Exploratory Data Analysis

using base R

Why Exploratory Data Analysis?

- better understand data
- detect data issues and bad data
- find patterns in data
- iterate exploration process to get better insight

Investigate variables

```
head(airquality)
str(airquality)
summary(airquality$Temp)
hist(airquality$Temp)

# count observations by month
table(airquality$Month)

# check missing values
table(is.na(airquality$Ozone))
barplot(table(is.na(airquality$Ozone)), main = "Ozone NA")
```

Investigate variables in batch

```
vars <- c("Ozone","Solar.R","Wind","Temp")
# use 2x2 plot grid as single plot
par(mfrow = c(2, 2))
# loop over variables and plot histogram
for (var in vars) {
  hist(airquality[[var]], main = var, xlab = "")
}
# restore default plot grid
par(mfrow = c(1, 1))

# count missing values in each column
sapply(airquality, function(x) sum(is.na(x)))
```

Cleaning bad data

```
# we assume Ozone sensors to have upper bound at 125 ppb
plot(airquality$Ozone)
abline(h = 125)

# we define 'bad data' for Ozone > 125
airquality[airquality$Ozone > 125 &
!is.na(airquality$Ozone), ]

# we define fix for bad data as value 125
airquality[airquality$Ozone > 125 &
!is.na(airquality$Ozone), "Ozone"] <- 125
```

Data transformation

```
# convert temperature from Fahrenheit to Celsius
FtoC <- function(x) (x - 32) * (5/9)
airquality$Temp <- FtoC(airquality$Temp)

# add month names
month.name
airquality$Month.Name <- month.name[airquality$Month]
# keep month name as factor type to get proper order on plot
airquality$Month.Name <- factor(airquality$Month.Name,
levels = month.name)
airquality$Month.Name <- droplevels(airquality$Month.Name)
```

Data insight - seasonality

```
# mean solar radiation by month
aggregate(Solar.R ~ Month.Name, data = airquality, FUN =
mean)
boxplot(Solar.R ~ Month.Name, data = airquality, xlab =
"Month", ylab = "Solar radiation (lang)")

# mean temperature by month
aggregate(Temp ~ Month.Name, data = airquality, FUN = mean)
boxplot(Temp ~ Month.Name, data = airquality, xlab =
"Month", ylab = "Temperature (C)")
```

Data insight - Ozone variable

```
plot(airquality$Temp, airquality$Ozone)
plot(airquality$Temp, airquality$Ozone, xlab="Temp (C)",
ylab="Ozone (ppb)", main="NYC May-Sep '73 air quality")
ozone_temp_lm <- lm(airquality$Ozone ~ airquality$Temp)
summary(ozone_temp_lm)
abline(ozone_temp_lm, col = "red")

# check other variables
pairs(Ozone ~ Temp + Wind + Solar.R, data = airquality)
```

Data insight - Ozone variable batch lm

```
# add linear model to pairs plot
pointslm <- function(x, y, ...){
  model <- lm(y ~ x)
  points(x, y)
  abline(a = model$coefficients[1], b =
model$coefficients[2], col = "red")
}
pairs(Ozone ~ Temp + Wind + Solar.R + Month, data =
airquality, lower.panel = pointslm, upper.panel = NULL)
```

Homework: read about and play with

- `sapply` and `lapply` functions to hide loops and make code cleaner
- subsetting data frames by integer and logical types
- `plot` and `pairs` functions
- graphic options like `par(mfrow = c(2, 1))`

Questions?

Jan Gorecki: github.com/jangorecki

Contact: **jan61ji@gmail.com**