Statistical Inference Course Project Part2

Naqeeb Asif

14 February 2018

Synopsis

This experiment explores the ToothGrowth data set. Exploratory analysis is first performed on the dataset. After the analysis the hypothesis test is performed according to which "dose" is the most significant in predicting the "length".

Loading the data

In this section "Tooth Growth" data is loaded.

```
data("ToothGrowth")
head(ToothGrowth)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

Exploratoty analysis

In this section exploratory analysis is performed on the data

Checking the structure of the data

```
str(ToothGrowth)
```

```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

Checking the unique entries in feature 'dose'

```
unique(ToothGrowth$dose)
```

```
## [1] 0.5 1.0 2.0
```

As there are only three unique values in this feature so lets change the calss of the this feature to factor.

```r
ToothGrowth$dose <- factor(ToothGrowth$dose)
```

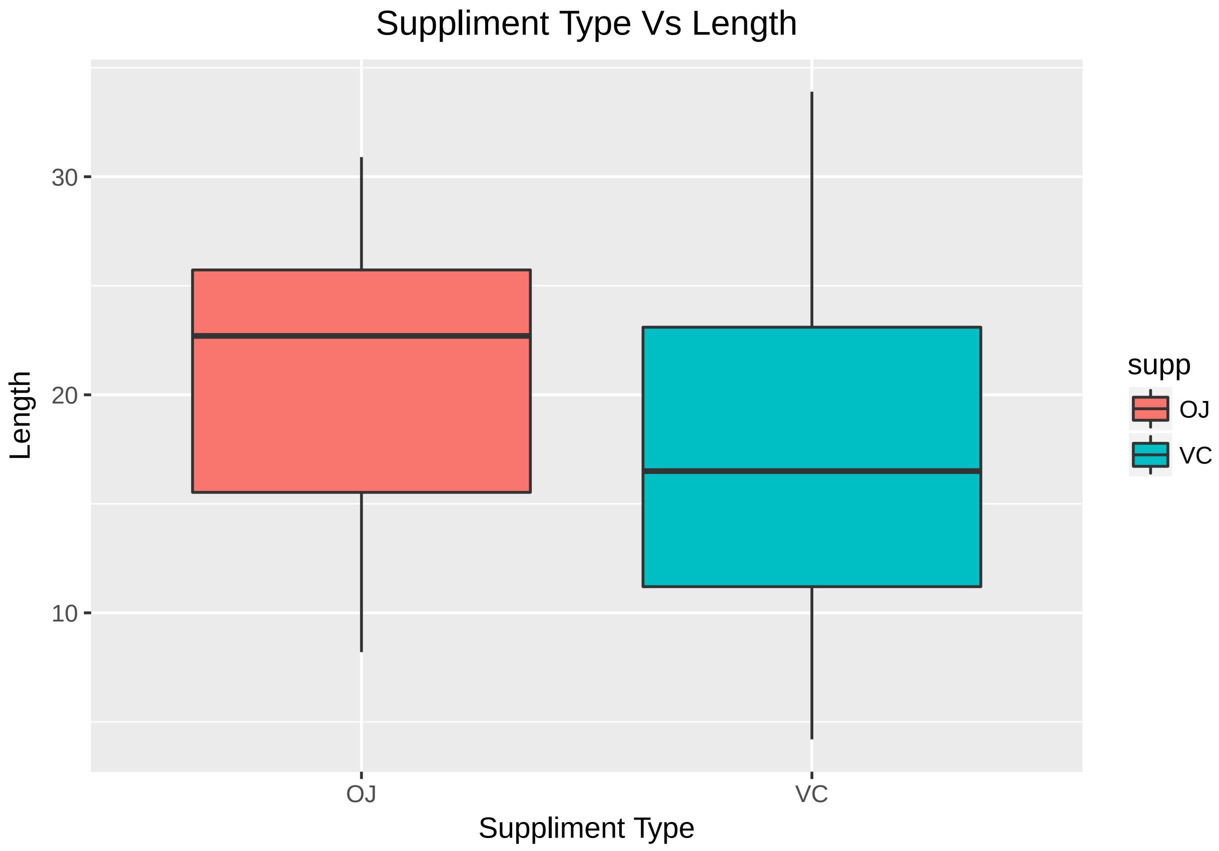
Boxplot of features

In this section boxplot is plotted between different features of the data.

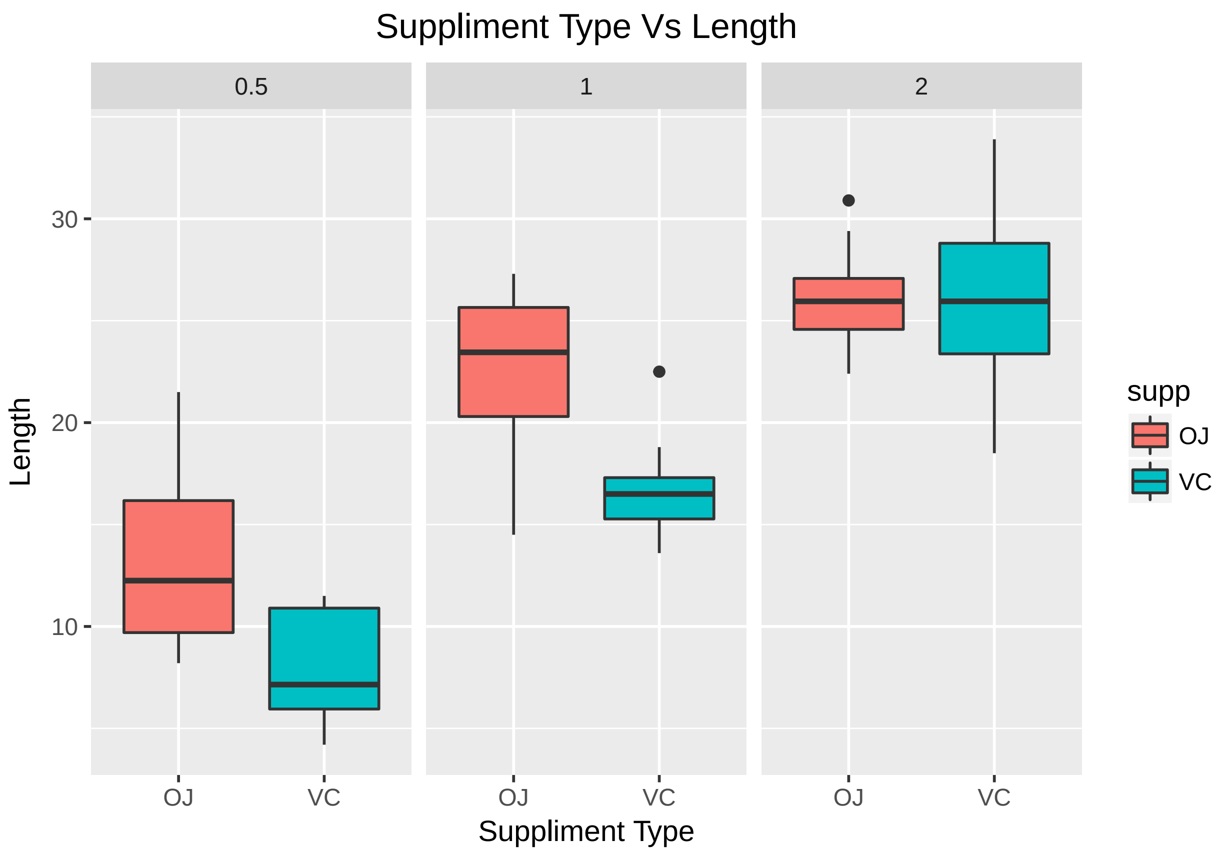
Boxplot between 'Suppliment type' and 'length'

In this section boxplot is plotted between length of tooth and supplement type.

```r
library(ggplot2)
g <- ggplot(data = ToothGrowth,aes(supp,len))
g + geom_boxplot(aes(fill=supp)) +
  labs(title="Suppliment Type Vs Length")+
  theme(plot.title = element_text(hjust = 0.5))+
  xlab("Suppliment Type")+ylab("Length")
```



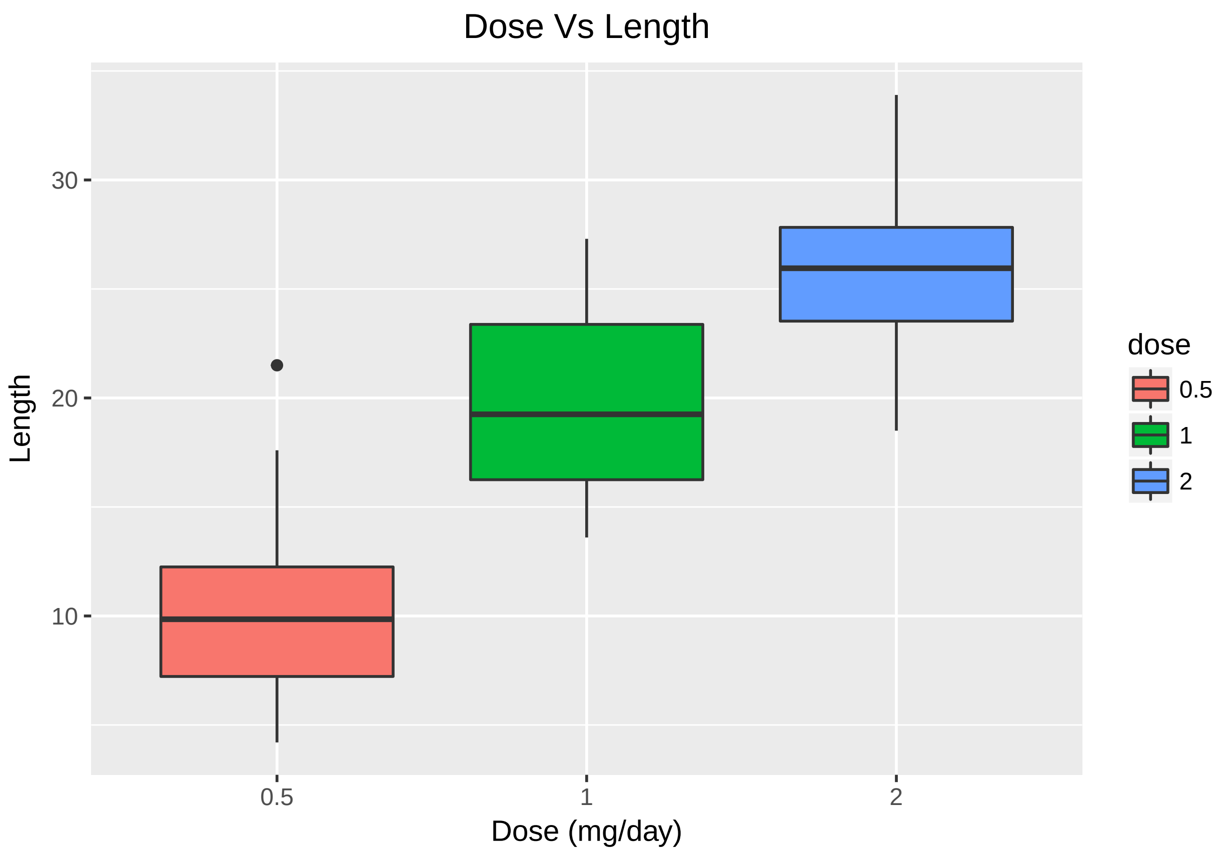
```r
library(ggplot2)
g <- ggplot(data = ToothGrowth,aes(supp,len))
g + geom_boxplot(aes(fill=supp)) +facet_grid(.~dose)+
  labs(title="Suppliment Type Vs Length")+
  theme(plot.title = element_text(hjust = 0.5))+
  xlab("Suppliment Type")+ylab("Length")
```



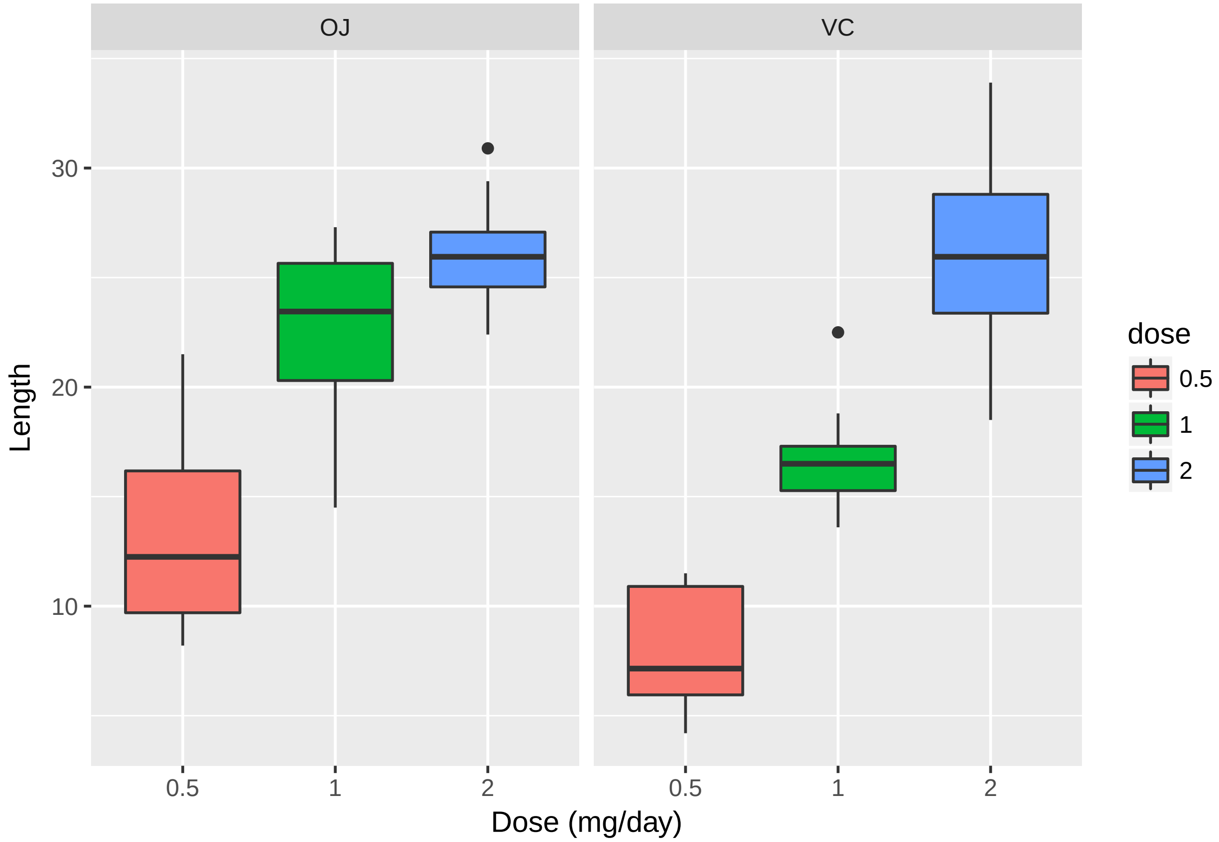
Boxplot between Dose and length

In this section boxplot is plotted between length of tooth and dose.

```
library(ggplot2)
g <- ggplot(data = ToothGrowth,aes(dose,len))
g + geom_boxplot(aes(fill=dose))+
  labs(title="Dose Vs Length")+
  theme(plot.title = element_text(hjust = 0.5))+
  xlab("Dose (mg/day)")+ylab("Length")
```



```
library(ggplot2)
g <- ggplot(data = ToothGrowth,aes(dose,len))
g + geom_boxplot(aes(fill=dose)) + facet_grid(.~supp)+
  xlab("Dose (mg/day)")+ylab("Length")
```



Results

As it can be seen from the plots that length depends upon the dose and suppliment. Each supllement type has different mean according to its type. Same is the case with different doeses.

Summary

In this section summary of the data set is presented.

```
md <-lm(len ~dose,data = ToothGrowth)
summary(md)
```

```
##
## Call:
## lm(formula = len ~ dose, data = ToothGrowth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6000 -3.2350 -0.6025  3.3250 10.8950
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.6050     0.9486  11.180 5.39e-16 ***
## dose1         9.1300     1.3415   6.806 6.70e-09 ***
## dose2        15.4950     1.3415  11.551  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.242 on 57 degrees of freedom
## Multiple R-squared:  0.7029, Adjusted R-squared:  0.6924
## F-statistic: 67.42 on 2 and 57 DF,  p-value: 9.533e-16
```

Hypothesis Tests

In this section hyothesis test is performed on the data set.

Test on length and suppliment type

In this section hypothesis test on length and suppliment type is performed.

```r
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
```

As it can be seen that p-value is smaller than 0.05 therefore the feature “Suppliment Type” is less significant.

Test on length and dose

In this section hypothesis test on length and dose is performed. As hypothesis test can be performed on the data with two groups or levels and ‘dose’ has three levels therefore the test is performed by subsetting the ‘dose’ in two levels.

1. Tesing for dose equals to 0.5 or 1

```r
library(data.table)
library(dplyr)
ToothGrowth <- data.table(ToothGrowth)
t.test(len~dose,data=filter(ToothGrowth,dose %in% c(0.5,1)))
```

```
##
##  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
```

2. Testing for dose equals to 1 or 2

```r
library(data.table)
library(dplyr)
ToothGrowth <- data.table(ToothGrowth)
t.test(len~dose,data=filter(ToothGrowth,dose %in% c(2,1)))
```

```
##
##  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
```

3. Testing for dose equals to 0.5 or 2

```r
library(data.table)
library(dplyr)
ToothGrowth <- data.table(ToothGrowth)
t.test(len~dose,data=filter(ToothGrowth,dose %in% c(0.5,2)))
```

```
##
##  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
```

AS it can be seen from the above results that p-value is less than 0.05 therefore feature "dose" is most significant in predicting the length of tooth.

Conclusion

We can conclude from the above results that length in the toothgrowth dataset depends on the supplement type and dose of the vitamin. But dose is the most significant feature in predicting the length whereas the feature supplement type is less significant in prediction because its p-value for supplement type is slightly less than 0.05.