Take data from SQL database:

```
/* Used this to find what columns look like */
SELECT *
  FROM city_list
 LIMIT 10;

/* Used this to double check that my city is listed */
SELECT *
  FROM city_list
 WHERE city = ‘Columbus’;

/* Used this to pull up Columbus temperature data */
SELECT *
  FROM city_data
 WHERE city = 'Columbus';

/* Used this to pull up Global temperature data */
SELECT *
  FROM global_data;

/* Used this to pull up Tunis temperature data */
SELECT *
  FROM city_data
 WHERE city = 'Tunis';

/* Used this to pull up Berlin temperature data */
SELECT *
  FROM city_data
 WHERE city = 'Berlin';

/* Used this to pull up Belgrade temperature data */
SELECT *
  FROM city_data
 WHERE city = 'Belgrade';
```

Begin Analysis in R:

```
#Prepare the workspace

library(dyplyr)
library(readr)
library(ggplot2)
library(Hmisc)

setwd("C:/Users/Skywind/Desktop/R/data")

list.files()
[1] "Belgrade_data.csv"  "Berlin_data.csv"    "Columbus_data.csv"  "Global_data.csv"
"team_standings.csv"
```

```
[6] "Tunis_data.csv"

Columbus <- read_csv("Columbus_data.csv")
Global <- read_csv("Global_data.csv")
Berlin <- read_csv("Berlin_data.csv")
Belgrade <- read_csv("Belgrade_data.csv")
Tunis <- read_csv("Tunis_data.csv")

#Check structure of data

glimpse(Columbus)
Observations: 271
Variables: 4
$ year     <int> 1743, 1744, 1745, 1746, 1747, 1748, 1749, 1750, 1751, 1752, 1753, 1754,
1755, 1756, 1757, 1758, 1759...
$ city     <chr> "Columbus", "Columbus", "Columbus", "Columbus", "Columbus", "Columbus",
"Columbus", "Columbus", "Col...
$ country  <chr> "United States", "United States", "United States", "United States",
"United States", "United States"...
$ avg_temp <dbl> 7.46, 15.73, 6.91, NA, NA, NA, NA, 14.62, 15.36, 8.30, 14.00, 14.11,
11.66, 14.24, 13.62, 12.55, 13....

glimpse(Global)
Observations: 266
Variables: 2
$ year     <int> 1750, 1751, 1752, 1753, 1754, 1755, 1756, 1757, 1758, 1759, 1760, 1761,
1762, 1763, 1764, 1765, 1766...
$ avg_temp <dbl> 8.72, 7.98, 5.78, 8.39, 8.47, 8.36, 8.85, 9.02, 6.74, 7.99, 7.19, 8.77,
8.61, 7.50, 8.40, 8.25, 8.41...

glimpse(Berlin)
Observations: 271
Variables: 4
$ year     <int> 1743, 1744, 1745, 1746, 1747, 1748, 1749, 1750, 1751, 1752, 1753, 1754,
1755, 1756, 1757, 1758, 1759...
$ city     <chr> "Berlin", "Berlin", "Berlin", "Berlin", "Berlin", "Berlin", "Berlin",
"Berlin", "Berlin", "Berlin", ...
$ country  <chr> "Germany", "Germany", "Germany", "Germany", "Germany", "Germany",
"Germany", "Germany", "Germany", "...
$ avg_temp <dbl> 6.33, 10.36, 1.43, NA, NA, NA, NA, 9.83, 9.75, 4.84, 8.72, 8.49, 8.26,
9.62, 9.15, 8.25, 9.04, 8.99,...

glimpse(Belgrade)
Observations: 271
Variables: 4
$ year     <int> 1743, 1744, 1745, 1746, 1747, 1748, 1749, 1750, 1751, 1752, 1753, 1754,
1755, 1756, 1757, 1758, 1759...
$ city     <chr> "Belgrade", "Belgrade", "Belgrade", "Belgrade", "Belgrade", "Belgrade",
"Belgrade", "Belgrade", "Bel...
$ country  <chr> "Serbia", "Serbia", "Serbia", "Serbia", "Serbia", "Serbia", "Serbia",
"Serbia", "Serbia", "Serbia", ...
```

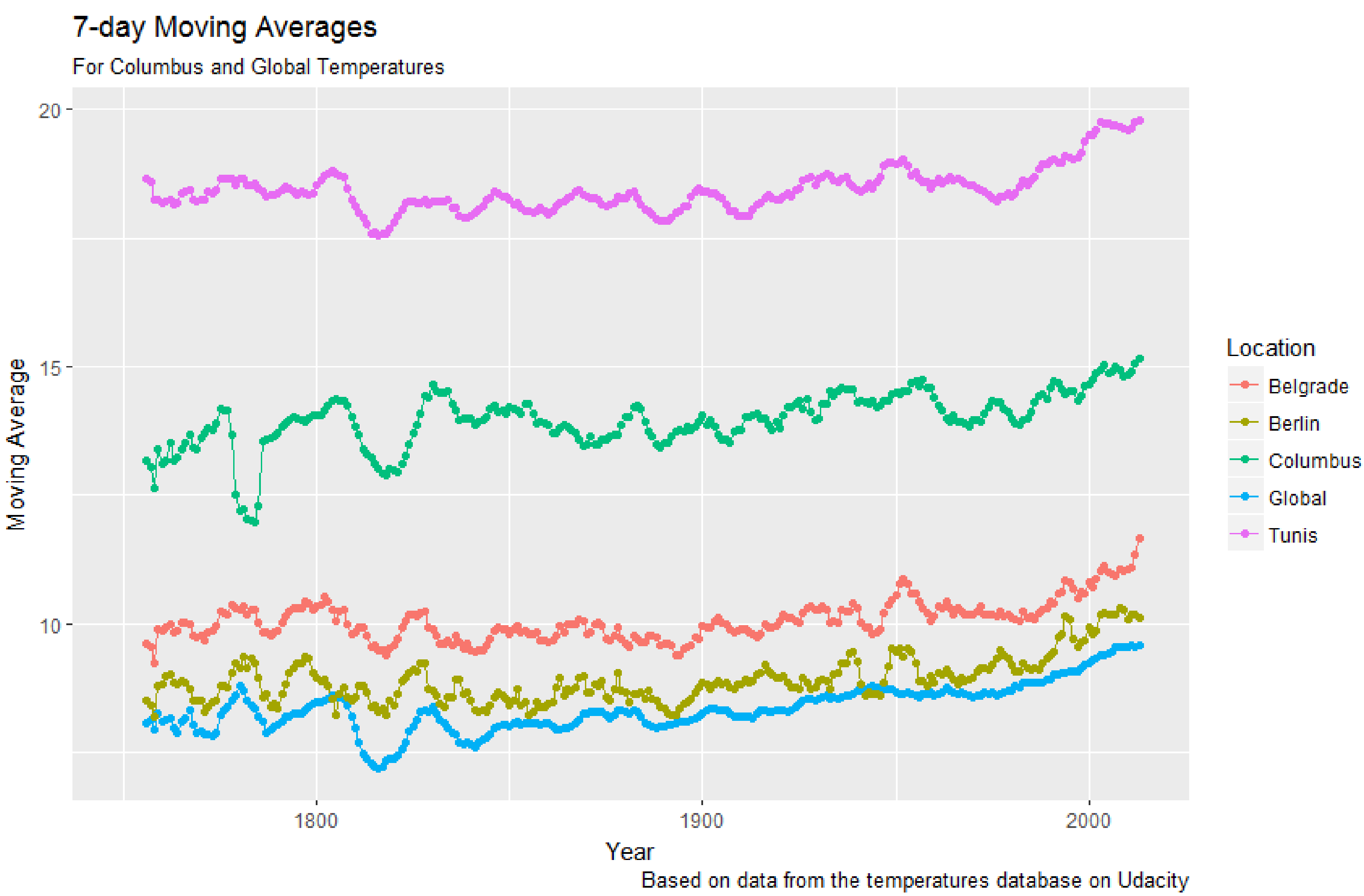
```
$ avg_temp <dbl> 5.53, 11.83, 2.57, NA, NA, NA, NA, 10.70, 11.12, 5.32, 10.00, 9.93,
9.72, 10.50, 10.18, 8.97, 10.01,...

glimpse(Tunis)
Observations: 271
Variables: 4
$ year     <int> 1743, 1744, 1745, 1746, 1747, 1748, 1749, 1750, 1751, 1752, 1753, 1754,
1755, 1756, 1757, 1758, 1759...
$ city     <chr> "Tunis", "Tunis", "Tunis", "Tunis", "Tunis", "Tunis", "Tunis", "Tunis",
"Tunis", "Tunis", "Tunis", "...
$ country  <chr> "Tunisia", "Tunisia", "Tunisia", "Tunisia", "Tunisia", "Tunisia",
"Tunisia", "Tunisia", "Tunisia", "...
$ avg_temp <dbl> 14.72, 19.66, 11.82, NA, NA, NA, NA, 18.83, 19.40, NA, 18.45, 18.47,
18.21, 18.67, 18.43, 17.25, 18....

tail(Columbus)
# A tibble: 6 x 4
   year     city       country avg_temp
  <int>    <chr>         <chr>    <dbl>
1  2008 Columbus United States    14.46
2  2009 Columbus United States    14.46
3  2010 Columbus United States    14.64
4  2011 Columbus United States    15.24
5  2012 Columbus United States    15.91
6  2013 Columbus United States    16.05

tail(Global)
# A tibble: 6 x 2
   year avg_temp
  <int>    <dbl>
1  2010     9.70
2  2011     9.52
3  2012     9.51
4  2013     9.61
5  2014     9.57
6  2015     9.83

tail(Berlin)
# A tibble: 6 x 4
   year   city country avg_temp
  <int>  <chr>   <chr>    <dbl>
1  2008 Berlin Germany    10.66
2  2009 Berlin Germany    10.06
3  2010 Berlin Germany     8.61
4  2011 Berlin Germany    10.56
5  2012 Berlin Germany     9.96
6  2013 Berlin Germany    10.12

tail(Belgrade)
# A tibble: 6 x 4
   year     city country avg_temp
```

```
  <int>    <chr>   <chr>    <dbl>
1  2008 Belgrade  Serbia    11.85
2  2009 Belgrade  Serbia    11.53
3  2010 Belgrade  Serbia    11.07
4  2011 Belgrade  Serbia    10.91
5  2012 Belgrade  Serbia    11.55
6  2013 Belgrade  Serbia    12.84

tail(Tunis)
# A tibble: 6 x 4
   year  city country avg_temp
  <int> <chr>   <chr>    <dbl>
1  2008 Tunis Tunisia    19.76
2  2009 Tunis Tunisia    19.64
3  2010 Tunis Tunisia    19.76
4  2011 Tunis Tunisia    19.53
5  2012 Tunis Tunisia    20.12
6  2013 Tunis Tunisia    20.00

#We see that the starting points and end points are not the same. Let's make them the same.

Columbus2 <- Columbus[-1:-7,]
Global2 <- Global[1:264,]
Berlin2 <- Berlin[-1:-7,]
Belgrade2 <- Belgrade[-1:-7,]
Tunis2 <- Tunis[-1:-7,]

#Create function to compute 7 day moving average

moving_avg <- function(x) {     #x is a vector

#Intialize loop variables and moving average vector

i <- 1
j <- 1
moving_average <- c()

 for (j in 1:(length(x)-6)) {

#Stop loop if go over limit

  if (j > (length(x)-6)) {

  break

  }

#Initialize variables after every possible calculation of moving average

 sum <- 0
```

```
 avg <- 0
 missing <- 0
 total <- 7

  for (i in j:(j+6)) {

#Stop loop if go over limit

   if (i > (j+6)) {

   break

   }

#If the value is missing, do not include in average for the 7 days

   if (is.na(x[i])) {

   total <- total - 1
   missing <- missing + 1


#If there were 7 zeros, insert a NA in the vector, can deal with it after entire vector is made

    if (missing == 7) {

    moving_average <- c(moving_average, NA)
    break

    }

   }

#Only if the number is not NA, add it to the sum

  if (!is.na(x[i])) {

  sum <- sum + x[i]

  }

#During the last iteration of the loop, find the average and put into the moving average vector

   if (i == (j+6) ) {

   avg <- sum / total
   moving_average <- c(moving_average, avg)

   }
```

```
  }

 }

return(moving_average)

}

#Create moving average vector for Columbus Table

moving_average_columbus <- moving_avg(Columbus2$avg_temp)

#Add NA's to beginning of list so matches Columbus table

moving_average_columbus <- c(rep(NA, 6), moving_average_columbus)

#Repeat above steps for Global data

moving_average_global <- moving_avg(Global2$avg_temp)
moving_average_global <- c(rep(NA, 6), moving_average_global)

#Repeat for Berlin data

moving_average_berlin <- moving_avg(Berlin2$avg_temp)
moving_average_berlin <- c(rep(NA, 6), moving_average_berlin)

#Repeat for Belgrade data

moving_average_belgrade <- moving_avg(Belgrade2$avg_temp)
moving_average_belgrade <- c(rep(NA, 6), moving_average_belgrade)

#Repeat for Tunis data

moving_average_tunis <- moving_avg(Tunis2$avg_temp)
moving_average_tunis <- c(rep(NA, 6), moving_average_tunis)

#Create new vectors to create a new data frame

year <- c(rep(1750:2013,5))
moving_average <- c(moving_average_columbus, moving_average_global,
moving_average_berlin, moving_average_belgrade, moving_average_tunis)
location <- c(rep("Columbus", 264), rep("Global", 264), rep("Berlin", 264),
rep("Belgrade", 264), rep("Tunis", 264))

#Create new data frame

combined <- data.frame(year, location, moving_average)

#Create line plot
```

```
p <- ggplot(data = combined, aes(x = year, y = moving_average, group = location)) +
  geom_line(aes(color = location)) +
  geom_point(aes(color = location))

#Change title

p <- p + ggtitle("7-day Moving Averages", subtitle = "For Columbus and Global
Temperatures")

#Change x and y axis labels

p <- p + xlab("Year")
p <- p + ylab("Moving Average")

#Insert caption statement

p <- p + labs(caption = "Based on data from the temperatures database on Udacity")

#Change legend title

p <- p + labs(colour = "Location")

p
```



#Create shorter variable names, remove NA's, and make new year variable

```
columbus <- moving_average_columbus[-1:-6]
belgrade <- moving_average_belgrade[-1:-6]
berlin <- moving_average_berlin[-1:-6]
tunis <- moving_average_tunis[-1:-6]
global <- moving_average_global[-1:-6]
year <- 1750:2013
year <- year[-1:-6]
```

#Create data frame to compute correlation matrix

```
combined2 <- data.frame(year, columbus, belgrade, berlin, tunis, global)
```

#Create correlation matrix

```
my_data <- rcorr(as.matrix(combined2))
cor_matrix <- my_data$r
cor_matrix
```

```
              year  columbus  belgrade    berlin     tunis    global
year     1.0000000 0.6376094 0.5340216 0.5982758 0.5472689 0.7237428
columbus 0.6376094 1.0000000 0.5244902 0.4545766 0.5806746 0.6071708
belgrade 0.5340216 0.5244902 1.0000000 0.8964901 0.8762659 0.8362129
berlin   0.5982758 0.4545766 0.8964901 1.0000000 0.7951491 0.8052800
tunis    0.5472689 0.5806746 0.8762659 0.7951491 1.0000000 0.8921835
global   0.7237428 0.6071708 0.8362129 0.8052800 0.8921835 1.0000000
```

#Create a simple linear regression model. Response is columbus. Predictor is global.

```
model <- lm(columbus ~ global, data = combined2)
summary(model)

Call:
lm(formula = columbus ~ global, data = combined2)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.06857 -0.16915  0.04729  0.22784  0.80411 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.03569    0.48613   16.53   <2e-16 ***
global       0.71066    0.05812   12.23   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4306 on 256 degrees of freedom
Multiple R-squared:  0.3687,	Adjusted R-squared:  0.3662
```

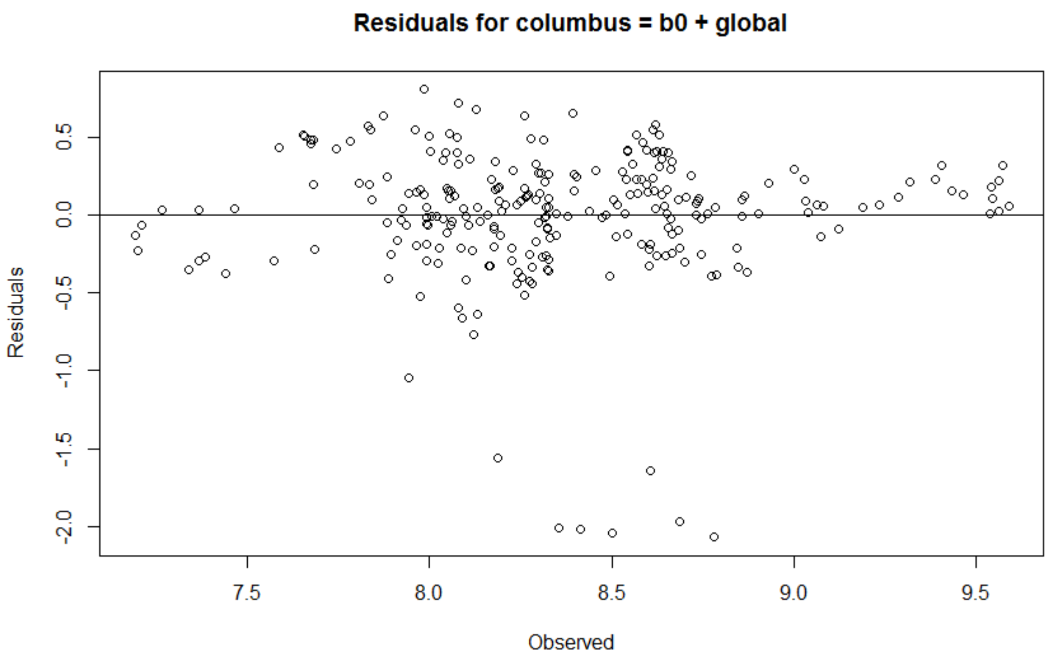
```
F-statistic: 149.5 on 1 and 256 DF,  p-value: < 2.2e-16
```

#Create another model. Response is global and predictor is year

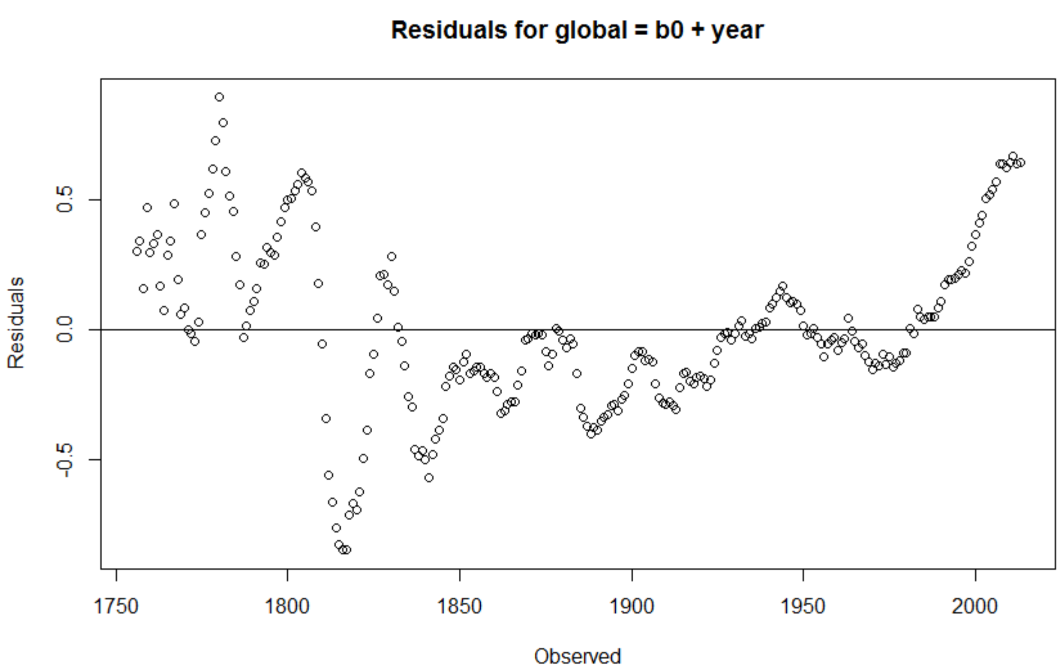
```
model2 <- lm(global ~ year, data = combined2)
summary(model2)

Call:
lm(formula = global ~ year, data = combined2)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.85234 -0.18095 -0.03272  0.17508  0.89758 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.0954114  0.5037194  -0.189     0.85    
year         0.0044819  0.0002671  16.781   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3195 on 256 degrees of freedom
Multiple R-squared:  0.5238,	Adjusted R-squared:  0.5219 
F-statistic: 281.6 on 1 and 256 DF,  p-value: < 2.2e-16
```

#Residual vs observed plot for model

```
model_res <- resid(model)

plot(combined2$global, model_res, ylab = "Residuals", xlab = "Observed", main = 
"Residuals for columbus = b0 + global")

abline(0,0)
```

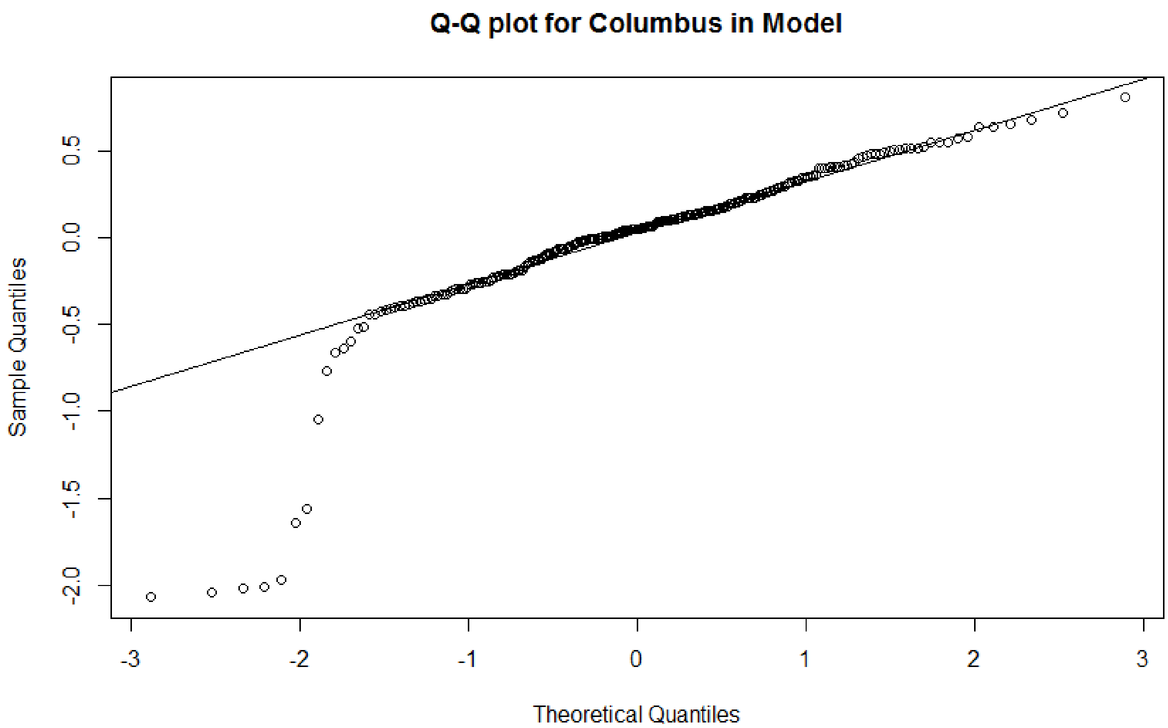



```
#Residual vs observed plot for model2

model2_res <- resid(model2)

plot(combined2$year, model2_res, ylab = "Residuals", xlab = "Observed", main =
"Residuals for global = b0 + year")

abline(0,0)
```



#QQ plot model 1

```
qqnorm(model_res, main = "Q-Q plot for Columbus in Model")
qqline(model_res)
```



Report/Analysis:

On the 7-day Moving Averages plot we see that my city, Columbus, on average is hotter compared to the Global average. We see that Tunis has the highest average temperature out of all five cities. Generally, the four cities that I have chosen are hotter than the Global average. Berlin sometimes was lower than the Global average around the year 1800 and 1945. Since the Columbus line never touches or goes below the Global line, the difference has been consistent over time. The changes in temperature in Columbus have been fairly consistent with changes in the Global average in general. We see that one of the biggest dips in average Global temperature occurs a little after the year 1800. We see a similar drop in average temperature in Columbus. The other cities on the plot have dropped around the same time as well. This is most likely due to a climate change that affected the entire world. The small peaks and dips that are present in the cities are just small differences from the year to year, some being hotter, and some being colder. In the overall trend we see that there are ups and downs, but it seems to increase in the long run. Comparing back to 1750, whether it's looking at the individual cities or globally, average temperature has gone up.

In the correlation matrix, we see a strong positive correlation between Global and Tunis with a value of 0.89. We see other high values as well such as between Tunis and Belgrade is 0.88, Berlin and Belgrade with 0.90, and Global and Belgrade with 0.84. We look at the rest of the correlation values between Belgrade, Berlin, Tunis, and Global, they are all high and related to each other. These high values are attributed due to the fact that they seem to peak and drop around the same time resulting in a high correlation. If we look at the lowest correlation value, we see Columbus and Berlin have a value of 0.45. We look at the overall moving average plot and see a massive drop in average temperature for Columbus between 1750 and 1800. Meanwhile, Belgrade, Berlin, Tunis, and Global increased around this time. With these small differences over the course of the years resulted in a lower correlation value. We also see a relatively high correlation value between year and Global with a value of 0.72. This is in line with my previous observation that overall, the average temperatures go up. Despite all these high correlation values, one does not cause the other. Time passing forward does not necessarily cause higher average Global temperature values. High temperature values in Belgrade, Berlin, Tunis, or Global does not influence the other. Most likely, there is a third variable confounding the results shown here. The likely culprit is the phenomenon called the greenhouse effect due to carbon dioxide emission. It would be related to average Global temperature it could be the cause of the average increase over the years. It would also be related to the year because technological advancements in the last several hundred years increased carbon dioxide and other greenhouse gasses production. But, we won't be able to make any conclusive evidence within this data set.

We can predict Columbus' temperature based on the Global temperature. In the simple linear regression model we have:

Columbus = 8.03569 + 0.71066Global   where Columbus is the predicted value of Columbus.

This model means that for every 1 degree Celsius increase in temperature in the Global temperature, we will observe an increase of 0.71066 degrees Celsius for Columbus.

The R^2 value for this model is 0.3662. This means that 36.62% of the variation observed in the average temperature in Columbus is explained by the variation in Global average temperature. This low value is expected because of that confounding variable. There are other variables that could be added to the model that could increase that R^2 value.

In the residual vs observed plot, we see that the plot satisfies the equal variances in our model assumption. The QQ plot indicates that our residuals are about normally distributed as well. However, the model does not satisfy the independence assumption. If any of these temperature data was plotted against time, such as seen in the residuals plot for model2, we see that it violates our assumptions. We also know that time and average temperatures are confounded by gas emission around the globe. Also, we have an observational study instead of experiment data so really none of these conclusions are really valid. In conclusion, average Global temperature is not independent from average Columbus temperature, to any other city's average temperature, or to year.

The model for the model2 senseless in itself. Instead of a simple linear regression model, a time series model would be better.

We could predict average temperature in Columbus based on average temperature of Global temperature or Global temperature based on the year if we wanted to. In the second model we have:

Global = -0.0954114 + 0.0044819year   where Global is the predicted value of Global

We can try to predict the average temperature in Columbus for year 3000 with our models.

Global(3000) = -0.0954114 + 0.0044819(3000) = 13.3503 Celsius

Columbus(13.3503) = 8.03569 + 0.71066(13.3503) = 17.5232 Celsius

Next, let's try to predict the average temperature in Columbus for year -4000 with our models (which is senseless since Columbus didn't exist back then, but the area still did).

Global(-4000) = -0.0954114 + 0.0044819(-4000) = -18.0230 Celsius

Columbus(-18.0230) = 8.03569 + 0.71066(-18.0230) = -4.7725 Celsius

We see that both of these values are so extreme and outrageous that it's definitely incorrect. We know from history that the last ice age was not 6000 years ago. These temperatures indicate that of an ice age is happening.

So, we could predict average temperature in Columbus based on the average temperature globally and the average temperature globally based on the year as you can see here, but with very inaccurate results. That comes with no surprise with the conclusions we made prior to attempting to make those predictions. Our model is not valid, so it's going to be outright wrong.