Lab 4

Kendall Dimson

1. Read in the Data

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

First download and then read in with data.table::fread()

```
if (!file.exists("met_all.gz"))
  download.file(
    url = "https://raw.githubusercontent.com/USCbiostats/data-science-data/master/02_met/met_
    destfile = "met_all.gz",
    method   = "libcurl",
    timeout  = 60
    )
met <- data.table::fread("met_all.gz")
```

2. Prepare the data

Remove temperatures less than -17C

```
met <- met[met$temp > -17, ]
```

Make sure there are no missing data in the key variables coded as 9999, 999, etc

```
summary(met)
```

```
    USAFID           WBAN           year          month        day        
 Min.   :690150   Min.   :  116   Min.   :2019   Min.   :8   Min.   : 1.00  
 1st Qu.:720867   1st Qu.: 3131   1st Qu.:2019   1st Qu.:8   1st Qu.: 8.00  
 Median :722550   Median :13829   Median :2019   Median :8   Median :16.00  
 Mean   :722909   Mean   :29158   Mean   :2019   Mean   :8   Mean   :16.01  
 3rd Qu.:724699   3rd Qu.:54743   3rd Qu.:2019   3rd Qu.:8   3rd Qu.:24.00  
 Max.   :726516   Max.   :94998   Max.   :2019   Max.   :8   Max.   :31.00  

     hour            min              lat             lon         
 Min.   : 0.00   Min.   : 0.00   Min.   :24.55   Min.   :-124.29  
 1st Qu.: 6.00   1st Qu.:20.00   1st Qu.:33.80   1st Qu.: -98.00  
 Median :11.00   Median :48.00   Median :37.87   Median : -90.65  
 Mean   :11.46   Mean   :39.19   Mean   :37.59   Mean   : -91.87  
 3rd Qu.:17.00   3rd Qu.:55.00   3rd Qu.:41.53   3rd Qu.: -82.57  
 Max.   :23.00   Max.   :59.00   Max.   :48.94   Max.   : -68.31  

     elev          wind.dir      wind.dir.qc        wind.type.code    
 Min.   : -13.0   Min.   :  3.0   Length:2200669     Length:2200669    
 1st Qu.:  95.0   1st Qu.:120.0   Class :character   Class :character  
 Median : 238.0   Median :180.0   Mode  :character   Mode  :character  
 Mean   : 407.2   Mean   :183.9                                        
 3rd Qu.: 392.0   3rd Qu.:260.0                                        
 Max.   :9999.0   Max.   :360.0                                        
                  NA's   :702097                                       
   wind.sp        wind.sp.qc          ceiling.ht    ceiling.ht.qc   
 Min.   : 0.000   Length:2200669     Min.   :    0   Min.   :1.000  
 1st Qu.: 0.000   Class :character   1st Qu.: 3048   1st Qu.:5.000  
 Median : 2.100   Mode  :character   Median :22000   Median :5.000  
 Mean   : 2.444                      Mean   :16164   Mean   :4.943  
 3rd Qu.: 3.600                      3rd Qu.:22000   3rd Qu.:5.000  
 Max.   :36.000                      Max.   :22000   Max.   :9.000  
 NA's   :31304                       NA's   :66396                  
```

```
 ceiling.ht.method    sky.cond            vis.dist      vis.dist.qc
 Length:2200669     Length:2200669     Min.   :     0   Length:2200669
 Class :character   Class :character   1st Qu.: 16093   Class :character
 Mode  :character   Mode  :character   Median : 16093   Mode  :character
                                       Mean   : 14904
                                       3rd Qu.: 16093
                                       Max.   :160000
                                       NA's   :30081
   vis.var           vis.var.qc             temp         temp.qc
 Length:2200669     Length:2200669     Min.   :-2.40   Length:2200669
 Class :character   Class :character   1st Qu.:20.00   Class :character
 Mode  :character   Mode  :character   Median :23.90   Mode  :character
                                       Mean   :23.81
                                       3rd Qu.:27.90
                                       Max.   :56.00

   dew.point      dew.point.qc         atm.press        atm.press.qc
 Min.   :-37.20   Length:2200669     Min.   : 960.5    Min.   :1.000
 1st Qu.: 14.00   Class :character   1st Qu.:1011.8    1st Qu.:5.000
 Median : 18.50   Mode  :character   Median :1014.1    Median :9.000
 Mean   : 17.21                      Mean   :1014.2    Mean   :7.694
 3rd Qu.: 22.00                      3rd Qu.:1016.4    3rd Qu.:9.000
 Max.   : 36.00                      Max.   :1059.9    Max.   :9.000
 NA's   :6115                        NA's   :1525297
       rh
 Min.   :  0.833
 1st Qu.: 55.684
 Median : 76.441
 Mean   : 71.593
 3rd Qu.: 90.703
 Max.   :100.000
 NA's   :6115
```

```r
met[met$elev==9999.0, ] <- NA

str(met$lon)
```

```
 num [1:2200669] -116 -116 -116 -116 -116 ...
```

Generate a date variable using the functions as.Date() (hint: You will need the following to create a date paste(year, month, day, sep = "-")).

```
year<- met$year
month<- met$month
day<- met$day

date <- as.Date(paste(year, month, day, sep ="-"))
```

Using the data.table::week function, keep the observations of the first week of the month.

```
met<- met[met$day <=7 ]
```

Compute the mean by station of the variables temp, rh, wind.sp, vis.dist, dew.point, lat, lon, and elev.

```
met_means <- met[, .(mean_temp=mean(temp, na.rm=TRUE),
                mean_rh= mean(rh, na.rm=TRUE),
                mean_wind.sp=mean(wind.sp, na.rm=TRUE),
                mean_vis.dist=mean(vis.dist, na.rm=TRUE),
                mean_dew.point=mean(dew.point, na.rm=TRUE),
                mean_lat=mean(lat, na.rm=TRUE),
                mean_lon=mean(lon, na.rm=TRUE),
                mean_elev=mean(elev, na.rm=TRUE)),
          by = USAFID]
```

Create a region variable for NW, SW, NE, SE based on lon = -98.00 and lat = 39.71 degrees

```
met_means$region <- ifelse(met_means$mean_lon<-98 & met_means$mean_lat >39.71, "NW",
                        ifelse(met_means$mean_lon< -98 & met_means$mean_lat<=39.71, "SW",
                        ifelse(met_means$mean_lon>=98 & met_means$mean_lat>39.71, "NE",
                                "SE")))
```

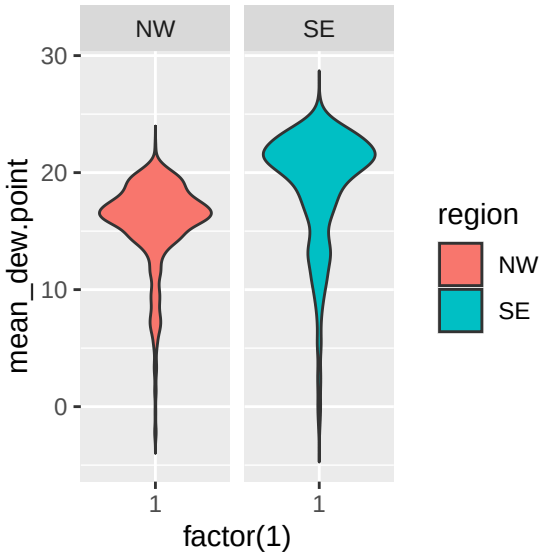
Create a categorical variable for elevation as in the lecture slides

```
met_means[, elev_cat := ifelse(mean_elev >252, "high", "low")]
```

3.Use geom_violin to examine the wind speed and dew point by region

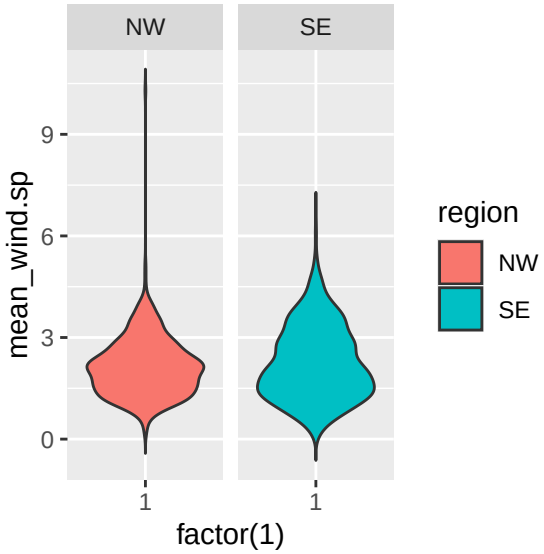
You saw how to use geom_boxplot in class. Try using geom_violin instead (take a look at the help). (hint: You will need to set the x aesthetic to 1)

```
ggplot(met_means[!is.na(elev_cat)], aes(x=factor(1), y=mean_dew.point, fill=region)) + geom_v
```




```
ggplot(met_means[!is.na(elev_cat)], aes(x=factor(1), y=mean_wind.sp, fill=region)) + geom_vio
```

```
Warning: Removed 12 rows containing non-finite outside the scale range
(`stat_ydensity()`).
```



Use facets Make sure to deal with NAs Describe what you observe in the graph

The graphs display distribution of wind speed and dewpoint in the NE and SE regions. For wind speed, it is between 0-5 m/s, and for dew point, the largest distribution is concentrated at dew.point=20.

4. Use geom_jitter with stat_smooth to examine the association between dew point and wind speed by region

Color points by region Make sure to deal with NAs Fit a linear regression line by region Describe what you observe in the graph

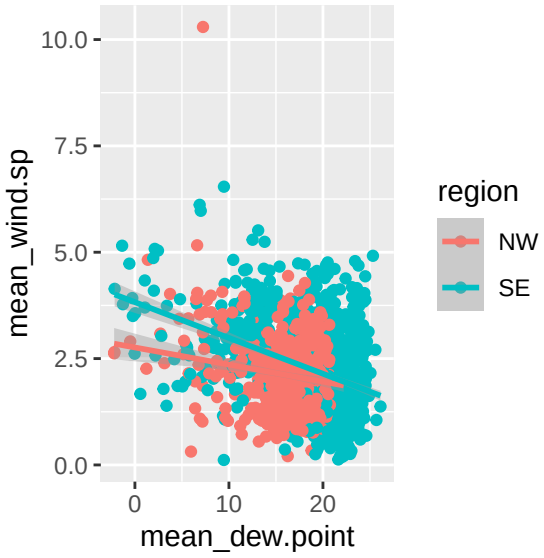
For both regions, as the average dew point increases, the average wind speed decreases.

```
ggplot(met_means, aes(x=mean_dew.point, y=mean_wind.sp, color=region))+
  geom_jitter()+
  stat_smooth(method="lm", aes(group=region))
```

```
`geom_smooth()` using formula = 'y ~ x'
```

```
Warning: Removed 12 rows containing non-finite outside the scale range
(`stat_smooth()`).
```

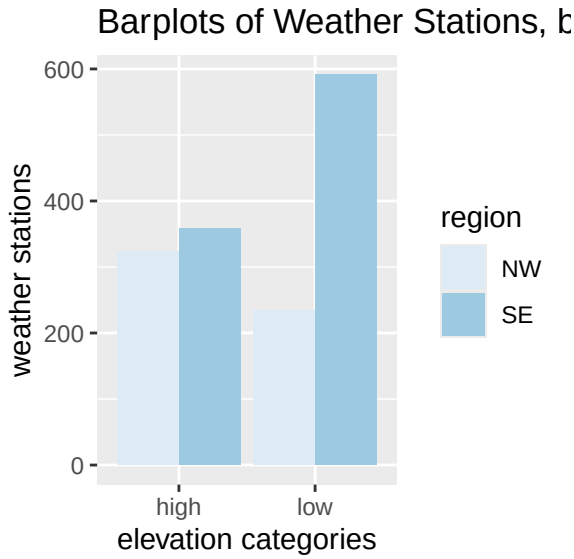
```
Warning: Removed 12 rows containing missing values or values outside the scale range
(`geom_point()`).
```



5. Use geom_bar to create barplots of the weather stations by elevation category colored by region

Bars by elevation category using position="dodge" Change colors from the default. Color by region using scale_fill_brewer see this Create nice labels on the axes and add a title Describe what you observe in the graph Make sure to deal with NA values

```
ggplot(met_means[!is.na(elev_cat)]) +
geom_bar(mapping=aes(x=elev_cat, fill=region), position="dodge") +
  scale_fill_brewer() +
  labs(title= "Barplots of Weather Stations, by elevation categories",
       x= "elevation categories",
       y="weather stations")
```



There are more weather stations located at lower elevations, in comparison to weather stations located at higher elevations.

6. Use stat_summary to examine mean dew point and wind speed by region with standard deviation error bars

Make sure to remove NAs

Use fun.data="mean_sdl" in stat_summary

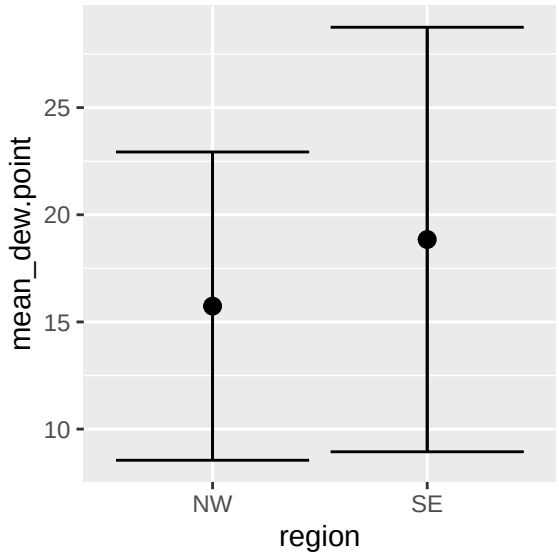
Add another layer of stats_summary but change the geom to "errorbar" (see the help).

Describe the graph and what you observe

```
met_means <- met_means[!is.na(mean_dew.point)]
met_means <- met_means[!is.na(mean_wind.sp)]

ggplot(data=met_means)+
  stat_summary(mapping=aes(x=region,y=mean_dew.point),
               fun.data='mean_sdl',
               geom='pointrange',
               position='dodge')+
  stat_summary(mapping=aes(x=region,y=mean_dew.point),
               fun.data='mean_sdl',
               geom='errorbar',
               position='dodge')
```

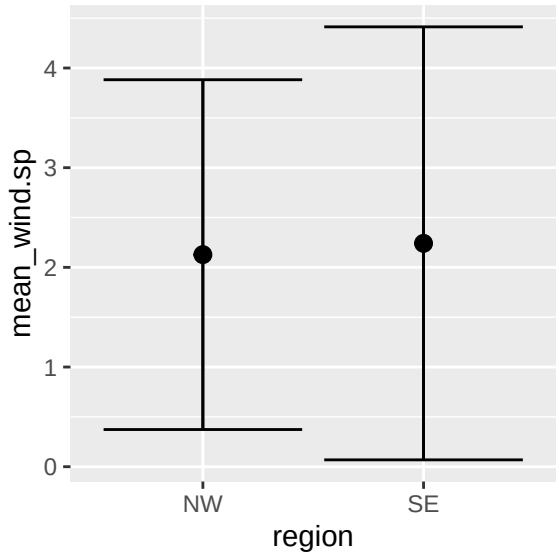
```
Warning: Width not defined
i Set with `position_dodge(width = ...)`
```




```
  ggplot(data=met_means)+
    stat_summary(mapping=aes(x=region,y=mean_wind.sp),
                 fun.data='mean_sdl',
                 geom='pointrange',
                 position='dodge')+
    stat_summary(mapping=aes(x=region,y=mean_wind.sp),
                 fun.data='mean_sdl',
```

```
                geom='errorbar',
                position='dodge')
```

```
Warning: Width not defined
i Set with `position_dodge(width = ...)`
```



Dew point is... average 15/16 in NW and 19 in SE.

Wind speed is... average 2 in NW and 2.25 in SE.

7. Make a map showing the spatial trend in relative humidity in the US

Make sure to remove NAs Use leaflet() Make a color palette with custom colors Use addMarkers to include the top 10 places in relative humidity (hint: this will be useful rank(-rh) <= 10) Add a legend Describe the trend in RH across the US

```
#library(leaflet)

#met<- met%>% filter (!is.na(met$rh) , !is.na(met$lon))

#temp.pal<- colorNumeric(c('blue','pink', 'green'),
                        #  domain=met_means$mean_rh)

#rh_top <- met_means %>% filter(rank(-mean_rh)<=10)
```

```
#met$lon <- as.numeric(as.character(met$lon))
#met_means$lon

#map <- leaflet(met) %>%
 # addProviderTiles('OpenStreetMap') %>%
  #addCircles(
 #   lng= ~lon,
   # lat= ~lat,
 #   color=~temp.pal(rh),
  #  fillOpacity = 0.5, radius=500
   # %>%
   # addMarkers(
   #   lng=~rh_top$lon,
   #   lat=~rh_top$lat,
   #   label
  #  ) %/%
  #    addLegend('bottomleft', pal=temp.pal, values= met_means$mean_rh,
  #              title='Relative Humidity', opacity=1)
 # )
#map

#wasn't able to get code to function correctly on last two questions :(
```

8. Use a ggplot extension

Pick an extension (except cowplot) from here and make a plot of your choice using the met data (or met_avg)

```
#library (ggplot2)
#install.packages(ggforce)
#library(gganimate)

#plot <- ggplot(met, aes(x=date, y=temp))+
  #geom_line(color="purple")+
 # label(title 'Temperature over Time')
#+transition_reveal(date)

#animate(plot, nframes=100, width=800, height=400)
```

Might want to try examples that come with the extension first (e.g. ggtech, gganimate, ggforce)