# Statistical Inference Course Project Part 2 - ToothGrowth Data - EDA and Hypothesis testing

Author: Ramesh Natarajan

## Overview

In this exercise, we're going to analyze the ToothGrowth data in the R datasets package by performing some exploratory data analysis on ToothGrowth data and performing Hypothesis testing to check mean diffrence in the growth for the two supplement methods.

## Question 1

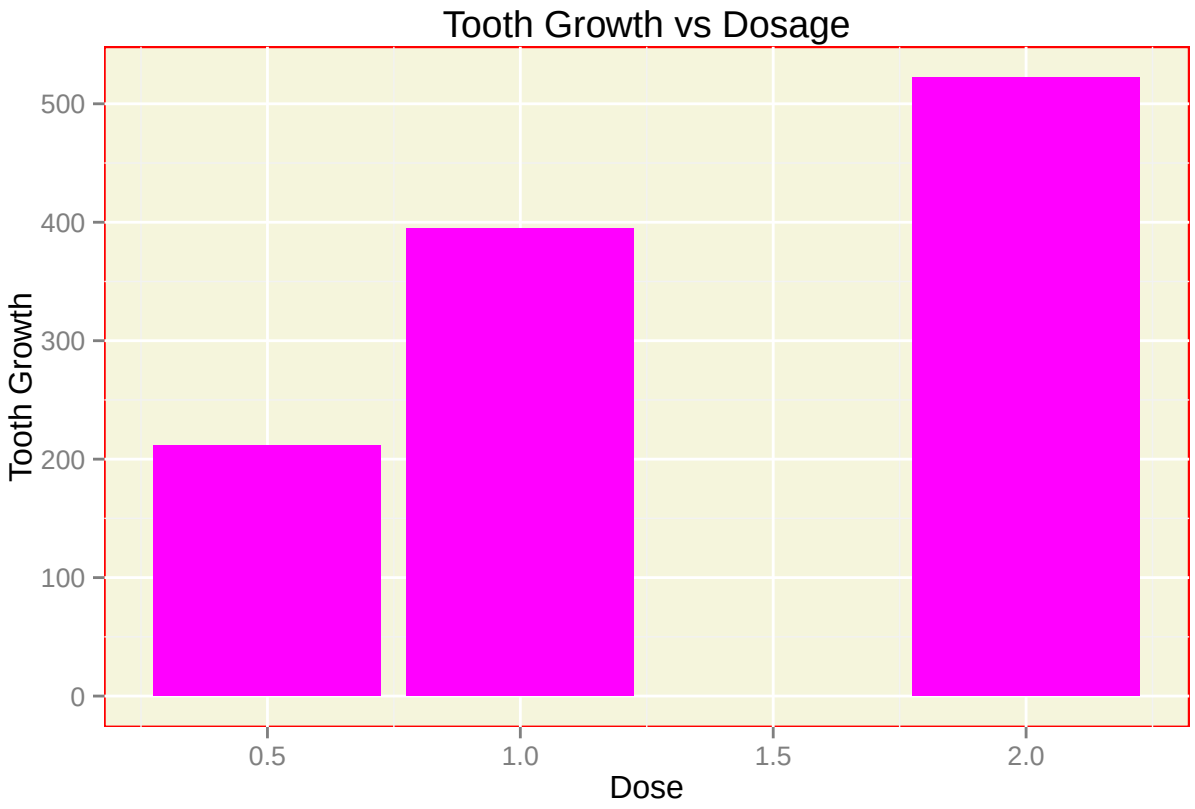
Perform some basic exploratory data analysis on ToothGrowth which contains a) len - Tooth Growth, b) supp - Delivery method of supplement (orange juice OJ and ascorbic acid (VC), c) dose - Dose level of Vitamin C

```
data(ToothGrowth)

library(ggplot2)

# Impact of Dosage on tooth growth
g <- ggplot(ToothGrowth, aes(dose, len), height=300, width=400)
g <- g + geom_bar(stat = "identity", fill = "magenta") +
  labs(title = "Tooth Growth vs Dosage",
       x = "Dose",
       y = "Tooth Growth") +
        theme(panel.background = element_rect(fill = 'beige', colour = 'red'))

print(g)
```

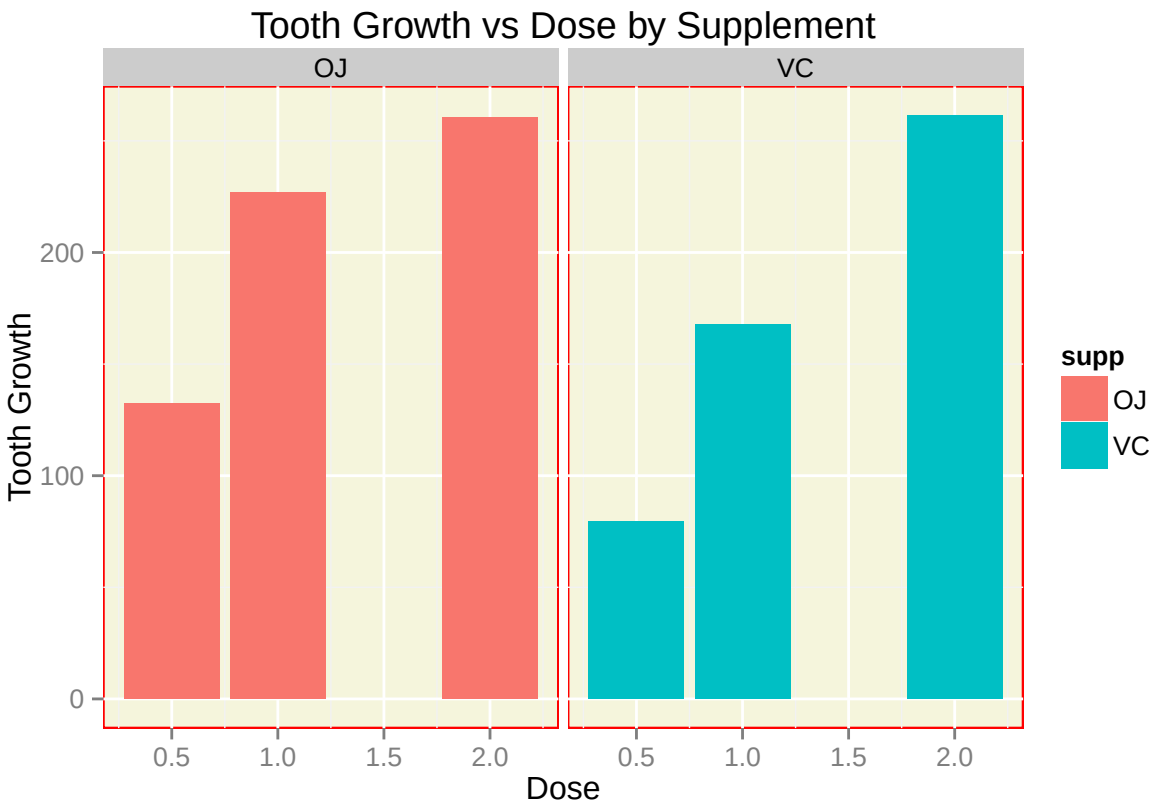


```r
# Impact of Dosage on tooth growth split by type of supplement

g <- ggplot(ToothGrowth, aes(dose, len, fill = supp), height=300, width=400)

g <- g + geom_bar(stat = "identity") +
  facet_grid(. ~ supp, scales = "free") +
  labs(title = "Tooth Growth vs Dose by Supplement",
       x = "Dose",
       y = "Tooth Growth") +
        theme(panel.background = element_rect(fill = 'beige', colour = 'red'))

print(g)
```

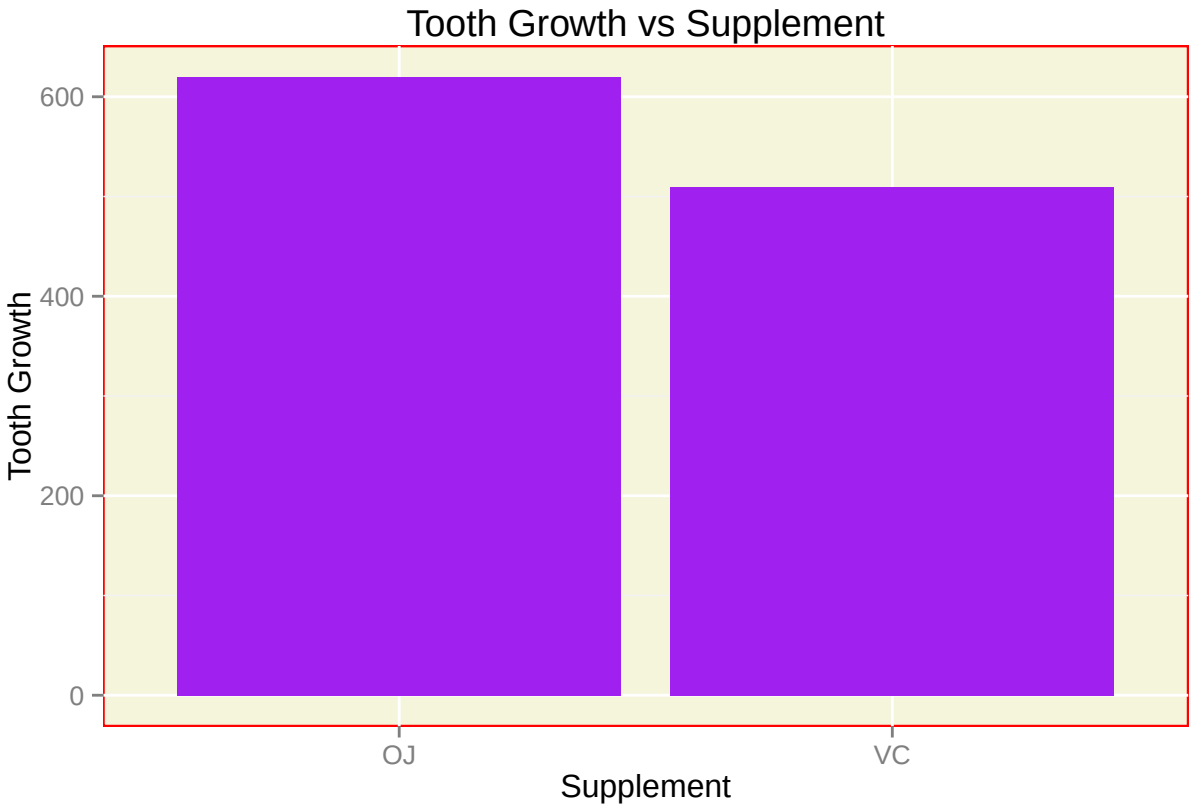



```r
# Impact of Supplement type on tooth growth

g <- ggplot(ToothGrowth, aes(supp, len), height=300, width=400)

g <- g + geom_bar(stat = "identity", fill = "purple") +
  labs(title = "Tooth Growth vs Supplement",
       x = "Supplement",
       y = "Tooth Growth") +
        theme(panel.background = element_rect(fill = 'beige', colour = 'red'))

print(g)
```

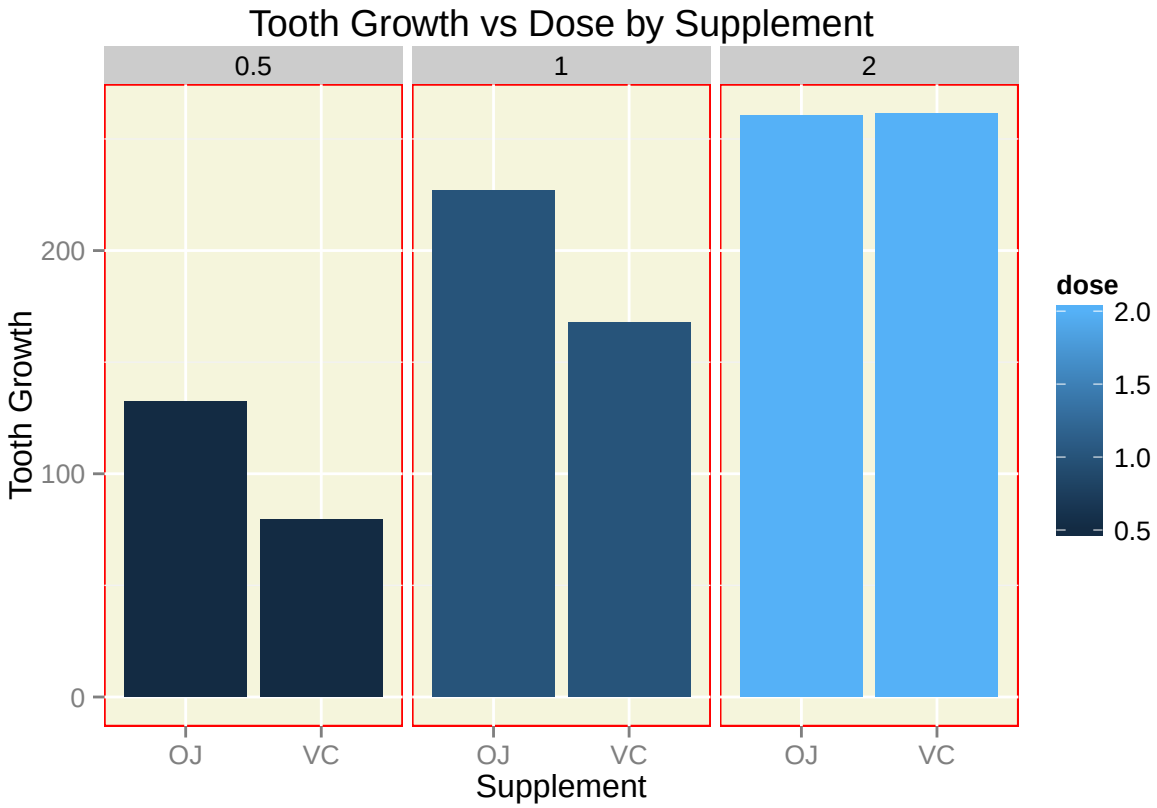



```
# Impact of Supplement type on tooth growth split by Dosage

g <- ggplot(ToothGrowth, aes(supp, len, fill = dose), height=300, width=400)

g <- g + geom_bar(stat = "identity") +
  facet_grid(. ~ dose, scales = "free") +
  labs(title = "Tooth Growth vs Dose by Supplement",
       x = "Supplement",
       y = "Tooth Growth") +
        theme(panel.background = element_rect(fill = 'beige', colour = 'red'))

print(g)
```




## Question 2

Let's get some summary information on ToothGrowth data

```
# summary of different fields and data in ToothGrowth data.
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
```

```
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 ...
```

```
summary(ToothGrowth)
```

```
##       len        supp         dose
##  Min.   : 4.20   OJ:30   Min.   :0.500
##  1st Qu.:13.07   VC:30   1st Qu.:0.500
##  Median :19.25           Median :1.000
##  Mean   :18.81           Mean   :1.167
##  3rd Qu.:25.27           3rd Qu.:2.000
##  Max.   :33.90           Max.   :2.000
```

## Question 3

Now we will perform some hypothesis testing on this data. In particular, we will focus on finding if the mean growth rate of tooth between two dose levels is the same across the two methods of supplements. We will use T Test since the number of observation is not high.

```
# we will first reshape the data by first grouping them into 10 sets of observations for each
# supplement method and dose levels
library(reshape2)
ToothGrowth <- cbind('subjectid' = 1:10, ToothGrowth)
wideTG <- dcast(ToothGrowth, subjectid + supp ~ dose, value.var = "len")
names(wideTG)[-(1 : 2)] <- paste("dose", names(wideTG)[-(1 : 2)], sep = "")

library(dplyr)
```

```
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:stats':
##
##     filter, lag
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# calculate growth as the difference in the length between 2 and 0.5 dosages
wideTG0.5to2 <- mutate(wideTG, growth = dose2 - dose0.5)

# perform T test to validate if the mean growth rate for the two supplements are the same
# for 0.5 and 2 dosages
t.test(growth ~ supp, paired = FALSE, var.equal = FALSE, data = wideTG0.5to2)
```

```
##
##  Welch Two Sample t-test
##
## data:  growth by supp
## t = -2.1114, df = 17.887, p-value = 0.04908
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -10.63603111  -0.02396889
## sample estimates:
## mean in group OJ mean in group VC
##            12.83            18.16
```

```
# calculate growth as the difference in the length between 2 and 1 dosages
wideTG1to2 <- mutate(wideTG, growth = dose2 - dose1)

# perform T test to validate if the mean growth rate for the two supplements are the same
# for 1 and 2 dosages
t.test(growth ~ supp, paired = FALSE, var.equal = FALSE, data = wideTG1to2)
```

```
##
##  Welch Two Sample t-test
##
## data:  growth by supp
## t = -2.4412, df = 17.997, p-value = 0.0252
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  -11.1823266  -0.8376734
## sample estimates:
## mean in group OJ mean in group VC
##             3.36             9.37
```

## Question 4

Based on both the T tests (difference in growth rate between 0.5 and 2 dosages and difference in growth rate between 1 and 2 dosages for the two supplement methods), we reject the hypothesis that the mean difference in the growth rate is zero since the t value is well below the cut-off for 95% CI. Note that we have assumed that the variance between the two supplements are not equal.