Kruskal-Wallace test for a categorical and continous variable.

The Kruskal–Wallis test is a non-parametric method that does not assume normalality, unlike the analogous one-way analysis of variance. It is assumed that the distribution of the population should not be necessarily normal and the variances should not be necessarily equal. The test can be implemented in R using the kruskal.test(x,g) function. The x parameter is a continuous (interval/ratio) variable. The g parameter is the categorical variable representing different groups to which the continuous values belong.The test does not identify where this stochastic dominance occurs or for how many pairs of groups stochastic dominance obtains. Since ANOVA assumes normal distribution let us use Kriskal-Eallis test. Even if we use ANOVA we get similar results.

The null hypothesis with a Kruskal-Wallace test is that all the different groups represented by the samples are very similar based on the median value.

```
df=read.csv("Marketing-Customer-Value-Analysis.csv")
df1 = subset(df, select = -c(Customer,Effective.To.Date) )
cat_var=sapply(df1,is.character)
data_matrix <- data.matrix(df1[cat_var])
colnames(data_matrix)
```

```
##  [1] "State"            "Response"         "Coverage"         "Education"
##  [5] "EmploymentStatus" "Gender"           "Location.Code"    "Marital.Status"
##  [9] "Policy.Type"      "Policy"           "Renew.Offer.Type" "Sales.Channel"
## [13] "Vehicle.Class"    "Vehicle.Size"
```

State

```
 kruskal.test(x =df$Customer.Lifetime.Value, g = as.factor(df$State))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  df$Customer.Lifetime.Value and as.factor(df$State)
## Kruskal-Wallis chi-squared = 5.0721, df = 4, p-value = 0.28
```

```
aggregate(df$Customer.Lifetime.Value ~ df$State, data = data.frame(df$Customer.Lifetime.Valu
e,df$State), FUN=mean, na.rm=T)
```

```
##     df$State df$Customer.Lifetime.Value
## 1    Arizona                   7861.341
## 2 California                   8003.648
## 3     Nevada                   8056.707
## 4     Oregon                   8077.901
## 5 Washington                   8021.472
```

There is 28% chance that the means are same.Therefore, we fail to reject the null hypothesis. There is no significant difference in CLV among the customers from different States.Thus State variable can be avoided in our model.

Response

```
kruskal.test(x =df$Customer.Lifetime.Value, g = as.factor(df$Response))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  df$Customer.Lifetime.Value and as.factor(df$Response)
## Kruskal-Wallis chi-squared = 0.42011, df = 1, p-value = 0.5169
```

```
aggregate(df$Customer.Lifetime.Value ~ df$Response, data = data.frame(df$Customer.Lifetime.Va
lue,df$Response), FUN=mean, na.rm=T)
```

```
##   df$Response df$Customer.Lifetime.Value
## 1          No                   8030.022
## 2         Yes                   7854.871
```

There is 51% chance that the means are same.Therefore, we fail to reject the null hypothesis. There is no significant difference in CLV among the customers based on their response to marketting calls.Thus Response variable can be avoided in our model.

Coverage

```
kruskal.test(x =df$Customer.Lifetime.Value, g = as.factor(df$Coverage))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  df$Customer.Lifetime.Value and as.factor(df$Coverage)
## Kruskal-Wallis chi-squared = 502.5, df = 2, p-value < 2.2e-16
```

```
aggregate(df$Customer.Lifetime.Value ~ df$Coverage, data = data.frame(df$Customer.Lifetime.Va
lue,df$Coverage), FUN=mean, na.rm=T)
```

```
##   df$Coverage df$Customer.Lifetime.Value
## 1       Basic                   7190.706
## 2    Extended                   8789.678
## 3     Premium                  10895.603
```

The p value here is << 0.05.Therefore, we reject the null hypothesis. There is significant difference in CLV among the customers with different policy coverages.Thus Coverage variable could be useful.

Education

```
kruskal.test(x =df$Customer.Lifetime.Value, g = as.factor(df$Education))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  df$Customer.Lifetime.Value and as.factor(df$Education)
## Kruskal-Wallis chi-squared = 12.234, df = 4, p-value = 0.01569
```

```
aggregate(df$Customer.Lifetime.Value ~ df$Education, data = data.frame(df$Customer.Lifetime.V
alue,df$Education), FUN=mean, na.rm=T)
```

```
##           df$Education df$Customer.Lifetime.Value
## 1             Bachelor                   7872.660
## 2              College                   7851.065
## 3               Doctor                   7520.345
## 4 High School or Below                   8296.709
## 5               Master                   8243.485
```

The p value here is < 0.05.Therefore, we reject the null hypothesis. There is significant difference in CLV among the customers with different Education levels.Thus Education of customers can be useful in predicting CLV.

EmploymentStatus

```
kruskal.test(x =df$Customer.Lifetime.Value, g = as.factor(df$EmploymentStatus))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  df$Customer.Lifetime.Value and as.factor(df$EmploymentStatus)
## Kruskal-Wallis chi-squared = 42.562, df = 4, p-value = 1.276e-08
```

```
aggregate(df$Customer.Lifetime.Value ~ df$EmploymentStatus, data = data.frame(df$Customer.Lif
etime.Value,df$EmploymentStatus), FUN=mean, na.rm=T)
```

```
##   df$EmploymentStatus df$Customer.Lifetime.Value
## 1            Disabled                   7847.889
## 2            Employed                   8219.118
## 3       Medical Leave                   7641.822
## 4             Retired                   7487.865
## 5          Unemployed                   7636.320
```

The p value here is < 0.05.Therefore, we reject the null hypothesis. There is significant difference in CLV among the customers with different Employment status.Thus Employment status of customers can be useful in predicting CLV.

Gender

```
kruskal.test(x =df$Customer.Lifetime.Value, g = as.factor(df$Gender))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  df$Customer.Lifetime.Value and as.factor(df$Gender)
## Kruskal-Wallis chi-squared = 0.48206, df = 1, p-value = 0.4875
```

```
aggregate(df$Customer.Lifetime.Value ~ df$Gender, data = data.frame(df$Customer.Lifetime.Valu
e,df$Gender), FUN=mean, na.rm=T)
```

```
##   df$Gender df$Customer.Lifetime.Value
## 1         F                   8096.602
## 2         M                   7909.551
```

The p value is > 0.05.Therefore, we fail to reject the null hypothesis. There is no significant difference in CLV among the customers based on Gender.Thus Gender can be avoided in our model.

Location code

```
kruskal.test(x =df$Customer.Lifetime.Value, g = as.factor(df$Location.Code))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  df$Customer.Lifetime.Value and as.factor(df$Location.Code)
## Kruskal-Wallis chi-squared = 2.4638, df = 2, p-value = 0.2917
```

```
aggregate(df$Customer.Lifetime.Value ~ df$Location.Code, data = data.frame(df$Customer.Lifeti
me.Value,df$Location.Code), FUN=mean, na.rm=T)
```

```
##   df$Location.Code df$Customer.Lifetime.Value
## 1            Rural                   7953.699
## 2         Suburban                   8004.457
## 3            Urban                   8064.133
```

The p value is > 0.05.Therefore, we fail to reject the null hypothesis. There is no significant difference in CLV among the customers based on their Location code.Thus Location Code can be avoided in our model.

Marital Status

```
kruskal.test(x =df$Customer.Lifetime.Value, g = as.factor(df$Marital.Status))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  df$Customer.Lifetime.Value and as.factor(df$Marital.Status)
## Kruskal-Wallis chi-squared = 20.896, df = 2, p-value = 2.901e-05
```

```
aggregate(df$Customer.Lifetime.Value ~ df$Marital.Status, data = data.frame(df$Customer.Lifet
ime.Value,df$Marital.Status), FUN=mean, na.rm=T)
```

```
##   df$Marital.Status df$Customer.Lifetime.Value
## 1          Divorced                   8241.239
## 2           Married                   8078.967
## 3            Single                   7714.837
```

The p value here is < 0.05.Therefore, we reject the null hypothesis. There is significant difference in CLV based on the Marital Status.Thus Marital Status can be useful in predicting CLV.

Policy Type

```
 kruskal.test(x =df$Customer.Lifetime.Value, g = as.factor(df$Policy.Type))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  df$Customer.Lifetime.Value and as.factor(df$Policy.Type)
## Kruskal-Wallis chi-squared = 4.6075, df = 2, p-value = 0.09988
```

```
aggregate(df$Customer.Lifetime.Value ~ df$Policy.Type, data = data.frame(df$Customer.Lifetim
e.Value,df$Policy.Type), FUN=mean, na.rm=T)
```

```
##   df$Policy.Type df$Customer.Lifetime.Value
## 1 Corporate Auto                   7814.410
## 2  Personal Auto                   8027.364
## 3   Special Auto                   8594.245
```

The p value is > 0.05.Therefore, we fail to reject the null hypothesis. There is no significant difference in CLV among the customers based on the Policy Type.Thus Policy Type can be avoided in our model.

Policy

```
 kruskal.test(x =df$Customer.Lifetime.Value, g = as.factor(df$Policy))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  df$Customer.Lifetime.Value and as.factor(df$Policy)
## Kruskal-Wallis chi-squared = 7.9444, df = 8, p-value = 0.4389
```

```
aggregate(df$Customer.Lifetime.Value ~ df$Policy, data = data.frame(df$Customer.Lifetime.Valu
e,df$Policy), FUN=mean, na.rm=T)
```

```
##      df$Policy df$Customer.Lifetime.Value
## 1 Corporate L1                   8474.928
## 2 Corporate L2                   7597.695
## 3 Corporate L3                   7707.722
## 4  Personal L1                   7989.762
## 5  Personal L2                   8054.909
## 6  Personal L3                   8023.912
## 7   Special L1                   8332.763
## 8   Special L2                   8326.906
## 9   Special L3                   9007.092
```

The p value is > 0.05.Therefore, we fail to reject the null hypothesis. There is no significant difference in CLV among the customers based on the Policy.Thus Policy can be avoided in our model.

Renew.Offer.Type

```
 kruskal.test(x =df$Customer.Lifetime.Value, g = as.factor(df$Renew.Offer.Type))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  df$Customer.Lifetime.Value and as.factor(df$Renew.Offer.Type)
## Kruskal-Wallis chi-squared = 168.9, df = 3, p-value < 2.2e-16
```

```
aggregate(df$Customer.Lifetime.Value ~ df$Renew.Offer.Type, data = data.frame(df$Customer.Lif
etime.Value,df$Renew.Offer.Type), FUN=mean, na.rm=T)
```

```
##   df$Renew.Offer.Type df$Customer.Lifetime.Value
## 1              Offer1                   8707.086
## 2              Offer2                   7396.754
## 3              Offer3                   7997.887
## 4              Offer4                   7179.947
```

The p value here is < 0.05.Therefore, we reject the null hypothesis. There is significant difference in CLV based on the Renew offer type.Thus Renew offer type can be useful in predicting CLV.

Sales.Channel

```
 kruskal.test(x =df$Customer.Lifetime.Value, g = as.factor(df$Sales.Channel))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  df$Customer.Lifetime.Value and as.factor(df$Sales.Channel)
## Kruskal-Wallis chi-squared = 4.4918, df = 3, p-value = 0.213
```

```
aggregate(df$Customer.Lifetime.Value ~ df$Sales.Channel, data = data.frame(df$Customer.Lifeti
me.Value,df$Sales.Channel), FUN=mean, na.rm=T)
```

```
##   df$Sales.Channel df$Customer.Lifetime.Value
## 1            Agent                   7957.709
## 2           Branch                   8119.712
## 3      Call Center                   8100.086
## 4              Web                   7779.788
```

The p value is > 0.05.Therefore, we fail to reject the null hypothesis. There is no significant difference in CLV among the customers based on the Sales channel.Thus Sales channel can be avoided in our model.

Vehicle.Class

```
 kruskal.test(x =df$Customer.Lifetime.Value, g = as.factor(df$Vehicle.Class))
```

```
## 
##  Kruskal-Wallis rank sum test
## 
## data:  df$Customer.Lifetime.Value and as.factor(df$Vehicle.Class)
## Kruskal-Wallis chi-squared = 1310.5, df = 5, p-value < 2.2e-16
```

```
aggregate(df$Customer.Lifetime.Value ~ df$Vehicle.Class, data = data.frame(df$Customer.Lifeti
me.Value,df$Vehicle.Class), FUN=mean, na.rm=T)
```

```
##   df$Vehicle.Class df$Customer.Lifetime.Value
## 1    Four-Door Car                   6631.727
## 2       Luxury Car                  17053.348
## 3       Luxury SUV                  17122.999
## 4       Sports Car                  10750.989
## 5              SUV                  10443.512
## 6     Two-Door Car                   6671.031
```

The p value here is < 0.05.Therefore, we reject the null hypothesis. There is significant difference in CLV based on the vehicle class the customers own.Thus Vehicle class of customers can be useful in predicting CLV.

Vehicle.Size

```
 kruskal.test(x =df$Customer.Lifetime.Value, g = as.factor(df$Vehicle.Size))
```

```
## 
##  Kruskal-Wallis rank sum test
## 
## data:  df$Customer.Lifetime.Value and as.factor(df$Vehicle.Size)
## Kruskal-Wallis chi-squared = 9.565, df = 2, p-value = 0.008375
```

```
aggregate(df$Customer.Lifetime.Value ~ df$Vehicle.Size, data = data.frame(df$Customer.Lifetim
e.Value,df$Vehicle.Size), FUN=mean, na.rm=T)
```

```
##   df$Vehicle.Size df$Customer.Lifetime.Value
## 1           Large                   7544.996
## 2         Medsize                   8050.662
## 3           Small                   8085.096
```

The p value here is < 0.05.Therefore, we reject the null hypothesis. There is significant difference in CLV based on the vehicles sizes.Thus Vehicle Sizes of customers can be useful in predicting CLV.

Therefore based on Kruskal Wallace test, the categorical variables that would help in predicting the CLV are:

- Vehicle.Size
- Vehicle.Class
- Renew.Offer.Type
- Marital.Status
- Coverage
- Education
- EmploymentStatus