New York Data Analysis

48 Stunden Prüfungsleistung - Hsin-Yu - Feb.1,2020

Overviw of dataset

The dataset includes data (total obs.=1280, July 2013 ~ Dec 2016) New York City in 3 categories :

- `Time`: DATE/ WEEKDAY/ HOLIDAY/ MONTHYEAR/ YEAR/ Month/ DAYMONTH/ WEEKEND
- `Weather`:
  - Numeric : AWND/ PRCP/ SNOW/ SNWD/ TAVG/ TMAX/ TMIN/ WSF2/ WSF5/ WDF2/ WDF5
  - Categorial : PRCP_LCL
  - Logical : WT01/ WT02/ WT03/ WT04/ WT06/ WT08/ WT09
- `Traffic` : BIKE/ TAXI/ GREEN/ TRAFFIC/ ACCIDENTS (TRAFFIC is continuous ; others are discrete)

We are curious ! **We want to know…**

- **impacts of weather conditions on traffic**
- **(changes of) traffic/weather patterns in New York over time.**

First, import the data set as `md`. Also, some useful packages like `dplyr` and `ggplot2` are loaded. (codes are not shown)

```
md <- read.csv("NY.csv")
```

Curiosity 1 : How do different levels of percipitation(PRCP_LVL) affect traffic?

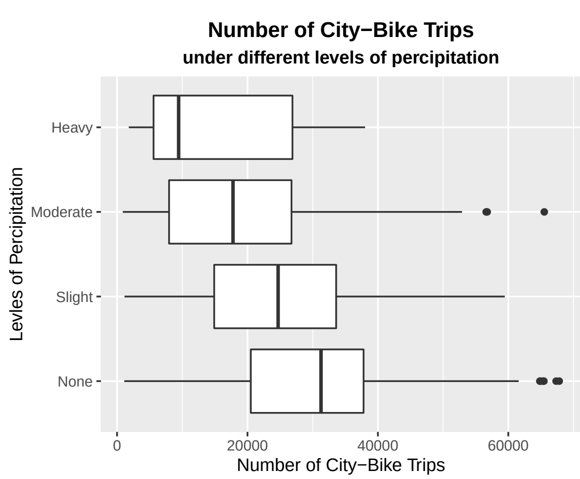
In the 'Weather' variables, only **PRCP_LVL** is categorial, indicating diffrent percipitation (perci.) levles in 4 groups:

```
levels(md$PRCP_LVL) # shows 4 levels: "Heavy","Moderate","None","Slight"
```

```
md$PRCP_LVL<-ordered(md$PRCP_LVL,levels=c("None","Slight","Moderate","Heavy")) #order in sequence of levels.
```

Start with **BIKE** to see its distribution under each levels of percipitation. (Note: 96 missing obs. of BIKE not included)

```
md %>%
  ggplot(aes(x=PRCP_LVL, y=BIKE))+
  geom_boxplot()+
  labs(x="Levles of Percipitation", y="Number of City-Bike Trips",
       title="Number of City-Bike Trips",subtitle="under different levels of percipitation")+
  theme( plot.title = element_text(hjust=0.5, face="bold"),
         plot.subtitle = element_text(hjust=0.5, face="bold"))+coord_flip()
```



The plot shows that overall higher level of perci. has lower numbers of city-bike trips. Let's do some tests to verify.

```
aov_BIKE <- aov(BIKE ~ PRCP_LVL, data=md)
Anova(aov_BIKE, type="III")  # type III for unbalanced design.
```

```
## Anova Table (Type III tests)
##
## Response: BIKE
##                 Sum Sq   Df F value    Pr(>F)
## (Intercept) 7.8240e+10    1 445.759 < 2.2e-16 ***
## PRCP_LVL    1.2167e+10    3  23.107 1.546e-14 ***
## Residuals   2.0712e+11 1180
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# The test shows indicats that different levels of perci. have different means in bike trip numbers.
```

Diagonsis of ANOVA

i) homogeneity of variance

```
leveneTest(BIKE ~ PRCP_LVL, data = md)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value Pr(>F)
## group    3   0.676 0.5668
##       1180
```

```
# p-value > 0.05, meaning we can assume the homogeneity of variances in the different groups.
```

ii) Normality

```
shapiro.test(resid(aov_BIKE)) # However, the residuals are not normally distributed(N.D)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(aov_BIKE)
## W = 0.98725, p-value = 1.197e-08
```

—–> Due to non-normality of the residulas, we need an *alternative non-parametric test* to the one-way ANOVA:

```
kruskal.test(BIKE ~ PRCP_LVL, data = md) #kruskal test shows same results as ANOVA.
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  BIKE by PRCP_LVL
## Kruskal-Wallis chi-squared = 70.282, df = 3, p-value = 3.714e-15
```

Thus, we can't deny that each level perci. related to different bike trip numbers.

Inter-group Differences -

Let's test to see between which groups have mean differences of bike-trip numbers.

```r
TukeyHSD(aov_BIKE)  # Differences are shown among these levels: (S-N), (M-N), (H-N), (M-S)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = BIKE ~ PRCP_LVL, data = md)
##
## $PRCP_LVL
##                        diff        lwr        upr     p adj
## Slight-None       -5033.896  -7565.965 -2501.8263 0.0000022
## Moderate-None     -9673.495 -13470.016 -5876.9746 0.0000000
## Heavy-None       -13394.245 -23737.009 -3051.4807 0.0049248
## Moderate-Slight   -4639.599  -8894.508  -384.6904 0.0262458
## Heavy-Slight      -8360.349 -18880.019  2159.3207 0.1723318
## Heavy-Moderate    -3720.750 -14614.126  7172.6264 0.8159062
```

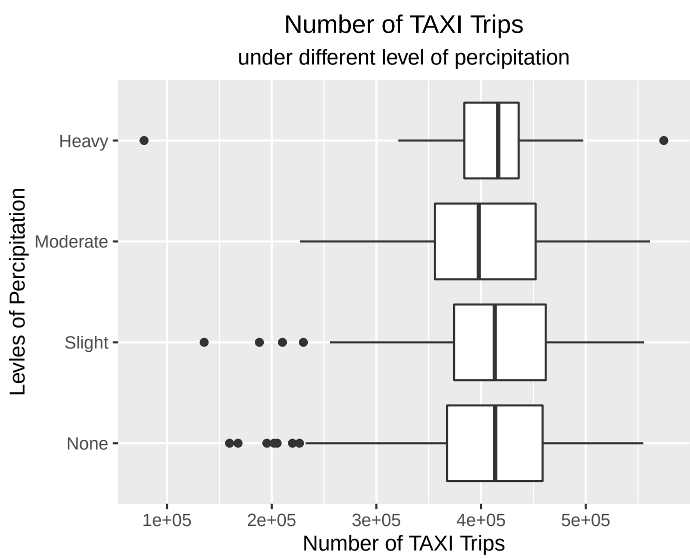
What about TAXI, GREEN & TRAFFIC?

How do each levels of percipitation correspond to number of limousine trips (GREEN) and number of taxi trips (Taxi)?

I repeated the above smae process to all other 'Weather' variables.

It turns out that all the other 'Weather' variables don't appear differently in each levels of perci.

(Here I show only graph and test for **TAXI** as an reference for other 2 'Weather' variables)




```
## Anova Table (Type III tests)
##
## Response: TAXI
##                 Sum Sq   Df   F value Pr(>F)
## (Intercept) 2.4875e+13    1 5673.7130 <2e-16 ***
## PRCP_LVL    1.0007e+10    3    0.7608 0.5161
## Residuals   5.5943e+12 1276
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Curiosity 2 : What are general underlying relations among variables?

Among 25 variables, I would like to have a simple idea of underlying relations among them before further modelling.

Look at only *numeric* variables with PCA

Due to missing values, I cleaned the data to include only values with TRAFFIC after row 257. Then do PCA.

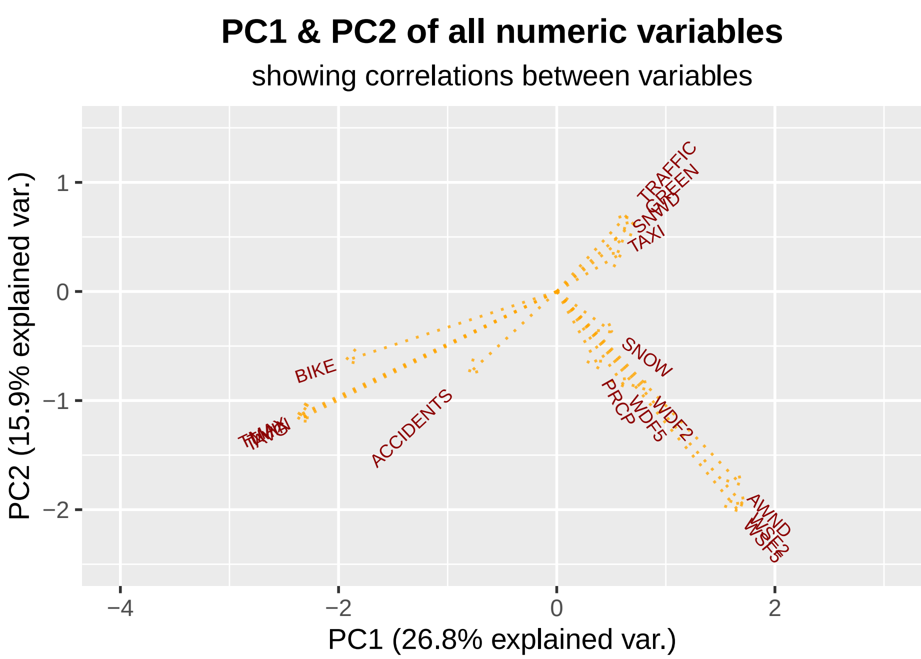
```
t.PCA <- prcomp(num_Total_na, center = TRUE, scale. = TRUE)
```

```
summary(prcomp)
```

The summary of *t.PCA* model shows that 41% of variance is explained by PC1 & PC2.

Let's take a look at the biplot:

```
ggbiplot2(t.PCA, obs.scale =1 , var.scale=1, labels="", varname.size = 2.5, varname.adjust = 1.5,
          color = "orange", linetype = 3, alpha_arrow = 0.8)+ xlim(-4,3)+ylim(-2.5, 1.5)+
  labs(title="PC1 & PC2 of all numeric variables" , subtitle ="showing correlations between variables")+
  theme(plot.title = element_text(hjust=0.5,face="bold"), plot.subtitle =element_text(hjust=0.5))
```




```
t.PCA$rotation  # shows relationshipbetween the initial variables and the principal components
```

Combining all information above, we see 3 groups of underlying relations between numeric variables.

- AWND, WSF2, WSF5, WDF2, WDF5, SNOW, PRCP
- SNWD, GREEN, TRAFFIC, TAXI
- TAVG, TMANX, TMIN, ACCIDENT,BIKE

Later, I would consider to first try modelling within the same and negatively-correlated groups.

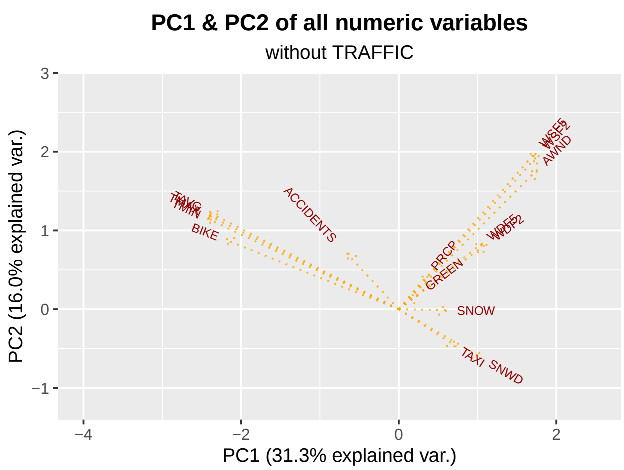
Look at *numeric* variables but without TRAFFIC with PCA

Because *TRAFFIC* has many missing values, I did another PCA that excludes TRAFFIC and see the relations with more observations.(only 156 instead of 857 observations are omitted)

```r
p.PCA  <- prcomp(num_pTotal_na, center = TRUE, scale. = TRUE)
```

The summary of *p.PCA* model shows that 47% of variance is explained by PC1 & PC2.

Take a look at the biplot of p.PCA :



Fliter out **Traffic** to include more observation doesn't change much of the variance explained (from t.PCA to p.PCA). Both shows similar groups, except *SNOW* is more in the PC1 direction in p.PCA.

———————— **Look at *logical* variables (WT01 ~ WT09, no WT05 & WT07)** ————————

```
## # A tibble: 2 x 8
##   Log.  WT01    WT02    WT03    WT04    WT06    WT07    WT09
##   <chr> <table> <table> <table> <table> <table> <table> <table>
## 1 FALSE 1225    1231    1270    1274    1274    1196    1274
## 2 TRUE    55      49      10       6       6      84       6
```

The above table shows (TURE/FALSE) of all logical variables. We can see that:

- *WT04*, *WT06*, *WT09* have the same structure (F:1274 / T:6), together with *WT03*, all 4 of them have very less TURE compared to FALSE.
- *WT01*, *WT02*, *WT08* have comparably more TRUE. –> **I would consider to include them for later modelling.**

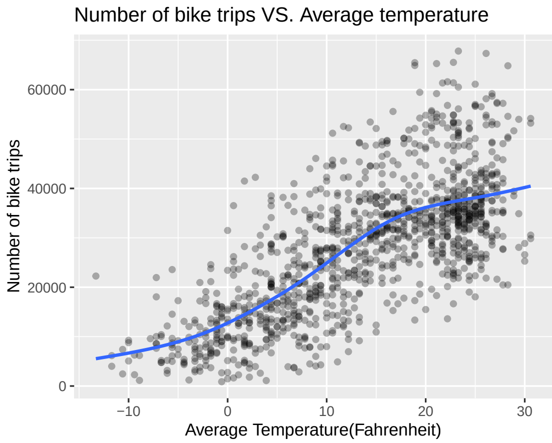
Curiosity 3 : How do weather conditions affect number of bike rides?

We see early on PCA biplots that **BIKE** is close to **TAVG** , while in negative direction of **SNWD** and **SNOW**.

Let's visualize and try to do a regression of these variables on BIKE.

```r
ggplot(md, aes(x = TAVG, y = BIKE))+
  geom_point(alpha=.3)+
  stat_smooth(se=FALSE)+
  labs(title="Number of bike trips VS. Average temperature",
       x="Average Temperature(Fahrenheit)", y="Number of bike trips")
```

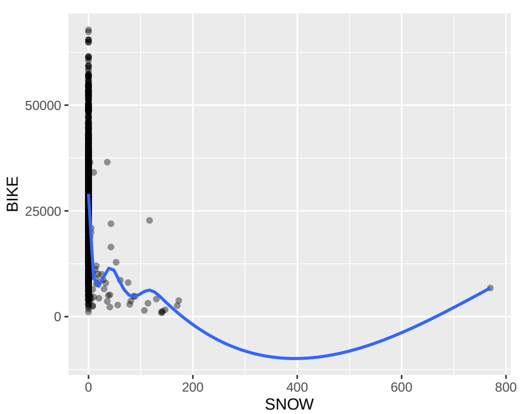
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
# Days with higher temperture have more bike trips
```

```
par(mfrow = c(1, 2))
ggplot(md, aes(x = SNOW, y = BIKE))+
  geom_point(alpha=.4)+
  geom_smooth(se=FALSE)
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
# We see an outlier when SNOW > 600. Do the similar for SNWD, we see outliers when SNWD >300
```

We exclude both the outliers in *SNWD* and *SNOW* and do a linear regression together with *TAVG* for *BIKE*.

```
md_BIKE_normSnow <- filter(md,SNOW<600 & SNWD<300)
mdl_BIKE_0 <- lm(BIKE ~ SNOW + TAVG + SNWD, data=md_BIKE_normSnow)
mdl_BIKE_0$coefficients
```

```
## (Intercept)        SNOW        TAVG        SNWD
## 15600.38251   -60.99540   943.81640   -28.25889
```

```
# Summary of the model shows that all 3 variables are significant
# +/- of coefficients suggests that there is more bike trip with...
  # i)less snow fall / ii) lower snow depth / iii)hotter temperature.
```

Diagnosis of linear regression

```r
shapiro.test(resid(mdl_BIKE_0)) # P-value < 0.05, Residuals are NOT normally distributed
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(mdl_BIKE_0)
## W = 0.97729, p-value = 1.194e-12
```

```r
# --> We should use other model.
```

Try GLM

```r
mdl_BIKE_glm_0 <- glm(BIKE ~ SNOW + TAVG + SNWD, data=md_BIKE_normSnow, poisson(link=log))
```

```r
summary(mdl_BIKE_glm_0)
# Summary of the model shows 3 variables are all significant(p<0.05)
```

Check multicollinearity of predictors

```r
vif(mdl_BIKE_glm_0)  # vif<4 :It is valid to use glm.
```

```
##     SNOW     TAVG     SNWD
## 1.052745 1.110204 1.122312
```

Consider logical 'Weather' variables with ANOVA

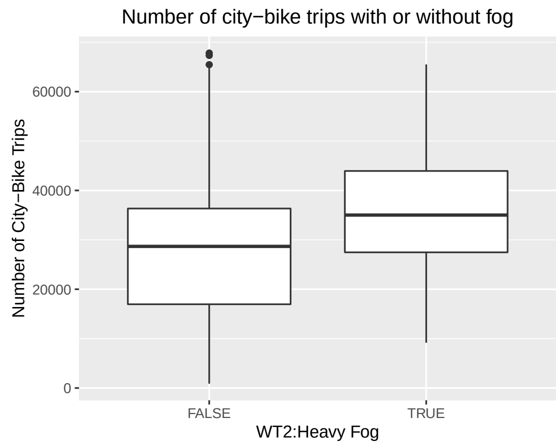
As said early, I will fist check **WT01**, **WT02**, **WT08** (because the others have very less TRUE comparably)

```r
mod_BIKE_logi<- aov(BIKE ~ WT01 + WT02 + WT08 + PRCP_LVL ,data=md)
summary(mod_BIKE_logi)  # WT02 and PRCP_LCL are significant, while WT01 and WT08 are not.
```

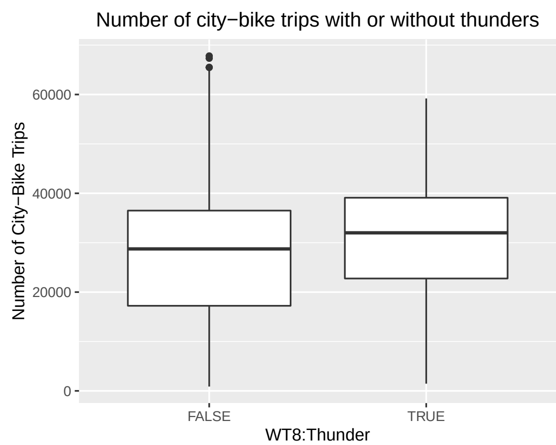
```
##               Df    Sum Sq   Mean Sq F value   Pr(>F)
## WT01           1 3.821e+08 3.821e+08   2.255   0.1334
## WT02           1 3.347e+09 3.347e+09  19.758 9.62e-06 ***
## WT08           1 5.684e+08 5.684e+08   3.355   0.0672 .
## PRCP_LVL       3 1.558e+10 5.195e+09  30.663  < 2e-16 ***
## Residuals   1177 1.994e+11 1.694e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 96 observations deleted due to missingness
```

Let's look at it visually:

```r
md %>%
  ggplot(aes(x=WT02, y=BIKE))+ geom_boxplot()+
  labs(x="WT2:Heavy Fog", y="Number of City-Bike Trips",
       title="Number of city-bike trips with or without fog")+
  theme( plot.title = element_text(hjust=0.5), plot.subtitle = element_text(hjust=0.5))
```



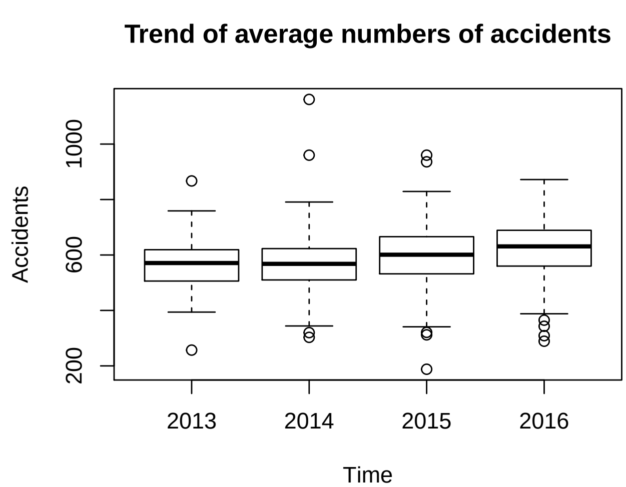

```
# Do the same for WT08. Visually, we see there are differences in bike trips for WT02 but not WT08.
# --> having fog(WT02) makes difference in numbers of bike trips, while having thunder(WT08) doesn't.
```



Overall, besides levels of percipitation(PRCP_LVL), **temperature(TAVG) positively correlated to bike trips**, while **heavy or freezing fog (WTO2), snow fall & depth(SNOW, SNWD) give rise to less bike trips.**

Curiosity 4 : What was some patterns in the number of average accidents in New York?

```
plot(md$YEAR2, md$ACCIDENTS, xlab="Time", ylab="Accidents", main="Trend of average numbers of accidents")
```



i) Although not obvious in the boxplot, *average accident numbers grew slightly from 2014 to 2016.*

ii) Also, we see there was **an outlier day of very high accidents (>1000) in 2014.** What days was that?

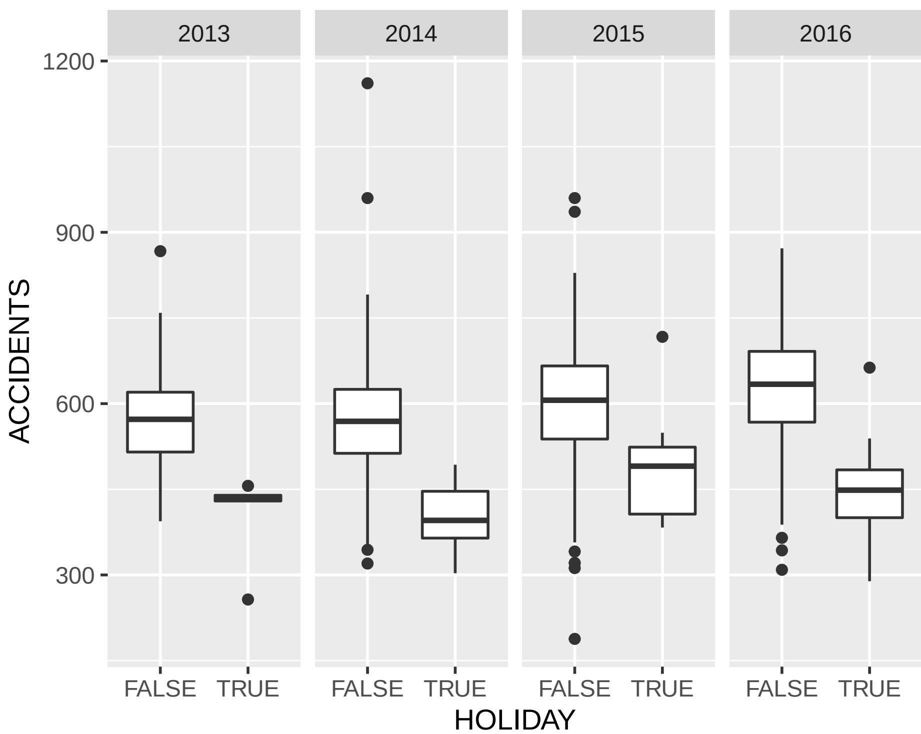
```
High_Accid <- filter(md, md$ACCIDENTS >1000)
High_Accid[,c(1,2,3,11,12,13)]
```

```
##         DATE WEEKDAY HOLIDAY PRCP_LVL SNOW SNWD
## 1 2014-01-21 Tuesday   FALSE   Slight  173    0
```

```
# That was a normal working day with slight rain, some snow fall (but no snow depth) on 21 Jan. 2014
```

iii) Do we have more accidents on holidays? If yes, does that change with year?

```
group_by(md,md$YEAR) %>%
  ggplot(aes(x=HOLIDAY,y=ACCIDENTS))+ geom_boxplot()+ facet_grid(~YEAR)
```



No matter each year, **counter-intuitively, there are less accidents on holidays.**

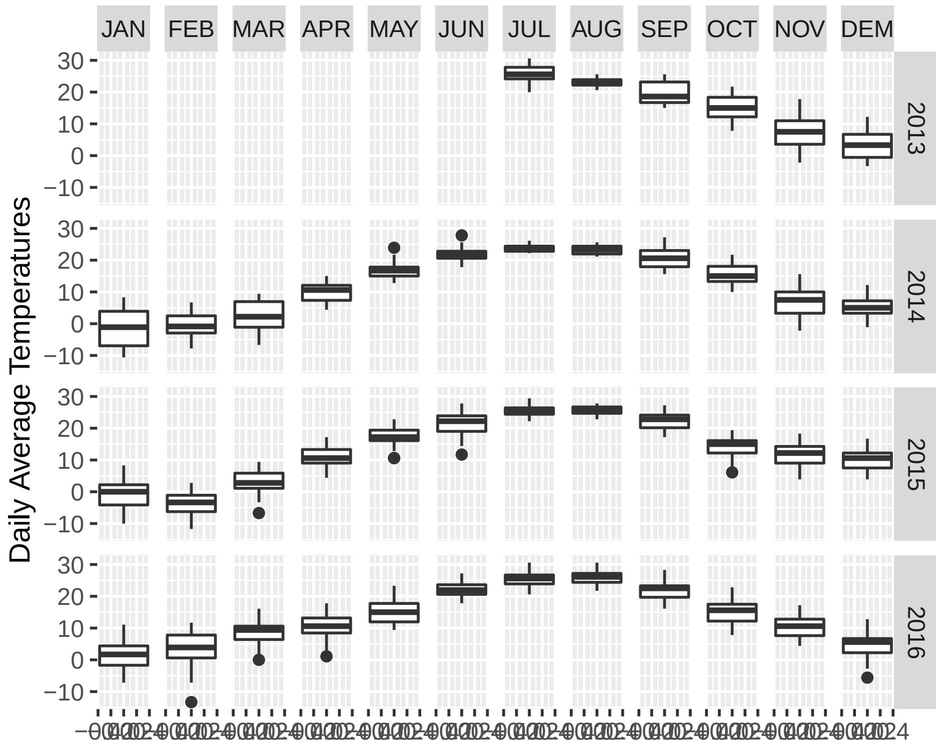
Curiosity 5 : What are some patterns in temperature for New York?

i) What is the overall patterns of temperature by months in each year?

```
md$YEAR2 <- as.factor(md$YEAR)
md$MONTH <- ordered(md$MONTH,
           levels = c("January","February","March","April","May","June","July",
                      "August","September","October","November","December"))

levels(md$MONTH) <- c("JAN","FEB","MAR","APR","MAY","JUN","JUL","AUG","SEP","OCT","NOV","DEM")

group_by(md,md$YEAR2,md$MONTH) %>%
  ggplot(aes(y=TAVG))+ geom_boxplot()+ facet_grid(YEAR~MONTH)+ ylab("Daily Average Temperatures")
```

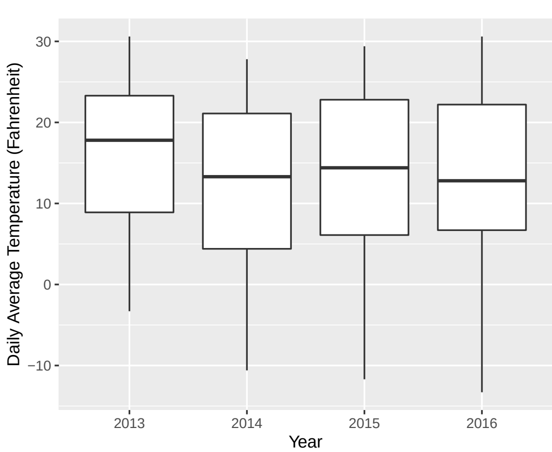


This graph indicates that…

- First of all, there is no data from JAN to JUN for 2013
- Each year there were similar yearly pattern for daily average temperture(D.A.T) :
  - JUL, AUG were the hottest / DEC, JAN, FEB were the coldest. (look at overall height of a single boxplot)
- Bigger range of the box-whisker suggests that the range of D.A.T was:
  - smaller in summer(JULY, AUG) & bigger in winter(JAN, FEB).

ii) Did daily average temperature gwow higher during 2013 to 2016 (maybe due to global warming) ?

```
group_by(md,md$YEAR2) %>%
  ggplot(aes(x=YEAR2,y=TAVG))+ geom_boxplot()+ labs(y="Daily Average Temperature (Fahrenheit)", x="Year")
```



First of all, since we only half year data for 2013, we should just **ignore the boxplot in 2013** here.

Comparing 2014 ~ 2016, it seems there was **no obvious trends of growing daily average tempertures(D.A.T)**.

However, **look at the range of a single boxplot**, the range of D.A.T got wider from 2013 to 2016 ! This indicates that **D.A.T got more extreme in 2016 compared to 2013 !**