Problem Set 4

Use the 2014 daily GHCN temperature data to obtain monthly means of the daily maximum temperature, and of the daily minimum temperature. Compute the ranges of these two sets of 12 values for each station.

SAS Code as follows:

```
##Set filename for file downloaded
libname mydata '~/506/';
filename ghcnd_gz pipe "gzip -dc 2014.csv.gz" lrecl=80;

##Read in temperature data and format for use
data ghcnd;
	infile ghcnd_gz delimiter=",";
	input station $ date : yymmdd8. obstype $ obsval;
	format date mmddyy10.;
	month = month(date);
	if obstype="TMAX" or obstype="TMIN";
	obsval=obsval/10;

##Read in station data for use, exclude unused colomns
data stations;
	infile "ghcnd-stations.txt";
	input station $ 1-11 lat 13-20 lon 22-30 elev 32-37 state $ 39-40;

proc sort data=ghcnd out=ghcnd2;
	by station;

proc sort data=stations out=stations2;
	by station;

data ghcnd3;
	merge ghcnd2(in=x) stations2(in=y);
	by station;
	if x=1 and y=1;

##Compute mean temperature for each station for each month
proc summary data=ghcnd3 nway;
	class station month obstype;
	output out=ghcnd4
		mean(obsval)=mntmp;

##Use the temperature type as variables
proc transpose data=ghcnd4(drop=_TYPE_ _FREQ_) out=ghcnd5;
```

```
        by station month;
        id obstype;
        var mntmp;

##Compute the range
data ghcnd6(drop=_NAME_ elev state);
        merge ghcnd5(in=x) stations2(in=y);
        by station;
        if x=1 and y=1;
        tmprange=TMAX-TMIN;

##Print out the data for review
proc print data=ghcnd6;

##Save the final data
data mydata.finaldata;
        set ghcnd6;

run;
```

a. Calculate the difference of the two ranges for each station (i.e. the range of maximum temperatures minus the range of minimum temperatures). Identify the stations which have the least and greatest values for this difference.

```
##Using the data we get from preparation.
libname mydata '~/506/';
data df;
        set mydata.finaldata;
##Use sql to get the range
proc sql;
        create table tmp_range as
        select station,
            range(TMAX) as tmax_range,
            range(TMIN) as tmin_range,
                range(TMAX)-range(TMIN) as range_diff
        from df
        group by station;
quit;

proc print data=tmp_range;

##Get the stations with the max and min range.
proc summary data=tmp_range;
        output out=maxrange(drop=_TYPE_ _FREQ_)
                maxid(range_diff(station))=station
```

```
            max(range_diff)=max_range;

proc summary data=tmp_range;
        output out=minrange(drop=_TYPE_ _FREQ_)
                minid(range_diff(station))=station
                min(range_diff)=min_range;

proc print data=maxrange;
proc print data=minrange;

data mydata.tmp_range;
        set tmp_range;
run;
```

```
We got the following stations with the max and min temperate range.
Obs    station   max_range
1     CA006059    32.0495
Obs    station   min_range
1     USS0006H    -266.395
```

b. In SAS, produce a reduced dataset with indicator variables indicating the stations that are in the bottom 10% of the distribution of values for each of the two ranges. Drop the stations that are not in the bottom 10% for either range. Export the indicator variables and the geographic coordinates of these stations to a text file using proc export.

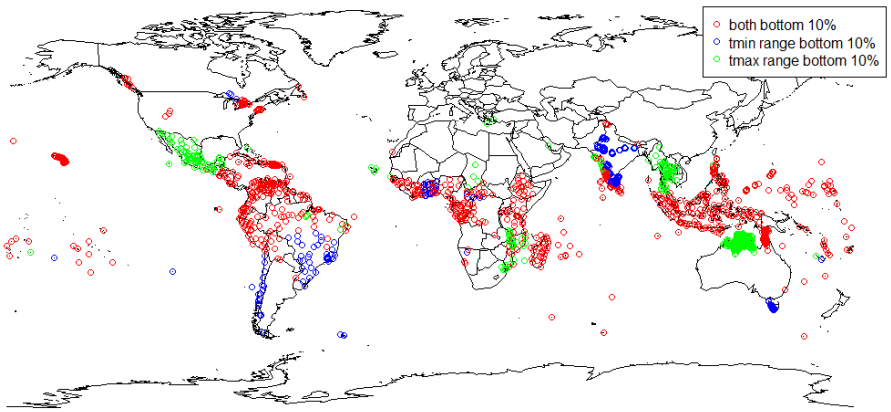
```
libname mydata '~/506/';

data tmp_range;
        set mydata.tmp_range;

## Get the quantiles for 10%
proc univariate data=tmp_range;
        var tmax_range tmin_range;
        output out=range10 pctlpts=10 pctlpre=tmax_range tmin_range;

proc print range10;

data stations;
        infile "ghcnd-stations.txt";
        input station $ 1-11 lat 13-20 lon 22-30 elev 32-37 state $ 39-40;

proc sort data=stations out=stations2;
        by station;

data df1;
```

```
        merge tmp_range(in=x) stations2(in=y);
        by station;
        if x=1 and y=1;
## Indicating for each row. Delete the ones in neither bottom 10%.
data df2;
        set df1;
        if tmax_range ge 6.68495 and tmin_range ge 6.54762 then delete;
        else if tmax_range ge 6.68495 and tmin_range lt 6.54762 then indicator=2;
        else if tmax_range lt 6.68495 and tmin_range ge 6.54762 then indicator=3;
        else indicator=1;
##Export the data to a txt file for later analysis in R
proc export data=df2 dbms=tab outfile='bottom10.txt' replace;
run;
```

Load these coordinates into R, and produce a map showing where these stations are located on a map of the Earth. Use three different colors so it is clear which stations are in the bottom 10% of the distribution for either range value, or for both of them. Briefly comment on your findings.

```
df=read.table(file="C:/Users/Heathtasia/Desktop/506/bottom10.txt",header=TRUE,sep="\t")
df$indicator=as.factor(df$indicator)
levels(df$indicator)=c("both bottom 10%","tmin range bottom 10%","tmax range bottom 10%")
library(mapproj)
library(maptools)
coord=mapproject(df$lon,df$lat)
map()
##Draw points on the map
points(coord,col=c("red","blue","green")[df$indicator])
legend(x="topright", legend = levels(df$indicator), col=c("red","blue","green"), pch=1)
```

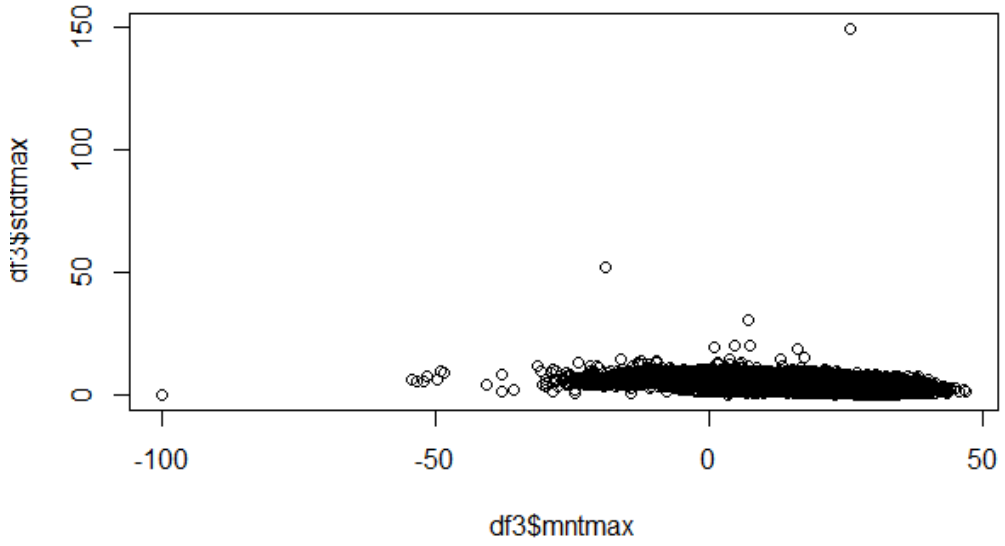
We get the map of temperature range in the bottom 10% as follows:



From the map we find that the regions near the equator has both tmax range and tmin range in the bottom 10%, which means the temperature is more stable in those areas.
And also a majority of areas near the ocean have either more stable tmin or tmax. The ocean play an important part in the climate nearby.

2. Using the 2014 daily maximum GHCN temperature data, calculate the mean and standard deviation of the values within each month for each station.

```
libname mydata '~/506/';
filename ghcnd_gz pipe "gzip -dc 2014.csv.gz" lrecl=80;

data ghcnd(rename=(obsval=tmax));
        infile ghcnd_gz delimiter=",";
        input station $ date : yymmdd8. obstype $ obsval;
        format date mmddyy10.;
        month = month(date);
        if obstype="TMAX";
        obsval=obsval/10;
##Get the mean and std of tmax
proc sql;
        creat table tmax_dist as
        select station,month, mean(tmax) as mntmax, std(tmax) as stdtmax
        from ghcnd
        group by station,month;
quit;
##Export the data to a txt file for later analysis in R
proc export data = tmax_dist outfile = "tmax_dist.txt" dbms=tab replace;

run;

##Read in data in R
df2=read.table(file="C:/Users/Heathtasia/Desktop/506/tmax_dist.txt",header=TRUE,sep="\t")
##Delete the missing values
df3=na.omit(df2)
##Calculate the simple correlation of mean and std of tmax
cor(df3$mntmax,df3$stdtmax)
#-0.5063753          the mean and std of tmax are negatively correlated
## Look at a simple scatter plot
plot(df3$mntmax,df3$stdtmax)
```

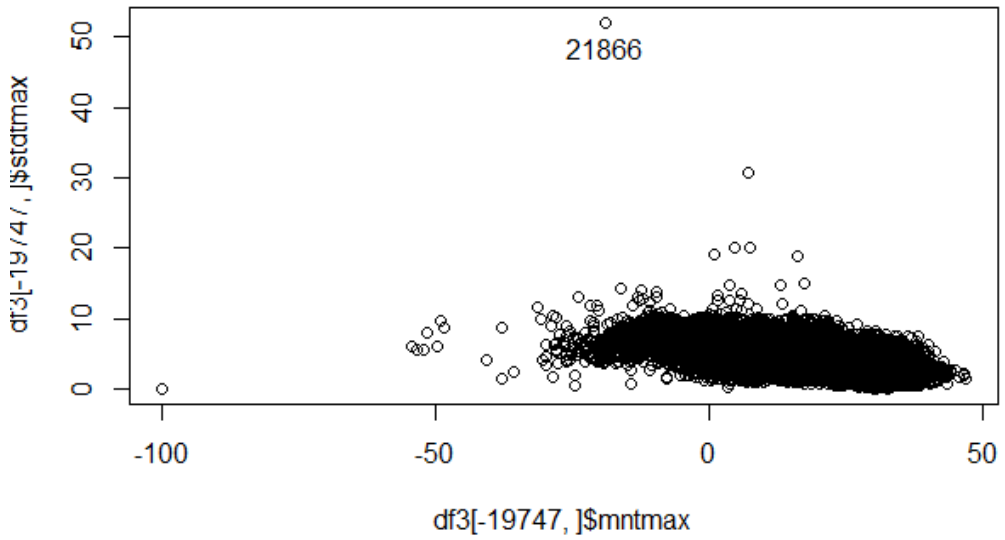


Identify outliers

```
identify(df3$mntmax,df3$stdtmax)
#[1] 19747
```

Remove the outlier to see more in detail

```
plot(df3[-19747,]$mntmax,df3[-19747,]$stdtmax)
```



3. Use the GHCN data to obtain the 2013 and 2014 mean values for daily maximum temperature for each day in January at each station. Calculate the difference between these two means (e.g. 2014 January mean minus 2013 January mean).

```
libname mydata '~/506/';
filename ghcnd_gz pipe "gzip -dc 2014.csv.gz" lrecl=80;
```

```
data ghcnd14(rename=(obsval=tmax14));
        infile ghcnd_gz delimiter=",";
        input station $ date : yymmdd8. obstype $ obsval;
        format date mmddyy10.;
        month = month(date);
        if obstype="TMAX" and month=1;
        obsval=obsval/10;

proc sql;
        create table mntmax_14 as
        select station,mean(tmax14) as mntmax14
        from ghcnd14
        group by station;
quit;

filename ghcnd_gz pipe "gzip -dc 2013.csv.gz" lrecl=80;

data ghcnd13(rename=(obsval=tmax13));
        infile ghcnd_gz delimiter=",";
        input station $ date : yymmdd8. obstype $ obsval;
        format date mmddyy10.;
        month = month(date);
        if obstype="TMAX" and month=1;
        obsval=obsval/10;

proc sql;
        create table mntmax_13 as
        select station,mean(tmax13) as mntmax13
        from ghcnd13
        group by station;
quit;

data meantmax;
        merge mntmax_14(in=x) mntmax_13(in=y);
        by station;
        if x=1 and y=1;
        tmax_diff=mntmax14-mntmax13;

proc export data=meantmax dbms=tab outfile="meantmax.txt" replace;

proc univariate data=meantmax;
        var tmax_diff;
        output out=tmax_diff_quantile pctlpts=10,90 pctlpre=tmax_diff;
```

```
proc print data=tmax_diff_quantile;

data meantmax_10;
        set meantmax;
        if tmax_diff ge -4.36793 and tmax_diff lt 3.70990 then delete;
        else if tmax_diff lt -4.36793 then indicator=1;
        else if tmax_diff ge 3.70990 then indicator=2;

data stations;
        infile "ghcnd-stations.txt";
        input station $ 1-11 lat 13-20 lon 22-30 elev 32-37 state $ 39-40;

proc sort data=stations out=stations2;
        by station;

data meantmax_10_map;
        merge meantmax_10(in=x) stations2(in=y);
        by station;
        if x=1 and y=1;
```

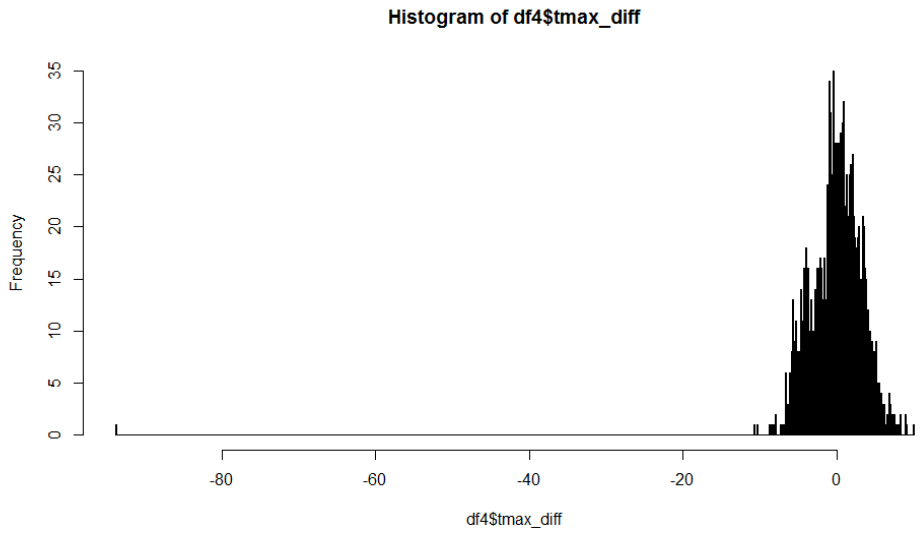
##Export the data to a txt file for later analysis in R

```
proc export data=meantmax_10_map dbms=tab outfile="meantmaxmap.txt" replace;run;
```

Then (i) export all the mean differences to R and make a histogram of them, (ii) export the geographic coordinates of the stations in the top and bottom 10% of the distribution, then use R to make a map of these two sets of points. Briefly comment on your findings.

(i)

```
df4=read.table(file="C:/Users/Heathtasia/Desktop/506/meantmax.txt",header=TRUE,sep="\t")
hist(df4$tmax_diff,breaks=1000)
```

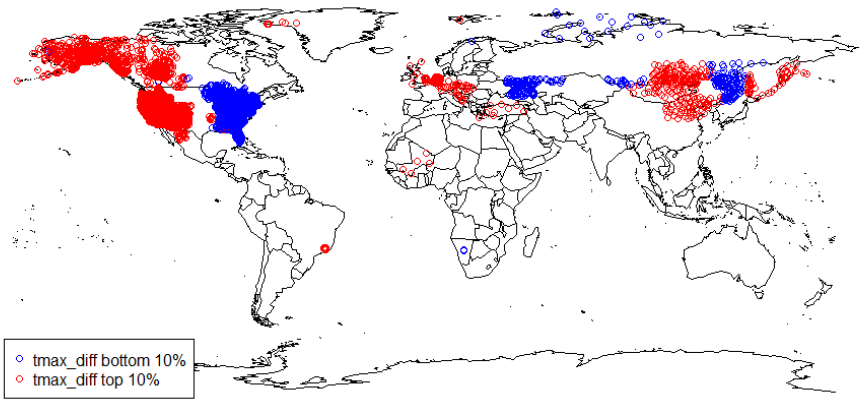


(ii)

```
##Calculate the quantiles
proc univariate data=meantmax;
        var tmax_diff;
        output out=tmax_diff_quantile pctlpts=10,90 pctlpre=tmax_diff;

proc print data=tmax_diff_quantile;
##Get the bottom 10% and top 10% with indicators
data meantmax_10;
        set meantmax;
        if tmax_diff ge -9.366667 and tmax_diff lt 8.904706 then delete;
        else if tmax_diff lt -9.366667 then indicator=1;
        else if tmax_diff ge 8.904706 then indicator=2;

data stations;
        infile "ghcnd-stations.txt";
        input station $ 1-11 lat 13-20 lon 22-30 elev 32-37 state $ 39-40;

proc sort data=stations out=stations2;
        by station;

data meantmax_10_map;
        merge meantmax_10(in=x) stations2(in=y);
        by station;
        if x=1 and y=1;
##Export the data to a txt file for later analysis in R
proc export data=meantmax_10_map dbms=tab outfile="meantmaxmap.txt" replace;
```

In R we plot the map for the top 10% and bottom 10% for tmax_diff

```
df5=read.table(file="C:/Users/Heathtasia/Desktop/506/meantmaxmap.txt",header=TRUE,sep="\t
")
df5$indicator=as.factor(df5$indicator)
levels(df5$indicator)=c("tmax_diff bottom 10%","tmax_diff top 10%")
map()
coord=mapproject(df5$lon,df5$lat)
points(coord,col=c("blue","red")[df5$indicator])
legend(x="bottomleft", legend = levels(df5$indicator), col=c("blue","red"), pch=1)
```



From the map we know that the areas with change(from 2013 to 2014) in tmax in the top 10% and bottom 10% are away from the equator and the polar. The blue ones are the bottom 10% tmax changing area and the red ones are the top 10% tmax changing area.