Tukey HSD Tests for Cases per Neighborhood

This notebook is used to conduct if there is significant difference between COVID-19 cases and rate between neighborhood clusters as detected earlier in Python.

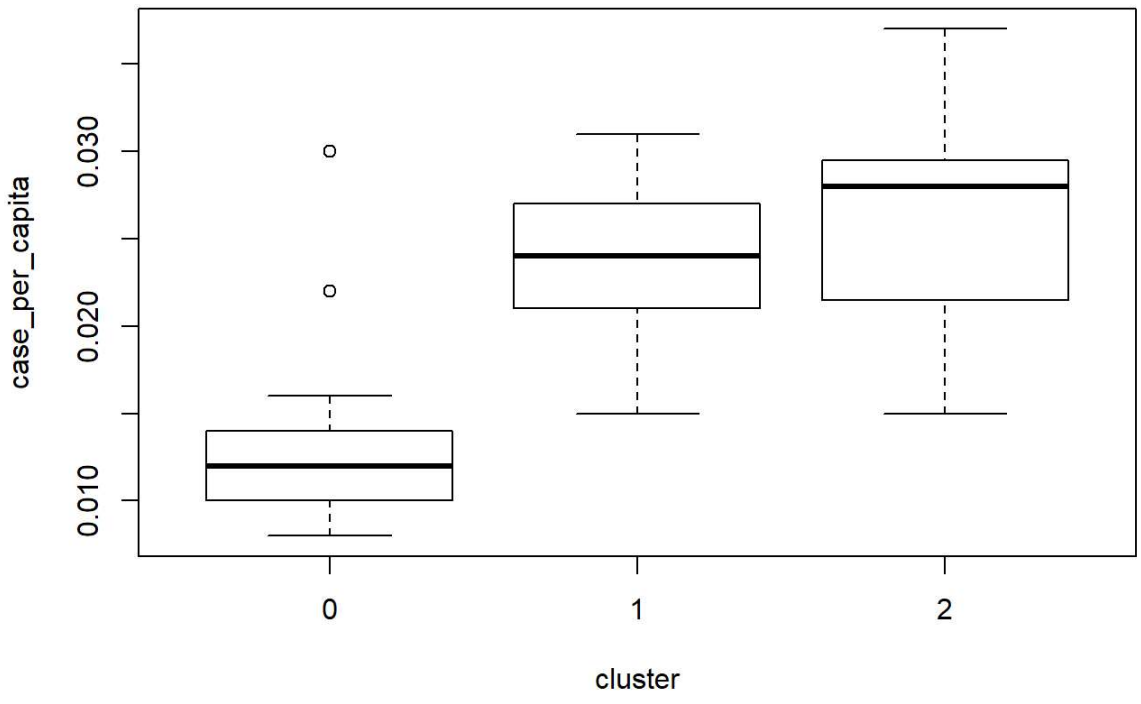
```
# Read in the data
data = read.csv("data/case_per_neighborhood.csv", head = TRUE, sep = ",")

data$cluster = factor(data$cluster)

# Show the first few rows of data
head(data, 3)
```

	borough <fctr>	uhf34_neigh <fctr>	uhf42_neigh <fctr>	population <int>
1	Queens	Bayside Little Neck-Fresh Meadows	Bayside - Little Neck	87423
2	Queens	Bayside Little Neck-Fresh Meadows	Fresh Meadows	95537
3	Brooklyn	Bedford Stuyvesant - Crown Heights	Bedford Stuyvesant - Crown Heights	316269

3 rows | 1-5 of 13 columns

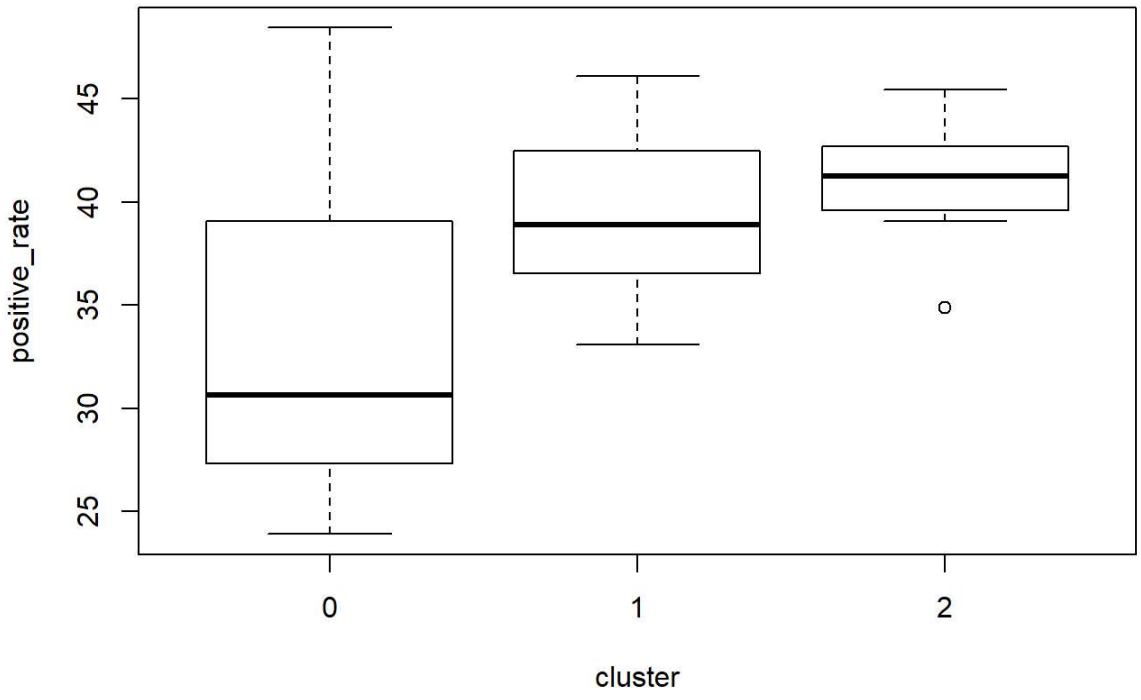
Boxplot of *case_per_capita* and *cluster*, with *case_per_capita* on the vertical axis:

```
boxplot(case_per_capita~cluster,data=data,
        xlab="cluster", ylab="case_per_capita")
```



There seems to be noticeable difference in means of *case_per_capita* across the clusters. Let's see about *positive_rate*:

```
boxplot(positive_rate~cluster,data=data,
        xlab="cluster", ylab="positive_rate")
```



First, one-way ANOVA tests will be conducted on *case_per_capita* to confirm if the means across clusters are statistically significant. The null hypothesis is that all means are equal.

```
case_per_capita_aov = aov(case_per_capita~cluster,data=data)

summary(case_per_capita_aov)
```

```
##             Df   Sum Sq   Mean Sq F value   Pr(>F)
## cluster      2 0.001095 0.0005474   16.77 5.55e-06 ***
## Residuals   39 0.001273 0.0000326
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

At α-level of 0.01,F-test p-value here is much smaller than that. Therefore, we reject the null hyphothesis of equal *case_per_capita* means between *cluster*. At least a pair of clusters has statistically significantly different *case_per_capita* means.

```
TukeyHSD(case_per_capita_aov, conf.level = 0.99)
```

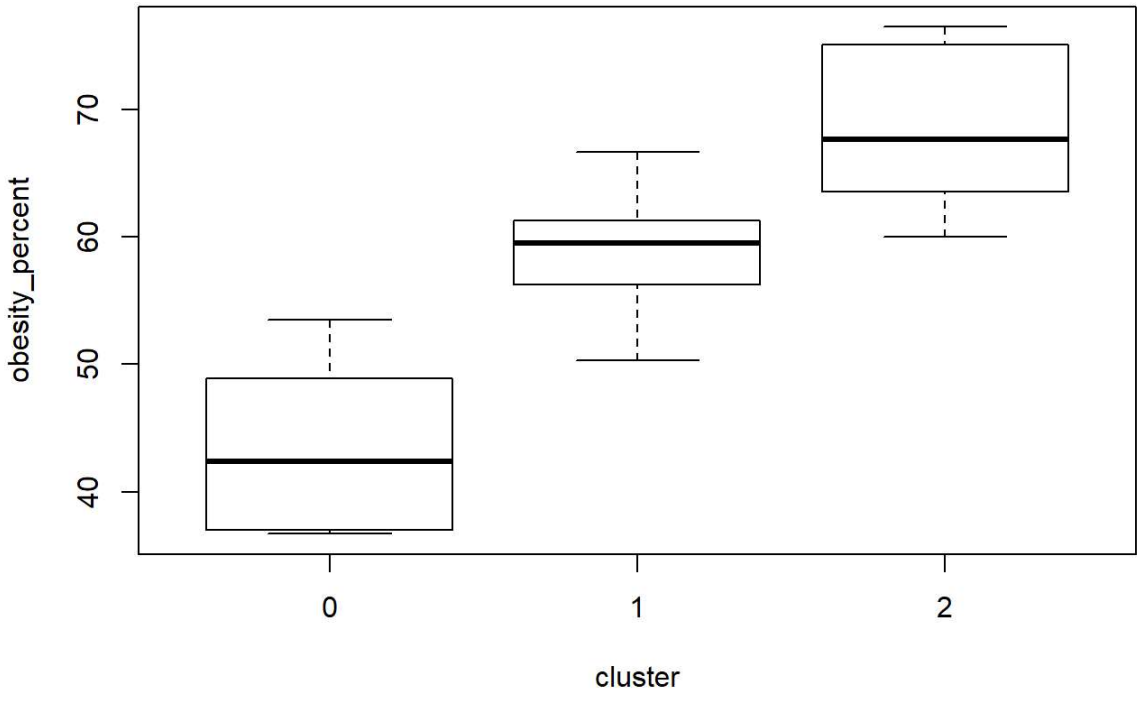
```
##   Tukey multiple comparisons of means
##     99% family-wise confidence level
##
## Fit: aov(formula = case_per_capita ~ cluster, data = data)
##
## $cluster
##            diff          lwr         upr     p adj
## 1-0 0.009484163  0.002974251 0.015994074 0.0001705
## 2-0 0.012474359  0.005401120 0.019547597 0.0000087
## 2-1 0.002990196 -0.003671649 0.009652041 0.3568170
```

As Tukey HSD test confirms, there is statistically signficant difference between Cluster 1 and Cluster 0 and also between Cluster 1 and Cluster 2 while the test do not reject the null hypothesis for difference between Cluster 2 and Cluster 0. Cluster 1 has the lowest *case_per_capita* mean.

Now, even if Cluster 2 and Cluster 0 has probably equal means in *case_per_capita*. They might be different in health issues that led to the division into 2 clusters.

Boxplot of *obesity_percent* and *cluster*, with *obesity_percent* on the vertical axis:

```
boxplot(obesity_percent~cluster,data=data,
        xlab="cluster", ylab="obesity_percent")
```




```
obesity_aov = aov(obesity_percent~cluster,data=data)

summary(obesity_aov)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## cluster      2   3983  1991.6   57.27 2.48e-12 ***
## Residuals   39   1356    34.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA test does confirm the difference in *obesity_percent* means across clusters.

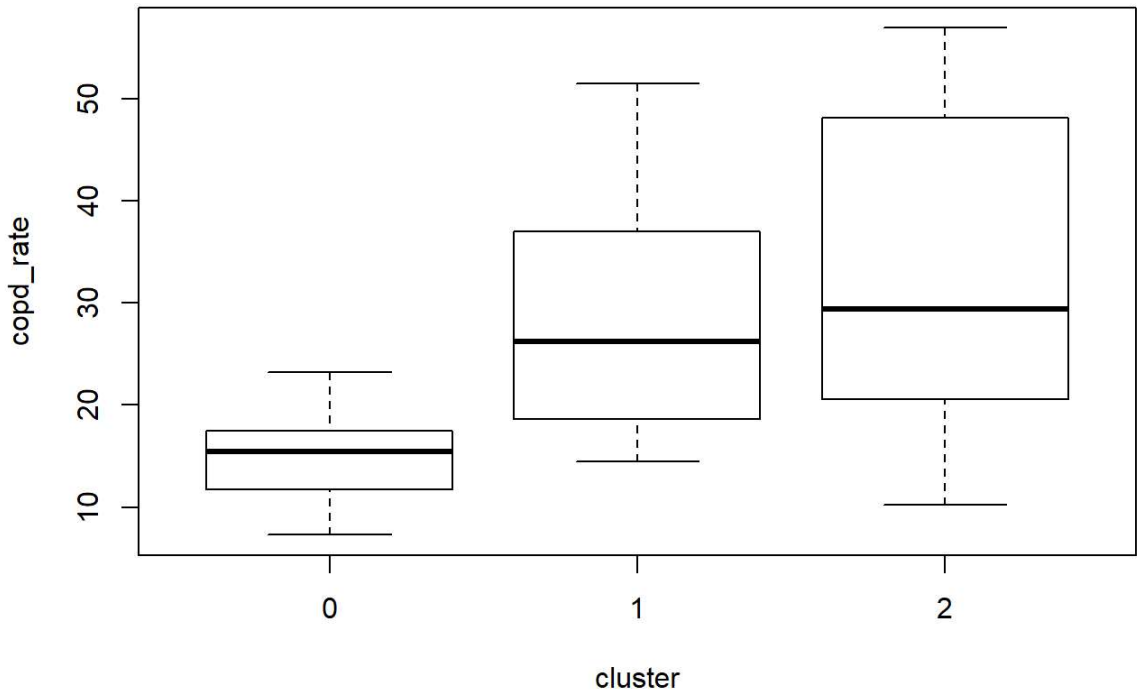
```
TukeyHSD(obesity_aov, conf.level = 0.99)
```

```
##   Tukey multiple comparisons of means
##     99% family-wise confidence level
##
## Fit: aov(formula = obesity_percent ~ cluster, data = data)
##
## $cluster
##          diff       lwr      upr     p adj
## 1-0 15.384163  8.664779 22.10355 0.0000000
## 2-0 24.882692 17.581855 32.18353 0.0000000
## 2-1  9.498529  2.622323 16.37474 0.0003475
```

All clusters have statiscally different *obesity_aov* with Cluster 1 still has the lowest while Cluster 2 has the highest mean.

Boxplot of *copd_rate* and *cluster*, with *copd_rate* on the vertical axis:

```
boxplot(copd_rate~cluster,data=data,
        xlab="cluster", ylab="copd_rate")
```



```
copd_aov = aov(copd_rate~cluster,data=data)

summary(copd_aov)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## cluster      2   2395  1197.3   8.931 0.000641 ***
## Residuals   39   5228   134.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA test confirms the difference in *copd_rate* means across clusters.

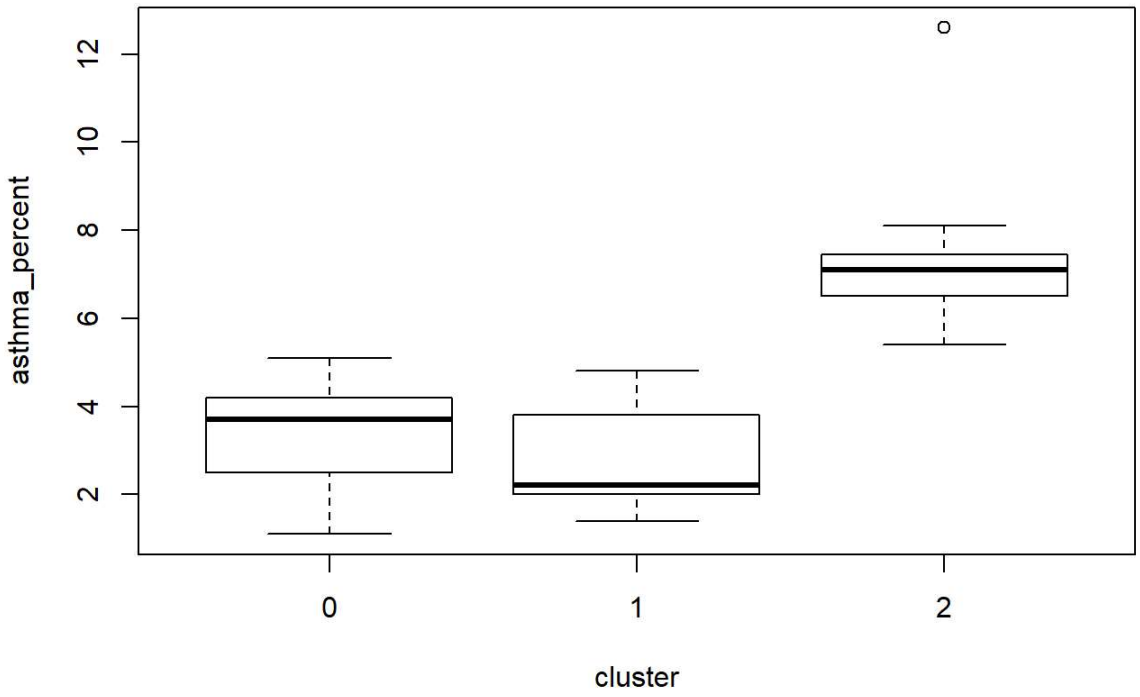
```
TukeyHSD(copd_aov, conf.level = 0.99)
```

```
##   Tukey multiple comparisons of means
##     99% family-wise confidence level
##
## Fit: aov(formula = copd_rate ~ cluster, data = data)
##
## $cluster
##          diff        lwr      upr    p adj
## 1-0 13.562443  0.3688863 26.75600 0.007956
## 2-0 18.698718  4.3634731 33.03396 0.000709
## 2-1  5.136275 -8.3652039 18.63775 0.474022
```

While again, Cluster 1 demonstrates a significant lower *copd_rate* mean between other clusters, Cluster 0 and Cluster 2 does not have any statistically difference in this health aspect.

Boxplot of *asthma_percent* and *cluster*, with *asthma_percent* on the vertical axis:

```
boxplot(asthma_percent~cluster,data=data,
        xlab="cluster", ylab="asthma_percent")
```




```
asthma_aov = aov(asthma_percent~cluster,data=data)

summary(asthma_aov)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## cluster      2 159.97   79.99   40.87 2.69e-10 ***
## Residuals   39  76.32    1.96
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA test confirms the difference in *asthma_percent* means across clusters.

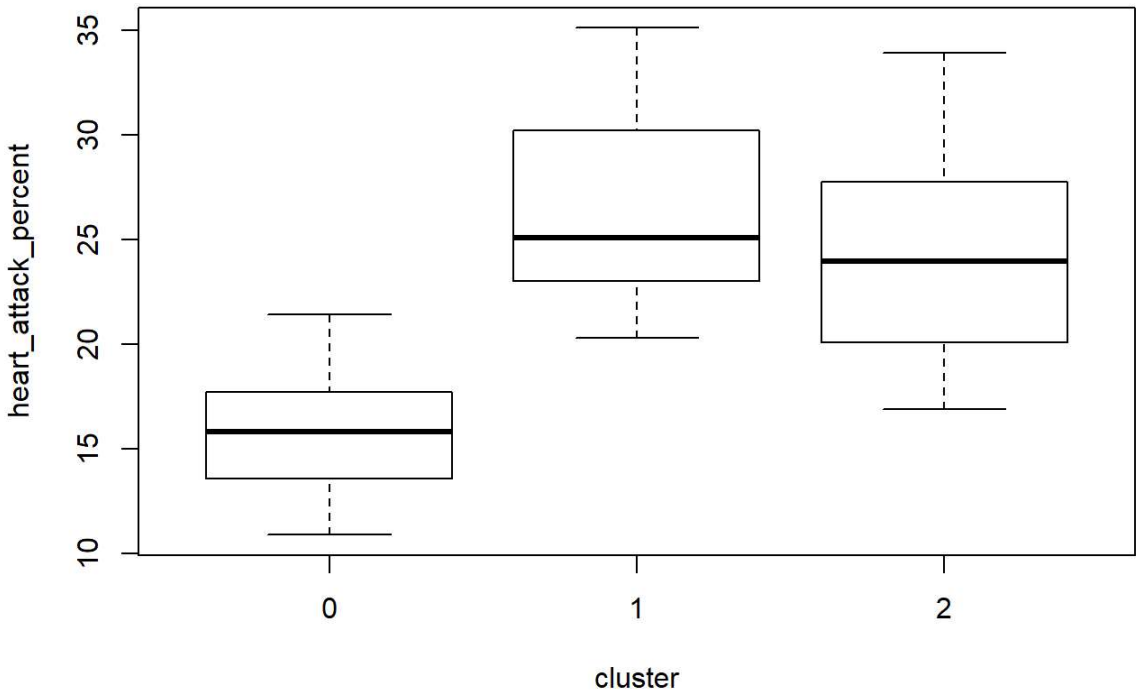
```
TukeyHSD(asthma_aov, conf.level = 0.99)
```

```
##   Tukey multiple comparisons of means
##     99% family-wise confidence level
##
## Fit: aov(formula = asthma_percent ~ cluster, data = data)
##
## $cluster
##           diff       lwr       upr     p adj
## 1-0 -0.6379852 -2.232039 0.9560681 0.4385190
## 2-0  3.9179487  2.185956 5.6499414 0.0000001
## 2-1  4.5559340  2.924677 6.1871905 0.0000000
```

Interestingly, only Cluster 2 exhibits a statiscally higher *asthma_percent* mean than other clusters. Cluster 1 and Cluster 0 does not have a significant difference between their *asthma_percent*.

Boxplot of *heart_attack_percent* and *cluster*, with *heart_attack_percent* on the vertical axis:

```
boxplot(heart_attack_percent~cluster,data=data,
        xlab="cluster", ylab="heart_attack_percent")
```




```
heart_aov = aov(heart_attack_percent~cluster,data=data)

summary(heart_aov)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## cluster      2  916.6   458.3   22.74 2.85e-07 ***
## Residuals   39  786.1    20.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA test confirms the difference in *heart_attack_percent* means across clusters.

```
TukeyHSD(heart_aov, conf.level = 0.99)
```

```
##   Tukey multiple comparisons of means
##     99% family-wise confidence level
##
## Fit: aov(formula = heart_attack_percent ~ cluster, data = data)
##
## $cluster
##          diff       lwr       upr     p adj
## 1-0 10.815837  5.700055 15.931619 0.0000003
## 2-0  8.658974  3.100505 14.217444 0.0000649
## 2-1 -2.156863 -7.392040  3.078315 0.4179907
```

We see something similar again in *heart_attack_percent*, Cluster 1 again has signficantly lower *heart_attack_percent* mean than the other 2 while between Cluster 0 and Cluster 2 they do not have a statistically difference.

In conclusion, the lower *case_per_capita* in Cluster 1 could be driven by the fact that it has lower *obesity_percent*, *copd_rate*, and *heart_attack_percent* than other clusters.

Though Cluster 0 and Cluster 2 do not have a statistically difference in *case_per_capita* means, they are different because of their difference in *obesity_percent* and *asthma_percent* .