Dataset Analysis Report

Aleksandra Viktorova

February 2, 2020

Library

```
library(dplyr)
library(tidyr)
library(modEvA)
library(ggplot2)
```

Some information about the dataset

The New York dataset resembles data on both weather (20 attributes) and traffic (5 attributes) in New York from July 2013 till December 2016. The traffic data has been gathered from various sources.

Number of Instances: 1280 days.
Number of Attributes: 32.
Missing Attributes:

WDF5: 30 days are missing
WSF5: 30 days are missing
BIKE: 96 days are missing
TRAFFIC: 669 days are missing
GREEN: 32 days are missing

Based on the available information, the following questions were formulated:

1. Do variables (except DATA,WEEKDAY,MONTHYEAR,MONTH,DAYMONTH,PRCP_LVL) affect the number of (registered) accidents?

2. Is there a relationship between average traffic speed and precipitation?

3. Is there a relationship between average temperature and the number of CitiBike trips?

4. In which months and days the week on average the largest and smallest number of accidents?

5. Does the choice of a vehicle (bike, taxi, limousine) depend on the precipitation group?

Load the data set into R from the file

```
DataSet_NY<-read.csv(file=file.choose(),header=TRUE,sep=",")
```

The answer to the first question

First, we select the necessary variables for analysis. Then we combine the “Holiday” and “Weekend” variables with a logical “OR” to treat these variables as days when people do not work. We also combine the logical variables “Weather Type” to treat the data as the presence of some kind of weather interference.Then we calculate the correlation matrix.

```
#Select the necessary variables
AccIDENTS_Set<-as_tibble(select(DataSet_NY[rowSums(is.na(DataSet_NY)) == 0,],-
DATE,-WEEKDAY,-MONTHYEAR,-MONTH,-DAYMONTH,-PRCP_LVL))

#Combine the logical variables
AccIDENTS_Set$NoWD<-as.numeric(AccIDENTS_Set$HOLIDAY | AccIDENTS_Set$WEEKEND)
AccIDENTS_Set$NoSR<-as.numeric(AccIDENTS_Set$WT01 | AccIDENTS_Set$WT02 |
                               AccIDENTS_Set$WT03 | AccIDENTS_Set$WT04 |
                               AccIDENTS_Set$WT06 | AccIDENTS_Set$WT08 |
                               AccIDENTS_Set$WT09)

#Calculate the correlation matrix
cor(select(AccIDENTS_Set, -HOLIDAY, -WEEKEND, -WT01,
                          -WT02, -WT03, -WT04, -WT06, -WT08, -WT09))

##                   YEAR         AWND         PRCP        SNOW         SNWD
## YEAR       1.000000000  0.141155821 -0.008402141  0.05796933  0.126472245
## AWND       0.141155821  1.000000000  0.180194863  0.17178806 -0.022256829
## PRCP      -0.008402141  0.180194863  1.000000000  0.42475500  0.021729552
## SNOW       0.057969334  0.171788056  0.424755000  1.00000000  0.138310390
## SNWD       0.126472245 -0.022256829  0.021729552  0.13831039  1.000000000
## TAVG      -0.211231770 -0.236897133 -0.019979364 -0.12227702 -0.232457287
## TMAX      -0.203953327 -0.266842606 -0.046357574 -0.13361832 -0.235519623
## TMIN      -0.212076865 -0.234890173  0.019597833 -0.10765745 -0.235471590
## WDF2       0.049108993  0.213754124 -0.094808815 -0.07213958  0.060054357
## WDF5       0.056359462  0.146074023 -0.106193032 -0.07030679  0.067303775
## WSF2       0.141741247  0.828585260  0.200738619  0.11537459 -0.005052601
## WSF5       0.118301668  0.836639820  0.200692955  0.12220806 -0.012399429
## BIKE       0.142623805 -0.243092061 -0.187929019 -0.11814481 -0.201298205
## TAXI      -0.262702616  0.025096989 -0.151638654 -0.27091578 -0.208603204
## GREEN     -0.274267255  0.014717083 -0.061664222 -0.17260046 -0.107793231
## TRAFFIC    0.006626483  0.041959086 -0.084918305 -0.09097799  0.067844516
## ACCIDENTS -0.001902431 -0.060803587  0.044292224 -0.11461765 -0.003599948
## NoWD       0.015682706  0.022023357 -0.020227170  0.06419410  0.019033946
## NoSR      -0.059016917 -0.008281548  0.285744534  0.09794127  0.011199156
##                  TAVG        TMAX        TMIN          WDF2         WDF5
## YEAR      -0.21123177 -0.20395333 -0.21207686  0.0491089932  0.056359462
## AWND      -0.23689713 -0.26684261 -0.23489017  0.2137541243  0.146074023
## PRCP      -0.01997936 -0.04635757  0.01959783 -0.0948088150 -0.106193032
## SNOW      -0.12227702 -0.13361832 -0.10765745 -0.0721395815 -0.070306788
## SNWD      -0.23245729 -0.23551962 -0.23547159  0.0600543569  0.067303775
## TAVG       1.00000000  0.97917775  0.98472058 -0.0993474218 -0.059428649
## TMAX       0.97917775  1.00000000  0.95441374 -0.0912107869 -0.033898100
## TMIN       0.98472058  0.95441374  1.00000000 -0.1278670966 -0.092077403
## WDF2      -0.09934742 -0.09121079 -0.12786710  1.0000000000  0.707450524
## WDF5      -0.05942865 -0.03389810 -0.09207740  0.7074505242  1.000000000
## WSF2      -0.21566279 -0.22247524 -0.21895910  0.2715767642  0.212274884
## WSF5      -0.19726376 -0.20167324 -0.19981437  0.2613420697  0.217785253
## BIKE       0.60008516  0.59725388  0.57679040 -0.0687122510 -0.033056479
## TAXI      -0.19128420 -0.18893129 -0.19255667  0.0871570296  0.052337869
## GREEN     -0.18796338 -0.19398597 -0.18713317  0.0703374050  0.050929471
## TRAFFIC   -0.19861488 -0.19774398 -0.19228292  0.0489556011 -0.016605967
## ACCIDENTS  0.25347058  0.25209751  0.24707883 -0.0003304574  0.037484981
## NoWD      -0.06366172 -0.06366943 -0.06687485  0.0147081229  0.004614653
## NoSR       0.10655615  0.10686023  0.14238488  0.0106450176  0.009072539
##                   WSF2         WSF5        BIKE        TAXI        GREEN
## YEAR       0.141741247  0.118301668  0.14262381 -0.26270262 -0.274267255
## AWND       0.828585260  0.836639820 -0.24309206  0.02509699  0.014717083
## PRCP       0.200738619  0.200692955 -0.18792902 -0.15163865 -0.061664222
## SNOW       0.115374589  0.122208062 -0.11814481 -0.27091578 -0.172600456
## SNWD      -0.005052601 -0.012399429 -0.20129820 -0.20860320 -0.107793231
## TAVG      -0.215662790 -0.197263758  0.60008516 -0.19128420 -0.187963384
## TMAX      -0.222475239 -0.201673238  0.59725388 -0.18893129 -0.193985969
## TMIN      -0.218959099 -0.199814368  0.57679040 -0.19255667 -0.187133168
## WDF2       0.271576764  0.261342070 -0.06871225  0.08715703  0.070337405
## WDF5       0.212274884  0.217785253 -0.03305648  0.05233787  0.050929471
## WSF2       1.000000000  0.961299210 -0.25727795  0.08269125  0.037359276
## WSF5       0.961299210  1.000000000 -0.23257045  0.05859578  0.020328010
## BIKE      -0.257277950 -0.232570445  1.00000000 -0.14685274 -0.218969562
## TAXI       0.082691253  0.058595782 -0.14685274  1.00000000  0.685149481
## GREEN      0.037359276  0.020328010 -0.21896956  0.68514948  1.000000000
## TRAFFIC    0.021524423  0.005556249 -0.20378224 -0.16147557  0.057423382
## ACCIDENTS -0.015505209 -0.014793910  0.24786280  0.23771801 -0.126955142
## NoWD      -0.026121410 -0.027640345 -0.03986487 -0.08551096  0.482870925
## NoSR       0.172187382  0.175271439 -0.06812906 -0.01576763 -0.003684072
##                TRAFFIC     ACCIDENTS         NoWD         NoSR
## YEAR       0.006626483 -0.0019024308  0.015682706 -0.059016917
## AWND       0.041959086 -0.0608035870  0.022023357 -0.008281548
## PRCP      -0.084918305  0.0442922244 -0.020227170  0.285744534
## SNOW      -0.090977993 -0.1146176484  0.064194101  0.097941273
## SNWD       0.067844516 -0.0035999477  0.019033946  0.011199156
## TAVG      -0.198614881  0.2534705829 -0.063661720  0.106556147
## TMAX      -0.197743979  0.2520975052 -0.063669428  0.106860229
## TMIN      -0.192282924  0.2470788306 -0.066874851  0.142384879
## WDF2       0.048955601 -0.0003304574  0.014708123  0.010645018
## WDF5      -0.016605967  0.0374849812  0.004614653  0.009072539
## WSF2       0.021524423 -0.0155052091 -0.026121410  0.172187382
## WSF5       0.005556249 -0.0147939104 -0.027640345  0.175271439
## BIKE      -0.203782245  0.2478627981 -0.039864871 -0.068129061
## TAXI      -0.161475573  0.2377180073 -0.085510956 -0.015767629
## GREEN      0.057423382 -0.1269551423  0.482870925 -0.003684072
## TRAFFIC    1.000000000 -0.4246108362  0.308053731 -0.057414130
## ACCIDENTS -0.424610836  1.0000000000 -0.616256282  0.024740567
## NoWD       0.308053731 -0.6162562824  1.000000000 -0.024289911
## NoSR      -0.057414130  0.0247405667 -0.024289911  1.000000000
```

As we can see, according to the Sheddock scale, the ACCIDENTS variable has a weak connection with the TRAFFIC variable and a moderate connection with the NoWD variable, which indicates non-working days. Thus, the ACCIDENTS variable (number of (registered) accidents) has relationship with other variables, but they do not have a significant effect.

The answer to the second question

Returning to the previously obtained correlation matrix, it can be noted that there is no relationship between average traffic speed and the variables that are responsible for weather conditions.

The answer to the third question

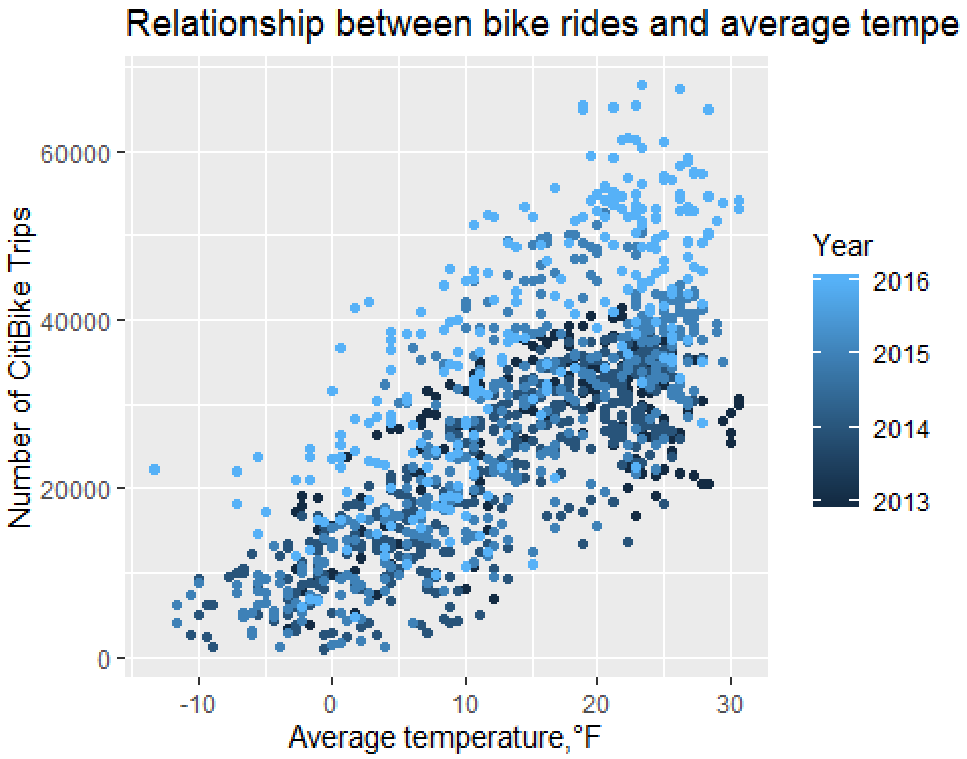
Returning to the previously obtained correlation matrix, it can be noted that there is a moderate relationship between the variables. Let's try to analyze the data separately from the rest of the array.

```
#Select the necessary variables
BIKE_Set<-as_tibble(select(DataSet_NY,AWND,BIKE,TAVG,YEAR))

#Calculate the new correlation matrix
cor(BIKE_Set[rowSums(is.na(BIKE_Set)) == 0,])

##             AWND       BIKE        TAVG       YEAR
## AWND  1.00000000 -0.2830786 -0.27652065 0.02934848
## BIKE -0.28307856  1.0000000  0.72091798 0.31797351
## TAVG -0.27652065  0.7209180  1.00000000 0.01006802
## YEAR  0.02934848  0.3179735  0.01006802 1.00000000

#Build chart the number of of CitiBike Trips and average temperature by years
ggplot(data=BIKE_Set[rowSums(is.na(BIKE_Set)) == 0,])+
      geom_point(aes(x =TAVG, y =BIKE,color =YEAR))+
      labs(x ="Average temperature,°F",y ="Number of CitiBike Trips",color
="Year")+
      ggtitle("Relationship between bike rides and average temperature")
```



It can be seen from the results that BIKE and TAVG have a correlation coefficient greater than that obtained previously. In addition, the graph shows that there is a significant relationship between the variables. Thus, there is a relationship between average temperature and the number of CitiBike trips.

The answer to the fourth question

We select the data for the variables. Then we group the data by month and day of the week, after which we calculate the average number of accidents.We display the months and days of delays for which the average value assumes the maximum and minimum values.

```
#Select the necessary variables
MW_Accident<-as_tibble(select(DataSet_NY[DataSet_NY$YEAR !=
2013,],WEEKDAY,MONTH, ACCIDENTS))

#Group the data by month
M_GROUPS<-group_by(MW_Accident, MONTH) %>%
    summarise(AV_acc = mean(ACCIDENTS))

#Months with maximum and minimum values
M_GROUPS$MONTH[c(which.max(M_GROUPS$AV_acc),which.min(M_GROUPS$AV_acc))]

## [1] June     January
## 12 Levels: April August December February January July June March ...
September

#Group the data by day of the week
WD_GROUPS<-group_by(MW_Accident, WEEKDAY) %>%
  summarise(AV_acc = mean(ACCIDENTS))

#Days of the week with maximum and minimum values
WD_GROUPS$WEEKDAY[c(which.max(WD_GROUPS$AV_acc),which.min(WD_GROUPS$AV_acc))]

## [1] Friday Sunday
## Levels: Friday Monday Saturday Sunday Thursday Tuesday Wednesday
```

Thus, the largest average number of accidents occurs in June and Friday, and the minimum in January and Sunday.

The answer to the fifth question

We select the variables from the data set, after which we group by weather groups and summarize the number of trips for each type of transport. Then we form a table and conducts a Chi-square test.

```
#Select the necessary variables
Precipitation_TransportSet<-
as_tibble(select(DataSet_NY,PRCP_LVL,BIKE,TAXI,GREEN))

#Group the data by Precipitation group
Groups_PT<-
group_by(Precipitation_TransportSet[rowSums(is.na(Precipitation_TransportSet))
== 0,],
                       PRCP_LVL) %>%
            summarise(
                      sumBike = sum(BIKE),
                      sumTaxi = sum(TAXI),
                      sumGreen = sum(GREEN))
#Create data table
Groups_Set<-
as.table(rbind(Groups_PT$sumBike,Groups_PT$sumTaxi,Groups_PT$sumGreen))
dimnames(Groups_Set) <- list(Transports =
names(Precipitation_TransportSet[2:4]),
                                       Precipitation =
unique(Precipitation_TransportSet$PRCP_LVL))

#Chi-Square Test
chisq.test(Groups_Set)

##
##  Pearson's Chi-squared test
##
## data:  Groups_Set
## X-squared = 402729, df = 6, p-value < 2.2e-16
```

Thus,the choice of transport depends heavily on precipitation.