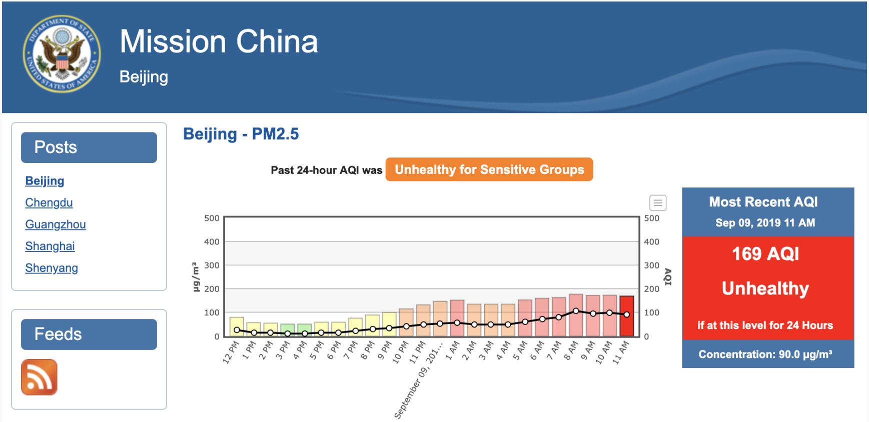
Exploring data #1

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



Mission China

Beijing

Posts

Beijing

Chengdu

Guangzhou

Shanghai

Shenyang

Feeds

Beijing - PM2.5

Past 24-hour AQI was **Unhealthy for Sensitive Groups**

Most Recent AQI

Sep 09, 2019 11 AM

169 AQI

Unhealthy

if at this level for 24 Hours

Concentration: 90.0 µg/m³

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

Example data—Beijing air quality

Download the data here.

Then you can read this data into your R session:

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

Example data—Beijing air quality

```r
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")
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
## 3 1/1/2017 2:00   466 Valid
```

Example data—Beijing air quality

This code will add the AQI categories:

```r
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
```