COURSE: STATISTICAL PROGRAMMING LANGUAGES

Term: WiSe 2016-2017

Humboldt University -Berlin (HU)

SEMINAR WORK: “Bike Sharing – the next thing”

Participant: - Isabel Lucia Chaquire, Matriculation Nr: 526796

Date: 27.03.2017

REPORT

I. INTRODUCTION

In recent years, urban bike rental program attracted more and more attention. They are a natural response to the desire of society as much as possible to use the bicycle as a vehicle in everyday life. After all, everyone knows that the bike has a less negative impact on the environment than any other transport. Initially, the concept of sharing bicycles revolutionary 60s evolved slowly until the emergence of new technologies has not stimulated the acceleration of its development.

At the time there are bikeshare systems around de world. One of them, Capital bikeshare system has over 350 stations in Washington, D. C. , Arlington, Alexandria, VA und Montgomery County and MD in the USA. Bike sharing systems are a new way of traditional bike rentals. The whole process from membership to rental and return back has become automatic. Datasets were generated by 500 bike-sharing programs and are available in some repositories.

II. THEORY AND DESIGN

For this work I represented the bike rental data of the city Seattle using descriptive statistics and later correlated the data using die explorative statistics (also called hypothesis-generating statistics, analytical statistics or data-mining). The latter is a mixed form of a descriptive and inductive procedures.

I used the following algorithms:

- mean,
- median
- variance
- histogram
- boxplots
- t_test
- least square coefficients
- F-test
- ANOVA test (with F- test)
- AIC

III. IMPLEMENTATION

a) Creating new columns variables

Main code:

```
#Dataset of tripdata: inserting a variable “Stadtschluessel “called “Seattle Pronto”
seattle_trip$Stadtschluessel<-"Seattle Pronto"

#giving new names to column variables:
names(seattle_trip)<-c("trip_id", "startdate", "stoptime", "bikeid", "tripduration",
"from_station_name", "to_station_name", "from_station_id", "to_station_id", "usertype",
"gender", "birthyear", "Stadtschluessel")
colnames(seattle_trip)   #checking new variables names

#Dataset weather:  giving new names to column variables
names(weatherSeattle14_16)<-c("PST", "Max.TemperaturC", "mittlereTemperaturC",
"Min.TemperaturC", "TaupunktC", "MeanDew PointC", "Min.DewpointC", "Max.Feuchtigkeit",
"Mean.Feuchtigkeit", "Min.Feuchtigkeit", "Max.Luftdruck_in_MeereshoehehPa",
"Mean.Luftdruck_in_MeereshoehehPa", "Min.Luftdruck_in_MeereshoehehPa",
"Max.SichtweiteKm", "Mean.SichtweiteKm", "Min.SichtweiteKm",
"Max.WindgeschwindigkeitKm.h", "Mean.WindgeschwindigkeitKm.h",
"Max.BoeengeschwindigkeitKm.h", "Niederschlagmm", "CloudCover",  "Ereignisse",
"WindDirDegrees", "Stadt")
colnames(weatherSeattle14_16)   #checking new variables names
```

b) Merging tripdata + weather into one merged single dataset.

Main code:

```
#merging datasets tripdata and weather  using two column keys:
seattle_merged<-merge(seattle_trip, weatherSeattle14_16, by.x=c("startdate",
"Stadtschluessel"), by.y=c("PST", "Stadt"))

#writing and saving dataframe seattle_merged as seattleMerged.csv :
fwrite(seattle_merged, "C:/Users/isabe/Documents/isabel/R_HU_Statistik/
Course_StatisticalProgramming/Projects_SPL/bikeRental/Rawdata_bikeRental/
seattleMerged.csv")
```

c) Plot in 3 dimensions with library “scatterplot3d”

Main Code:

```
#creation 3D Plot of Tripduration in function of mean Temperatur and Precipitation
# variable coloured: mean temperatur
library ("scatterplot3d")
layout(cbind(1:2, 1:2), heights = c(2, 1))
temp<-hsv((temp <-
0.7*seattle_data$mittlereTemperaturC/diff(range(seattle_data$mittlereTemperaturC)))-
min(temp) + 0.3)    #the colours code is given through variable temp
s3d<-scatterplot3d(seattle_data$Niederschlagmm, seattle_data$mittlereTemperaturC,
seattle_data$tripduration,
      pch=5, color=temp,
```

```
      main="Influence of precipitation and temperature on tripduration",
      xlab="precipitation, mm",
      ylab="mean temperature, °C",
      zlab="tripduration, min")
#Setting the parameters for graph edition:
par(mar=c(5, 3, 0, 3))
plot(seq(min(seattle_data$mittlereTemperaturC), max(seattle_data$mittlereTemperaturC),
length = 10), rep(0, 10), pch = 2,
      axes = FALSE, xlab = "color code of variable \"mean T°C\"", ylab = "",
      col = hsv(seq(0.3, 1, length = 10)))
axis(1, at = seq(-20, 25, 5))
```

d) Plot in 3D with library "3dPlot" (4 Variables)

Main code:

```
# creation 3D Plot of Tripduration vs mean Temperatur and Precipitation with a 4th coloured
# variable Weather
library(plot3D)
#
par(mfrow = c(1.0, 1.0))
panelfirst <- function(pmat) {
      zmin <- min(seattle_data$tripduration)
      XY <- trans3D(seattle_data$Niederschlagmm, seattle_data$mittlereTemperaturC,
      z = rep(zmin, nrow(seattle_data)), pmat = pmat)
      scatter2D(XY$x, XY$y,
      colvar = seattle_data$Ereignisse,
      pch = ".",
      cex = 2, add = TRUE, colkey = FALSE)
      xmin <- min(seattle_data$Niederschlagmm)
      XY <- trans3D(x =rep(xmin, nrow(seattle_data)), y=seattle_data$mittlereTemperaturC,
      z = seattle_data$tripduration, pmat = pmat)
     scatter2D(XY$x, XY$y, colvar = seattle_data$Ereignisse,
     pch = ".",
     cex = 2, add = TRUE, colkey = FALSE)
}     #Setting of graphs, title, colkey (Weather):
with(seattle_data, scatter3D(x=seattle_data$Niederschlagmm,
     y=seattle_data$mittlereTemperaturC, z=seattle_data$tripduration,
     colvar=seattle_data$Ereignisse,
     pch=8, cex=1.0, xlab="°Precipitation mm", ylab ="Mean T °C",
     zlab="Tripduration, min", clab=c("Weather"),
     main="People rent bikes more in raining than by clouding between 10-20°C\n
     City: Seattle\n Weather values: 1=cloud, 2=rain, 3=snow, 4=cloudy-rain, 5=rain-storm\n Note:
     0=NA \n",
     ticktype="detailed",
     panel.first=panelfirst, theta=15, d=2.0,
     colkey=list(length=0.5, width=0.5, cex.clab=0.75))
)
```

e) POISSON MODEL: tripduration ~ usertype + Gender + Ereignisse

```
> poisson.mod<-glm(seattleData$tripduration ~ seattleData$usertype+
  seattleData$gender+seattleData$Ereignisse, family=poisson,
  data=seattleData)
```

f) QUASIPOISSON: tripduration ~ usertype + gender + Ereignisse

```
> quasipoisson.mod<-glm(seattleData$tripduration ~  seattleData$usertype +
  seattleData$gender+seattleData$Ereignisse, family=quasipoisson,
  data=seattleData)
```

g) NEGATIVE BINOMIAL REGRESSION: tripduration ~ usertype+gender+Ereignisse

```
> negbinom.mod<-glm.nb(seattleData$tripduration~seattleData$usertype+
  seattleData$gender+seattleData$Ereignisse, data=seattleData, link=log)
```

h) LINEAR REGRESSION WITH TRANSFORMATION: Modell with 4 variables, 1 of them is log(y)

```
> lin.mod<-
  lm(seattleData$tripduration~seattleData$usertype+seattleData$gender+
  seattleData$Ereignisse, data=seattleData)
```

i) COMPARISON OF TWO TRANSFORMED LINEAR REGRESSION with ANOVA (Chisq- test)

```
> anova(lin.mod, xtransf_lin.mod, test="Chisq")
```

j) BACKWARD MODEL (working with ONLY log(y):

```
> lin.mod_back <- step(lin.mod, direction = "backward")

> summary(lin.mod_back)
```

COMPLEMENTARY CODE:

k) SUMMARY(MEAN,MEDIAN, QUANTILES)

```
> sd(seattle_data$tripduration) #standard deviation of tripduration
```

k) HISTOGRAMS:

```
> hist(seattle_data$birthyear)

> hist(seattle_data$mittlereTemperaturC)
```

l) DENSITY:

```
> density(seattle_data$tripduration) #density of tripduration
```

m) PLOTS:

```
> plot(seattle_data$birthyear, seattle_data$tripduration)

> plot(seattle_data$gender, seattle_data$tripduration)
```

IV. EMPIRICAL STUDY / TESTING

Table 1.- Original Tripdata Datasets

Sources of datasets	Beschreibung	Ort
https://www.divvybikes.com/system-data	2013 Q3 & Q4 DATA - 2016 Q1 & Q2 DATA	Chicago, IL
https://www.citibikenyc.com/system-data	2013 July - 2016 Sept	New York City, NY
https://s3.amazonaws.com/tripdata/index.html	2013 July - 2016 Sept	New York City, NY
https://data.chattlibrary.org/Transportation/Bike-Chattanooga-Trip-Data/8yba-nwv8	2012 Jan - 2015 Dez	Chattanooga, TN
http://hubwaydatachallenge.org/register/?next=/data-api/		Metro Boston / Brookline / Cambridge / Somerville, MA
http://www.capitalbikeshare.com/trip-history-data	2010 Q4 - 2016 Q3	Washington DC Metro Area
http://www.capitalbikeshare.com/system-data	2010 Q4 - 2016 Q3	Washington DC Metro Area
https://www.niceridemn.org/data/	2010 - 2015	Minneapolis-St. Paul, MN
http://www.bayareabikeshare.com/open-data	2013 August - 2016 August	San Francisco Bay Area, CA
http://www.prontocycleshare.com/data	2014 Oct - 2016 Aug	Seattle Prontoo

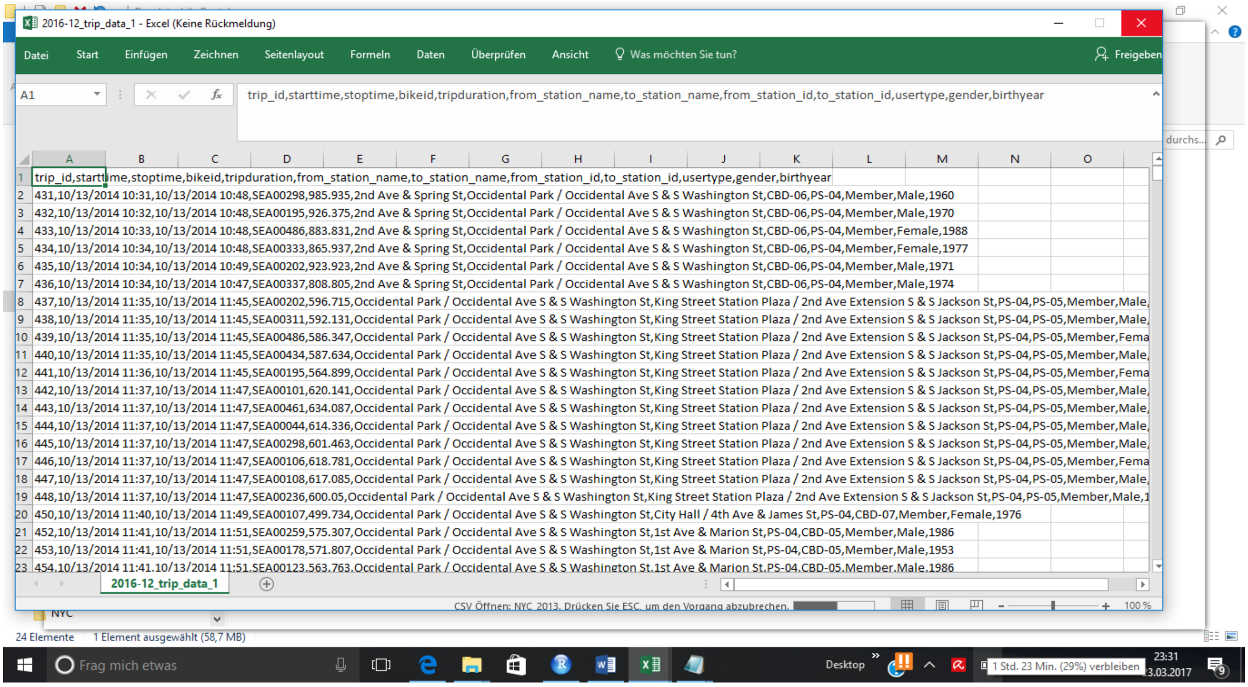
Weather Datasets were obtained trough homepages: - Wounderground.com

IV.1. Example city: Seattle Pronto

Aim: Merging of the datasets Tripdata and Weather

IV.1.1) INPUT: *2016-12_trip_data.csv* with 12 variables.

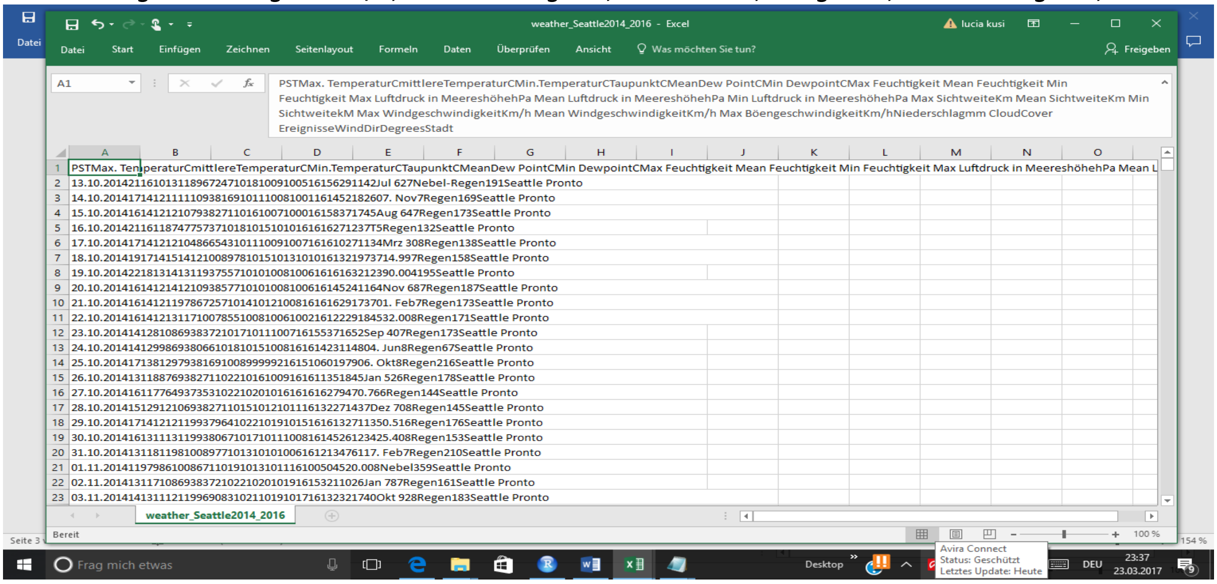
Column variables are:
trip_id, starttime, stoptime, bikeid, tripduration, from_station_name, to_station_name, from_station_id, to_station_id, usertype, gender, birthyear



IV.1.2.) INPUT: *weather_Seattle2014_2016.csv* with 24 variables

Column variables are:

PST, Max. TemperaturC, mittlereTemperaturC, Min.TemperaturC, TaupunktC, MeanDew PointC, Min DewpointC, Max Feuchtigkeit, Mean Feuchtigkeit, Min Feuchtigkeit, Max Luftdruck in MeereshoehehPa, Mean Luftdruck in MeereshoehehPa, Min Luftdruck in MeereshoehehPa, Max SichtweiteKm, Mean SichtweiteKm, Min SichtweitekM, Max WindgeschwindigkeitKm/h, Mean WindgeschwindigkeitKm/h, Max BoeengeschwindigkeitKm/h, Niederschlagmm, CloudCover, Ereignisse, WindDirDegrees, Stadt



IV.2. DATA PREPARATION:

IV.2.1) Conversion of Input tripdata to 13 variables. The inserted 13[th] column is *Stadtschluessel*. A renaming applied to *starttime* as *startdate.*

IV.2.2) Merging tripdata + weather into one merged single dataset.
New dataset produced: seattleMerged_reduced.csv with 21 variables.

- From the original file *2016-12_trip-data.csv* were considered 8 variables: *startdate, Stadtschluessel, trip_id, bikeid, tripduration, usertype, gender, birthyear.*

  Merging keys: *startdate* and *Stadtschluessel*

- From weather_Seattle2014_2016.csv were considered 15 variables: *PST , Max.TemperaturC, mittlereTemperaturC, Min.TemperaturC, TaupunktC, MeanDew PointC, Mean.Feuchtigkeit, Mean.Luftdruck_in_MeereshoehehPa, Mean.SichtweiteKm, Mean.WindgeschwindigkeitKm/h, Niederschlagmm, CloudCover, Ereignisse, WindDirDegrees, Stadt*

  The variables *PST* and *Stadt* were the secondary merging keys. That is why they do not appear with these names any more.

  PST contains the same variable as *startdate* and *Stadt* as *Stadtschluessel.*

  OUTPUT file *seattleMerged.csv*:

  startdate, Stadtschluessel, trip_id, bikeid, tripduration, usertype, gender, birthyear, Max.TemperaturC, mittlereTemperaturC, Min.TemperaturC, TaupunktC, MeanDew PointC, Mean.Feuchtigkeit, Mean.Luftdruck_in_MeereshoehehPa, Mean.SichtweiteKm, Mean.WindgeschwindigkeitKm/h, Niederschlagmm, CloudCover, Ereignisse, WindDirDegrees

seattleMerged_reduced.csv is as dataframe: seattle_data

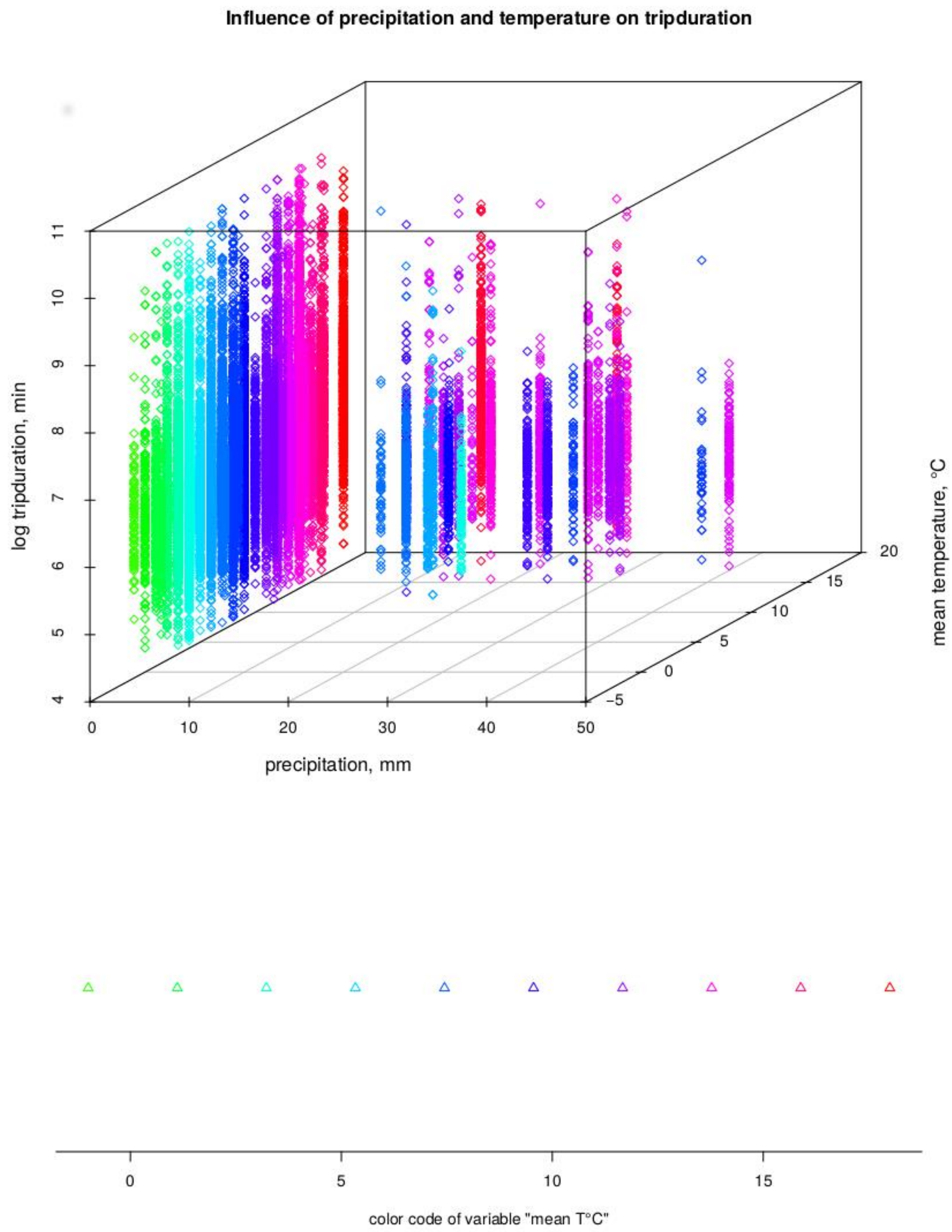
```
> str(seattle_data)
Classes ‘data.table’ and 'data.frame':236065 obs. of  21 variables:
 $ startdate                       : chr  "13.10.2014" "13.10.2014" "13.10.
2014" "13.10.2014" ...
 $ Stadtschluessel                 : chr  "Seattle Pronto" "Seattle Pronto"
"Seattle Pronto" "Seattle Pronto" ...
 $ trip_id                         : int  908 906 905 904 903 902 901 900 8
99 898 ...
 $ bikeid                          : chr  "SEA00230" "SEA00392" "SEA00341"
"SEA00117" ...
 $ tripduration                    : int  27583 388 863 289 494 306 125 198
784 643 ...
 $ usertype                        : chr  "casual" "subscriber" "casual" "s
ubscriber" ...
 $ gender                          : chr  "" "0" "" "0" ...
 $ birthyear                       : int  NA 1984 NA 1981 NA 1987 1981 1982
1988 1984 ...
 $ Max.TemperaturC                 : int  21 21 21 21 21 21 21 21 21 21 ...
 $ mittlereTemperaturC             : num  16 16 16 16 16 16 16 16 16 16 ...
 $ Min.TemperaturC                 : int  10 10 10 10 10 10 10 10 10 10 ...
 $ TaupunktC                       : int  13 13 13 13 13 13 13 13 13 13 ...
 $ MeanDew PointC                  : int  11 11 11 11 11 11 11 11 11 11 ...
 $ Mean.Feuchtigkeit               : int  72 72 72 72 72 72 72 72 72 72 ...
 $ Mean.Luftdruck_in_MeereshoehehPa: int  1009 1009 1009 1009 1009 1009 100
9 1009 1009 1009 ...
```

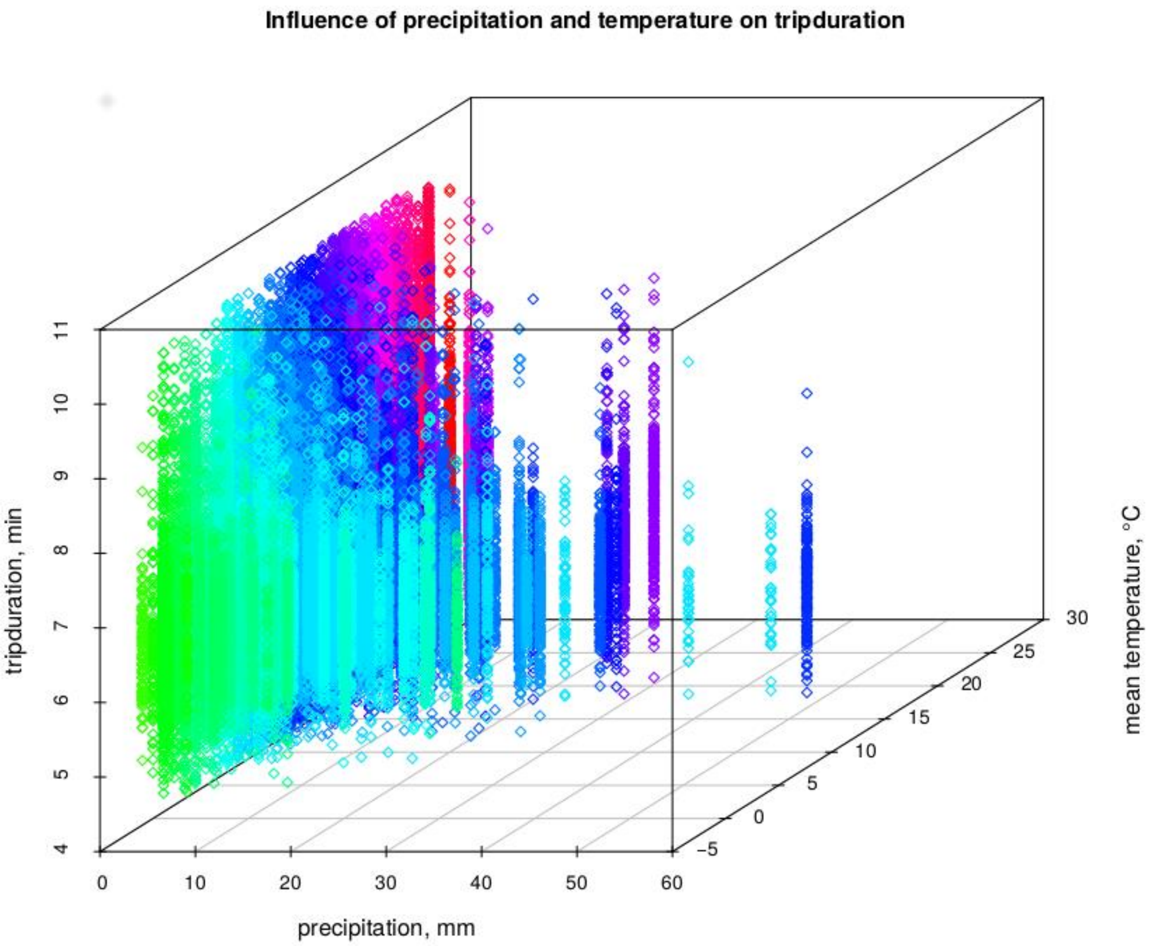
```
 $ Mean.SichtweiteKm                : int  15 15 15 15 15 15 15 15 15 15 ...
 $ Mean.WindgeschwindigkeitKm.h     : int  11 11 11 11 11 11 11 11 11 11 ...
 $ Niederschlagmm                   : num  7.62 7.62 7.62 7.62 7.62 7.62 7.6
2 7.62 7.62 7.62 ...
 $ CloudCover                       : int  7 7 7 7 7 7 7 7 7 7 ...
 $ Ereignisse                       : chr  "Nebel-Regen" "Nebel-Regen" "Nebe
l-Regen" "Nebel-Regen" ...
 $ WindDirDegrees                   : int  191 191 191 191 191 191 191 191 1
91 191 ...
 - attr(*, ".internal.selfref")=<externalptr>
```

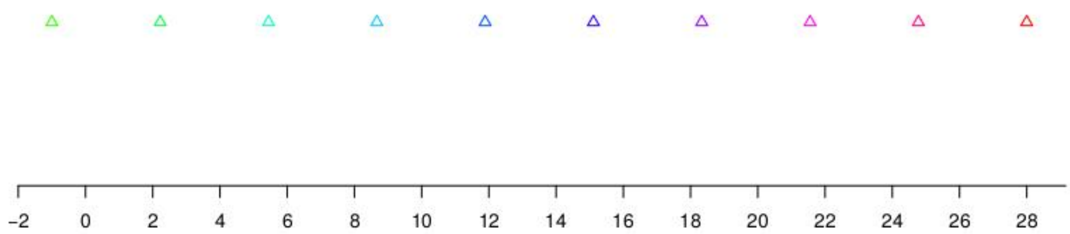
IV.3. DESCRIPTIVE PLOTS.

IV.3.1) OUTPUT Plot in 3D with library "scatterplot3d"

Aim : the representation of tripduration, mean temperature and precipitation.
The coloured parameter is mean temperature. They are respectively the following variables: *tripduration*, *mittlereTemperaturC, Niederschlagmm*

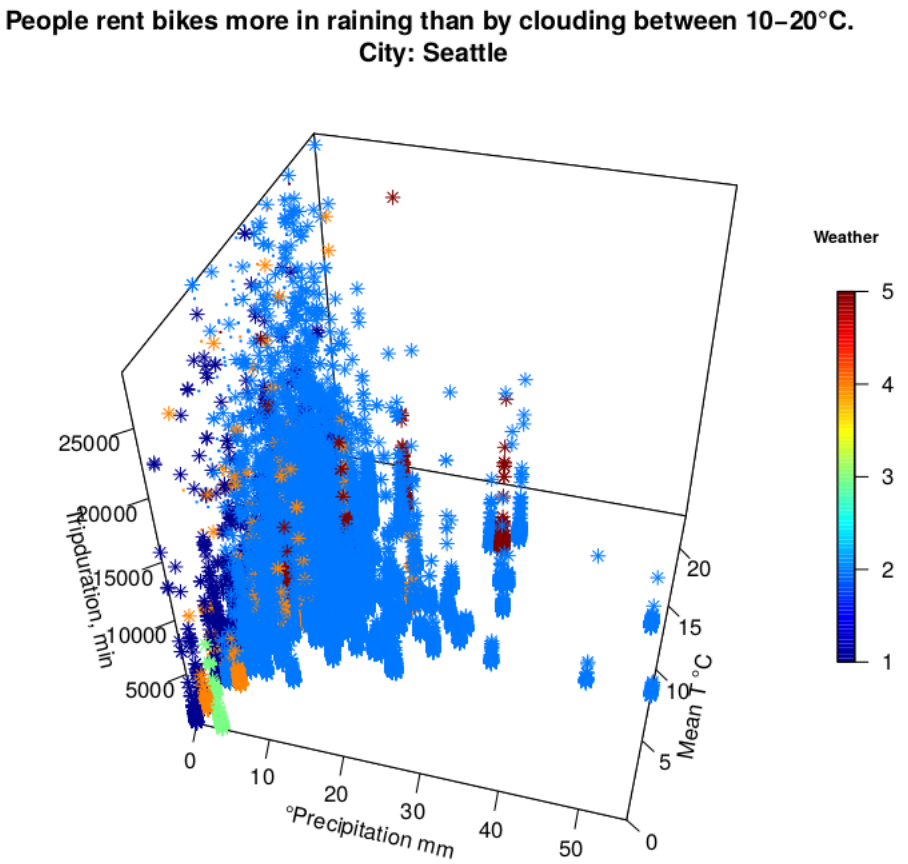



Influence of precipitation and temperature on tripduration
tripduration, min
4
5
6
7
8
9
10
11
precipitation, mm
0
10
20
30
40
50
60
mean temperature, °C
−5
0
5
10
15
20
25
30


−2
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28

IV.3.2) OUTPUT Plot in 3D with library “Plot3D”

Aim : the representation of tripduration, mean temperature and precipitation with weather as coloured parameter. They are respectively the following variables: t*ripduration, mittlereTemperaturC, Niederschlagmm, Ereignisse*



IV.4. DATA ANALYSIS.

IV.4.1) MODELL POISSON.

a) OUTPUT POISSON: tripduration ~ usertype + Ereignisse

Table 2.- Coefficientes of poisson modell

```
 (Intercept)         seattleData$usertypecasual      seattleData$Ereignisse
 611.8199306                    3.6576649                     0.9645039
```

Table 3: Summary of the poisson model

```
Call:
glm(formula = seattleData$tripduration ~ seattleData$usertype + 
    seattleData$Ereignisse, family = poisson, data = seattleData)

Deviance Residuals: 
   Min       1Q   Median       3Q      Max  
-62.62   -19.06    -8.27     3.45   403.96  

Coefficients:
                             Estimate  Std. Error  z value   Pr(>|z|)    
(Intercept)                 6.416e+00   1.190e-04  53915.1    <2e-16 ***
seattleData$usertypecasual  1.297e+00   1.304e-04   9946.2    <2e-16 ***
seattleData$Ereignisse     -3.614e-02   5.912e-05   -611.3    <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 371702207  on 235826  degrees of freedom
Residual deviance: 256867894  on 235824  degrees of freedom
  (238 observations deleted due to missingness)
AIC: 258849665

Number of Fisher Scoring iterations: 5
```

b) OUTPUT POISSON MODEL: tripduration ~ usertype + Gender + Ereignisse

TABLE 4.- Coefficients of model poisson

```
(Intercept)      seattleData$usertypecasual          seattleData$gender 
586.016923                  2.771781                     1.173598 

seattleData$Ereignisse 
 0.964996 
```

TABLE 5.- Summary regression poisson

```
Call:
glm(formula = seattleData$tripduration ~ seattleData$usertype + 
    seattleData$gender + seattleData$Ereignisse, family = poisson, 
    data = seattleData)

Deviance Residuals: 
   Min     1Q    Median      3Q      Max  
-62.61  -18.96    -8.11    3.48   406.81  
```

```
Coefficients:
                             Estimate Std. Error  z value   Pr(>|z|)    
(Intercept)                 6.373e+00  1.339e-04  47613.5   <2e-16 ***
seattleData$usertypecasual  1.019e+00  3.860e-04   2641.2   <2e-16 ***
seattleData$gender          1.601e-01  2.121e-04    754.6   <2e-16 ***
seattleData$Ereignisse     -3.563e-02  5.912e-05   -602.7   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 371702207  on 235826  degrees of freedom
Residual deviance: 256321608  on 235823  degrees of freedom
  (238 observations deleted due to missingness)
AIC: 258303381

Number of Fisher Scoring iterations: 5
```

IV.4.2) MODELL QUASIPOISSON.

a) OUTPUT QUASIPOISSON: tripduration ~ usertype + Ereignisse

TABLE 5.- Coefficients of quasipoisson regression

```
(Intercept)      seattleData$usertypecasual      seattleData$Ereignisse 
 611.8199306                3.6576649                   0.9645039 
```

TABLE 6.- Summary of modell quassipoisson

```
Call:
glm(formula = seattleData$tripduration ~ seattleData$usertype + 
    seattleData$Ereignisse, family = quasipoisson, data = seattleData)

Deviance Residuals: 
   Min       1Q   Median       3Q      Max  
-62.62   -19.06   -8.27      3.45   403.96  

Coefficients:
                            Estimate Std. Error  t value   Pr(>|t|)    
(Intercept)                 6.416438   0.005425  1182.65    <2e-16 ***
seattleData$usertypecasual  1.296825   0.005944   218.18    <2e-16 ***
seattleData$Ereignisse     -0.036141   0.002695   -13.41    <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 2078.287)

    Null deviance: 371702207  on 235826  degrees of freedom
Residual deviance: 256867894  on 235824  degrees of freedom
  (238 observations deleted due to missingness)
AIC: NA

Number of Fisher Scoring iterations: 5
```

b) OUTPUT QUASIPOISSON: tripduration ~ usertype + gender + Ereignisse

TABLE 6.- Coefficients of the quasipoisson model

```
(Intercept)         seattleData$usertypecasual          seattleData$gender
 586.016923                  2.771781                     1.173598

seattleData$Ereignisse
      0.964996
```

TABLE 7.- Summary of the quasipoisson model

```
Call:
glm(formula = seattleData$tripduration ~ seattleData$usertype +
    seattleData$gender + seattleData$Ereignisse, family = quasipoisson,
    data = seattleData)

Deviance Residuals:
   Min       1Q    Median      3Q      Max
-62.61    -18.96    -8.11     3.48    406.81


Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 6.373349   0.006099 1045.04   <2e-16 ***
seattleData$usertypecasual  1.019490   0.017586   57.97   <2e-16 ***
seattleData$gender          0.160074   0.009665   16.56   <2e-16 ***
seattleData$Ereignisse     -0.035631   0.002693  -13.23   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 2075.854)

    Null deviance: 371702207  on 235826  degrees of freedom
Residual deviance: 256321608  on 235823  degrees of freedom
  (238 observations deleted due to missingness)
AIC: NA

Number of Fisher Scoring iterations: 5
```

IV.4.3) OUTPUT NEGATIVE BINOMIAL REGRESSION: tripduration ~ usertype+gender+Ereignisse

TABLE 8- Coefficients of the negative binomial regression

```
               (Intercept)  seattleData$usertypecasual         seattleData$gender
              579.410289                    2.740876                     1.183037
    seattleData$Ereignisse
                0.975630
```

TABLE 9.- Summary of the negative binomial regression

```
Call:
glm.nb(formula = seattleData$tripduration ~ seattleData$usertype +
    seattleData$gender + seattleData$Ereignisse, data = seattleData,
    link = log, init.theta = 1.736212596)

Deviance Residuals:
    Min        1Q    Median       3Q       Max
-3.0194   -0.9017   -0.4010   0.1681   12.6026
```

```
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                  6.362011   0.004405 1444.18   <2e-16 ***
seattleData$usertypecasual  1.008278   0.013533   74.50   <2e-16 ***
seattleData$gender          0.168085   0.007063   23.80   <2e-16 ***
seattleData$Ereignisse     -0.024672   0.002454  -10.05   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.7362) family taken to be 2.84
7669)

    Null deviance: 430350  on 235826  degrees of freedom
Residual deviance: 257528  on 235823  degrees of freedom
  (238 observations deleted due to missingness)
AIC: 3680597

Number of Fisher Scoring iterations: 1
```

IV.4.4) LINEAR REGRESSION WITH TRANSFORMATION:

a) OUTPUT Modell with 4 variables, 1 of them is log(y) = log(*tripduration*)

TABLE 10.- Adjusted R-squared coeff.of a linear regression with transformat ion

```
adjusted.R.squared
         0.3216422
```

TABLE 11.- Summary of the linear regression with transformation

```
Call:
lm(formula = seattleData$tripduration ~ seattleData$usertype + 
    seattleData$gender + seattleData$Ereignisse, data = seattleData)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.1295 -0.4504 -0.0240  0.3945  4.1322 

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                  6.150872   0.002545 2416.49   <2e-16 ***
seattleData$usertypecasual  0.774117   0.007822   98.97   <2e-16 ***
seattleData$gender          0.149422   0.004081   36.61   <2e-16 ***
seattleData$Ereignisse     -0.021801   0.001418  -15.37   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.741 on 235823 degrees of freedom
  (238 observations deleted due to missingness)
Multiple R-squared:  0.3217,	Adjusted R-squared:  0.3216 
F-statistic: 3.727e+04 on 3 and 235823 DF,  p-value: < 2.2e-16
```

b) OUTPUT Modell with completed transformation

Both are transformed: log(x), log(y)

TABLE 12.- Adjusted R-squared coef.of a XY transformed linear regression

```
adjusted.R.squared
         0.3216422
```

```
Note: continuing in Backward regression (***) or (IV.4.4.b)
```

c) COMPARISON OF TWO TRANSFORMED LINEAR REGRESSION with ANOVA (Chi-square test)

TABLE 13.- Analysis of Variance Table

```
Model 1: seattleData$tripduration ~ seattleData$usertype +
                              seattleData$gender + seattleData$Ereignisse

Model 2: seattleData$tripduration ~ seattleData$usertype +
                              seattleData$gender + seattleData$Ereignisse
  Res.Df    RSS Df   Sum of Sq Pr(>Chi)
1 235823 129495
2 235823 129495  0 -9.8953e-10
```

IV.4.5) SHRINKAGE MODELS

A) OUTPUT BACKWARD MODEL (working with ONLY log(y):

TABLE 13.- AIC coefficients of the backward modell

```
Start:  AIC=-141360.6
seattleData$tripduration ~ seattleData$usertype + seattleData$gender +
    seattleData$Ereignisse

                         Df Sum of Sq    RSS     AIC
<none>                                129495 -141361
- seattleData$Ereignisse  1     129.8 129624 -141126
- seattleData$gender      1     736.0 130231 -140026
- seattleData$usertype    1    5378.9 134874 -131765
```

TABLE 14.- Summary of the backward modell

```
Call:
lm(formula = seattleData$tripduration ~ seattleData$usertype +
    seattleData$gender + seattleData$Ereignisse, data = seattleData)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1295 -0.4504 -0.0240  0.3945  4.1322

Coefficients:

                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 6.150872   0.002545 2416.49   <2e-16 ***
seattleData$usertypecasual  0.774117   0.007822   98.97   <2e-16 ***
seattleData$gender          0.149422   0.004081   36.61   <2e-16 ***
seattleData$Ereignisse     -0.021801   0.001418  -15.37   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.741 on 235823 degrees of freedom
  (238 observations deleted due to missingness)
```

```
Multiple R-squared:  0.3217,	Adjusted R-squared:  0.3216 
F-statistic: 3.727e+04 on 3 and 235823 DF,  p-value: < 2.2e-16
```

B) OUTPUT BACKWARD MODEL: working with log(x), log(y):

#Continuing from (***) or (IV.5.2):

TABLE 15.- AIC coefficients of the backward modell for a double transformed linear regression

```
Start:  AIC=-141360.65
seattleData$tripduration ~ seattleData$usertype + seattleData$gender + 
    seattleData$Ereignisse

                         Df Sum of Sq       RSS        AIC
<none>                                  129494.62 -141360.65
- seattleData$Ereignisse  1  129.7896 129624.41 -141126.40
- seattleData$gender      1  735.9710 130230.59 -140026.14
- seattleData$usertype    1 5378.8960 134873.52 -131764.94
```

TABLE 16.- Summary of the double transformed linear regression

```
Call:
lm(formula = seattleData$tripduration ~ seattleData$usertype + 
    seattleData$gender + seattleData$Ereignisse, data = seattleData)

Residuals:
       Min         1Q     Median         3Q        Max 
-3.1294881 -0.4504280 -0.0240024  0.3944781  4.1321537 


Coefficients:
                          Estimate   Std. Error    t value             Pr(>|t|)    
(Intercept)            6.150871601  0.002545378 2416.48671 < 0.0000000000000000222 
***
seattleData$usertype   1.116814274  0.011284106   98.97233 < 0.0000000000000000222 
***
seattleData$gender     0.149422193  0.004081479   36.60982 < 0.0000000000000000222 
***
seattleData$Ereignisse -0.021800689  0.001418022  -15.37401 < 0.0000000000000000222 
***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7410249 on 235823 degrees of freedom
  (238 observations deleted due to missingness)

Multiple R-squared:  0.3216508,	Adjusted R-squared:  0.3216422 

F-statistic: 37273.17 on 3 and 235823 DF,  p-value: < 0.0000000000000002220
4
```

IV.5. ADDITIONAL DESCRIPTIVE:

```
> summary(seattle_data$tripduration)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
     60     392     633    1202    1145   28790 

> sd(seattle_data$tripduration) #standard deviation of tripduration
[1] 2066.425

> density(seattle_data$tripduration) #density of tripduration

Call:
	density.default(x = seattle_data$tripduration)

Data: seattle_data$tripduration (236065 obs.);	Bandwidth 'bw' = 42.59

       x                 y            
 Min.   :  -67.78   Min.   :1.400e-09  
 1st Qu.: 7179.61   1st Qu.:3.047e-07  
 Median :14427.00   Median :1.003e-06  
 Mean   :14427.00   Mean   :3.446e-05  
 3rd Qu.:21674.39   3rd Qu.:5.125e-06  
 Max.   :28921.78   Max.   :1.150e-03  
```

V. CONCLUSION

Table.- Summary of the modell regressions

Modell	AIC	R^2
Poisson Regression (2 independent Variables)	258849665	
(3 independent Variables)	258303381	
Quasipoison Regression (2 ind. Var.)	No value	
(3 ind. Var.)	No value	
Negative Binomial Regression (3 ind.var.)	3680597	
Linear regression with transformation on y (3 ind.var.)		0.3216422
Linear regression with transformation on X and on Y (3 ind.var.)		0.3216422
Backward Regression of Y-transformed lineal regression (3 ind. Var.) : Ereignisse	-131765	0.3216
Gender	-140026	
Usertype	-131765	
Backward Regression of XY-transformed lineal regression (3 ind. Var.) : Ereignisse	-141126.40	0.3216422
Gender	-140026.14	
Usertype	-131764.94	

1.- For the Poisson regression it is observed less AIC value for the regression with 3 independent X variables (AIC=258303381) than for the regression with 2 ind. Variables.

2.- Looking at the Backward regression models both with 3 independent X variables, of them the first one has a transformed logY . Its AIC values are smaller than the AIC ones of the second regression with logX and logY.

3. |AIC(Y-transformed lineal)| < |AIC(XY-transformed lineal)|, therefore better.

4. Related to the the 4 linear regression models there is no difference in its quality, because R^2 for the 4 regressions is 0.3216422

5. Because R^2 values are too far from 1.0, it is deduced, that the data do not follow a linear regression.

6. Regression Poisson, Quassi-poisson and negative polynomial values give us test values, that say, that the data do not fit very well in these models. With this affirmation, we are close to the conclusion given by the Washington Citybike System. See links: https://www.kaggle.com/c/bike-sharing-demand/data.

7. Regarding New York City data, tough I did not mention it before, it is important to say, it was a huge dataset of 8 Millions of GB. It was not possible to be analysed it further through my other group fellows. Also I tried to present my advances with NYC now, but because of lack of time I am not going to present this now.

Recomendation:

1. Repeat the data analysis considering more than 3 independent variables. It could be considered 5 variables.
2. Decide for the new one, if R^2 is greater and AIC value is smaller than the actual value.
3. New York city data should be handled with R and mainly a bigger database such as SQL.

VI. APPENDICES

- **Scripts: see extra file Quantlets**
- **Graphs : see in Ordners: Graphs Descriptive and Graphs Modells**