Statistical Inference - EDA

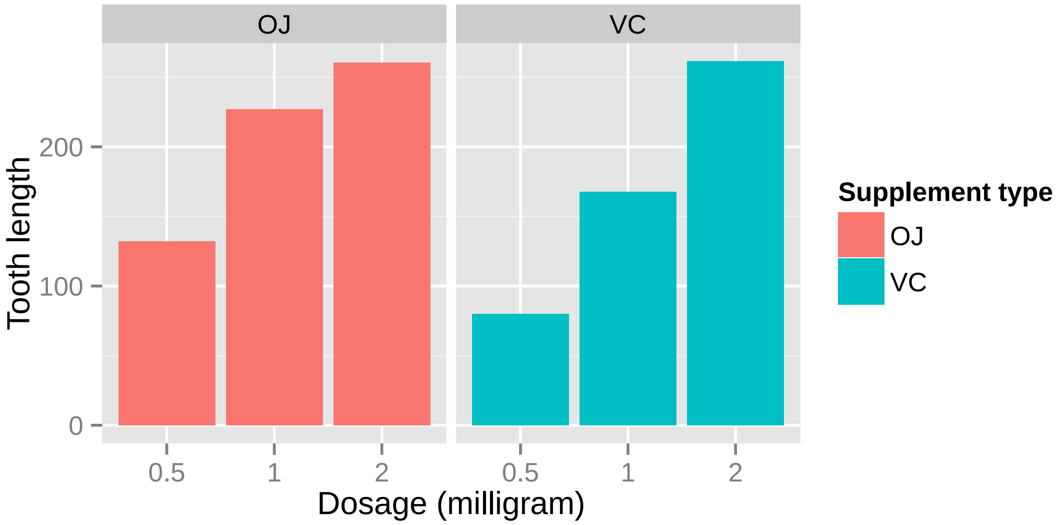
Gaurav Tripathi

Loading of and EDA using ToothGrowth data

Following code imports the data and performs EDA on it to study the impact of Supplement Type and Dosage (in milligram) on Tooth Length.

```
library(datasets)
library(ggplot2)

# Plot Tooth length vs. Dosage (mg) split by Supplement Type (OJ or VC)
ggplot(data=ToothGrowth, aes(x=as.factor(dose), y=len, fill=supp)) +
     geom_bar(stat="identity",) + facet_grid(. ~ supp) + xlab("Dosage (milligram)") +
     ylab("Tooth length") + guides(fill=guide_legend(title="Supplement type"))
```



The following observations can be made:

- There is a positive correlation between Dosage (mg) and Tooth length. Hence, as Dosage of supplement increases, there is an observed increase in Tooth Length (Note, that this is not Causative - We cannot say for sure that Tooth Length is higher due to Higher Doses)
- In case of OJ supplement, the increase in Tooth Length when dosage changes from 0.5 to 1mg is almost same as that in case of VC supplement. But the increase in Tooth Length caused by increase in OJ dosage from 1 to 2mg is lesser than the increase in Tooth Length caused by same dosage change for VC supplement

We can further build a linear regression model to Validate the correlation results, as follows:

```
# Validate the correlation results with a linear model and summarize results
fit <- lm(len ~ dose + supp, data = ToothGrowth)
summary(fit)
```

```
##
## Call:
## lm(formula = len ~ dose + supp, data = ToothGrowth)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.600 -3.700  0.373  2.116  8.800
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.2725     1.2824   7.231 1.31e-09 ***
## dose          9.7636     0.8768  11.135 6.31e-16 ***
## suppVC       -3.7000     1.0936  -3.383   0.0013 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.236 on 57 degrees of freedom
## Multiple R-squared:  0.7038, Adjusted R-squared:  0.6934
## F-statistic: 67.72 on 2 and 57 DF,  p-value: 8.716e-16
```

Observations:

- Adjusted R-square = 0.693 - Which means 69.3% of the data is explained by this model
- P-value for both 'dose' and 'suppVC' variables < 0.05, which means both are significant
- Coefficient for 'dose' is 9.76, which validates the correlation observed before (Increase in dose by 1mg would increase the Tooth Length by 9.76 units)
- Coefficient for 'suppVC' is -3.70, which indicates that if dosage is same between OJ and VC, administering VC would have a negative impact on Tooth Length by 3.7 units

95% Confidence Intervals

The following code gets the 95% Confidence Intervals for the variables and the intercept:

```
confint(fit)
```

```
##                 2.5 %    97.5 %
## (Intercept)  6.704608 11.840392
## dose         8.007741 11.519402
## suppVC      -5.889905 -1.510095
```

Interpretation: It means that if we collect different random set of data and estimate parameters of the linear model each time, 95% of the times, the coefficient estimates will be in these ranges. For each variable, intercept, dose and suppVC, the null hypothesis H0 is that the coefficients are zero, meaning that Tooth Length variation is not explained by these variables. But since all p-values are less than 0.05, we reject H0.