Visualizing the nature of data sets

Abhijit Dasgupta, PhD

The nature of a data set

Data characteristics

Some of the things we care about in a data set are

- Nature of each column
- Missing data patterns
- Correlation patterns

The **visdat** package and the **naniar** package help us with visualizing these.

Without visualization

```
summary(airquality)
```

```
     Ozone           Solar.R          Wind      
 Min.   :  1.00   Min.   :  7.0   Min.   : 1.700  
 1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400  
 Median : 31.50   Median :205.0   Median : 9.700  
 Mean   : 42.13   Mean   :185.9   Mean   : 9.958  
 3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500  
 Max.   :168.00   Max.   :334.0   Max.   :20.700  
 NA's   :37       NA's   :7                       
     Month            Day      
 Min.   :5.000   Min.   : 1.0  
 1st Qu.:6.000   1st Qu.: 8.0  
 Median :7.000   Median :16.0  
 Mean   :6.993   Mean   :15.8  
 3rd Qu.:8.000   3rd Qu.:23.0  
 Max.   :9.000   Max.   :31.0  
```

```
glimpse(airquality, width=40)
```

```
Rows: 153
Columns: 6
$ Ozone   <int> 41, 36, 12, 18, NA, 2…
$ Solar.R <int> 190, 118, 149, 313, N…
$ Wind    <dbl> 7.4, 8.0, 12.6, 11.5,…
$ Temp    <int> 67, 72, 74, 62, 56, 6…
$ Month   <int> 5, 5, 5, 5, 5, 5, 5, …
$ Day     <int> 1, 2, 3, 4, 5, 6, 7, …
```

These give us a variable-by-variable view.

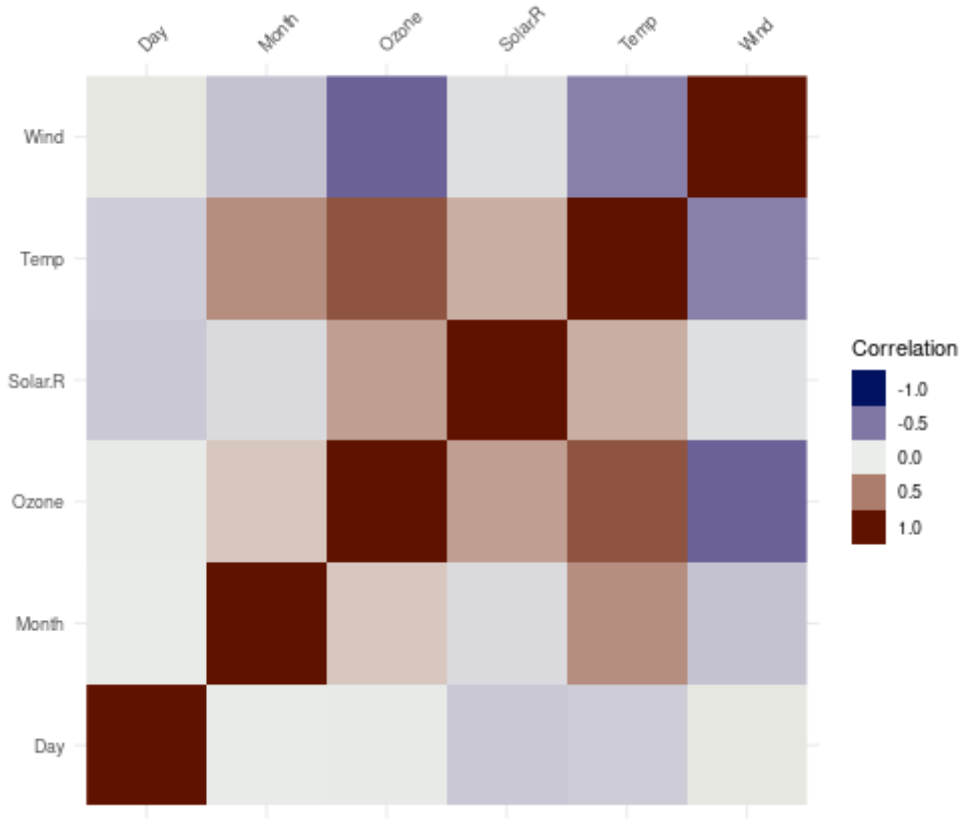
Visualizing a dataset

```
visdat::vis_dat(airquality)
```

- What kinds of variables are in the dataset
- Which elements are missing
- A sense of missing patterns

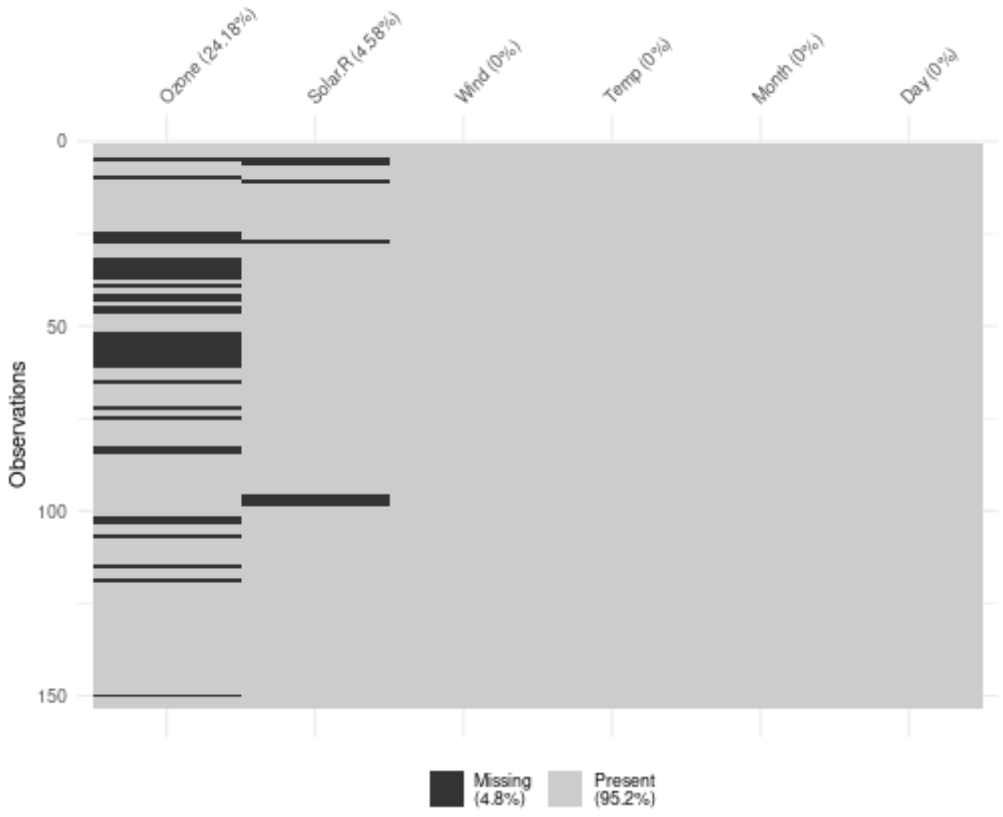
Correlation patterns

```
visdat::vis_cor(airquality)
```



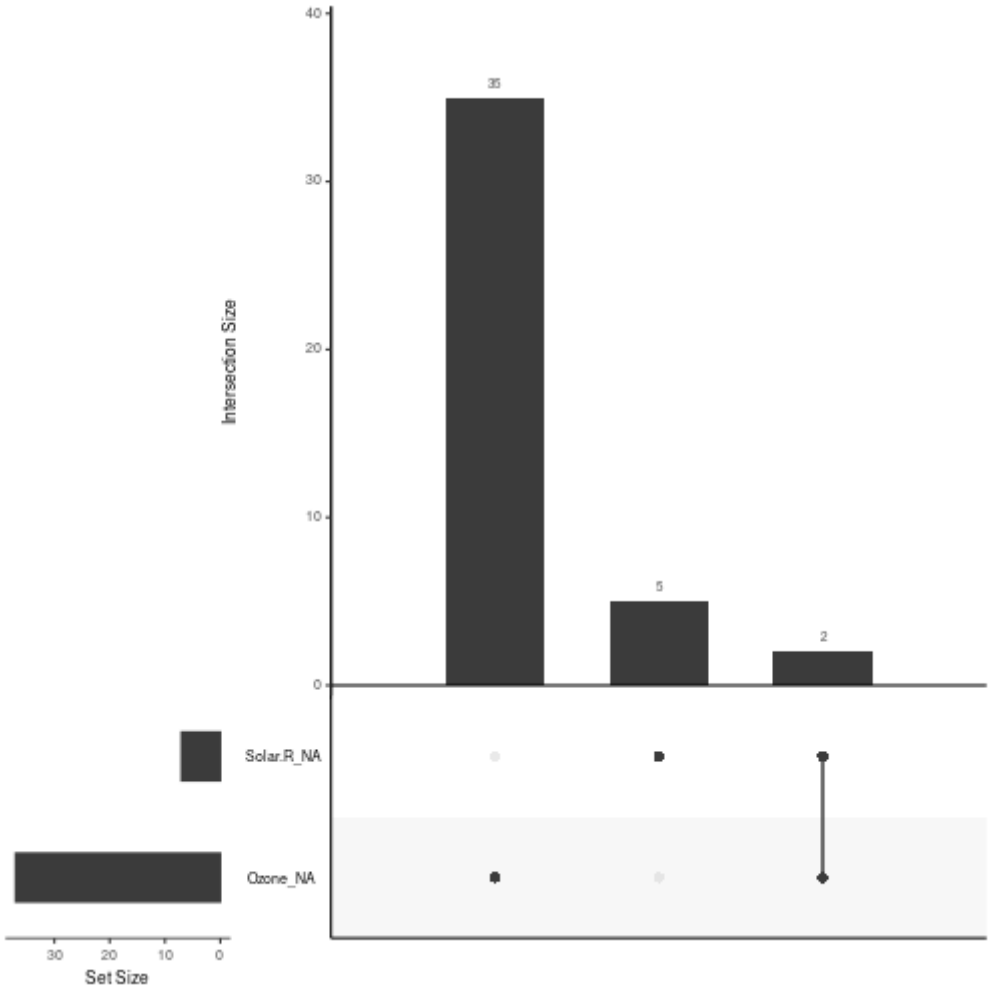
Focus on missing data patterns

```
visdat::vis_miss(airquality)
```

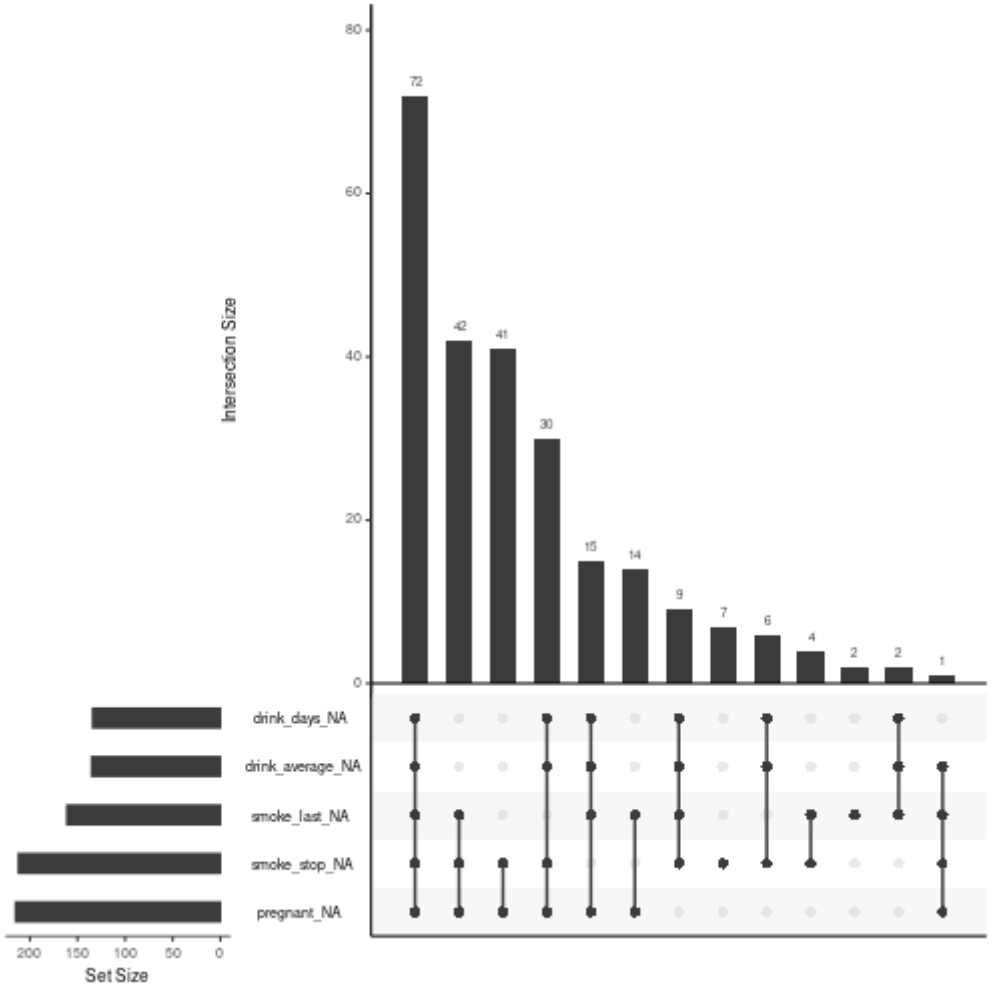


A deeper look at missing data

```
library(naniar)
gg_miss_upset(airquality)
```



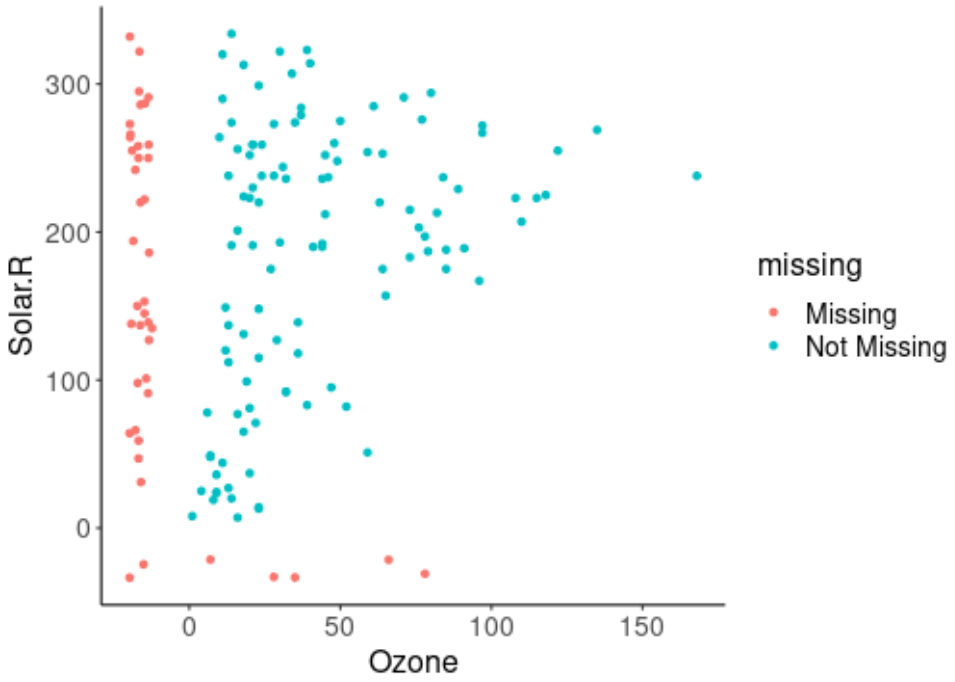
```
gg_miss_upset(riskfactors)
```



Missing at random?

Does missingness in one variable depend on values of another variable?

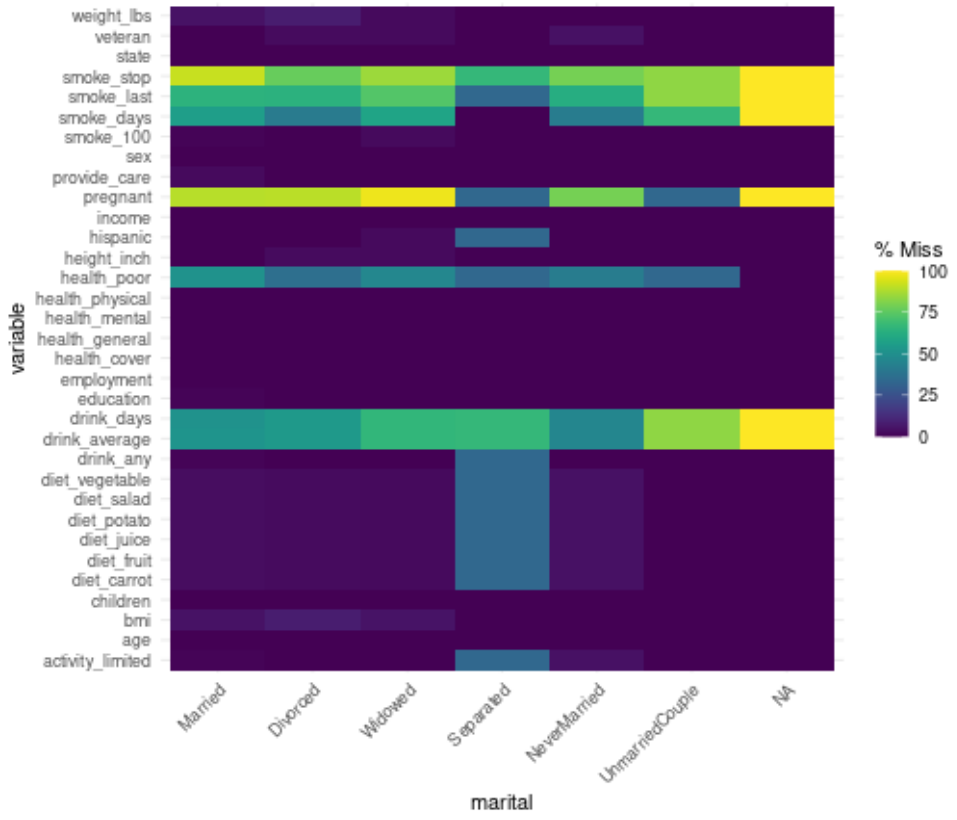
```
ggplot(airquality,
       aes(Ozone, Solar.R))+
  geom_miss_point()
```



The red points are the values of one variable when the other variable is missing

Missing at random?

```
gg_miss_fct(x = riskfactors, fct=marital)
```



Percent missing in each variable by levels of a factor

What you're looking for is relatively even colors across

Further exploration

1. The **naniar** website