Chicago Weather Data Preparation and Analysis

Advaith Cheruvu

8/31/2021

library(rmarkdown) render("W2L2Cheruvu.Rmd", output_format = "pdf_document") ## Abstract

This report analyzes the weather data for Chicago from January 1987 to December 2005. The measured weather parameters were daily temperature (degrees Celsius), dewpoint (degrees Celsius), amount of PM2.5, amount of PM10, amount of ozone, and amount of nitrogen trioxide. Data munging, cleaning and normalization were done in order to prepare the dataset for further analysis. Histograms of numerical data were made.

Clean up and Set up

Before importing the data, the workspace must be cleared and the working directory must be set up.

```
# Clearing the workspace and setting the working directory
rm(list=ls())
setwd("/Users/advai/Documents/DSFS")
```

Obtaining Functions and Installing Packages

"myfunctions.R" is necessary for some of the later code to run. "tidyverse" has useful tools that help with plotting and transforming data.

```
# install and load libraries
source("myfunctions.R")
install.packages("tidyverse")
library(tidyverse)
```

Loading and Overviewing the Weather Data

Before cleaning the data, we must know what the dataset includes. The "names" command shows the names of the columns. The "summary" command gives a short summary of each column. The "str" command shows the type of data in each column. The "dim" command shows the number of columns and rows in the dataset. The "class" command shows that weather is a data frame.

```
# load weather data
weather <- read.csv(file = "C:\\Users\\advai\\Documents\\DSFS\\chicago.csv",h
eader=T)
# overview of the data set
names(weather)
```

```
## [1] "indx"       "city"       "tmpd"       "dptp"       "date"
## [6] "pm25tmean2" "pm10tmean2" "o3tmean2"   "no2tmean2"
```

```
summary(weather)
```

```
##       indx          city                tmpd             dptp
##  Min.   :   1   Length:6940        Min.   :-16.00   Min.   :-25.62
##  1st Qu.:1736   Class :character   1st Qu.: 35.00   1st Qu.: 27.00
##  Median :3470   Mode  :character   Median : 51.00   Median : 39.88
##  Mean   :3470                      Mean   : 50.31   Mean   : 40.34
##  3rd Qu.:5205                      3rd Qu.: 67.00   3rd Qu.: 55.75
##  Max.   :6940                      Max.   : 92.00   Max.   : 78.25
##                                    NA's   :1        NA's   :2
##      date             pm25tmean2      pm10tmean2        o3tmean2
##  Length:6940        Min.   : 1.70   Min.   :  2.00   Min.   : 0.1528
##  Class :character   1st Qu.: 9.70   1st Qu.: 21.50   1st Qu.:10.0729
##  Mode  :character   Median :14.66   Median : 30.28   Median :18.5218
##                     Mean   :16.23   Mean   : 33.90   Mean   :19.4355
##                     3rd Qu.:20.60   3rd Qu.: 42.00   3rd Qu.:27.0010
##                     Max.   :61.50   Max.   :365.00   Max.   :66.5875
##                     NA's   :4447    NA's   :242
##    no2tmean2
##  Min.   : 6.158
##  1st Qu.:19.654
##  Median :24.556
##  Mean   :25.232
##  3rd Qu.:30.139
##  Max.   :62.480
##
```

```
str(weather)
```

```
## 'data.frame':    6940 obs. of  9 variables:
##  $ indx      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ city      : chr  "chic" "chic" "chic" "chic" ...
##  $ tmpd      : num  31.5 33 33 29 32 40 34.5 29 26.5 32.5 ...
##  $ dptp      : num  31.5 29.9 27.4 28.6 28.9 ...
##  $ date      : chr  "1/1/87" "1/2/87" "1/3/87" "1/4/87" ...
##  $ pm25tmean2: num  NA NA NA NA NA NA NA NA NA NA ...
##  $ pm10tmean2: num  34 NA 34.2 47 NA ...
##  $ o3tmean2  : num  4.25 3.3 3.33 4.38 4.75 ...
##  $ no2tmean2 : num  20 23.2 23.8 30.4 30.3 ...
```

```
dim(weather)
```

```
## [1] 6940    9
```

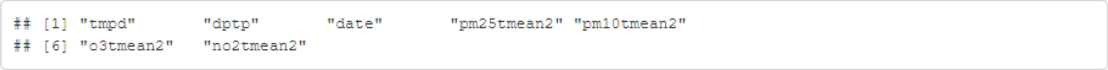
```
class(weather)
```

```
## [1] "data.frame"
```

Removing Columns

The index and city data in this dataset is unecessary so we must remove those columns.

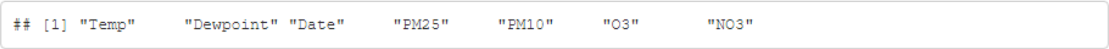
```
# removing index and city columns
weather <- select(weather, -1:-2)
names(weather)
```



```
## [1] "tmpd"       "dptp"       "date"       "pm25tmean2" "pm10tmean2"
## [6] "o3tmean2"   "no2tmean2"
```

Renaming Columns Some of the columns are ambigous or hard to read. Using these commands, we can rename them to something more readeable.

```
# renaming columns
colnames(weather)<-c("Temp", "Dewpoint","Date","PM25","PM10","O3","NO3")
names(weather)
```



```
## [1] "Temp"     "Dewpoint" "Date"     "PM25"     "PM10"     "O3"       "NO3"
```

Changing the Date Format

The date format is a character instead of a number. using these commands, we can change the date to a number, as well as make it more readable.

```
# Changing the date format
weather$Date <- as.Date(weather$Date, format= "%m/%d/%y")
```

Normalizing the PM data

The PM data is missing some values and it should be normalized (on a scale of 0-1) to prepare it for data analysis. This applies to both PM2.5 and PM10.

```
# Normalizing PM25 Data
weather$PM25 <- ifelse(is.na(weather$PM25), round(mean(weather$PM25, na.rm=TR
UE),3), weather$PM25)
PM25.normalize <- weather$PM25/max(weather$PM25)
weather$PM25 <- PM25.normalize
# Normalizing PM10 Data
weather$PM10 <- ifelse(is.na(weather$PM10), round(mean(weather$PM10, na.rm=TR
UE),3), weather$PM10)
PM10.normalize <- weather$PM10/max(weather$PM10)
weather$PM10 <- PM10.normalize
```

Taking out numeric data

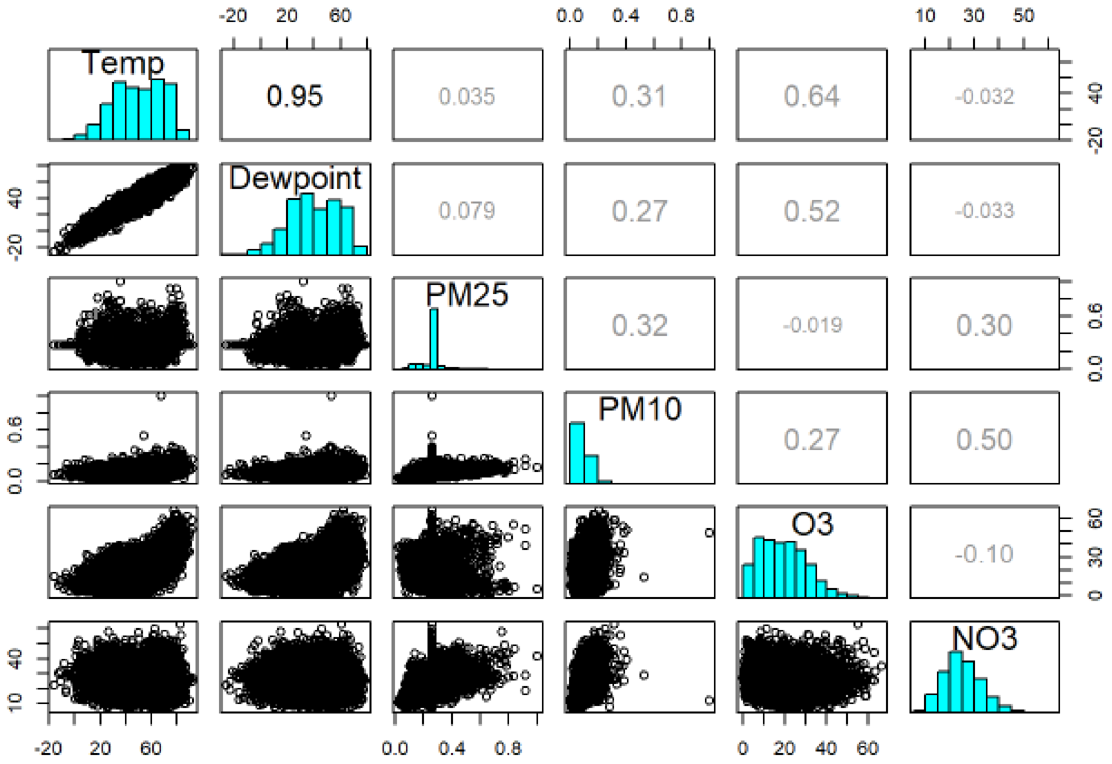
To prepare the data for analysis, we must provide the columns with numeric data. These commands seperate the numerica data (all the columns except for the date).

```
# Naming and removing numeric data
numerics <- weather[,c(1:2,4:7)]
```

Observing Correlation

This is pairwise correlation which helps to find out which data pairs have recognizeable patterns and are worth analyzing further.

```
# Looking at Correlation
pairs(numerics,upper.panel = panel.cor,diag.panel = panel.hist)
```

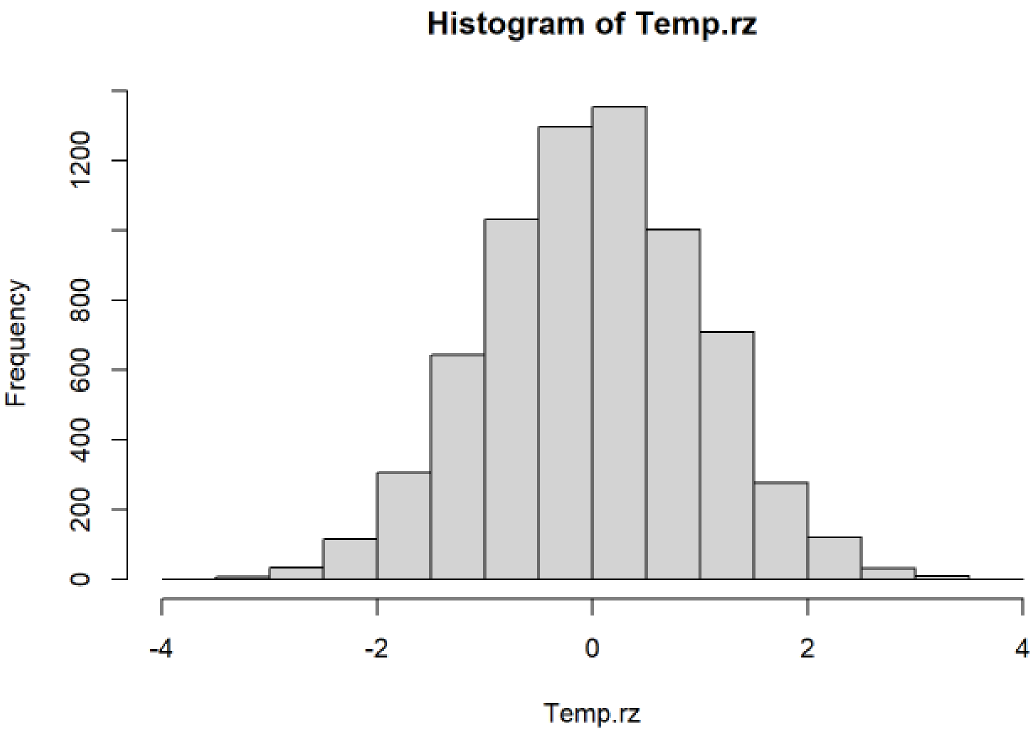


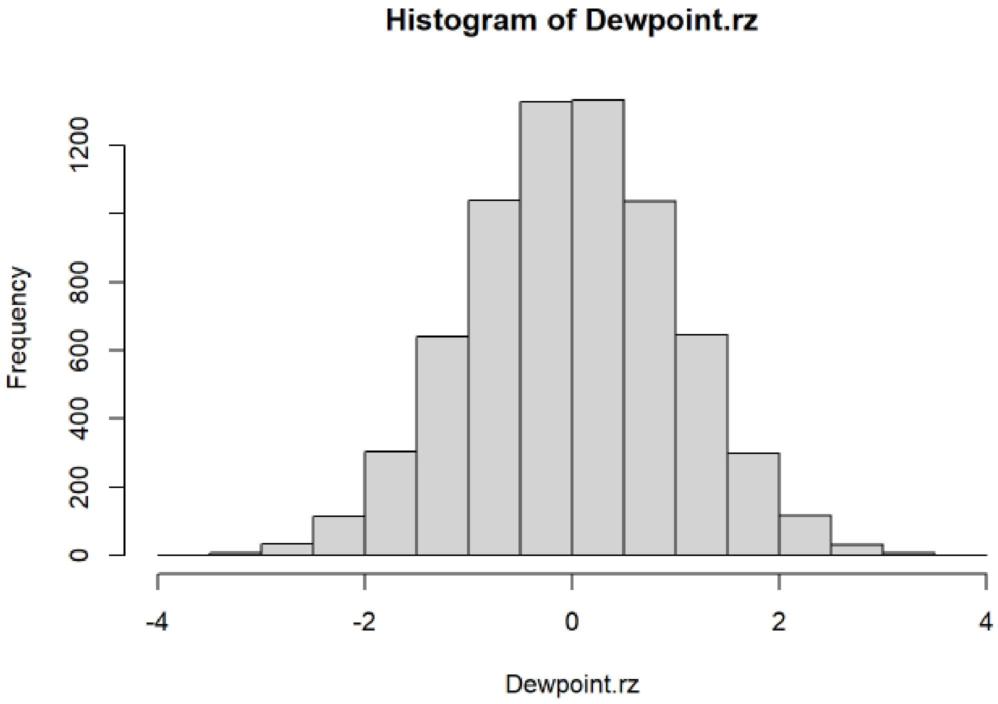
Histograms

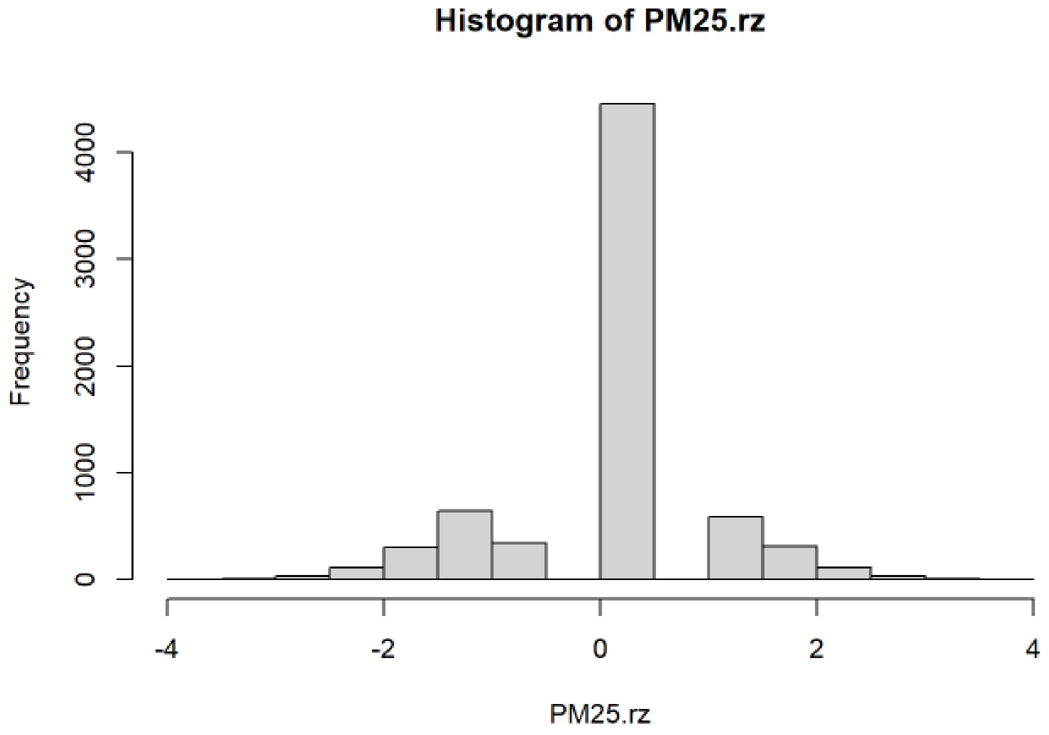
As part of Exploratory Data Analysis (EDA), transformations must be made to the data to normalize it and compare it to other data. This transformation is the rank z transformation which results in a normal distribution. This can be seen for all graphs except for PM25 where the correlation is not as strong.

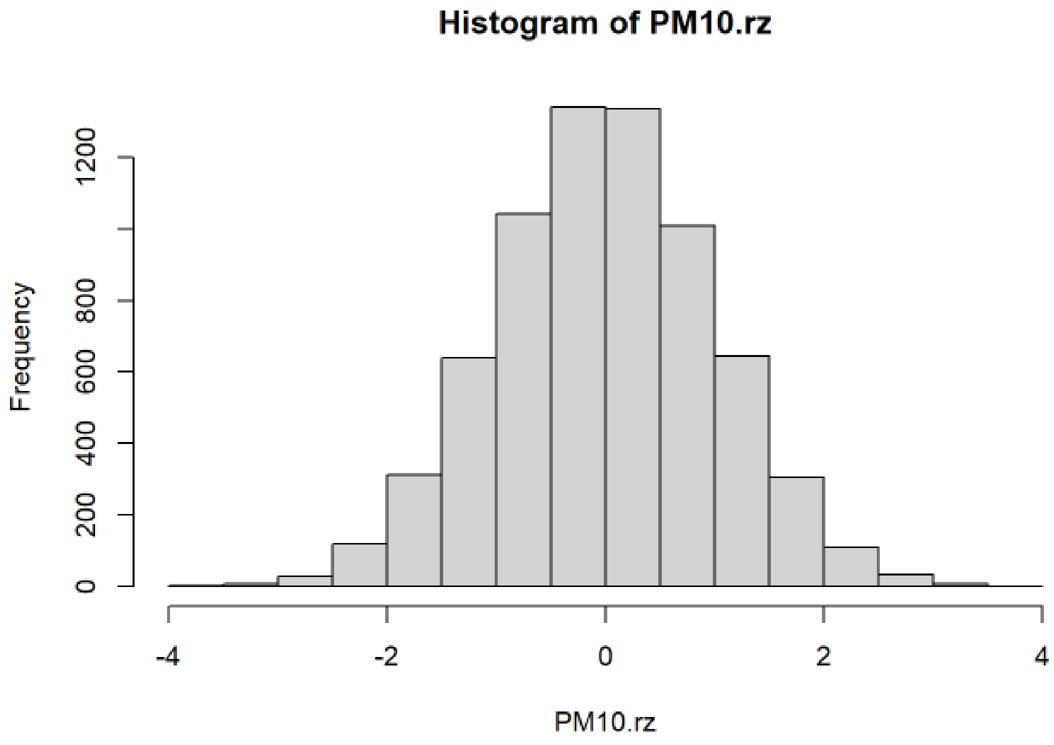
```
# Transformations and histogram plots
Temp.rz <- rz.transform(weather$Temp)
hist(Temp.rz)
Dewpoint.rz <- rz.transform(weather$Dewpoint)
hist(Dewpoint.rz)
PM25.rz <- rz.transform(weather$PM25)
hist(PM25.rz)
PM10.rz <- rz.transform(weather$PM10)
hist(PM10.rz)
O3.rz <- rz.transform(weather$O3)
hist(O3.rz)
```

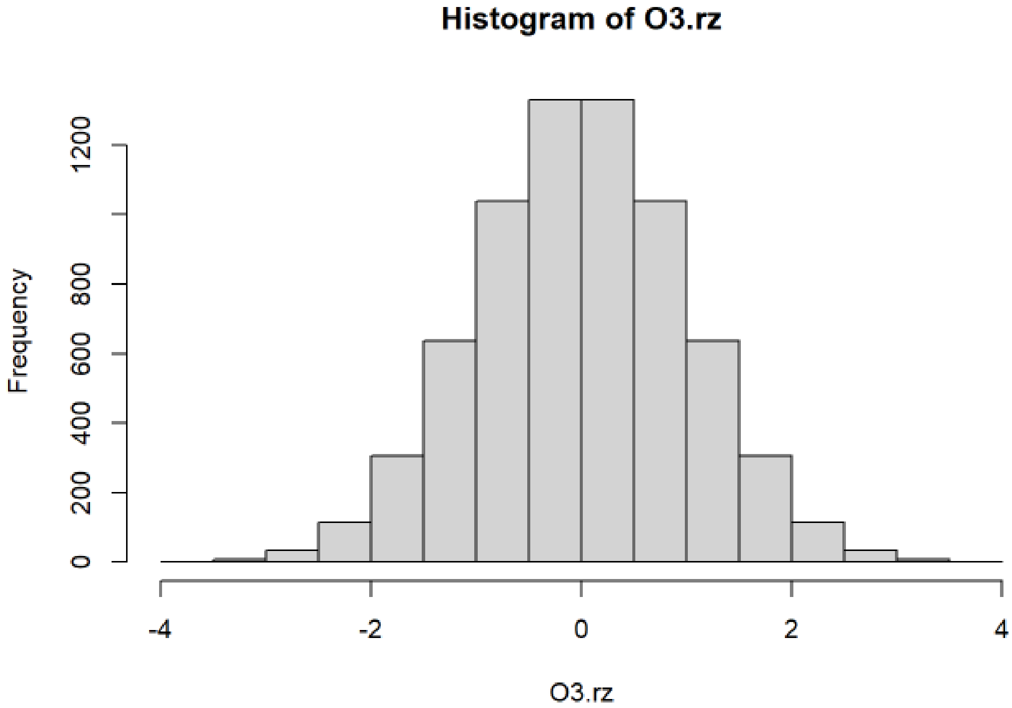
```
NO3.rz <- rz.transform(weather$NO3)
hist(NO3.rz)
```

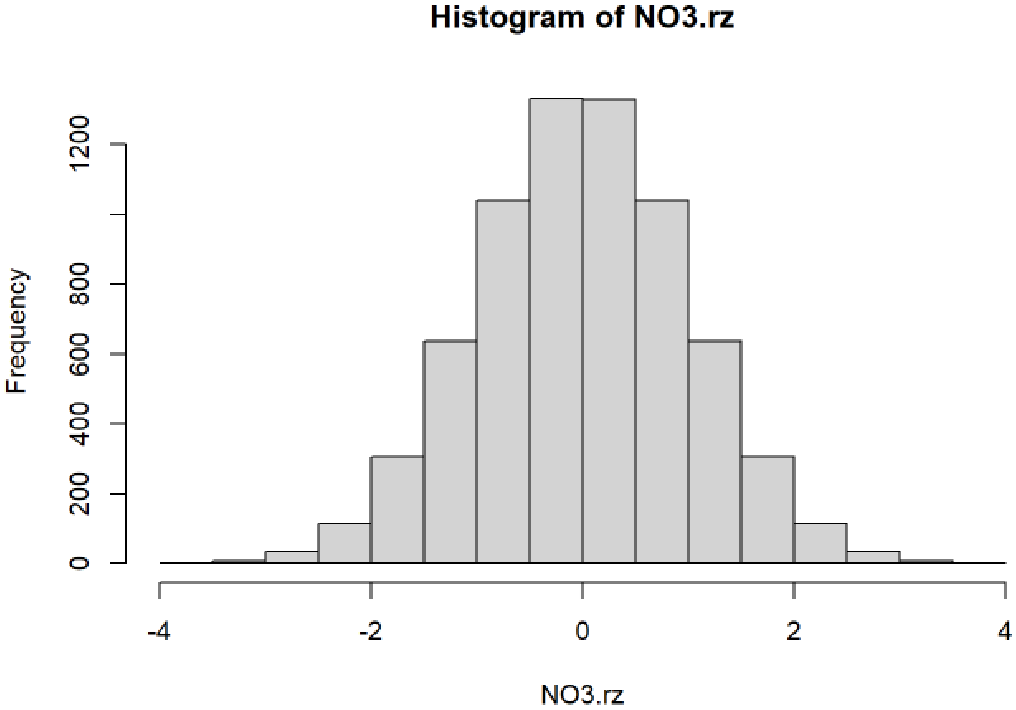





Histogram of PM25.rz
Frequency
0
1000
2000
3000
4000
-4
-2
0
2
4
PM25.rz


Histogram of PM10.rz
Frequency
0
200
400
600
800
1200
-4
-2
0
2
4
PM10.rz


Histogram of O3.rz
Frequency
0
200
400
600
800
1200
-4
-2
0
2
4
O3.rz


Histogram of NO3.rz
Frequency
0
200
400
600
800
1200
-4
-2
0
2
4
NO3.rz