Project: Data analysis for New York weather

Kaiwen Zhou 3170104135

2020/7/12

Introduction

The dataset we analyze in this project is 'airquality', which is a dataset in library "datasets". This dataset record the air condition and weather condition of New York in 4 month. The dataset have 6 variables, two of which are month and day, the other 4 are ozone, solar radiation, wind, temperature, seperately. Here is an overview of the dataset:

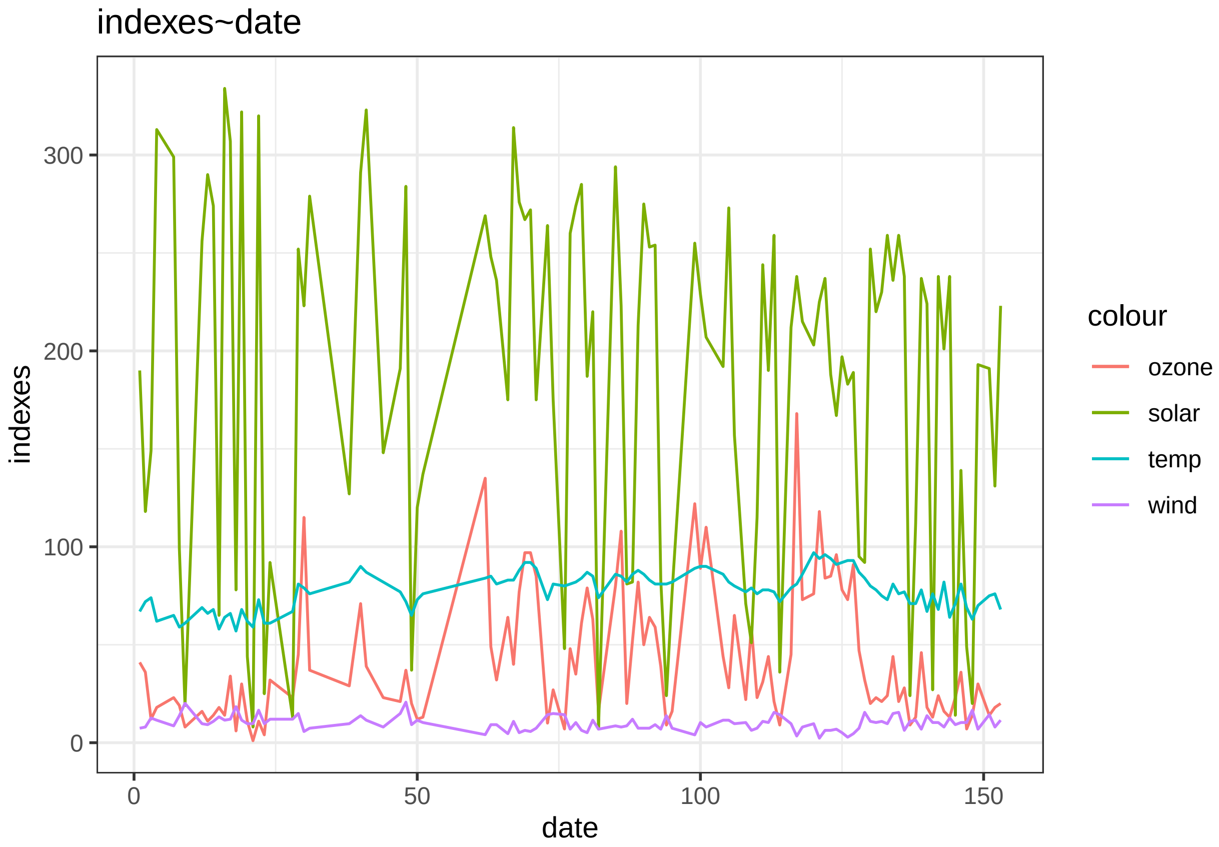
```
##   Ozone Solar.R Wind Temp Month Day
## 1    41     190  7.4   67     5   1
## 2    36     118  8.0   72     5   2
## 3    12     149 12.6   74     5   3
## 4    18     313 11.5   62     5   4
## 5    NA      NA 14.3   56     5   5
## 6    28      NA 14.9   66     5   6
```

Tidy data and data visualization

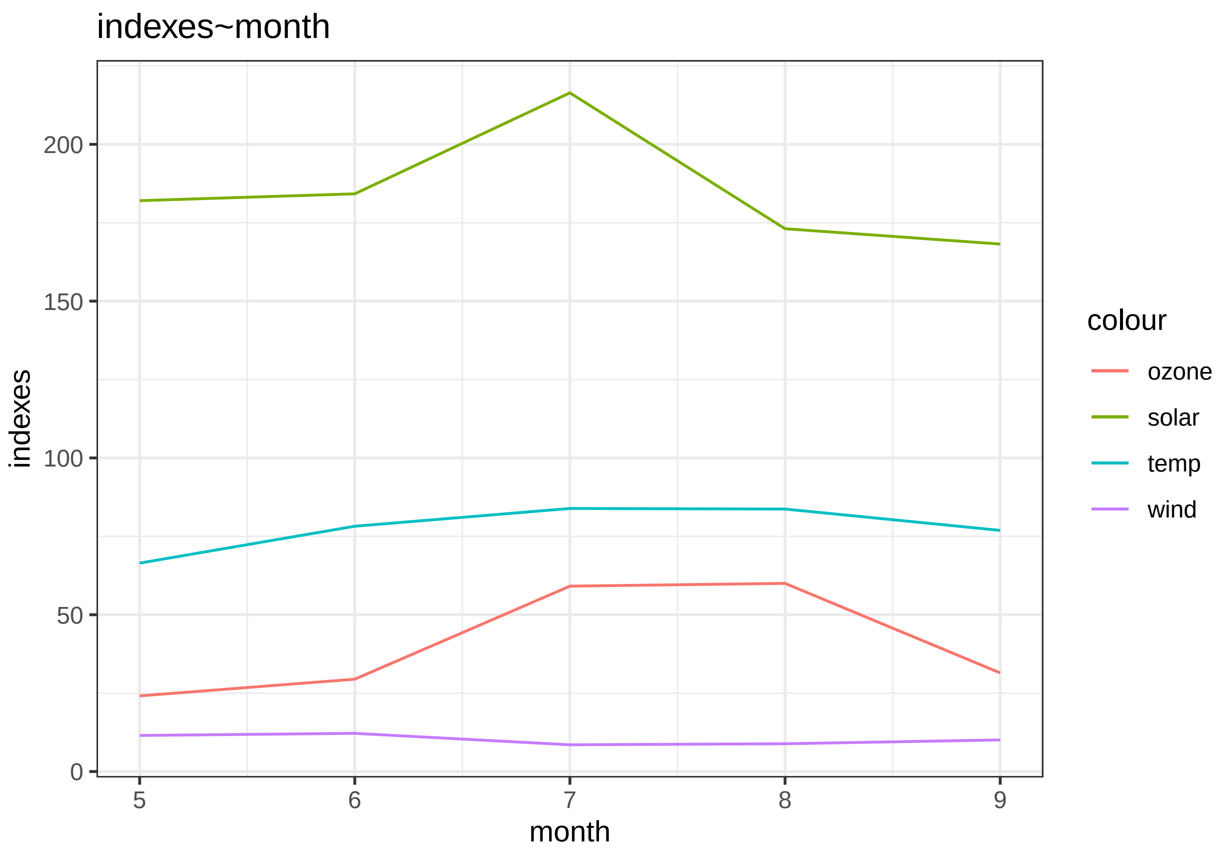
First, to make it more clearly to show the change of indexes in plot, we add a column showing the day. Second, we want to change the data type to dataframe, and remove the obervations that have NAs. Finally, we want to collect the month average of each index.

```
## -- Attaching packages ----------------------------- tidyverse 1.3.0 --

## v ggplot2 3.2.1     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   0.8.3
## v tidyr   1.1.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts -------------------------------- tidyverse_conflicts() --
## x tidyr::extract()   masks magrittr::extract()
## x dplyr::filter()    masks stats::filter()
## x dplyr::lag()       masks stats::lag()
## x purrr::set_names() masks magrittr::set_names()

##   Month ave_ozone ave_solar  ave_wind ave_temp
## 1     5  24.12500  182.0417 11.504167 66.45833
## 2     6  29.44444  184.2222 12.177778 78.22222
## 3     7  59.11538  216.4231  8.523077 83.88462
## 4     8  60.00000  173.0870  8.860870 83.69565
## 5     9  31.44828  168.2069 10.075862 76.89655
```

Then we want to use ggplot to make the features of the data more clearly. First, we want to show the change of 4 indexes with time in one plot.



Then we show the change of 4 indexes with month, to see the difference between different monthes directly.



We can see that the solar radiation raise from May to July, and decrease from July to September. Both temperature and ozone increase from May to August and decrease from August to September. The wind level doesn't change a lot from May to September.

Data analysis using statistic model

In this part, we do research on 3 problems on this dataset using linear regression and hypothetical test.

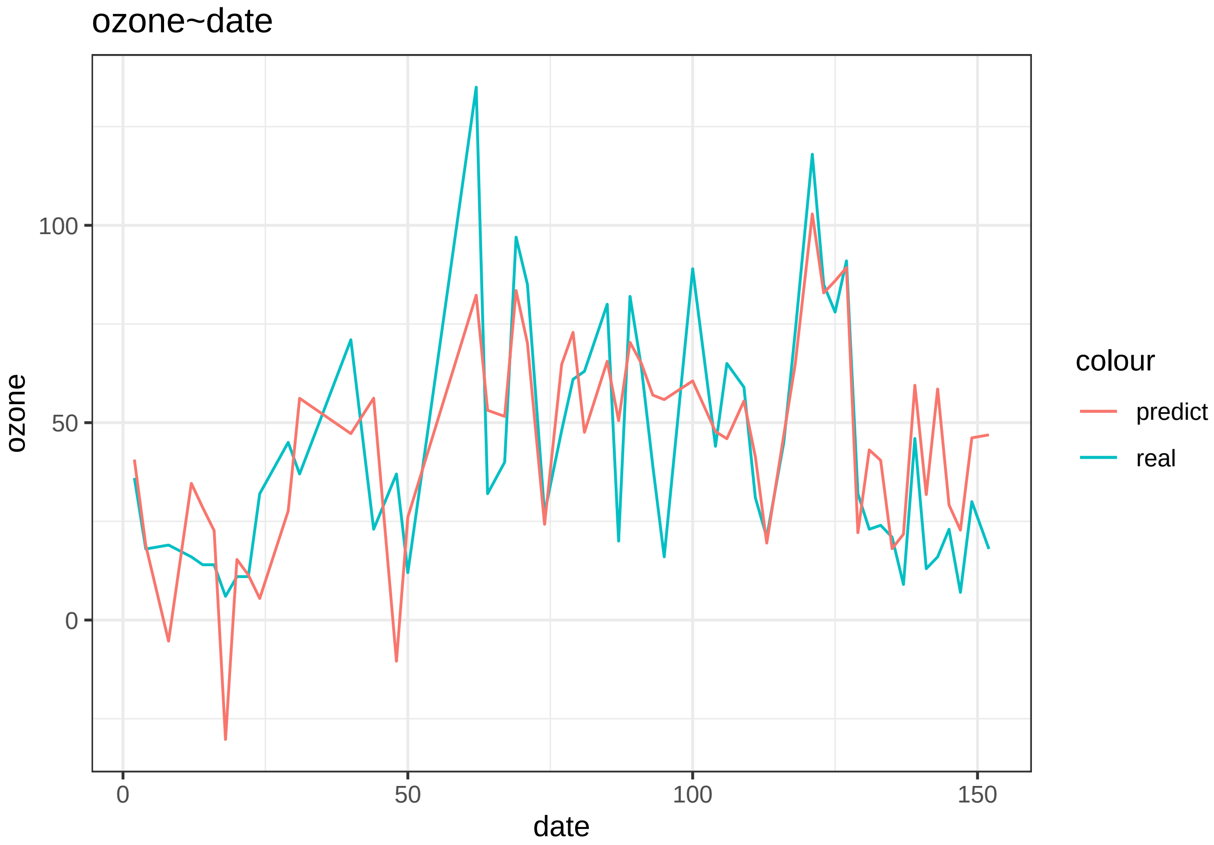
Problem1:The relation between ozone and other variables

```
##
## Call:
## lm(formula = Ozone ~ Temp + Solar.R + Wind, data = air)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.485 -14.219  -3.551  10.097  95.619
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -64.34208   23.05472  -2.791  0.00623 **
## Temp          1.65209    0.25353   6.516 2.42e-09 ***
## Solar.R       0.05982    0.02319   2.580  0.01124 *
## Wind         -3.33359    0.65441  -5.094 1.52e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 21.18 on 107 degrees of freedom
## Multiple R-squared:  0.6059, Adjusted R-squared:  0.5948
## F-statistic: 54.83 on 3 and 107 DF,  p-value: < 2.2e-16
```

We can see that there is a siginificant positive correlation between ozone and temperature, solar radiation, and a siginificant negative correlation between ozone and wind. From professional knowledge we know that high temperature, strong solar radiation and poor atmospheric diffusion conditions are helpful to the produce of ozone, which can help explain the strong correlations.

Then we try to use this model to predict ozone level, and plot the test data and prediction in one plot:

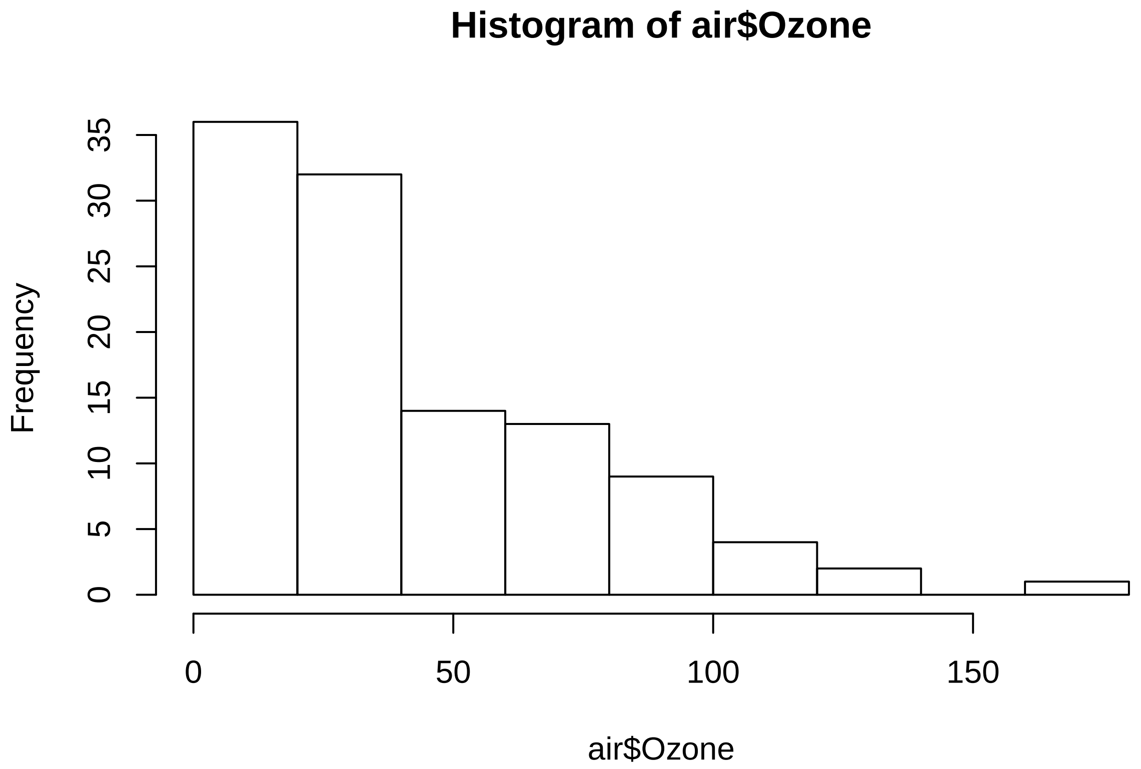


We can see that in most days, the prediction increase as real data increasing, but the exact value may be different. So, we can conclude that three predictors can explain the responsor, but there might be other factors which can influence the ozone level.

Problem2:Fit a distribution of ozone data

In this problem ,we will look into the distribution of the ozone level using nonparametric test.

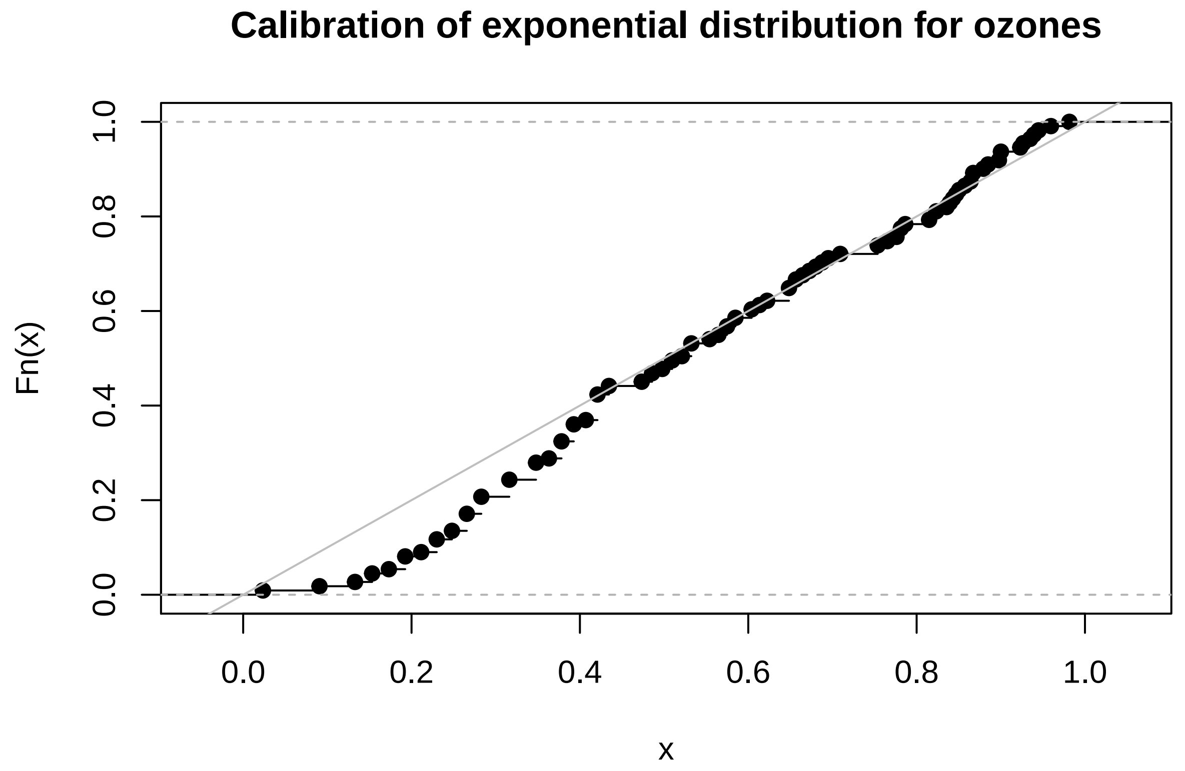
First, we have a look at the histogram of ozone.



We can see that the number of days decrease as the ozone level increase. Then we can try to fit a exponential distribution and a poisson distribution.

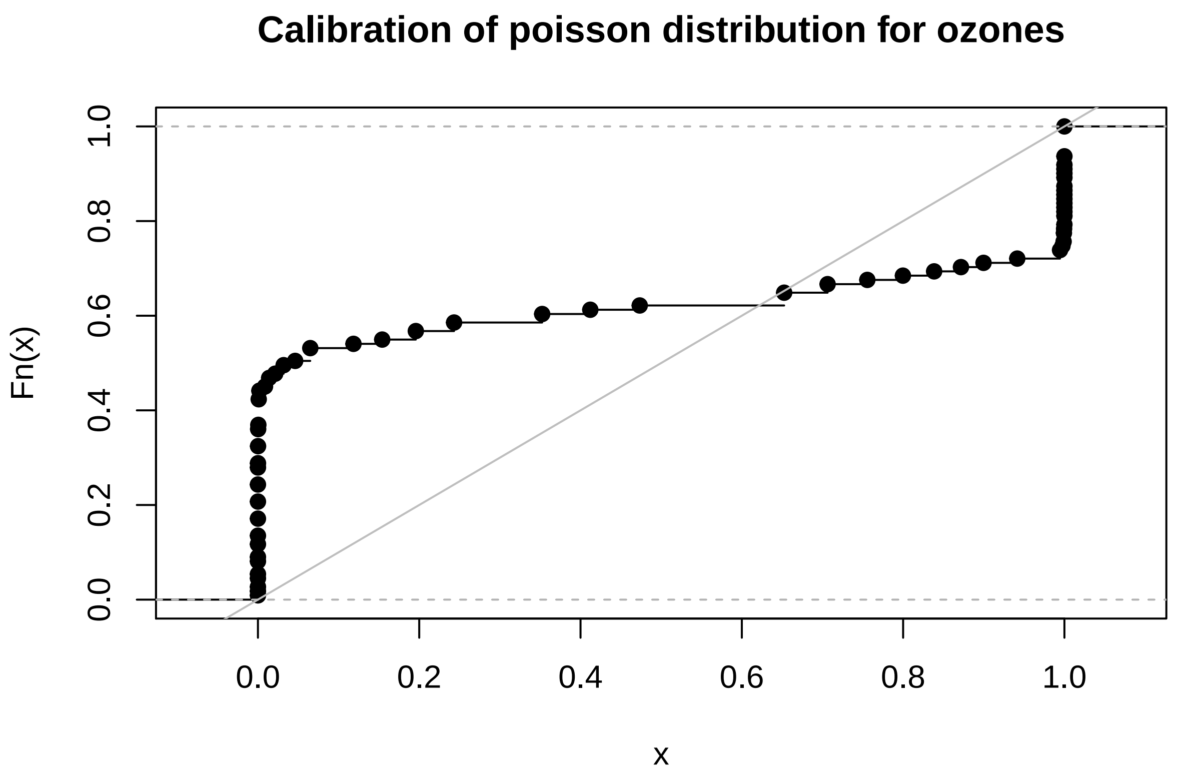
First, we try a exponential distribution, we use kolmogorov-Smirnov test and calibration plot to see if the fitting is right:

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  air$Ozone
## D = 0.13985, p-value = 0.02602
## alternative hypothesis: two-sided
```



Then, we try a poisson distribution:

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  air$Ozone
## D = 0.46663, p-value < 2.2e-16
## alternative hypothesis: two-sided
```



We can see from kolmogorov-Smirnov test and calibration plot that exponential ditribution is more reasonable.

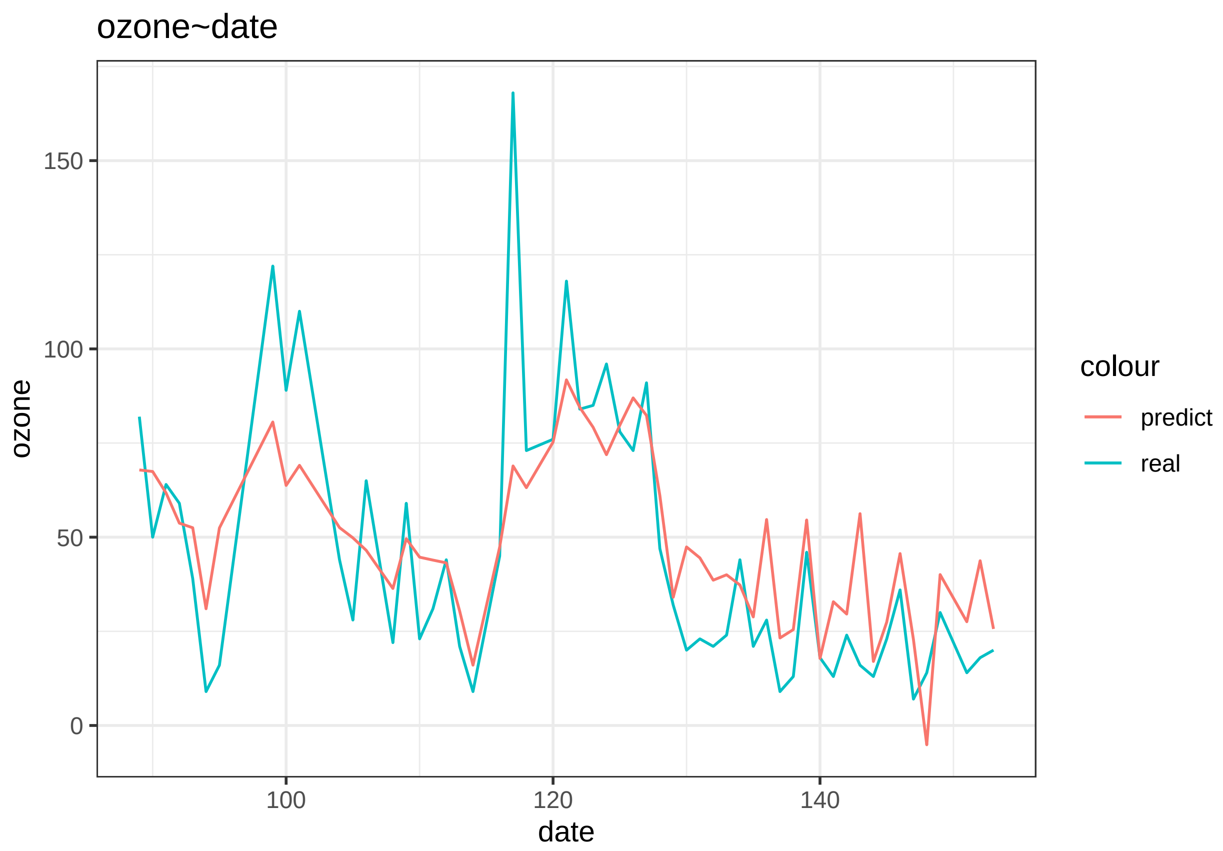
Problem3:The relationship between the quality of prediction and the selection of training set

In this part, we will use linear model based on problem1. and choose different training set to see how the selection of training set influence the quality of prediction. We will use a function to pack all the predction evaluation methods, who takes the training index vector as input, and output a list with mse and a plot. We will choose half of the data as training set.

```
evaluate<-function(Train){
  lm.ozone = lm(Ozone~Temp+Solar.R+Wind, data = air[Train,])
  predict.ozone = predict(lm.ozone, newdata = air[-Train,])

  plot = ggplot(data = air[-Train,]) +
  geom_line(aes(x = date, y = Ozone,colour = "real"))+
  geom_line(aes(x = date, y = predict.ozone,colour = 'predict'))+
  labs(title = "ozone~date",   y = "ozone",   x = "date")+
  theme_bw()
  mse = mean((air$Ozone[-Train]-predict.ozone)^2)
  return(list(mse,plot))
}
```

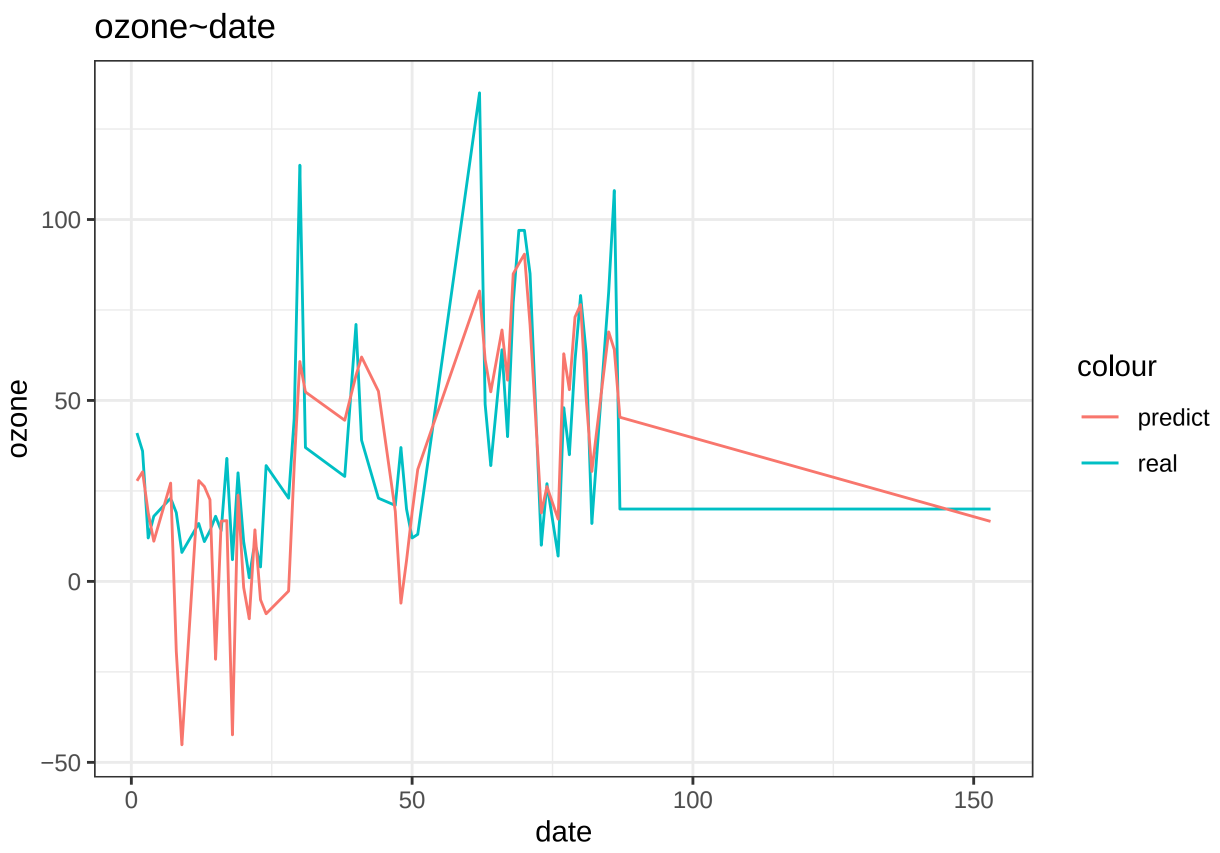
First, we use the first half of the data as training set:

```
## [[1]]
## [1] 482.4545
##
## [[2]]
```



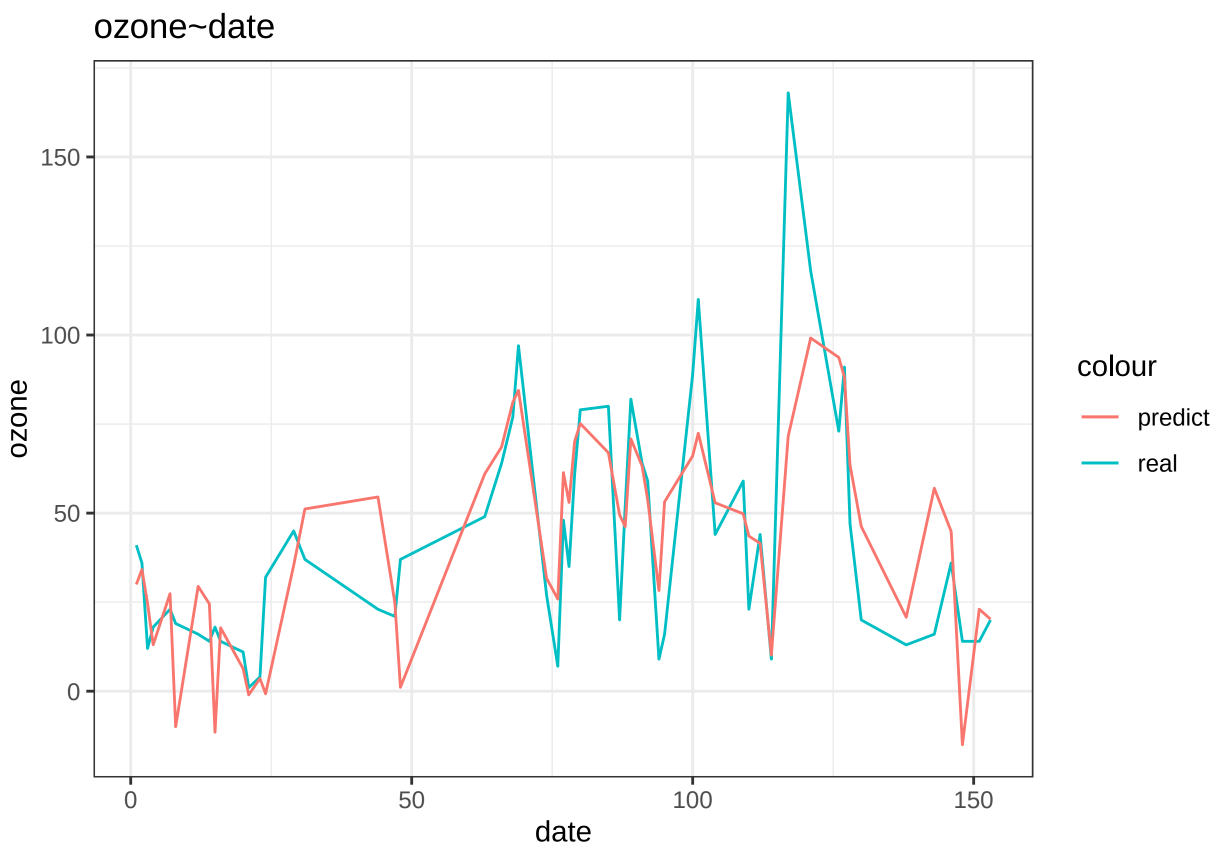
Second, we select the second half of the data:

```
## [1] 505.9083
```



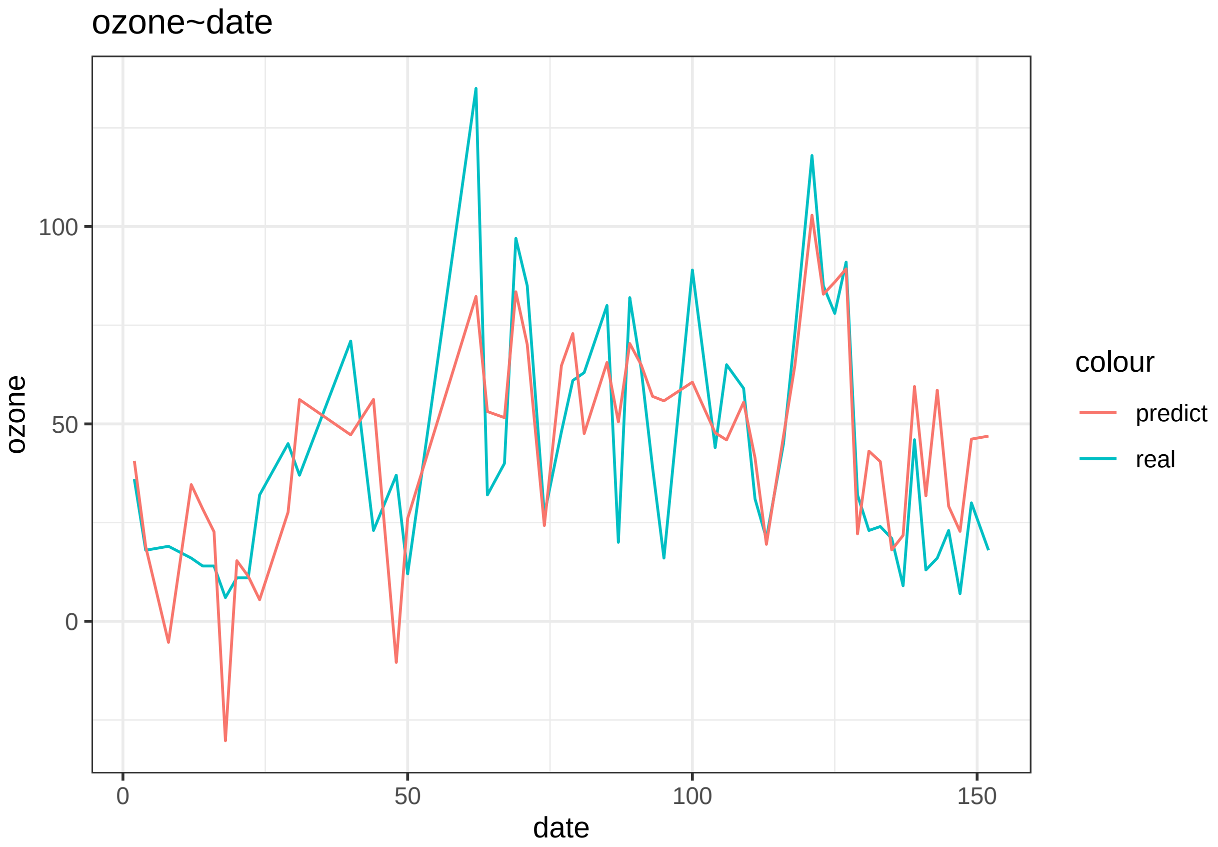
Then we randomly select half of the data:

```
## [1] 475.9802
```



Finally, we take the interval as one to select half of the data:

```
## [1] 400.6373
```



We can see that taking the interval as one to select half of the data have the lowest predicting MSE. Although the differnence between results of different ways of selection is small, we still need to be careful about the way to choose training set. The best way will be random selection or fixed interval selection, which prevent the model from being influenced by the difference of the data caused by the order. Like in this dataset, the data might be correlated with some other variables related with time, and this will cause bias in the model if we select training data unevenly.