Part 2: Basic Inferential Data Analysis

Endri Raco

4/11/2020

Importing data

```
# Load ToothGrowth data
library(datasets)
data(ToothGrowth)
```

Exploratory data analyses

Overview of the data

Let's understand the dimensions of the dataset, variable names, overall missing summary and data types of each variables

```
## overview of the data;
glimpse(ToothGrowth)
```

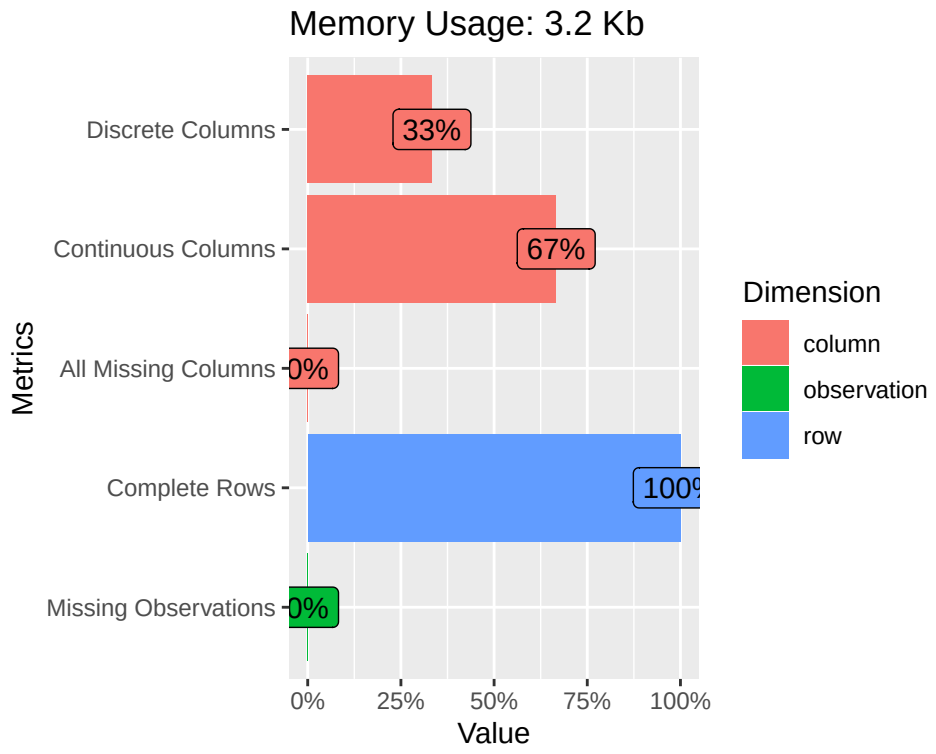
```
## Observations: 60
## Variables: 3
## $ len  <dbl> 4.2, 11.5, 7.3, 5.8, 6.4, 10.0, 11.2, 11.2, 5.2, 7.0, 16.5, 16...
## $ supp <fct> VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, VC...
## $ dose <dbl> 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 1.0, 1.0, 1....
```

```
## structure of the data
introduce(ToothGrowth)
```

```
##   rows columns discrete_columns continuous_columns all_missing_columns
## 1   60       3                1                  2                   0
##   total_missing_values complete_rows total_observations memory_usage
## 1                    0            60                180         3256
```

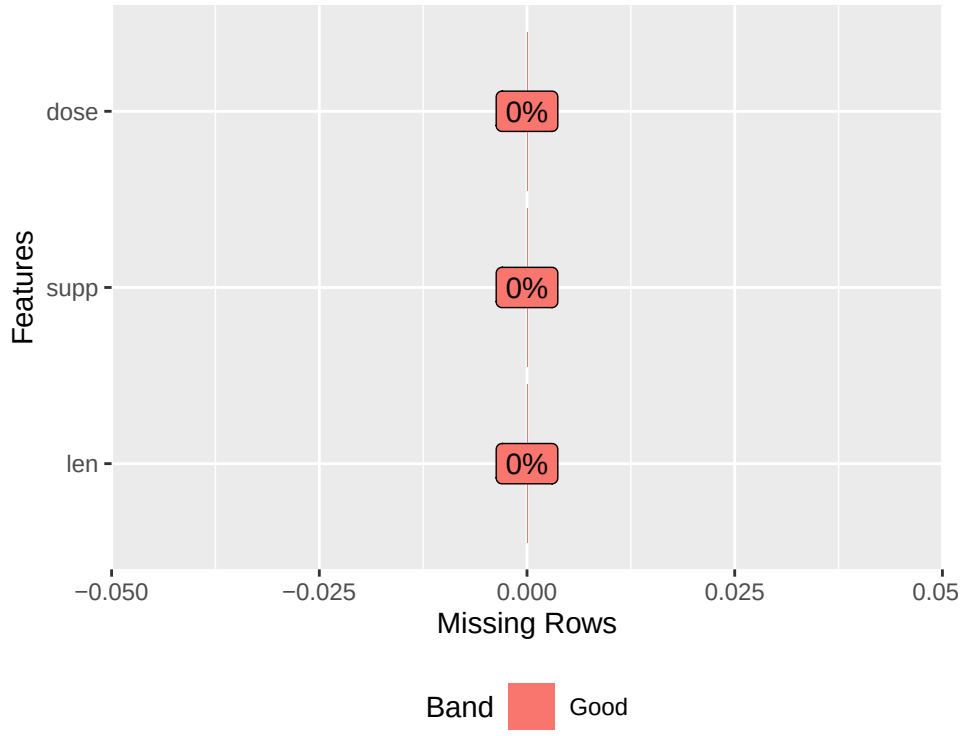
Let's show the result of **introduce** by plotting

```
plot_intro(ToothGrowth)
```



Now let's plot missing values

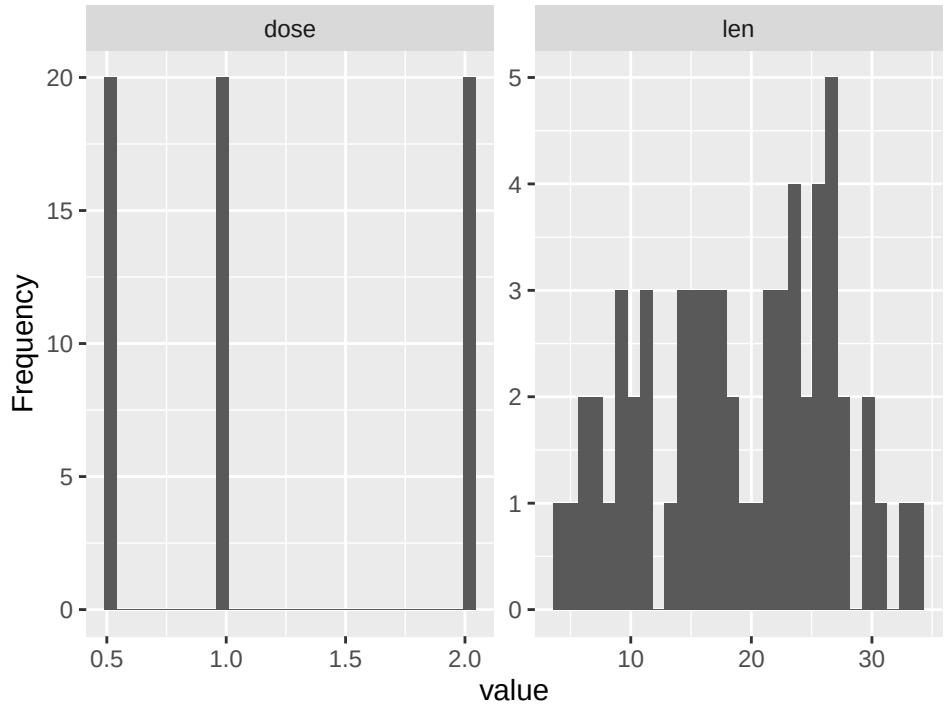
```
plot_missing(ToothGrowth)
```



Data Summary

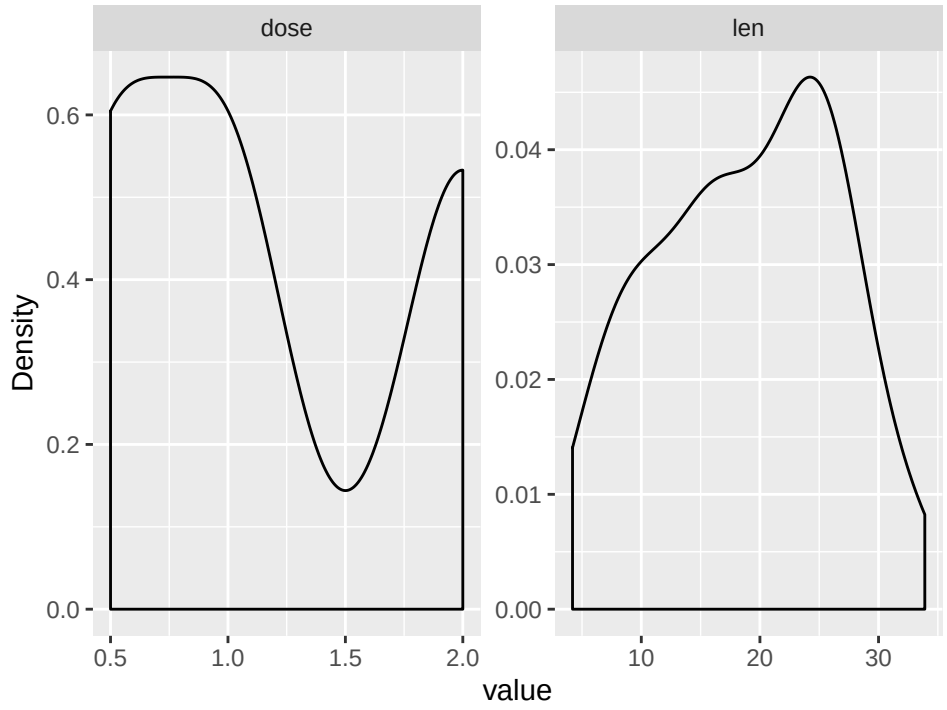
Let's understand data distribution

```
# Plot histogram
plot_histogram(ToothGrowth)
```



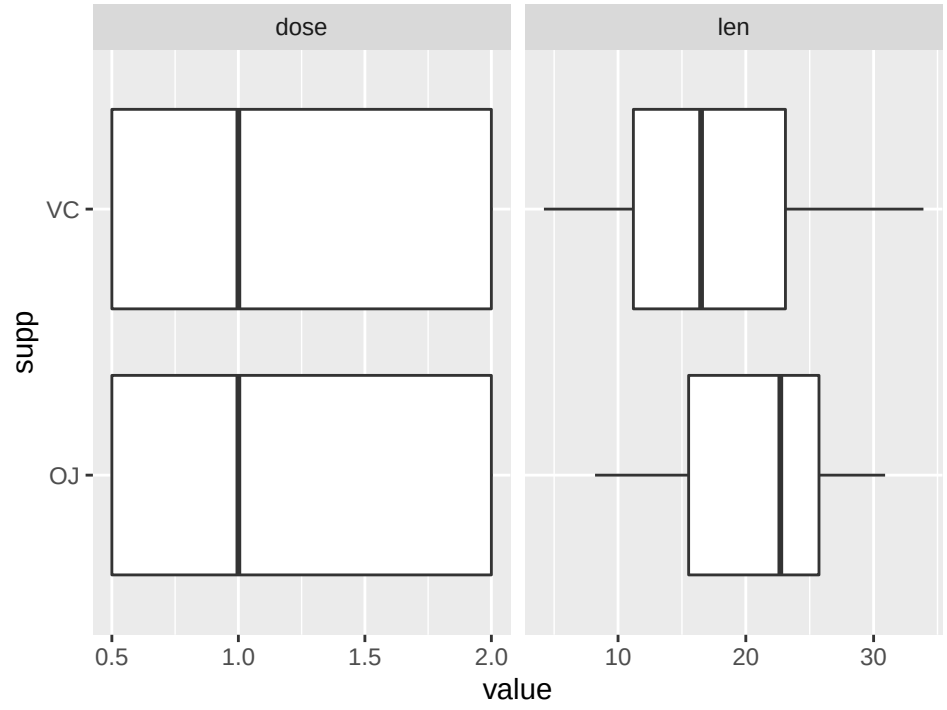
In similar way let's plot densities

```
# Plot densities
plot_density(ToothGrowth)
```



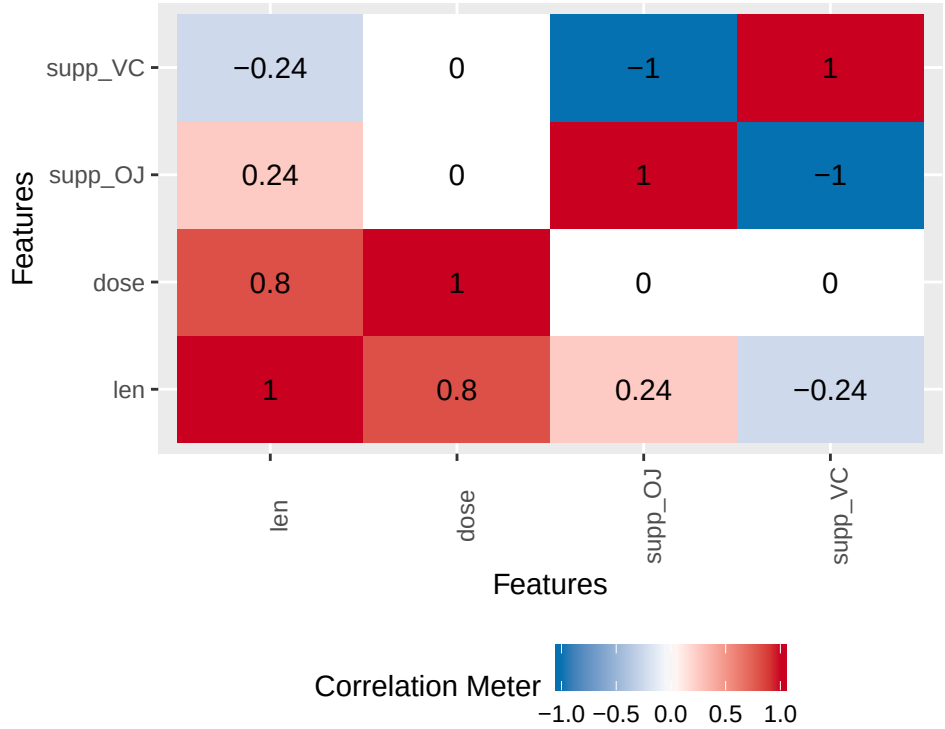
and boxplots:

```
plot_boxplot(ToothGrowth, by = "supp", ncol = 2)
```



Finally let's understand if variables are correlated:

```
# Correlation plot
plot_correlation(ToothGrowth, cor_args = list(use = "complete.obs"))
```



Comparision of tooth growth by supp and dose.

We will use **t.test** to check if there are group differences due to different supplement type. To perform this test we will assume initially unequal variances between the two groups

```
# T test
t.test(len ~ supp, data = ToothGrowth)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 1.9153, df = 55.309, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1710156  7.5710156
## sample estimates:
## mean in group OJ mean in group VC
##         20.66333         16.96333
```

We get a $p - value = 0.06063$, and a confidence interval $(-0.17101567.5710156)$ which contains zero.

This indicates that we can not reject the null hypothesis. It means that the different supplement types have no effect on tooth length.

Next step is to split the data into 3 subsets. Each subset corrensponds to one dosage.

```
# Subset 1
dosage_f <- subset(ToothGrowth, dose %in% c(0.5, 1))
# Subset 2
dosage_s <- subset(ToothGrowth, dose %in% c(0.5, 2))
# Subset 3
dosage_t <- subset(ToothGrowth, dose %in% c(1, 2))
```

Let's perform t-test for each of the subsets.

```
# T test for subset 1
t.test(len ~ dose, paired = F, var.equal = F, data = dosage_f)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -6.4766, df = 37.986, p-value = 1.268e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.983781  -6.276219
## sample estimates:
## mean in group 0.5   mean in group 1
##            10.605            19.735
```

The confidence interval $(-11.983781 - 6.276219)$ gives reason the rejection of the null hypothesis. It means there is significant correlation between tooth length and dose levels.

```
# T test for subset 2
t.test(len ~ dose, paired = F, var.equal = F, data = dosage_s)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -11.799, df = 36.883, p-value = 4.398e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.15617 -12.83383
## sample estimates:
## mean in group 0.5   mean in group 2
##            10.605            26.100
```

The confidence interval $(-18.15617 - 12.83383)$ gives reason the rejection of the null hypothesis. It means there is significant correlation between tooth length and dose levels.

```
# T test for subset 3
t.test(len ~ dose, paired = F, var.equal = F, data = dosage_t)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -4.9005, df = 37.101, p-value = 1.906e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.996481 -3.733519
## sample estimates:
## mean in group 1 mean in group 2
##          19.735          26.100
```

The confidence interval $(-8.996481 - 3.733519)$ gives reason the rejection of the null hypothesis. It means there is significant correlation between tooth length and dose levels.

Assumptions vs Conclusions

- We made the assumptions that populations were indipendent.

- We reached the conclusion that increase of dosage leads to an increase in tooth growth

- Supplement has no effect on tooth growth.