VIT®

Vellore Institute of Technology

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

Essentials of Data Analytics - (CSE3506)

Faculty – Lakshmi Pathi Jakkamputi **Sir**

Lab-4

Harshanth k Prakash

19BCE1293

L21-22 Slot

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

To understands and perform ANOVA test and determine the difference between the samples from the given population dataset.

Algorithm

1. Start
2. Read the given dataset into variable
3. Use dplyr library
4. Group the data with respect to color using the group_by command in dplyr library, summarize the count and mean for the column responses.
5. Generate the ANOVA model using the ANOVA command, display the summary
6. If the probability value is > 0.05 accept the null hypothesis and conclude that there is no statistical difference in means.
7. If the probability value is < 0.05 accept the alternate hypothesis and perform Tukey Multiple Comparisons Of Means.
8. From the test, find groups with statistically significant differences. If p value is less than 0.05, there are statistically significant differences in the group of colors.
9. Stop.

Result

```
> head(colore)
  block color response
1     a   red      1.9
2     b   red      2.6
3     c   red      3.4
4     d   red      0.8
5     e   red      5.3
6     f   red      1.5
```

```
> group_by(colore,color) %>% summarise(count = n(),mean = mean(response, na.rm = TRUE))
# A tibble: 3 x 3
  color count  mean
  <chr> <int> <dbl>
1 blue     24 10.6
2 green    24  8.53
3 red      24  2.49
```

```
> 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
```

```
> TukeyHSD(ANOVA)
  Tukey multiple comparisons of means
    95% family-wise confidence level

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

$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

We can infer that on performing ANOVA, we get the p-value as 4.44e-06 (which is less than 0.05) , so we can reject the null hypothesis while accepting the alternate hypothesis (that there is difference among group means). According to the Tukey multiple comparisons of means test, the difference between groups green and blue is not statistically significant, However, there are statistically significant differences in the following groups of colours with $p < 0.05$: red-blue and red-green from above attached screenshot its clear.

Program

```
rm(list=ls())

colore<-read.csv("D:/6th Sem Works/A2- EDA/LAB/Lab4/color.csv");

head(colore)

library(dplyr)

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

# ANOVA

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

summary(ANOVA)

# Tukey HSD (Tukey Honest Significant Differences)

TukeyHSD(ANOVA)
```