Part2

Mohit Anand

Overview

The aim of this report is to analyze the ToothGrowth DataSet in R. As per the help file of this dataset, the response is the length of odontoblasts (teeth) in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1, and 2 mg) with each of two delivery methods (orange juice or ascorbic acid). First we will load the ggplot library and get the summary of ToothGrowth Dataset.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.3
```

```
data("ToothGrowth")
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 ...
```

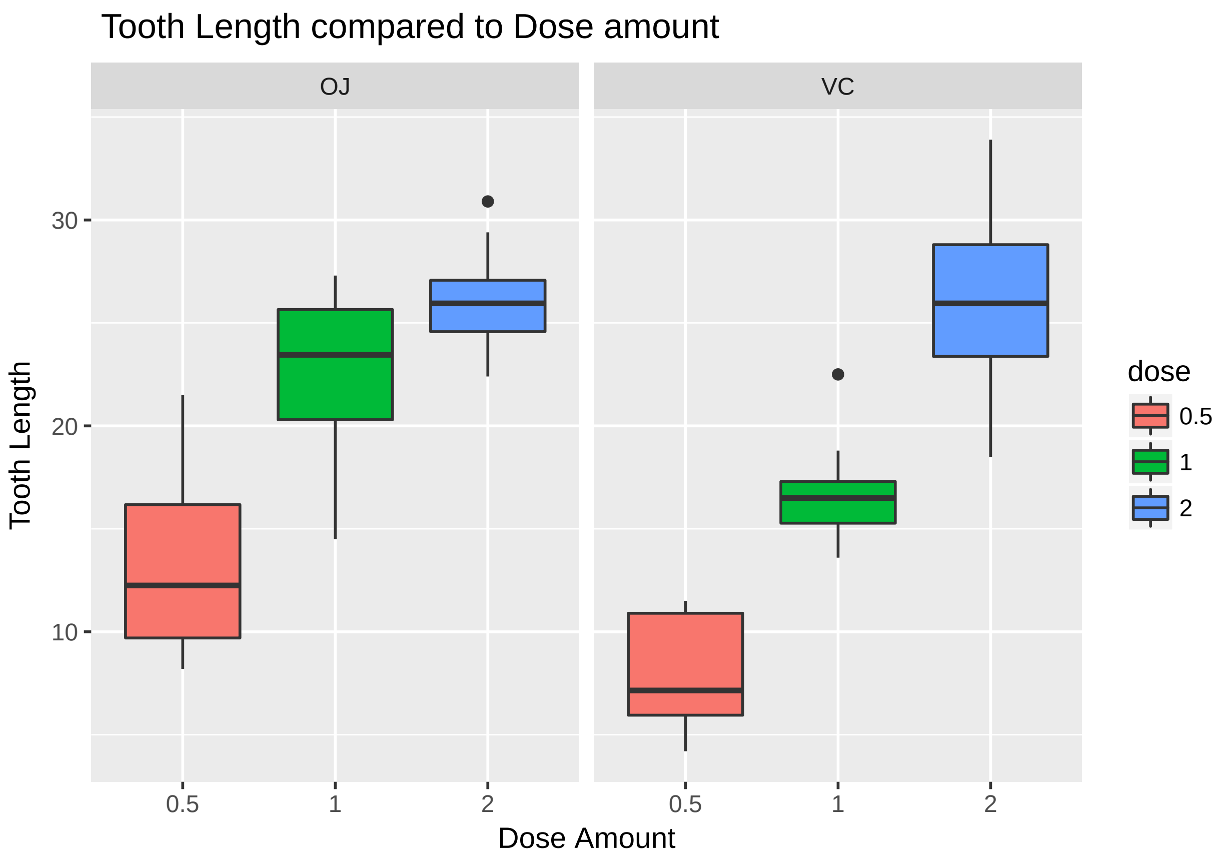
Since dose only has three values which are essentially categorical hence we convert it to factor.

```
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
```

Exploratory Analysis

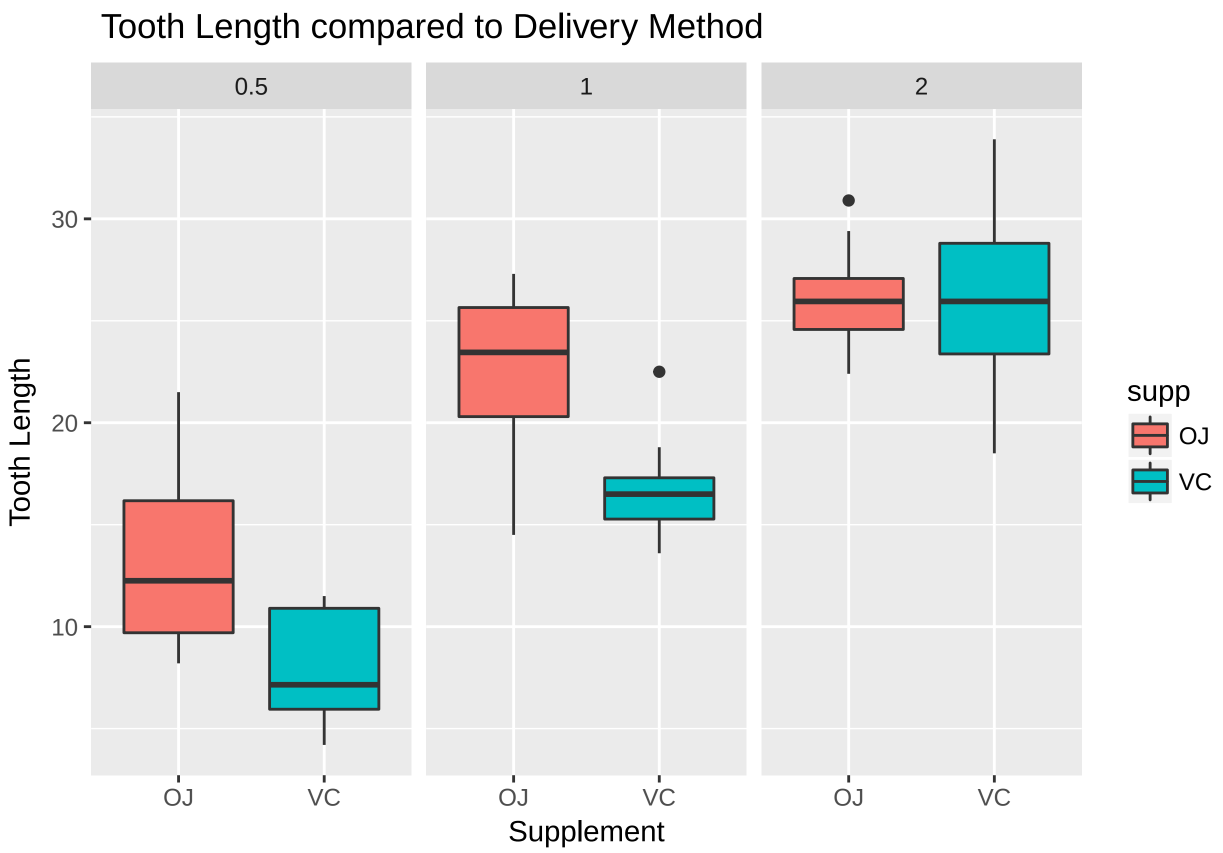
Now we will create some plots to explore the data

```
bxplt <- ggplot(aes(x=dose, y=len),data = ToothGrowth)+ geom_boxplot(aes(fill=dose)) + facet_grid(.~supp
print (bxplt)
```



We can also explore on the basis of Dose amount

```
bxplt2 <- ggplot(aes(x=supp, y=len),data = ToothGrowth)+ geom_boxplot(aes(fill=supp)) + facet_grid(.~dos
print (bxplt2)
```



Hyposthesis Testing

Now we will perform Hypothesis testing to determine if the categorical variables affect the result i.e. the length of Tooth.

Supplement as a Factor

For this we will use t test. First we will run the test on the basis of supplement chosen.

```
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 can see that the p-value is more than 0.05 and the confidence interval has zero in it. Therefore, we can conclude that supplement types don't have impact on the Tooth growth in this test.

Dosage as a Factor

Now we shall run a t test to see if the Dosage is a Factor or not. First we will split the Dosage data into subsets of two

```
TG_dose1 <- subset (ToothGrowth, dose %in% c(0.5, 1.0))
TG_dose2 <- subset (ToothGrowth, dose %in% c(0.5, 2.0))
TG_dose3 <- subset (ToothGrowth, dose %in% c(1.0, 2.0))
```

Checking for dose levels (0.5,1.0)

```
t.test(len~dose,data=TG_dose1)
```

```
##
##  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
```

Checking for dose levels (0.5,2.0)

```
t.test(len~dose,data=TG_dose2)
```

```
##
##  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
```

Checking for dose levels (2.0,1.0)

```
t.test(len~dose,data=TG_dose3)
```

```
##
##  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
```

As we can see for all the three t tests the p-value is much less than 0.05 and the confidence interval does not contain zero. Based on this we can make a conclusion that tooth length is directly related to the dosage. Hence we reject the null hypothesis.

Assumptions

1. The sample is representative of the population of guinea pigs
2. For the t-tests, the variances are assumed to be different for the two groups being compared.
3. The distribution of sample means follows Central Limit Theorem