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  192M    0 33116    0     0  33839      0  1:39:15 --:--:--  1:39:15 33826
  0  192M    0  540k    0     0   276k      0  0:11:52  0:00:01  0:11:51  276k
  1  192M    1 2626k    0     0   889k      0  0:03:41  0:00:02  0:03:39  888k
  3  192M    3 6430k    0     0  1633k      0  0:02:00  0:00:03  0:01:57 1633k
  6  192M    6 11.9M    0     0  2482k      0  0:01:19  0:00:04  0:01:15 2481k
 10  192M   10 19.6M    0     0  3383k      0  0:00:58  0:00:05  0:00:53 4044k
 14  192M   14 28.3M    0     0  4182k      0  0:00:47  0:00:06  0:00:41 5715k
 19  192M   19 36.9M    0     0  4767k      0  0:00:41  0:00:07  0:00:34 7065k
 23  192M   23 45.7M    0     0  5247k      0  0:00:37  0:00:08  0:00:29 8094k
 28  192M   28 54.8M    0     0  5648k      0  0:00:34  0:00:09  0:00:25 8774k
 33  192M   33 64.6M    0     0  6053k      0  0:00:32  0:00:10  0:00:22 9222k
 38  192M   38 74.6M    0     0  6406k      0  0:00:30  0:00:11  0:00:19 9492k
 43  192M   43 84.4M    0     0  6681k      0  0:00:29  0:00:12  0:00:17 9716k
 49  192M   49 94.2M    0     0  6928k      0  0:00:28  0:00:13  0:00:15 9932k
 54  192M   54  103M    0     0  7117k      0  0:00:27  0:00:14  0:00:13  9.8M
 59  192M   59  113M    0     0  7310k      0  0:00:26  0:00:15  0:00:11  9.8M
 64  192M   64  123M    0     0  7468k      0  0:00:26  0:00:16  0:00:10  9.7M
 69  192M   69  133M    0     0  7619k      0  0:00:25  0:00:17  0:00:08  9.8M
 74  192M   74  142M    0     0  7731k      0  0:00:25  0:00:18  0:00:07 9971k
 79  192M   79  153M    0     0  7860k      0  0:00:25  0:00:19  0:00:06  9.8M
 84  192M   84  163M    0     0  7982k      0  0:00:24  0:00:20  0:00:04  9.8M
 90  192M   90  173M    0     0  8082k      0  0:00:24  0:00:21  0:00:03  9.9M
 95  192M   95  182M    0     0  8167k      0  0:00:24  0:00:22  0:00:02  9.8M
100  192M  100  192M    0     0  8248k      0  0:00:23  0:00:23 --:--:--  9.9M
##   % 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  189M    0  151k    0     0   133k      0  0:24:15  0:00:01  0:24:14  133k
  0  189M    0 1370k    0     0   651k      0  0:04:57  0:00:02  0:04:55  651k
  2  189M    2 4284k    0     0  1390k      0  0:02:19  0:00:03  0:02:16 1390k
  4  189M    4 9362k    0     0  2304k      0  0:01:24  0:00:04  0:01:20 2304k
  8  189M    8 16.3M    0     0  3312k      0  0:00:58  0:00:05  0:00:53 3382k
 13  189M   13 25.9M    0     0  4389k      0  0:00:44  0:00:06  0:00:38 5373k
 18  189M   18 35.7M    0     0  5179k      0  0:00:37  0:00:07  0:00:30 7098k
 23  189M   23 44.0M    0     0  5594k      0  0:00:34  0:00:08  0:00:26 8195k
 27  189M   27 52.0M    0     0  5880k      0  0:00:32  0:00:09  0:00:23 8785k
```

```
 31  189M   31 59.9M    0     0  6100k      0  0:00:31  0:00:10  0:00:21 8922k
 35  189M   35 67.8M    0     0  6276k      0  0:00:30  0:00:11  0:00:19 8561k
 40  189M   40 75.9M    0     0  6449k      0  0:00:30  0:00:12  0:00:18 8242k
 44  189M   44 84.2M    0     0  6600k      0  0:00:29  0:00:13  0:00:16 8221k
 48  189M   48 91.2M    0     0  6643k      0  0:00:29  0:00:14  0:00:15 8026k
 52  189M   52 99.1M    0     0  6732k      0  0:00:28  0:00:15  0:00:13 8001k
 56  189M   56  107M    0     0  6832k      0  0:00:28  0:00:16  0:00:12 8065k
 61  189M   61  116M    0     0  7010k      0  0:00:27  0:00:17  0:00:10 8365k
 66  189M   66  126M    0     0  7147k      0  0:00:27  0:00:18  0:00:09 8577k
 71  189M   71  134M    0     0  7211k      0  0:00:26  0:00:19  0:00:07 8804k
 75  189M   75  142M    0     0  7264k      0  0:00:26  0:00:20  0:00:06 8872k
 79  189M   79  150M    0     0  7315k      0  0:00:26  0:00:21  0:00:05 8862k
 83  189M   83  158M    0     0  7372k      0  0:00:26  0:00:22  0:00:04 8607k
 89  189M   89  168M    0     0  7488k      0  0:00:25  0:00:23  0:00:02 8718k
 94  189M   94  178M    0     0  7595k      0  0:00:25  0:00:24  0:00:01 9065k
 99  189M   99  188M    0     0  7690k      0  0:00:25  0:00:25 --:--:-- 9402k
100  189M  100  189M    0     0  7701k      0  0:00:25  0:00:25 --:--:-- 9691k
##   % 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  107k    0     0   113k      0  0:16:22 --:--:--  0:16:22  113k
  1  109M    1 1131k    0     0   592k      0  0:03:08  0:00:01  0:03:07  592k
  3  109M    3 3560k    0     0  1224k      0  0:01:31  0:00:02  0:01:29 1224k
  6  109M    6 7271k    0     0  1870k      0  0:00:59  0:00:03  0:00:56 1870k
 11  109M   11 12.3M    0     0  2583k      0  0:00:43  0:00:04  0:00:39 2583k
 17  109M   17 19.6M    0     0  3412k      0  0:00:32  0:00:05  0:00:27 4043k
 26  109M   26 28.7M    0     0  4283k      0  0:00:26  0:00:06  0:00:20 5699k
 35  109M   35 38.6M    0     0  5014k      0  0:00:22  0:00:07  0:00:15 7227k
 44  109M   44 48.4M    0     0  5586k      0  0:00:20  0:00:08  0:00:12 8475k
 53  109M   53 58.4M    0     0  6050k      0  0:00:18  0:00:09  0:00:09 9439k
 62  109M   62 67.9M    0     0  6396k      0  0:00:17  0:00:10  0:00:07 9909k
 71  109M   71 78.0M    0     0  6725k      0  0:00:16  0:00:11  0:00:05  9.8M
 80  109M   80 88.1M    0     0  7001k      0  0:00:15  0:00:12  0:00:03  9.8M
 89  109M   89 97.5M    0     0  7189k      0  0:00:15  0:00:13  0:00:02  9.7M
 96  109M   96  105M    0     0  7280k      0  0:00:15  0:00:14  0:00:01 9708k
100  109M  100  109M    0     0  7297k      0  0:00:15  0:00:15 --:--:-- 9508k
```

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 20159048 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 8959k    0 54159    0     0  68081      0  0:02:14 --:--:--  0:02:14 68038
  4 8959k    4  404k    0     0   224k      0  0:00:39  0:00:01  0:00:38  224k
 15 8959k   15 1412k    0     0   516k      0  0:00:17  0:00:02  0:00:15  516k
 40 8959k   40 3615k    0     0   970k      0  0:00:09  0:00:03  0:00:06  970k
 81 8959k   81 7271k    0     0  1537k      0  0:00:05  0:00:04  0:00:01 1537k
100 8959k  100 8959k    0     0  1754k      0  0:00:05  0:00:05 --:--:-- 2066k
```

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.

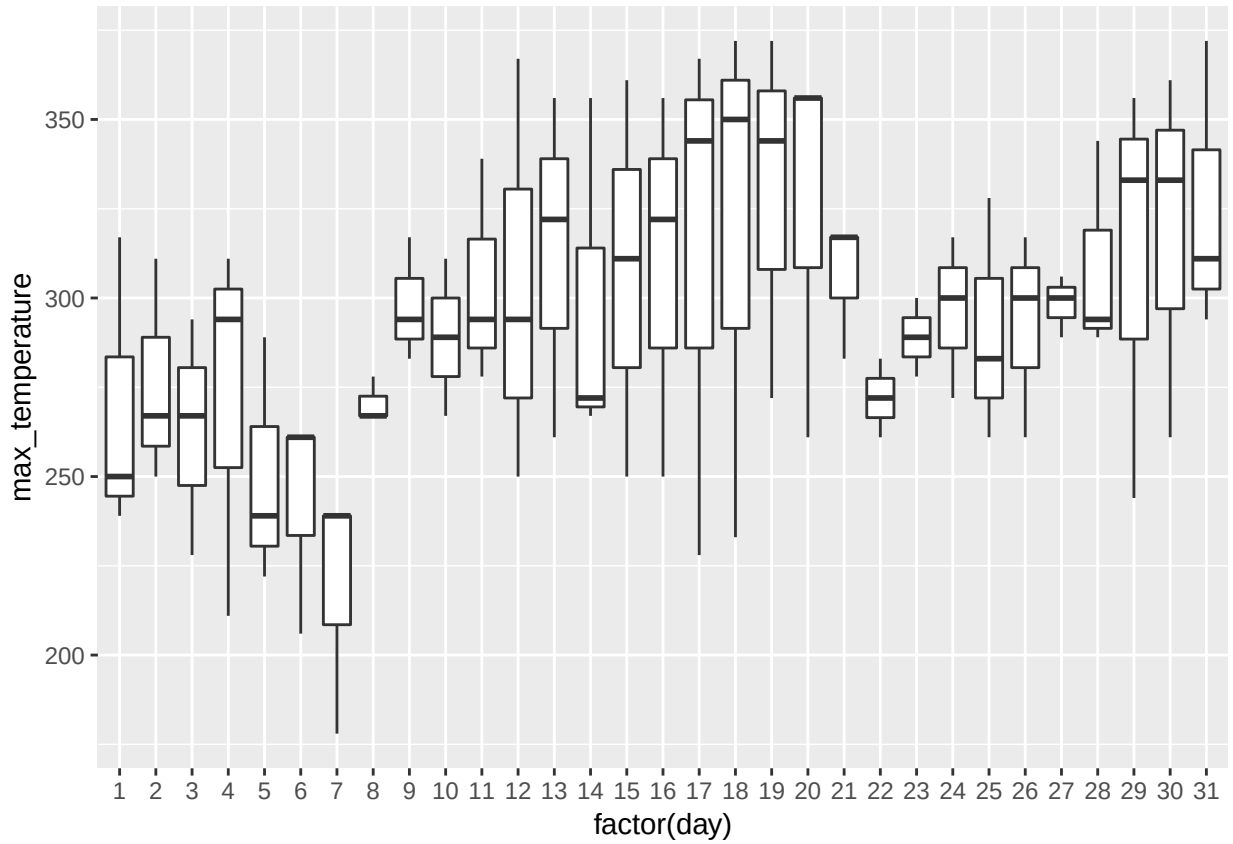
```
ID=$(grep 'DEATH VALLEY' ghcnd-stations.txt | cut -d' ' -f1)
echo $ID
for i in {2016..2018};do
   grep $ID ${i}.csv | grep TMAX | grep ${i}03 >> TMAX.txt
done
```

```
## 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 ${2} | grep -h "${3}${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 62148    0     0  74032      0  0:45:22 --:--:--  0:45:22 73985
  0  192M    0  568k    0     0   305k      0  0:10:45  0:00:01  0:10:44  304k
  1  192M    1 2162k    0     0   756k      0  0:04:20  0:00:02  0:04:18  756k
  2  192M    2 4912k    0     0  1274k      0  0:02:34  0:00:03  0:02:31 1274k
  4  192M    4 9224k    0     0  1904k      0  0:01:43  0:00:04  0:01:39 1903k
  7  192M    7 15.3M    0     0  2681k      0  0:01:13  0:00:05  0:01:08 3118k
 12  192M   12 23.2M    0     0  3473k      0  0:00:56  0:00:06  0:00:50 4657k
 16  192M   16 31.1M    0     0  4068k      0  0:00:48  0:00:07  0:00:41 5968k
 20  192M   20 38.6M    0     0  4484k      0  0:00:43  0:00:08  0:00:35 6967k
 24  192M   24 46.3M    0     0  4821k      0  0:00:40  0:00:09  0:00:31 7650k
 28  192M   28 54.3M    0     0  5133k      0  0:00:38  0:00:10  0:00:28 8007k
 32  192M   32 62.3M    0     0  5387k      0  0:00:36  0:00:11  0:00:25 8007k
 35  192M   35 68.9M    0     0  5501k      0  0:00:35  0:00:12  0:00:23 7751k
 39  192M   39 76.3M    0     0  5646k      0  0:00:34  0:00:13  0:00:21 7698k
 43  192M   43 83.1M    0     0  5739k      0  0:00:34  0:00:14  0:00:20 7546k
 47  192M   47 90.7M    0     0  5871k      0  0:00:33  0:00:15  0:00:18 7473k
 51  192M   51 98.6M    0     0  6003k      0  0:00:32  0:00:16  0:00:16 7464k
 55  192M   55  106M    0     0  6115k      0  0:00:32  0:00:17  0:00:15 7689k
 59  192M   59  114M    0     0  6224k      0  0:00:31  0:00:18  0:00:13 7823k
 63  192M   63  122M    0     0  6307k      0  0:00:31  0:00:19  0:00:12 7990k
 67  192M   67  130M    0     0  6409k      0  0:00:30  0:00:20  0:00:10 8114k
 71  192M   71  138M    0     0  6485k      0  0:00:30  0:00:21  0:00:09 8109k
 76  192M   76  146M    0     0  6558k      0  0:00:30  0:00:22  0:00:08 8139k
 81  192M   81  155M    0     0  6699k      0  0:00:29  0:00:23  0:00:06 8489k
 85  192M   85  164M    0     0  6790k      0  0:00:28  0:00:24  0:00:04 8705k
 89  192M   89  172M    0     0  6852k      0  0:00:28  0:00:25  0:00:03 8700k
 95  192M   95  182M    0     0  6969k      0  0:00:28  0:00:26  0:00:02 9081k
100  192M  100  192M    0     0  7072k      0  0:00:27  0:00:27 --:--:-- 9428k
## 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
  1 8959k    1  175k    0     0   167k      0  0:00:53  0:00:01  0:00:52  168k
```

```
 12 8959k   12 1143k    0     0   573k      0  0:00:15  0:00:01  0:00:14  573k
 44 8959k   44 4003k    0     0  1337k      0  0:00:06  0:00:02  0:00:04 1337k
100 8959k  100 8959k    0     0  2251k      0  0:00:03  0:00:03 --:--:-- 2251k
## 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  6068    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0
100  6068  100  6068    0     0  12017      0 --:--:-- --:--:-- --:--:-- 12015
```

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   7795      0 --:--:-- --:--:-- --:--:--  7808
## 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 26.6M    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0
  1 26.6M    1  274k    0     0   181k      0  0:02:30  0:00:01  0:02:29  181k
  5 26.6M    5 1516k    0     0   609k      0  0:00:44  0:00:02  0:00:42  609k
 14 26.6M   14 3977k    0     0  1140k      0  0:00:23  0:00:03  0:00:20 1140k
 28 26.6M   28 7860k    0     0  1747k      0  0:00:15  0:00:04  0:00:11 1747k
 50 26.6M   50 13.4M    0     0  2518k      0  0:00:10  0:00:05  0:00:05 2769k
 81 26.6M   81 21.6M    0     0  3428k      0  0:00:07  0:00:06  0:00:01 4416k
100 26.6M  100 26.6M    0     0  3878k      0  0:00:07  0:00:07 --:--:-- 5666k
## 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
  0  1086    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0
100  1086  100  1086    0     0   2104      0 --:--:-- --:--:-- --:--:--  2100
## 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 8959k    0 81127    0     0  87798      0  0:01:44 --:--:--  0:01:44 87799
  6 8959k    6  555k    0     0   289k      0  0:00:30  0:00:01  0:00:29  289k
 25 8959k   25 2290k    0     0   785k      0  0:00:11  0:00:02  0:00:09  785k
 62 8959k   62 5571k    0     0  1429k      0  0:00:06  0:00:03  0:00:03 1429k
100 8959k  100 8959k    0     0  1937k      0  0:00:04  0:00:04 --:--:-- 1937k
## 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
  0     0    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0
100   270  100   270    0     0    563      0 --:--:-- --:--:-- --:--:--   562
## 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
  1 3707k    1 75263    0     0  89561      0  0:00:42 --:--:--  0:00:42 89492
 14 3707k   14  534k    0     0   301k      0  0:00:12  0:00:01  0:00:11  300k
 65 3707k   65 2440k    0     0   881k      0  0:00:04  0:00:02  0:00:02  881k
100 3707k  100 3707k    0     0  1171k      0  0:00:03  0:00:03 --:--:-- 1171k
## 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  47156      0 --:--:-- --:--:-- --:--:-- 47233
## 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     0    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0
100 31860  100 31860    0     0  56099      0 --:--:-- --:--:-- --:--:-- 56091
```