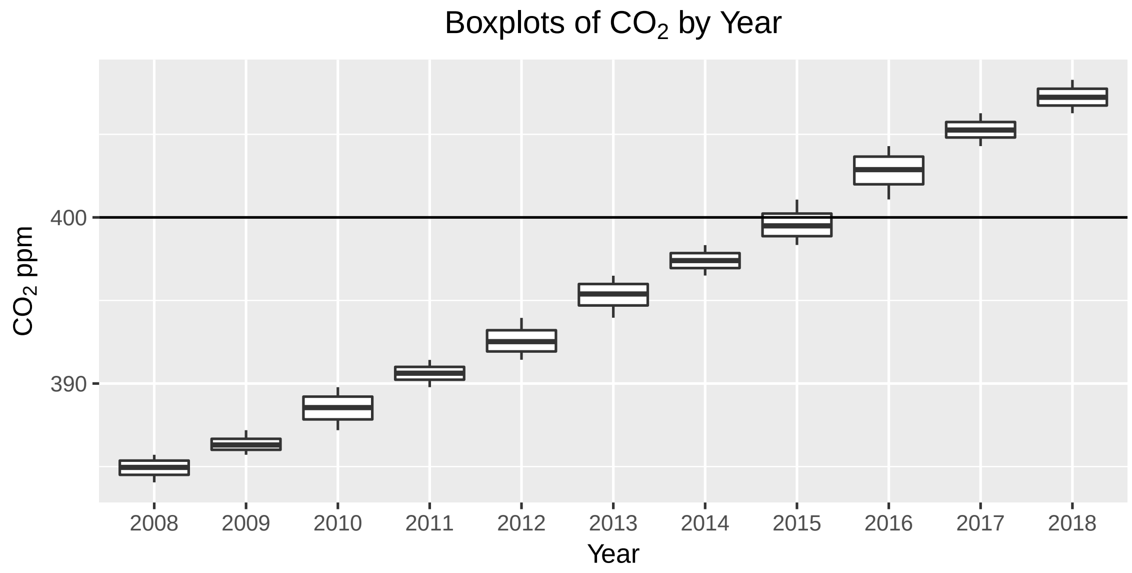
```
attach(Weekly)
Weekly$Change1Yr <-  CO2ppm - X1yr_ago
Weekly$Change10Yr <- CO2ppm - X10yr_ago
WeeklyNew <- Weekly %>% filter(decimal >= 2000, CO2ppm > 0, X1yr_ago > 0, X10yr_ago > 0)
# summary(WeeklyNew)

ggplot(WeeklyNew) + geom_line(aes(x = decimal, y = Change1Yr), colour = "red") + geom_line(aes(x = deci
```



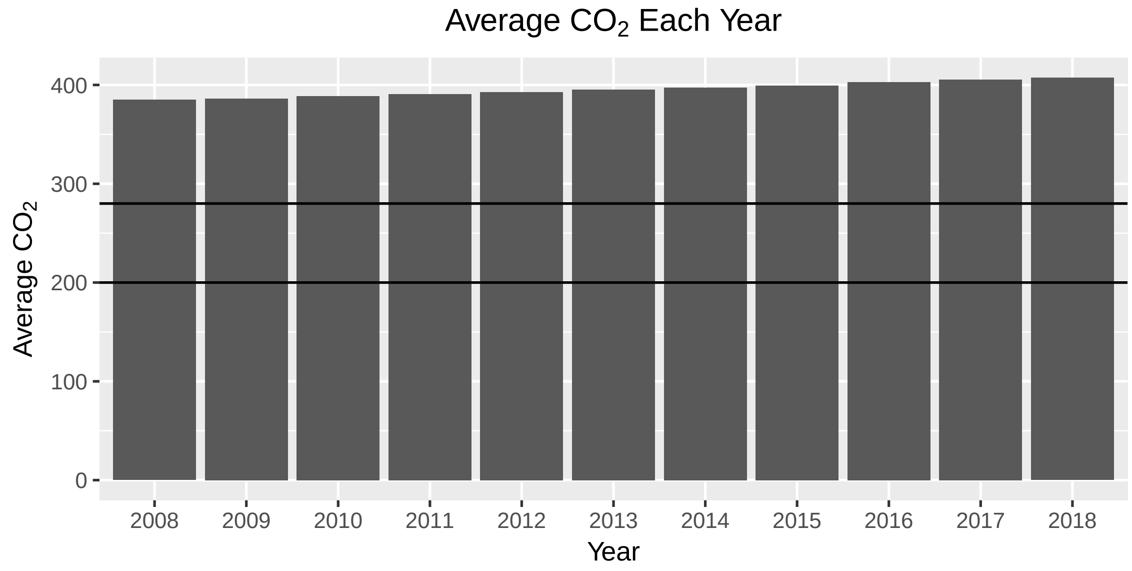
- Use the `grid.arrange()` function from the `gridExtra` package to create a display that includes the following three plots stacked on top of one another. Exclude the incomplete 2018 data from all plots.
  - Using daily global data, create side-by-side box plots of CO_2 by year. Include a horizontal line at 400ppm, which is considered by many to be a symbolic threshold CO_2 value for global warming/climate change.
  - Using daily global data, create a bar plot showing the mean CO_2 for each year. Include a horizontal line at 280ppm (approximate pre-industrial CO_2 average) and at 200ppm (approximate ice age CO_2 average).
  - Using the global growth data, create a bar plot of growth rates for the same time period shown in the other two plots, including error bars to indicate the degree of uncertainty in the estimates.

```
attach(Daily)

one <- ggplot(Daily, aes(x = factor(year), y = trend)) + geom_boxplot() + geom_hline(yintercept = 400)


data <- Daily %>% group_by(year) %>% summarise(mean = mean(trend))
two <- ggplot(data, aes(x = factor(year), y = mean)) + geom_bar(stat = "identity") + geom_hline(yinterc


attach(Growth)
GrowthNew <- Growth %>% filter(year >= 2008)
three <- ggplot(GrowthNew, aes(x = factor(year), y = anninc)) + geom_bar(stat = "identity") + geom_erro
```

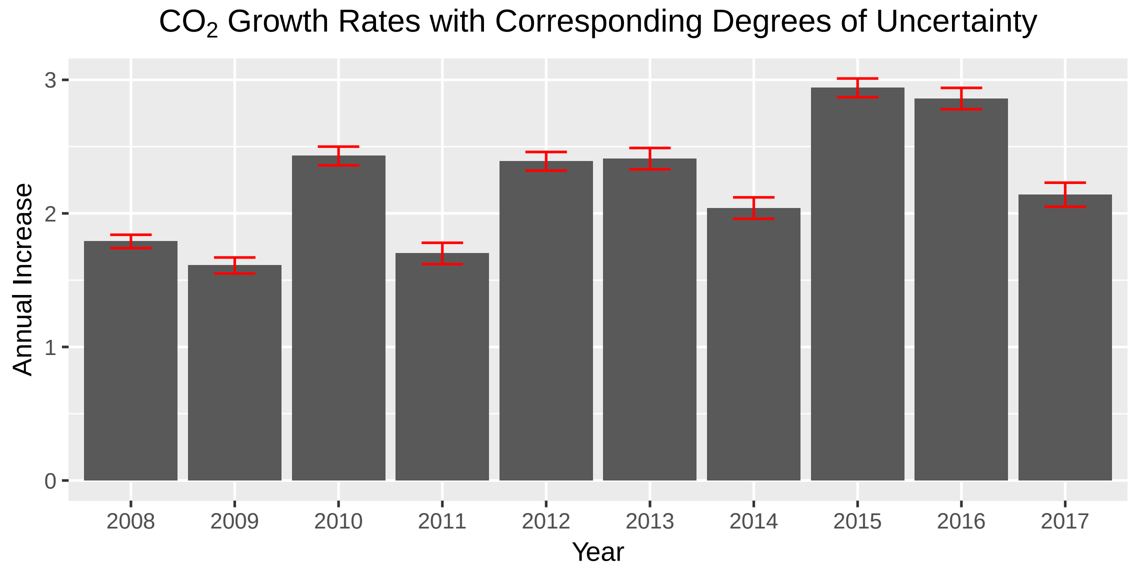
```
grid.arrange(one, two, three, ncol = 1, nrow = 3, heights=c(3,3,3))
```



Boxplots of CO_2 by Year



Average CO_2 Each Year



CO_2 Growth Rates with Corresponding Degrees of Uncertainty

QUESTIONS

1) What trend(s) or patterns do you observe with regard to CO_2 concentration over time?

ANSWER: CO_2 concentration has increased generally in a straight line over the past 5 years.

2) In what way could these analyses be used to support the theory of anthropogenic (man-made) climate change?

ANSWER: This analyses could be used to support anthropogenic climate change, but more research would need to be collected to make a causal claim. Currently with these analyses, we can only make a correlational claim that as more CO_2 has entered the atmosphere, it has changed the earth by hurting the ozone layer over Antarctica and creating a greenhouse effect by trapping energy from the sun inside the atmosphere instead of allowing it to escape as normal.

3) Why are data and graphs such as these *evidence* rather than *proof* of anthropogenic climate change?

ANSWER: These data and graphs are evidence, not proof, because they are not collected in a true experimental fashion. To do a true experiment, the data would need to be collected in a closed and controlled environment so you can clearly see that the increase in CO_2 directly effects the envrironment in a specific way. If this were able to be done, a direct causal claim could be made. However, there could be extra unseen variables contributing to the greenhouse effect and anthropogenic climate change that these analyses do not account for.