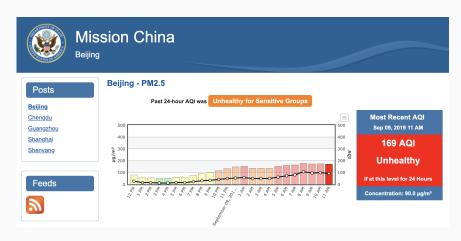# Exploring data #1

Exploring data

- How to explore depends on data type / class
- Data exploration includes simple statistics (max, mean, min, standard deviation)
- Data exploration include plots

Example data—Beijing air quality

Source: http://www.stateair.net/web/post/1/1.HTML

The U.S. embassy in Beijing has an air-quality monitoring station that tracks the level of certain pollutants in China's notoriously smoggy capital — and then broadcasts results via Twitter. Most tweets from the sober-minded scientists behind @BeijingAir look like this:

11-17-2010; 10:00; PM2.5; 154.0; 204; Very Unhealthy // Ozone; 0.2; 0

But yesterday a new reading was pronounced, one not listed on the US EPA's usual air-quality index:

11-19-2010; 02:00; PM2.5; 562.0; 500; Crazy Bad

Source: https://foreignpolicy.com/2010/11/19/beijing-air-crazy-bad/

Example data—Beijing air quality

Find out more:

https://www.wired.com/2015/03/opinion-us-embassy-beijing-tweeted-clear-air/

https://www.theguardian.com/environment/blog/2010/nov/19/crazy-bad-beijing-air-pollution

https://www.sciencemag.org/news/2018/04/rooftop-sensors-us-embassies-are-warning-world-about-crazy-bad-air-pollution

Example data—Beijing air quality

Download the data here.

Then you can read this data into your R session:

```r
library("readr")

beijing_pm_raw <- read_csv("data/Beijing_2017_HourlyPM25.csv",
                          skip = 3)
```

Example data—Beijing air quality

```
head(beijing_pm_raw, n = 3)

## # A tibble: 3 x 11
##   Site  Parameter `Date (LST)`  Year Month   Day
##   <chr> <chr>     <chr>        <dbl> <dbl> <dbl>
## 1 Beij~ PM2.5     1/1/2017 0:~  2017     1     1
## 2 Beij~ PM2.5     1/1/2017 1:~  2017     1     1
## 3 Beij~ PM2.5     1/1/2017 2:~  2017     1     1
## # ... with 5 more variables: Hour <dbl>,
## #   Value <dbl>, Unit <chr>, Duration <chr>, `QC
## #   Name` <chr>
```

Example data—Beijing air quality

Let's clean this up a bit:

```r
library("dplyr")

## Warning: package 'dplyr' was built under R version
## 3.5.2

beijing_pm <- beijing_pm_raw %>%
  rename(sample_time = `Date (LST)`,
         value = Value,
         qc = `QC Name`) %>%
  select(sample_time, value, qc)
head(beijing_pm, n = 3)

## # A tibble: 3 x 3
##   sample_time    value qc
##   <chr>          <dbl> <chr>
## 1 1/1/2017 0:00    505 Valid
## 2 1/1/2017 1:00    485 Valid
```

Example data—Beijing air quality

This code will add the AQI categories:

```
beijing_pm <- beijing_pm %>%
  mutate(aqi = cut(value,
                   breaks = c(0, 50, 100, 150, 200,
                              300, 500, Inf),
                   labels = c("Good", "Moderate",
                              "Unhealthy for Sensitive Groups",
                              "Unhealthy", "Very Unhealthy",
                              "Hazardous", "Beyond Index")))
head(beijing_pm, n = 2)

## # A tibble: 2 x 4
##   sample_time    value qc    aqi
##   <chr>          <dbl> <chr> <fct>
## 1 1/1/2017 0:00    505 Valid Beyond Index
## 2 1/1/2017 1:00    485 Valid Hazardous
```