2 - Prediction

J. Williams

23 feb 2022

Classification and choice of K

On a first instance we will be exploring our Minneapolis police incidents dataset in a more systematic way and construct visualizations that explore the effect of the value of K in our KNN model.

Let's start by loading the dataset

```
mn.police.tbl <- read_csv("~/Mscs 341 S22/Class/Data/police_incidents.mn.csv")
```

And remember the steps that we took last time, namely:

1. Divide our dataset into training and testing datasets.
2. Construct a model based on the training dataset.
3. Evaluate the fit of the model by calculating the MSE on the testing dataset

These steps can be done as follows:

```
# Divide dataset into testing and training
train.mn.police.tbl <- mn.police.tbl %>%
  filter(year==2016)
test.mn.police.tbl <- mn.police.tbl %>%
  filter(year==2017)

# Build the KNN model
kNear=7
knn.model <- knnreg(tot~week, data=train.mn.police.tbl,k=kNear)

# Evaluate the MSE of the KNN model in the testing dataset
test.pred <- predict(knn.model, test.mn.police.tbl)
(mse.test <- mean ((test.mn.police.tbl$tot-test.pred)^2))
```

```
## [1] 155.8458
```

We are interested in calculating systematically the MSE as we iterate over the parameter k by doing the following steps:

1. Create a function `calc_MSE(kNear, train.tbl, test.tbl)` that trains a KNN model with parameter `kNear` on `train.tbl` and then applies the model on `test.tbl` and calculates the MSE. Test your function with k=7 and k=35 using our testing and training datasets.

```
calc_MSE <- function(kNear, train.tbl, test.tbl) {
  knn.model <- knnreg(tot~week, data=train.tbl, k=kNear)

  test.pred <- predict(knn.model, test.tbl)
  mean((test.tbl$tot - test.pred)^2)
}
```

```
calc_MSE(7, train.mn.police.tbl, test.mn.police.tbl)
```

```
## [1] 155.8458
```

```
calc_MSE(35, train.mn.police.tbl, test.mn.police.tbl)
```

```
## [1] 149.0255
```

2. We would like to create a vector with the MSE values for our testing dataset. Notice it only makes sense to look at values of k in increments of 7 (why?). Use a for loop in R(Look at the syntax of for loops in https://rafalab.github.io/dsbook/programming-basics.html#for-loops) to create the mse for $k = 7, 14, 21, ..., 364$

```
MSE.test <- vector(length = 52)
MSE.train <- vector(length = 52)
for (i in 1:52) {
  MSE.test[i] <- calc_MSE(i*7, train.mn.police.tbl, test.mn.police.tbl)
  MSE.train[i] <- calc_MSE(i*7, train.mn.police.tbl, train.mn.police.tbl)
}

MSE.test
```

```
##  [1] 155.8458 150.4184 151.1748 148.0340 149.0255 148.9372 149.4296 149.7515
##  [9] 149.7698 149.4829 149.5480 147.5733 147.9949 147.8568 147.7932 147.1485
## [17] 146.8099 147.5607 147.7083 148.0340 147.5022 148.2566 148.0311 147.6000
## [25] 146.7298 146.9207 146.0966 146.5842 145.9356 146.2218 145.0544 146.0451
## [33] 145.2785 146.7315 146.2946 148.0077 147.8052 149.6433 150.0378 152.3477
## [41] 152.7568 155.0693 155.8649 158.4674 159.6662 161.8858 163.8238 166.2826
## [49] 167.5178 169.4999 171.2018 172.9603
```

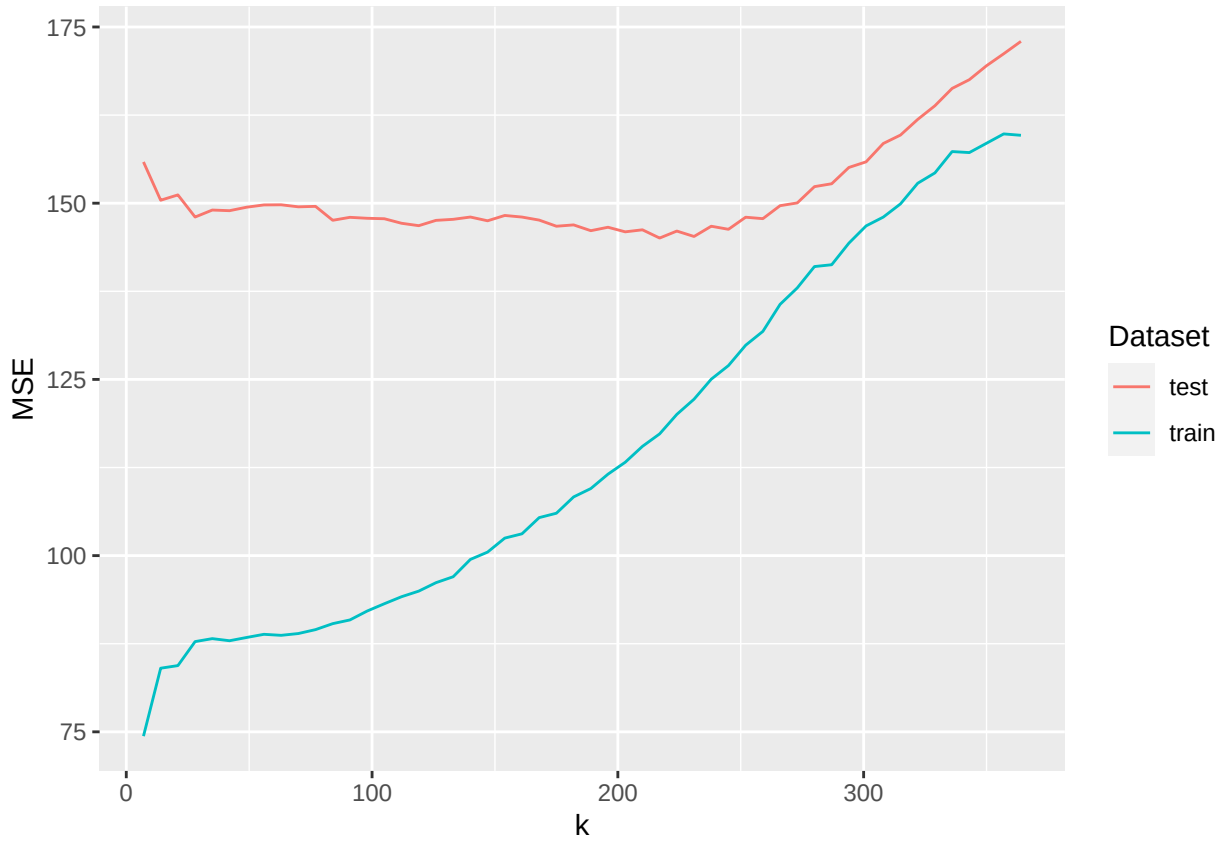
```
MSE.train
```

```
##  [1]  74.38533  84.02415  84.39819  87.80070  88.22261  87.92317  88.38951
##  [8]  88.83341  88.68916  88.94294  89.49829  90.34616  90.87219  92.13889
## [15]  93.17287  94.16757  94.96160  96.15722  97.00190  99.45887 100.49841
## [22] 102.48378 103.09478 105.38817 106.01265 108.32493 109.51318 111.55615
## [29] 113.24546 115.49922 117.27036 120.03796 122.19204 125.02854 126.96264
## [36] 129.84923 131.81150 135.64431 137.97729 141.00899 141.26744 144.33255
## [43] 146.77651 148.02694 149.90044 152.82722 154.29698 157.32222 157.17755
## [50] 158.50259 159.82773 159.62983
```

3. Generate a graph depicting the MSE as a function of `k` for both testing and training datasets. What is the optimal value for `k` based on the testing dataset? Can you find this value in a systematic way? (*Hint*: Check the documentation for function `slice_min`). Compare this graph against figures 2.9 and 2.10 from your book. What would be the equivalent of the parameter `flexibility` in your KNN model?

```
MSE.tbl <- tibble(k = 7*(1:52),
                  train = MSE.train,
                  test = MSE.test) %>%
  pivot_longer(2:3, names_to = "Dataset", values_to = "MSE")

ggplot(MSE.tbl, aes(x = k, y = MSE, color = Dataset)) +
  geom_line()
```




```
#optimal k = 145
MSE.tbl %>%
  filter(Dataset == "test") %>%
  slice_min(MSE, with_ties="false")
```

```
## # A tibble: 1 x 3
##       k Dataset   MSE
##   <dbl> <chr>   <dbl>
## 1   217 test     145.
```

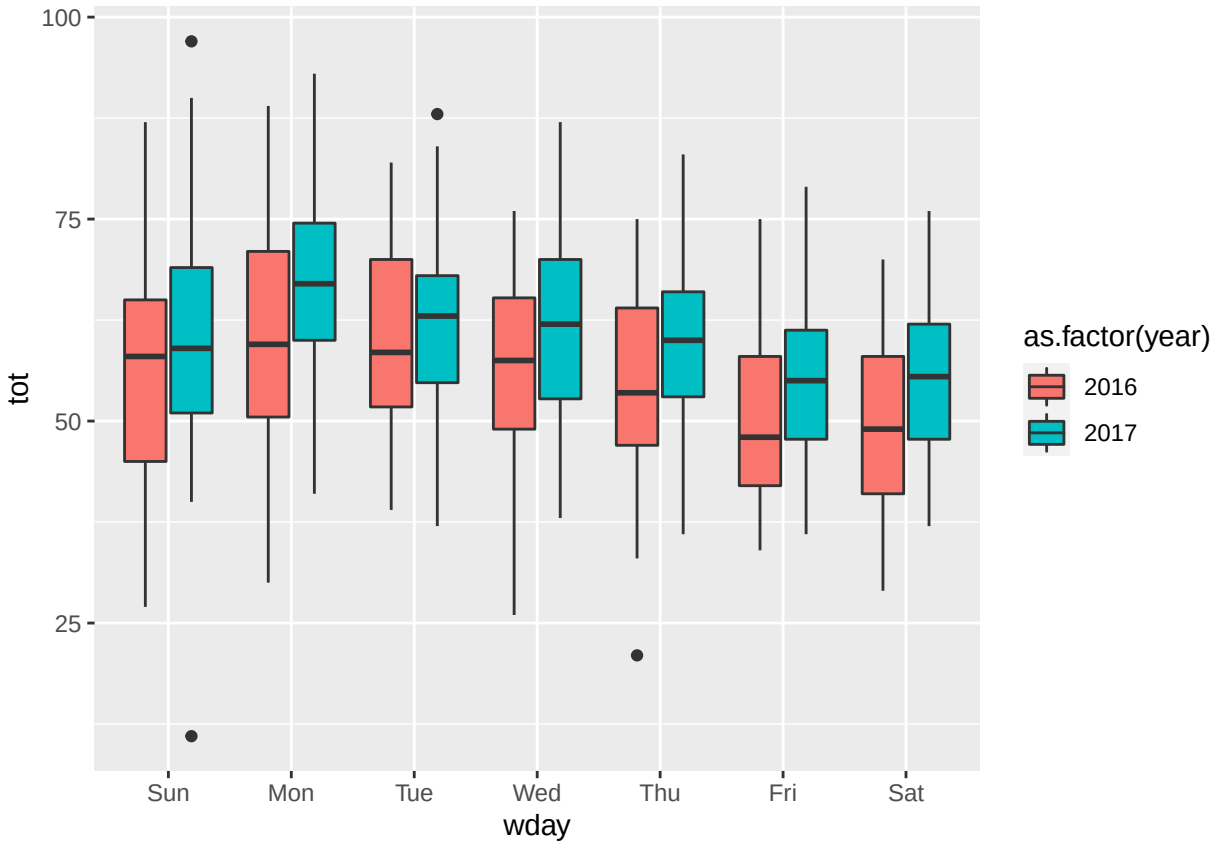
Improving your model

One way to improve the performance of a model is to use the information provided by other variables. In the following exercises we will explore this in more detail:

4. Does the distribution of number of incidents across the days of the week? Generate a boxplot to explore this question and check if this behavior is consistent across the years

```
days = unique(mn.police.tbl$wday)

mn.police.tbl %>%
  mutate(year = as.factor(year),
         wday = factor(wday, levels=days)) %>%
ggplot(aes(x = wday, y = tot, fill=as.factor(year))) +
  geom_boxplot()
```




```
# incidents cycle toward a peak on Mondays
```

5. It seems police incidents are higher on Monday as opposed to other days. Subset you dataset to only Mondays and construct a KNN model using `week` as input variable and train it using the data from 2016. Plot a graph with the MSE for all different choices of k and select a k that minimizes the MSE on the testing. Is the MSE smaller using this model than our original model?

```
police.monday.tbl <- mn.police.tbl %>%
  filter(wday == "Mon", year==2016)

monday.test.tbl <- mn.police.tbl %>%
  filter(wday == "Mon", year==2017)

police.monday.tbl #training set
```

```
## # A tibble: 52 x 4
##    wday   week  year   tot
##    <chr> <dbl> <dbl> <dbl>
##  1 Mon       1  2016    53
##  2 Mon       2  2016    42
##  3 Mon       3  2016    30
##  4 Mon       4  2016    70
##  5 Mon       5  2016    33
##  6 Mon       6  2016    49
##  7 Mon       7  2016    40
##  8 Mon       8  2016    51
##  9 Mon       9  2016    34
```

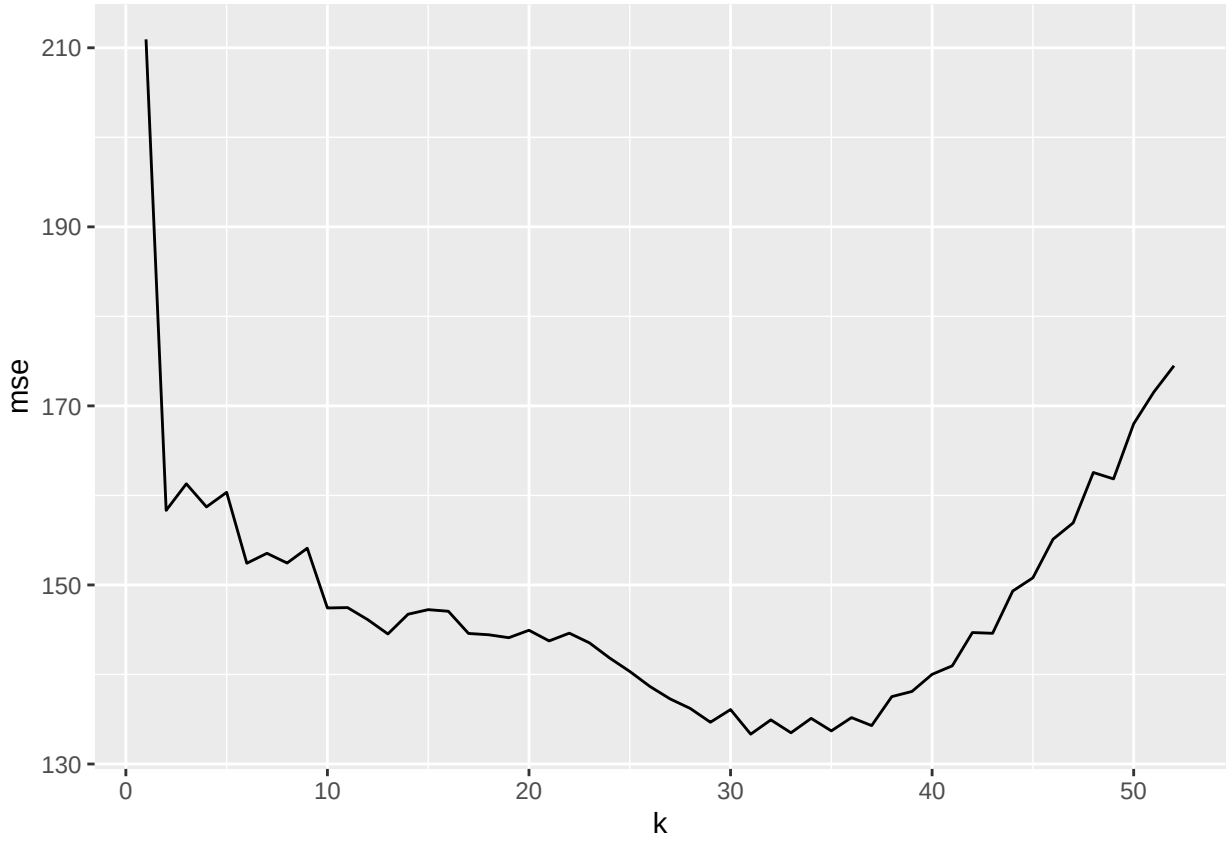
```
## 10 Mon      10  2016    54
## # ... with 42 more rows
```

```
# create
MSE.monday <- vector(length = 52)
for (i in 1:52) {
  MSE.monday[i] <- calc_MSE(i, police.monday.tbl, monday.test.tbl)
}
monday.tbl <- tibble (k = 1:52,
                      mse = MSE.monday)

monday.tbl
```

```
## # A tibble: 52 x 2
##        k   mse
##    <int> <dbl>
##  1     1  211.
##  2     2  158.
##  3     3  161.
##  4     4  159.
##  5     5  160.
##  6     6  152.
##  7     7  154.
##  8     8  152.
##  9     9  154.
## 10    10  147.
## # ... with 42 more rows
```

```
# plot
ggplot(monday.tbl, aes(x = k, y = mse)) +
  geom_line()
```




```
best.k <- monday.tbl %>%
  slice_min(order_by = mse, with_ties = FALSE)

best.k # best k is 31, mse of 133
```

```
## # A tibble: 1 x 2
##       k   mse
##   <int> <dbl>
## 1    31  133.
```

Studying the COVID pandemic

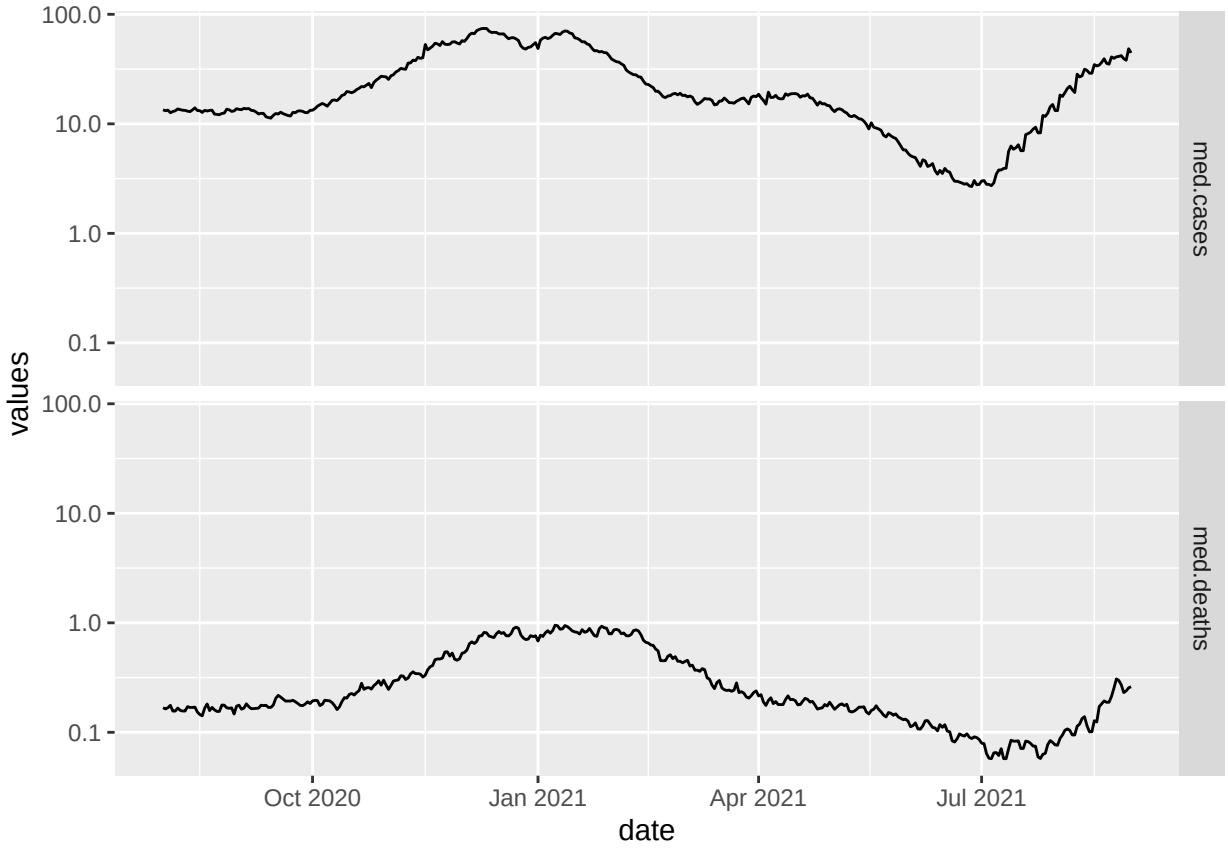
For the next set of exercises we will be using the US covid dataset procured from CovidActNow. We'll start by loading the dataset. Notice that I also loaded the library `lubridate` which allows for convenient use of dates.

```
library(lubridate)
covid.tbl <- read_csv("~/Mscs 341 S22/Class/Data/covid.csv")
```

6. Subset your dataset from August 2020 to August 2021 and plot the median number of cases and the median number of deaths (per 100,000). *Hint*: You can filter dates using `filter` and you will need to create a reference date with `as.Date`

```
covid.tbl %>%
  filter(date >= as.Date("2020-08-01"), date <= as.Date("2021-08-31")) %>%
  group_by(date) %>%
```

```
  mutate(med.cases = median(cases.100k),
         med.deaths = median(deaths.100k)) %>% # summarize(med.cases = median(cases.100k), med.deaths =
  pivot_longer(6:7, names_to="type", values_to="values") %>%
    ggplot(aes(date,values)) +
      geom_line() +
    facet_grid(type ~.)+
  scale_y_continuous(trans = 'log10')
```

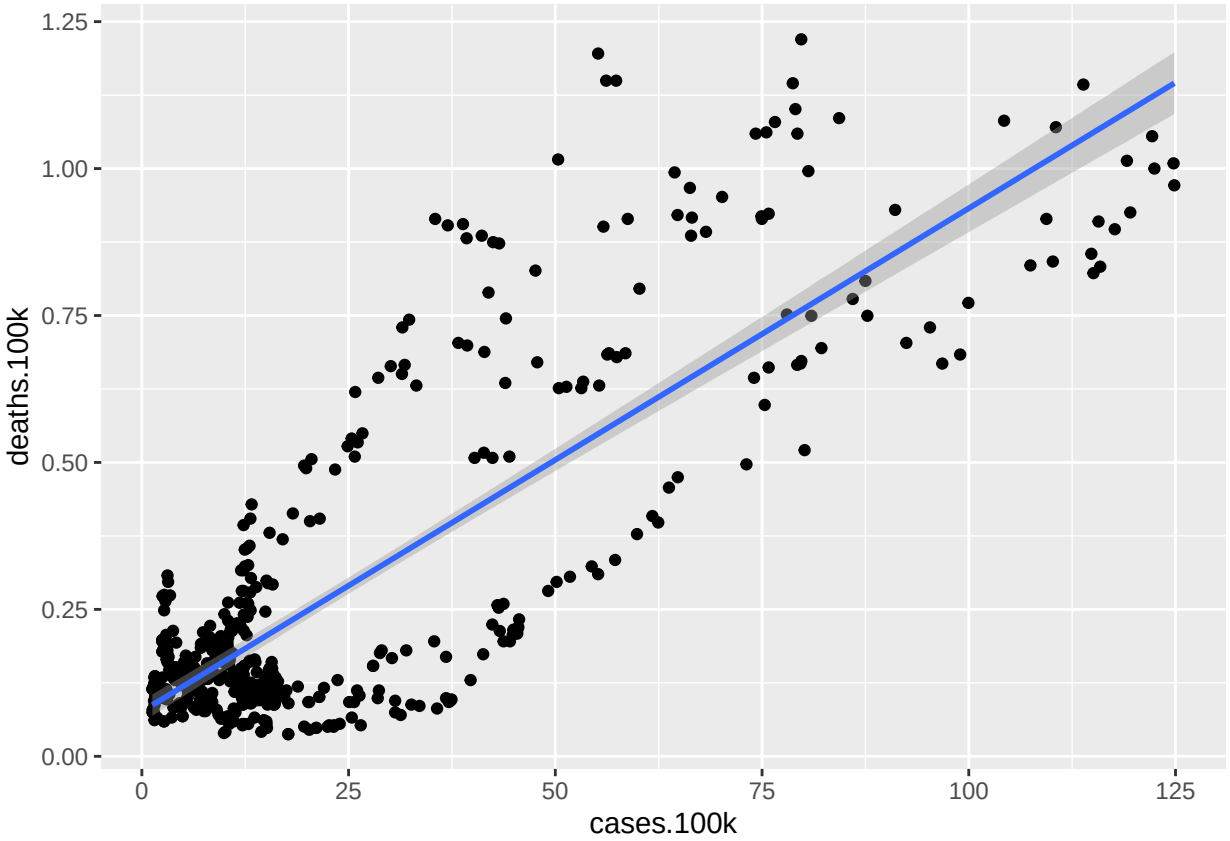


7. We would like to create a model that would be able to predict the number of deaths based on the number of cases. To do that let's create training and testing datasets from neighboring states, let's say WI and MN. Plot the number of cases against the number of deaths for your training dataset and include a linear trend in your plot

```
covid.train.tbl <- covid.tbl %>%
  filter(state == "WI") %>%
  select(cases.100k, deaths.100k)

covid.test.tbl <- covid.tbl %>%
  filter(state == "MN") %>%
  select(cases.100k, deaths.100k)

ggplot(covid.train.tbl, aes(x = cases.100k, y = deaths.100k)) +
  geom_point() +
  geom_smooth(method = "lm")
```



8. Create a linear model using `lm()` using the data from WI (training) and evaluate how well it does in MN (testing).

```
lm.model <-  lm(deaths.100k ~ cases.100k, data = covid.train.tbl)

death.pred <- predict(lm.model, covid.test.tbl)
mse.death.pred <- mean ((covid.test.tbl$deaths.100k-death.pred)^2)


mse.death.pred # MSE of 0.03550234
```

```
## [1] 0.03550234
```

9. To improve our model, let's make use of the fact that the number of covid cases is a good predictor of the number of deaths a couple of weeks afterwards. Create a function `calc_MSE_lag(time.lag, train.tbl, test.tbl)` that calculates the MSE on the testing dataset by using a linear model where the deaths lag the number of cases by `time.lag` *Hint* Make use of the functions `lead/lag` from the tidyverse.

```
calc_MSE_lag <- function(time.lag, train.tbl, test.tbl) {
  train.tbl <- train.tbl %>%
    mutate(lag.cases = lag(train.tbl$cases.100k, time.lag)) %>%
    filter(!is.na(lag.cases))

  test.tbl <- test.tbl %>%
    mutate(lag.cases = lag(test.tbl$cases.100k, time.lag)) %>%
    filter(!is.na(lag.cases))
```

```
  model = lm(deaths.100k ~ lag.cases, data = train.tbl)
  pred.test <- predict(model, test.tbl)

  mean((test.tbl$deaths.100k - pred.test)^2)
}

calc_MSE_lag(7, covid.train.tbl, covid.test.tbl)
```

```
## [1] 0.02468632
```

```
calc_MSE_lag(14, covid.train.tbl, covid.test.tbl)
```

```
## [1] 0.01998069
```

```
calc_MSE_lag(21, covid.train.tbl, covid.test.tbl)
```

```
## [1] 0.02165807
```

10. Plot lag versus MSE on the testing dataset and find the optimal parameter of lag. How do you interpret this optimal lag? Using this value of lag, plot the lagged number of cases versus the deaths and the linear trend line for the testing dataset.

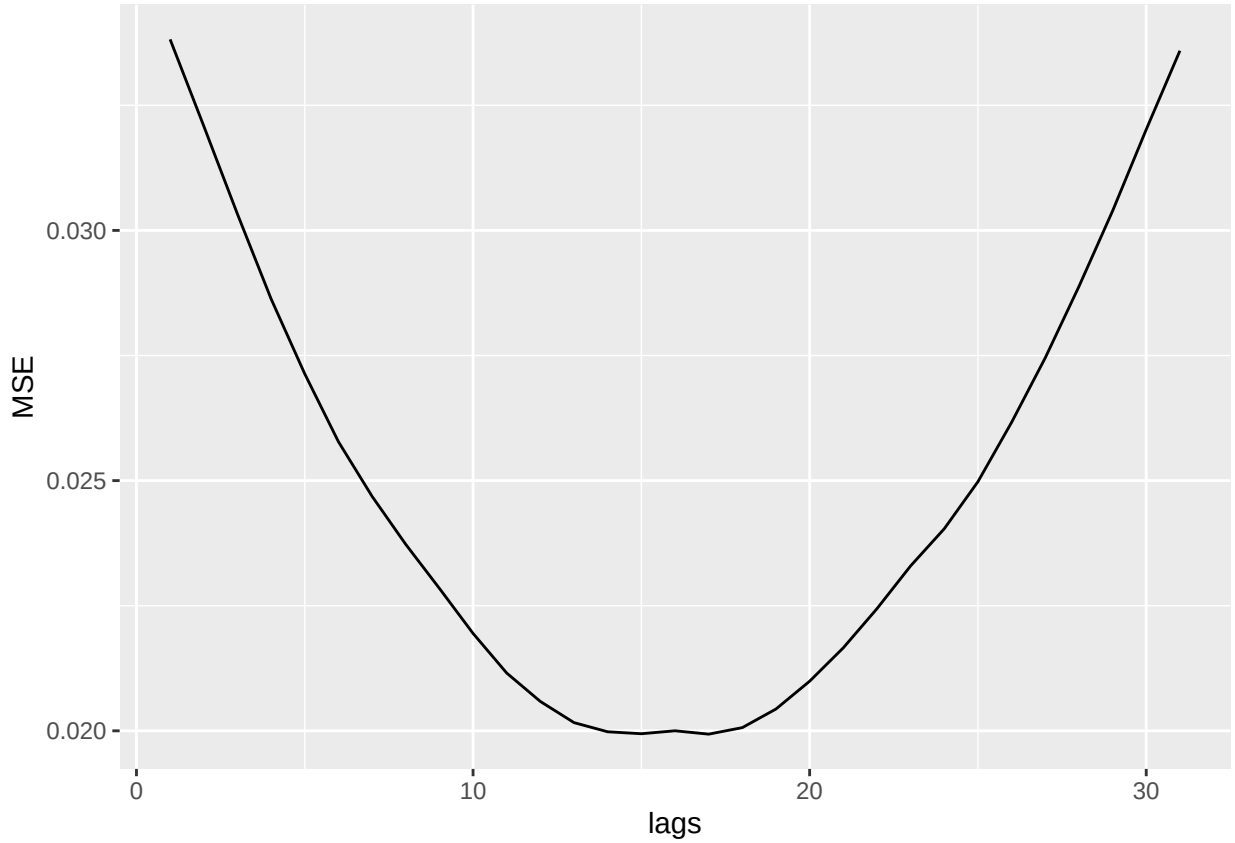
```
test.mse <- vector(length = 31)
for (i in 1:31) {
  test.mse[i] <- calc_MSE_lag(i, covid.train.tbl, covid.test.tbl)
}
test.mse
```

```
##  [1] 0.03381827 0.03208577 0.03032746 0.02863154 0.02713340 0.02577796
##  [7] 0.02468632 0.02372597 0.02284667 0.02194681 0.02115490 0.02058671
## [13] 0.02016346 0.01998069 0.01994182 0.02000036 0.01993318 0.02006407
## [19] 0.02043582 0.02099120 0.02165807 0.02244098 0.02329612 0.02403824
## [25] 0.02497813 0.02616486 0.02745443 0.02887880 0.03038843 0.03201660
## [31] 0.03359132
```

```
lags.tbl <- tibble (MSE = test.mse) %>%
  mutate(lags = 1:31)
lags.tbl
```

```
## # A tibble: 31 x 2
##       MSE  lags
##     <dbl> <int>
##  1 0.0338     1
##  2 0.0321     2
##  3 0.0303     3
##  4 0.0286     4
##  5 0.0271     5
##  6 0.0258     6
##  7 0.0247     7
##  8 0.0237     8
##  9 0.0228     9
## 10 0.0219    10
## # ... with 21 more rows
```

```
lags.tbl %>%
  ggplot(aes(x = lags, y = MSE)) +
  geom_line()
```

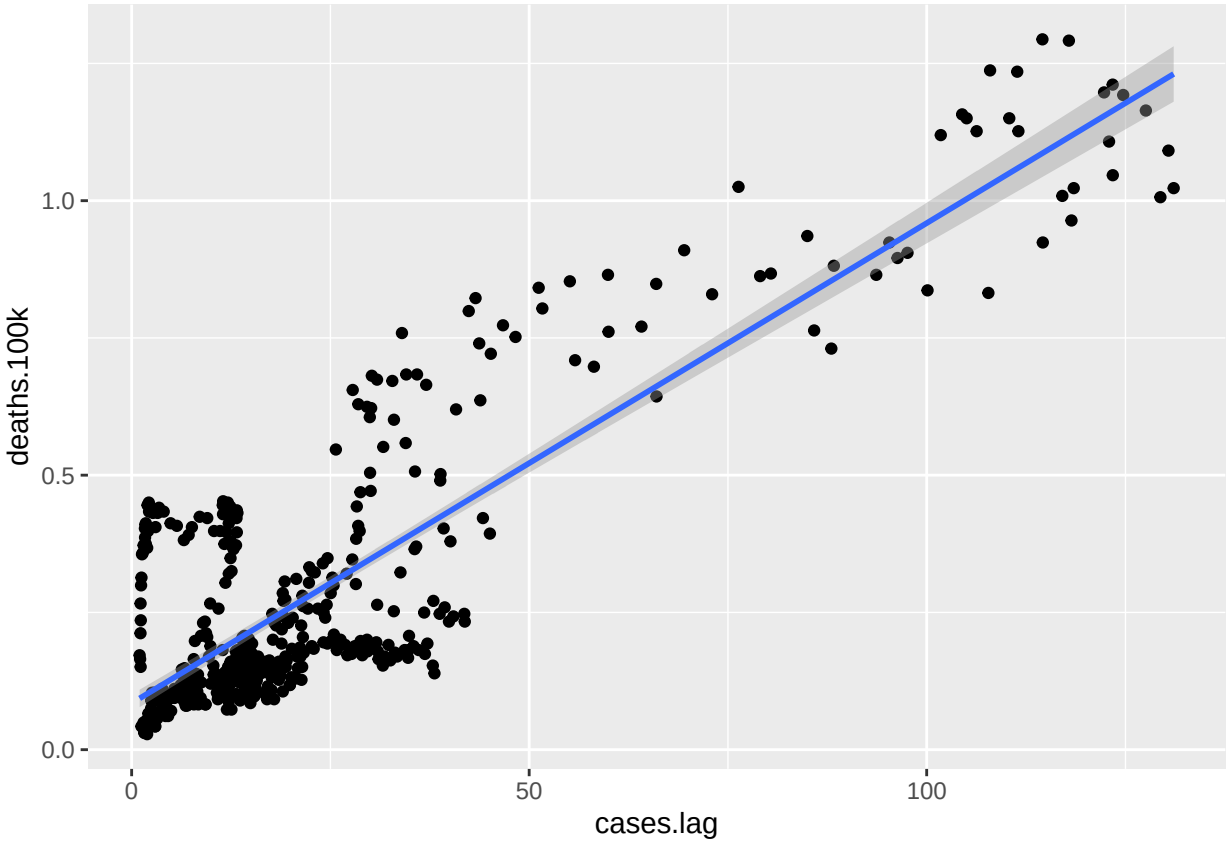



```
best.lag <- lags.tbl %>%
  slice_min(order_by = MSE, with_ties = FALSE)
best.lag
```

```
## # A tibble: 1 x 2
##      MSE  lags
##    <dbl> <int>
## 1 0.0199    17
```

```
# best lag is 17

covid.test.tbl %>%
  mutate(cases.lag = lag(covid.test.tbl$cases.100k, 17)) %>%
  filter(!is.na(cases.lag)) %>%
  ggplot(aes(x = cases.lag, y = deaths.100k)) +
  geom_point() +
  geom_smooth(method = "lm")
```


deaths.100k
1.0
0.5
0.0
0
50
100
cases.lag