VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

ABHINAV VIJAYAKUMAR

19BCE1311

CSE3506 – ESSENTIALS OF DATA ANALYTICS
LAB-4

DR. LAKSHMI PATHI JAKKAMPUTI (L21 + L22)

Tasks for Week-4: Analysis of Variance (ANOVA)

Perform ANOVA test and determine the statistical differences between the means of individual groups given in the data

Aim: Perform ANOVA test and determine the statistical differences between the means of individual groups given in the data.

Algorithm:

1. Import the dataset and load the dplyr library.

2. Group the data using the group_by function based on color.

3. Apply ANOVA using response with respect to color and generate summary.

4. If Pr(>F) value< 0.05, then perform the Tukey HSD test.

5. If the p-adjusted value of the pair is less than 0.05 then they are significantly different else they are not.

Statistics:

1. Applying group by:

```
group_by(data,color) %>% summarise(count = n(),mean =
mean(response, na.rm = TRUE))
```

	color	count	mean
	<chr>	<int>	<dbl>
1	blue	24	10.6
2	green	24	8.53
3	red	24	2.49

2. Summary of ANOVA:

```
ANOVA <- aov(response~color, data = data)

summary(ANOVA)
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
color        2  857.2   428.6   14.81 4.44e-06 ***
Residuals   69 1996.4    28.9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

3. Conducting Tukey HSD Test

```
TukeyHSD(ANOVA)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = response ~ color, data = data)

$color
                 diff        lwr       upr     p adj
green-blue -2.101667  -5.821045  1.617711 0.3709119
red-blue   -8.140417 -11.859795 -4.421039 0.0000049
red-green  -6.038750  -9.758128 -2.319372 0.0006628
```

Inference:

1. As seen in the summary of ANOVA, the profit value (Pr(>F)) is less than 0.05, hence the null hypothesis is rejected and the Tukey HSD test is required.

2. As seen in the Tukey HSD test results,

i) green and blue are not significantly different since p adj is more than 0.05.

ii) red and blue are significantly different since p adj is less than 0.05.

iii) green and red are significantly different since p adj is less than 0.05.

Program:

```
# To clear the environment

rm(list=ls())


setwd("C:/Users/Abhinav Vijayakumar/Desktop/VIT Academics/Sem
6/EDA/LAB/LAB 4")

data <- read.csv("color-anova-example.csv")

library(dplyr) # To group the data

group_by(data,color) %>% summarise(count = n(),mean =
mean(response, na.rm = TRUE))
```

```
# ANOVA
ANOVA <- aov(response~color, data = data)
summary(ANOVA)
TukeyHSD(ANOVA)
```