Exploratory Data Analysis(EDA) on ToothGrowth Dataset

B Santhoshini

2023-03-27

EDA on ToothGrowth dataset

##Overview In this project we will perform EDA on ToothGrowth dataset and determine whether there is an actual change in the growth of tooth using these two types of supplements in subsequent dosages.We conduct Hypothesis tests and also construct Confidence Intervals for our hypotheses.

##Load Data

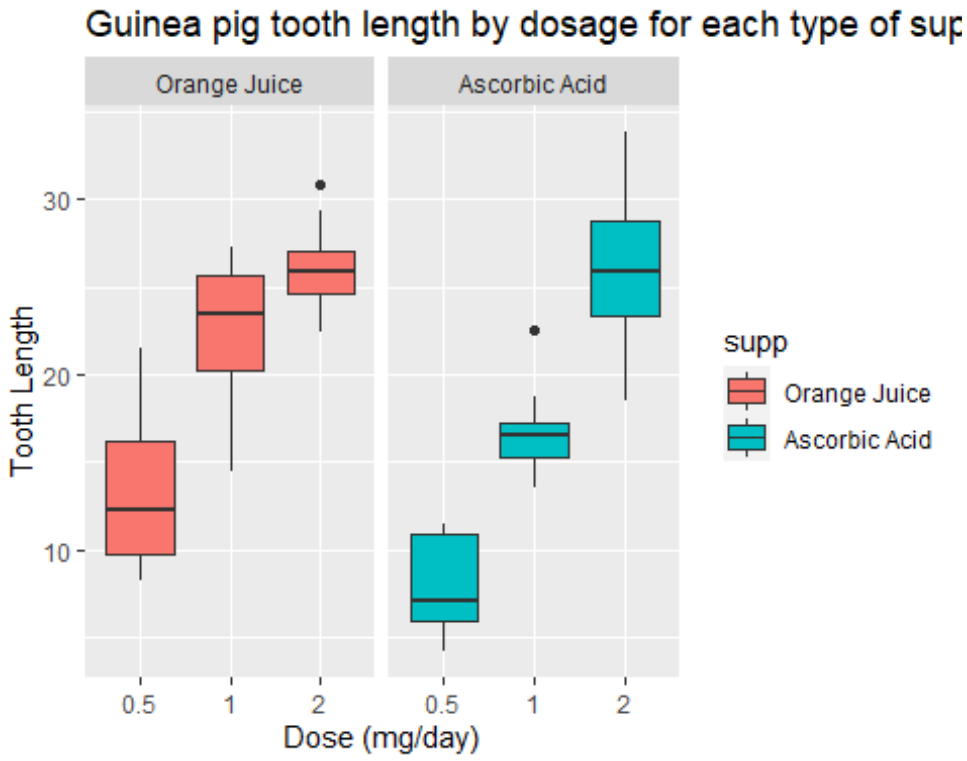
```
library(datasets)
data(ToothGrowth)
str(ToothGrowth)
head(ToothGrowth)
summary(ToothGrowth)
```

##Perform Exploratory data analysis

```
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.2.1

t_set = ToothGrowth
levels(t_set$supp) <- c("Orange Juice", "Ascorbic Acid")
ggplot(t_set, aes(x=factor(dose), y=len)) +
  facet_grid(.~supp) +
  geom_boxplot(aes(fill = supp), show.legend = TRUE) +
  labs(title="Guinea pig tooth length by dosage for each type of supplement",
    x="Dose (mg/day)",
    y="Tooth Length")
```



Summary of the Data

From the box plots it's observed that as the dosage increases the growth increases.Orange juice seems to be more effective within the dosage range of 0.5-1mg/day. And ascorbic acid seems slightly more effective than orange juice when the dosage is 2mg/day. mg = milligrams #Usage of Hypotheses tests and Confidence Intervals for analysing the growth by dosage and type of supplement.

##Hypothesis Test - 1 H_o = There is no difference between the growth of tooth by two types of supplements across the whole range.

```
hypothesis_test1 <- t.test(len ~ supp, data = t_set)
hypothesis_test1$p.value

## [1] 0.06063451

hypothesis_test1$conf.int

## [1] -0.1710156  7.5710156
## attr(,"conf.level")
## [1] 0.95
```

#Conclusion-1

As zero is included in the dataset and p value is also greater than 0.05 which states that we can't reject null hypothesis.

##Hypothesis test- 2

H_o = There is no difference between the growth of tooth by two types of supplements when dosage is 0.5mg/day. H_a = There is a significant difference between the growth of tooth by two types of supplements when dosage is 0.5mg/day.

```
hypothesis_test2 <- t.test(len~supp, data =subset(t_set,dose == 0.5))
hypothesis_test2$p.value

## [1] 0.006358607

hypothesis_test2$conf.int

## [1] 1.719057 8.780943
## attr(,"conf.level")
## [1] 0.95
```

#Conclusion-2 From the above statistics we can deduce that the p value is less than 0.05 and the confidence interval doesn't contain zero we can't reject alternative hypothesis.

#Hypothesis test - 3 H_o = There is no difference between the growth of tooth by two types of supplements when dosage is 1mg/day. H_a = There is a significant difference between the growth of tooth by two types of supplements when dosage is 1mg/day.

```
hypothesis_test3 <- t.test(len~supp, data =subset(t_set, dose == 1))
hypothesis_test3$p.value

## [1] 0.001038376

hypothesis_test3$conf.int

## [1] 2.802148 9.057852
## attr(,"conf.level")
## [1] 0.95
```

#Conclusion-3 As the p value is less than 0.05 and the confidence interval doesn't contain zero we can't reject alternative hypothesis.

##Hypothesis test-4 H_o = There is no difference between the growth of tooth by two types of supplements when dosage is 2mg/day. H_a = There is a significant difference between the growth of tooth by two types of supplements when dosage is 2mg/day.

```
hypothesis_test4 <- t.test(len~supp, data =subset(t_set, dose == 2))
hypothesis_test4$p.value

## [1] 0.9638516

hypothesis_test4$conf.int

## [1] -3.79807  3.63807
## attr(,"conf.level")
## [1] 0.95
```

Conclusion-4

As the p value is above 0.05 and the confidence interval does contain zero. so we can't reject null hypothesis.

##Final Conclusion Orange juice delivers more tooth growth than ascorbic acid for dosages 0.5 & 1.0.Orange juice and ascorbic acid deliver the same amount of tooth growth for dose amount 2.0 mg/day.Across the data set we cannot conclude orange juice is more effective that ascorbic acid.

Assumptions include that tooth growth follow gaussian distribution.