STAT-HW1

Sihan Chen

2018/09/01

3.First we need to download the files from 2016 to 2018

```
for i in {2016..2018};do
   curl -o ${i}.csv.gz https://www1.ncdc.noaa.gov/pub/data/ghcn/daily/by_year/${i}.csv.gz
done
```

```
##   % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
##                                  Dload  Upload   Total   Spent    Left  Speed
##
  0     0    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0
  0     0    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0
  0  192M    0  128k    0     0   110k      0  0:29:35  0:00:01  0:29:34  110k
  0  192M    0  979k    0     0   460k      0  0:07:07  0:00:02  0:07:05  460k
  1  192M    1 3159k    0     0  1013k      0  0:03:14  0:00:03  0:03:11 1013k
  3  192M    3 7018k    0     0  1705k      0  0:01:55  0:00:04  0:01:51 1704k
  6  192M    6 12.7M    0     0  2550k      0  0:01:17  0:00:05  0:01:12 2635k
 10  192M   10 20.9M    0     0  3501k      0  0:00:56  0:00:06  0:00:50 4290k
 15  192M   15 29.9M    0     0  4304k      0  0:00:45  0:00:07  0:00:38 5943k
 20  192M   20 39.1M    0     0  4940k      0  0:00:39  0:00:08  0:00:31 7389k
 25  192M   25 49.0M    0     0  5499k      0  0:00:35  0:00:09  0:00:26 8618k
 30  192M   30 58.6M    0     0  5940k      0  0:00:33  0:00:10  0:00:23 9410k
 35  192M   35 68.1M    0     0  6276k      0  0:00:31  0:00:11  0:00:20 9670k
 40  192M   40 78.1M    0     0  6600k      0  0:00:29  0:00:12  0:00:17 9870k
 45  192M   45 87.8M    0     0  6853k      0  0:00:28  0:00:13  0:00:15 9960k
 50  192M   50 97.5M    0     0  7078k      0  0:00:27  0:00:14  0:00:13 9964k
 55  192M   55  107M    0     0  7285k      0  0:00:27  0:00:15  0:00:12  9.7M
 60  192M   60  116M    0     0  7431k      0  0:00:26  0:00:16  0:00:10  9.7M
 66  192M   66  126M    0     0  7594k      0  0:00:25  0:00:17  0:00:08  9.7M
 71  192M   71  136M    0     0  7718k      0  0:00:25  0:00:18  0:00:07 9984k
 76  192M   76  146M    0     0  7845k      0  0:00:25  0:00:19  0:00:06  9.7M
 81  192M   81  156M    0     0  7954k      0  0:00:24  0:00:20  0:00:04 9977k
 86  192M   86  166M    0     0  8053k      0  0:00:24  0:00:21  0:00:03  9.8M
 91  192M   91  175M    0     0  8144k      0  0:00:24  0:00:22  0:00:02  9.7M
 96  192M   96  185M    0     0  8231k      0  0:00:23  0:00:23 --:--:--  9.8M
100  192M  100  192M    0     0  8284k      0  0:00:23  0:00:23 --:--:--  9.8M
##   % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
##                                  Dload  Upload   Total   Spent    Left  Speed
##
  0     0    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0
  0  189M    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0
  0  189M    0  292k    0     0   211k      0  0:15:17  0:00:01  0:15:16  211k
  0  189M    0 1503k    0     0   627k      0  0:05:08  0:00:02  0:05:06  627k
  1  189M    1 3487k    0     0  1025k      0  0:03:08  0:00:03  0:03:05 1025k
  3  189M    3 6167k    0     0  1407k      0  0:02:17  0:00:04  0:02:13 1407k
  5  189M    5 9987k    0     0  1860k      0  0:01:44  0:00:05  0:01:39 2052k
  7  189M    7 15.0M    0     0  2384k      0  0:01:21  0:00:06  0:01:15 2980k
 11  189M   11 22.1M    0     0  3078k      0  0:01:02  0:00:07  0:00:55 4261k
 16  189M   16 31.1M    0     0  3824k      0  0:00:50  0:00:08  0:00:42 5743k
```

```
 21  189M   21 41.1M    0     0  4509k      0  0:00:42  0:00:09  0:00:33 7245k
 26  189M   26 51.0M    0     0  5048k      0  0:00:38  0:00:10  0:00:28 8479k
 32  189M   32 60.9M    0     0  5476k      0  0:00:35  0:00:11  0:00:24 9500k
 37  189M   37 70.4M    0     0  5836k      0  0:00:33  0:00:12  0:00:21 9893k
 42  189M   42 80.4M    0     0  6165k      0  0:00:31  0:00:13  0:00:18  9.8M
 47  189M   47 90.3M    0     0  6437k      0  0:00:30  0:00:14  0:00:16  9.8M
 52  189M   52  100M    0     0  6674k      0  0:00:29  0:00:15  0:00:14  9.8M
 58  189M   58  109M    0     0  6880k      0  0:00:28  0:00:16  0:00:12  9.8M
 63  189M   63  119M    0     0  7059k      0  0:00:27  0:00:17  0:00:10  9.8M
 68  189M   68  129M    0     0  7201k      0  0:00:26  0:00:18  0:00:08 9969k
 73  189M   73  138M    0     0  7339k      0  0:00:26  0:00:19  0:00:07 9933k
 78  189M   78  148M    0     0  7461k      0  0:00:25  0:00:20  0:00:05 9877k
 83  189M   83  157M    0     0  7566k      0  0:00:25  0:00:21  0:00:04 9815k
 88  189M   88  167M    0     0  7651k      0  0:00:25  0:00:22  0:00:03 9706k
 93  189M   93  176M    0     0  7734k      0  0:00:25  0:00:23  0:00:02 9690k
 98  189M   98  185M    0     0  7812k      0  0:00:24  0:00:24 --:--:-- 9644k
100  189M  100  189M    0     0  7844k      0  0:00:24  0:00:24 --:--:-- 9643k
##   % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
##                                 Dload  Upload   Total   Spent    Left  Speed
##
  0     0    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0
  0  109M    0 33117    0     0  45849      0  0:41:45 --:--:--  0:41:45 45804
  0  109M    0  485k    0     0   286k      0  0:06:31  0:00:01  0:06:30  286k
  1  109M    1 2055k    0     0   768k      0  0:02:25  0:00:02  0:02:23  768k
  4  109M    4 4876k    0     0  1333k      0  0:01:24  0:00:03  0:01:21 1332k
  8  109M    8 9094k    0     0  1958k      0  0:00:57  0:00:04  0:00:53 1958k
 13  109M   13 14.8M    0     0  2686k      0  0:00:41  0:00:05  0:00:36 3073k
 20  109M   20 22.5M    0     0  3472k      0  0:00:32  0:00:06  0:00:26 4563k
 29  109M   29 32.0M    0     0  4296k      0  0:00:26  0:00:07  0:00:19 6194k
 38  109M   38 41.7M    0     0  4944k      0  0:00:22  0:00:08  0:00:14 7593k
 47  109M   47 51.6M    0     0  5480k      0  0:00:20  0:00:09  0:00:11 8750k
 55  109M   55 61.0M    0     0  5873k      0  0:00:19  0:00:10  0:00:09 9469k
 64  109M   64 70.4M    0     0  6145k      0  0:00:18  0:00:11  0:00:07 9637k
 72  109M   72 79.1M    0     0  6403k      0  0:00:17  0:00:12  0:00:05 9619k
 81  109M   81 89.0M    0     0  6679k      0  0:00:16  0:00:13  0:00:03 9678k
 90  109M   90 98.8M    0     0  6912k      0  0:00:16  0:00:14  0:00:02 9673k
 99  109M   99  108M    0     0  7106k      0  0:00:15  0:00:15 --:--:-- 9732k
100  109M  100  109M    0     0  7123k      0  0:00:15  0:00:15 --:--:-- 9989k
```

Then we need to unzip the files

```
gunzip *.csv.gz
```

(a).Now we need to use wc -l to count the observations in each year

```
for i in {2016..2018};do
   echo "${i} has$(wc -l < ${i}.csv) observations."
done
```

```
## 2016 has 35384539 observations.
## 2017 has 34748555 observations.
## 2018 has 20229121 observations.
```

(b).First we still need to download ghcnd-stations.txt

```
curl -o ghcnd-stations.txt https://www1.ncdc.noaa.gov/pub/data/ghcn/daily/ghcnd-stations.txt
```

```
##   % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
##                                  Dload  Upload   Total   Spent    Left  Speed
##
  0     0    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0
  0     0    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0
  2 8959k    2  209k    0     0   181k      0  0:00:49  0:00:01  0:00:48  181k
 16 8959k   16 1521k    0     0   707k      0  0:00:12  0:00:02  0:00:10  707k
 43 8959k   43 3873k    0     0  1229k      0  0:00:07  0:00:03  0:00:04 1229k
 79 8959k   79 7123k    0     0  1716k      0  0:00:05  0:00:04  0:00:01 1716k
100 8959k  100 8959k    0     0  1943k      0  0:00:04  0:00:04 --:--:-- 2033k
```

Then we need to get the code from the ghcnd-stations.txt, after that we can grep the lines we need from the three files and put them into a new file named TMAX.txt.

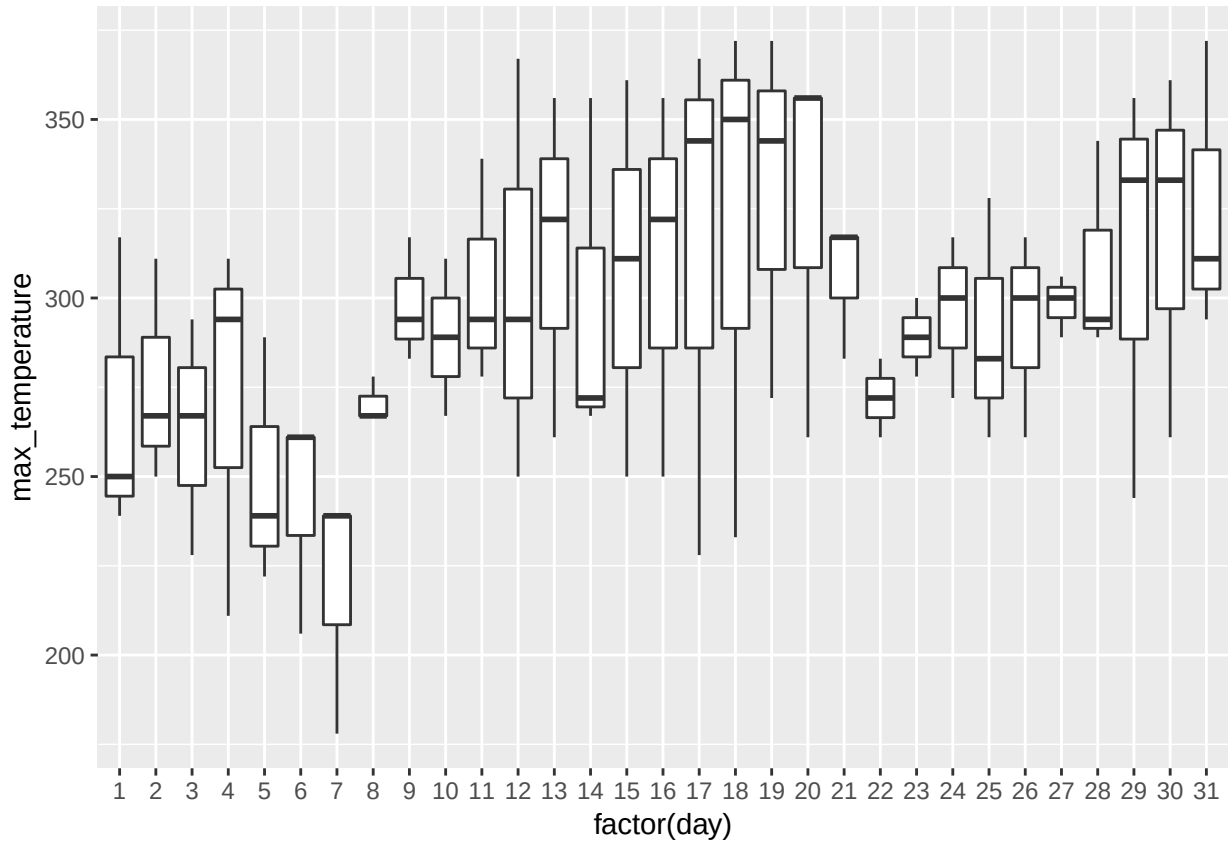
```
ID=$(grep 'DEATH VALLEY' ghcnd-stations.txt | cut -d' ' -f1)
echo $ID
for i in {2016..2018};do
   grep $ID ${i}.csv | grep TMAX | grep ${i}03 >> TMAX.txt
done
```

```
## USC00042319
```

After that we can grep the lines we need from the three files and put them into a new file named TMAX.txt

(c).I read the TMAX.txt file into data, and put the max temperature and day value into data_m, and put the same day in March from different years together and draw the side by side boxplot

```
library('ggplot2')
data=read.table("TMAX.txt",sep=",")
data_m=data[c(2,4)]
data_m[1]=data_m[1]%%1000
data_m=data_m[which(data_m[1]>=300) && which(data_m[1]<400),]
data_m[1]=data_m[1]%%100
data_m=data_m[order(data_m[1]),]
names(data_m)=c("day","max_temperature")
p<-ggplot(data=data_m, aes(x=factor(day),y=max_temperature))+geom_boxplot()
p
```



(d).First we need to define a function

```
function get_weather(){
    if [ "$1" == "-h" ];then
        echo "get_weather usage:get_weather location weather year month filename"
    else
        if [ "$#" != "5" ];then
            echo "Wrong arguments number"
        else
            for i in ${3};do
                curl -o ${i}.csv.gz https://www1.ncdc.noaa.gov/pub/data/ghcn/daily/by_year/${i}.csv.gz
                if [ ! -e ${i}.csv ];then
                    rm -f ${i}.csv
                fi
                gunzip ${i}.csv.gz
            done
            curl -o ghcnd-stations.txt https://www1.ncdc.noaa.gov/pub/data/ghcn/daily/ghcnd-stations.tx
            code=$(grep -h "${1}" ghcnd-stations.txt | cut -d' ' -f1)
            if [ -n $code ];then
                if [ -e $5 ];then
                    rm -f $5
                fi
                for i in ${3};do
                    grep $code ${i}.csv | grep -h ${2} | grep -h "${i}${4}" >> $5
                done
                if [ -s $5 ];then
                    echo "finished"
```

```
                else
                    echo "None fits"
                fi
            else
                echo 'Wrong location'
            fi
            for i in ${3};do
                rm -f ${i}.csv
                rm -f ${i}.csv.gz
            done
            rm -f ghcnd-stations.txt
        fi
    fi
}
get_weather 'DEATH VALLEY' TMAX 2016 03 TMAX1.txt
```

```
##   % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
##                                  Dload  Upload   Total   Spent    Left  Speed
##
  0     0    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0
  0  192M    0 78148    0     0  88087      0  0:38:07 --:--:--  0:38:07 88004
  0  192M    0  568k    0     0   303k      0  0:10:48  0:00:01  0:10:47  303k
  0  192M    0 1904k    0     0   666k      0  0:04:55  0:00:02  0:04:53  666k
  2  192M    2 4334k    0     0  1124k      0  0:02:55  0:00:03  0:02:52 1123k
  3  192M    3 7443k    0     0  1531k      0  0:02:08  0:00:04  0:02:04 1531k
  6  192M    6 11.7M    0     0  2052k      0  0:01:35  0:00:05  0:01:30 2403k
  9  192M    9 17.9M    0     0  2688k      0  0:01:13  0:00:06  0:01:07 3585k
 13  192M   13 25.7M    0     0  3355k      0  0:00:58  0:00:07  0:00:51 4893k
 18  192M   18 35.1M    0     0  4059k      0  0:00:48  0:00:08  0:00:40 6323k
 23  192M   23 44.4M    0     0  4614k      0  0:00:42  0:00:09  0:00:33 7615k
 28  192M   28 53.9M    0     0  5093k      0  0:00:38  0:00:10  0:00:28 8661k
 32  192M   32 63.3M    0     0  5473k      0  0:00:35  0:00:11  0:00:24 9292k
 37  192M   37 72.6M    0     0  5788k      0  0:00:33  0:00:12  0:00:21 9609k
 42  192M   42 80.7M    0     0  5958k      0  0:00:33  0:00:13  0:00:20 9302k
 43  192M   43 83.4M    0     0  5621k      0  0:00:35  0:00:15  0:00:20 7478k
 44  192M   44 84.8M    0     0  5465k      0  0:00:35  0:00:15  0:00:20 6269k
 45  192M   45 86.6M    0     0  5235k      0  0:00:37  0:00:16  0:00:21 4682k
 45  192M   45 88.3M    0     0  5061k      0  0:00:38  0:00:17  0:00:21 3200k
 47  192M   47 91.3M    0     0  4962k      0  0:00:39  0:00:18  0:00:21 2181k
 49  192M   49 94.5M    0     0  4874k      0  0:00:40  0:00:19  0:00:21 2441k
 50  192M   50 97.6M    0     0  4792k      0  0:00:41  0:00:20  0:00:21 2636k
 52  192M   52  100M    0     0  4728k      0  0:00:41  0:00:21  0:00:20 2977k
 54  192M   54  103M    0     0  4654k      0  0:00:42  0:00:22  0:00:20 3195k
 55  192M   55  107M    0     0  4597k      0  0:00:42  0:00:23  0:00:19 3221k
 57  192M   57  110M    0     0  4553k      0  0:00:43  0:00:24  0:00:19 3277k
 59  192M   59  114M    0     0  4523k      0  0:00:43  0:00:25  0:00:18 3405k
 61  192M   61  118M    0     0  4518k      0  0:00:43  0:00:26  0:00:17 3601k
 64  192M   64  123M    0     0  4528k      0  0:00:43  0:00:27  0:00:16 3951k
 66  192M   66  128M    0     0  4568k      0  0:00:43  0:00:28  0:00:15 4432k
 70  192M   70  135M    0     0  4641k      0  0:00:42  0:00:29  0:00:13 5077k
 74  192M   74  143M    0     0  4751k      0  0:00:41  0:00:30  0:00:11 5931k
 79  192M   79  152M    0     0  4896k      0  0:00:40  0:00:31  0:00:09 6926k
 83  192M   83  161M    0     0  5025k      0  0:00:39  0:00:32  0:00:07 7792k
 88  192M   88  170M    0     0  5156k      0  0:00:38  0:00:33  0:00:05 8545k
```

```
 93  192M   93  179M    0     0  5272k      0  0:00:37  0:00:34  0:00:03 9043k
 98  192M   98  188M    0     0  5395k      0  0:00:36  0:00:35  0:00:01 9371k
100  192M  100  192M    0     0  5432k      0  0:00:36  0:00:36 --:--:-- 9338k
## gunzip: 2016.csv already exists -- skipping
##   % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
##                                  Dload  Upload   Total   Spent    Left  Speed
##
  0     0    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0
  0 8959k    0 31751    0     0  47597      0  0:03:12 --:--:--  0:03:12 47531
  5 8959k    5  499k    0     0   305k      0  0:00:29  0:00:01  0:00:28  305k
 18 8959k   18 1632k    0     0   620k      0  0:00:14  0:00:02  0:00:12  620k
 39 8959k   39 3577k    0     0   990k      0  0:00:09  0:00:03  0:00:06  990k
 77 8959k   77 6929k    0     0  1498k      0  0:00:05  0:00:04  0:00:01 1498k
100 8959k  100 8959k    0     0  1762k      0  0:00:05  0:00:05 --:--:-- 2021k
## finished
```

4. First we need to extract the original html code from the website

```
curl https://www1.ncdc.noaa.gov/pub/data/ghcn/daily/ -o daily
```

```
##   % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
##                                  Dload  Upload   Total   Spent    Left  Speed
##
  0     0    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0
  0     0    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0
100  6068  100  6068    0     0  13119      0 --:--:-- --:--:-- --:--:-- 13105
```

Then we need to extract all the txt files from the html. We find that each txt files' names is in a pair of quotes, so we extract all the data in the quotes and then extract those contains .txt. Finally we use curl to download these files from the website.

```
files=$(egrep -o '".*?"' daily | grep ".txt" | tr -d '"')
for i in $files;do
    echo "Now we are downloading ${i}."
    curl -o ${i} https://www1.ncdc.noaa.gov/pub/data/ghcn/daily/${i}
done
rm -f daily
```

```
## Now we are downloading ghcnd-countries.txt.
##   % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
##                                  Dload  Upload   Total   Spent    Left  Speed
##
  0     0    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0
100  3670  100  3670    0     0   8127      0 --:--:-- --:--:-- --:--:--  8137
## Now we are downloading ghcnd-inventory.txt.
##   % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
##                                  Dload  Upload   Total   Spent    Left  Speed
##
  0     0    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0
  0     0    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0
  0 26.6M    0  190k    0     0   162k      0  0:02:47  0:00:01  0:02:46  162k
  4 26.6M    4 1151k    0     0   515k      0  0:00:52  0:00:02  0:00:50  515k
 10 26.6M   10 2956k    0     0   923k      0  0:00:29  0:00:03  0:00:26  923k
 21 26.6M   21 5948k    0     0  1421k      0  0:00:19  0:00:04  0:00:15 1421k
 38 26.6M   38 10.1M    0     0  2014k      0  0:00:13  0:00:05  0:00:08 2098k
 60 26.6M   60 16.2M    0     0  2697k      0  0:00:10  0:00:06  0:00:04 3291k
```

```
 93 26.6M   93 24.8M    0     0  3544k      0  0:00:07  0:00:07 --:--:-- 4916k
100 26.6M  100 26.6M    0     0  3711k      0  0:00:07  0:00:07 --:--:-- 5860k
## Now we are downloading ghcnd-states.txt.
##   % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
##                                 Dload  Upload   Total   Spent    Left  Speed
##
  0     0    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0
100  1086  100  1086    0     0   2364      0 --:--:-- --:--:-- --:--:--  2366
## Now we are downloading ghcnd-stations.txt.
##   % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
##                                 Dload  Upload   Total   Spent    Left  Speed
##
  0     0    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0
  0     0    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0
  2 8959k    2  266k    0     0   196k      0  0:00:45  0:00:01  0:00:44  196k
 15 8959k   15 1399k    0     0   592k      0  0:00:15  0:00:02  0:00:13  592k
 40 8959k   40 3633k    0     0  1084k      0  0:00:08  0:00:03  0:00:05 1083k
 81 8959k   81 7337k    0     0  1685k      0  0:00:05  0:00:04  0:00:01 1685k
100 8959k  100 8959k    0     0  1919k      0  0:00:04  0:00:04 --:--:-- 2092k
## Now we are downloading ghcnd-version.txt.
##   % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
##                                 Dload  Upload   Total   Spent    Left  Speed
##
  0     0    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0
100   270  100   270    0     0    566      0 --:--:-- --:--:-- --:--:--   567
## Now we are downloading mingle-list.txt.
##   % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
##                                 Dload  Upload   Total   Spent    Left  Speed
##
  0     0    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0
  0     0    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0
  6 3707k    6  226k    0     0   197k      0  0:00:18  0:00:01  0:00:17  196k
 36 3707k   36 1351k    0     0   615k      0  0:00:06  0:00:02  0:00:04  615k
 96 3707k   96 3585k    0     0  1128k      0  0:00:03  0:00:03 --:--:-- 1127k
100 3707k  100 3707k    0     0  1162k      0  0:00:03  0:00:03 --:--:-- 1161k
## Now we are downloading readme.txt.
##   % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
##                                 Dload  Upload   Total   Spent    Left  Speed
##
  0     0    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0
100 26498  100 26498    0     0  47023      0 --:--:-- --:--:-- --:--:-- 47065
## Now we are downloading status.txt.
##   % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
##                                 Dload  Upload   Total   Spent    Left  Speed
##
  0     0    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0
  0 31860    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0
100 31860  100 31860    0     0  49862      0 --:--:-- --:--:-- --:--:-- 49859
```