Untitled

Tata iles

2023-06-12

```
library(reticulate)
use_python("C:/Users/tatai/AppData/Local/Programs/Python/Python311/python.exe")
```

Libraries

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
sns.set_theme(style = "darkgrid")
```

Data

```
data = pd.read_csv("D:/Downloads/archive (1)/Mall_Customers.csv")
```

```
data = pyData
data %>% head()
```

```
##   CustomerID Gender Age Annual Income (k$) Spending Score (1-100)
## 1          1   Male  19                  15                      39
## 2          2   Male  21                  15                      81
## 3          3 Female  20                  16                       6
## 4          4 Female  23                  16                      77
## 5          5 Female  31                  17                      40
## 6          6 Female  22                  17                      76
```

EDA

```
data.info()
```

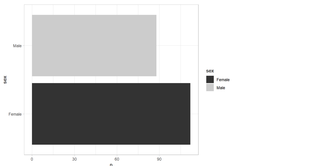
```
## <class 'pandas.core.frame.DataFrame'>
## RangeIndex: 200 entries, 0 to 199
## Data columns (total 5 columns):
##  #   Column                  Non-Null Count  Dtype 
## ---  ------                  --------------  ----- 
##  0   CustomerID              200 non-null    int64 
##  1   Gender                  200 non-null    object
##  2   Age                     200 non-null    int64 
##  3   Annual Income (k$)      200 non-null    int64 
##  4   Spending Score (1-100)  200 non-null    int64 
## dtypes: int64(4), object(1)
## memory usage: 7.9+ KB
```

```
data %>% summary()
```

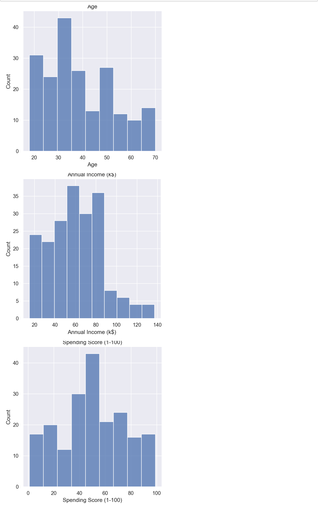
```
##    CustomerID        Gender               Age        Annual Income (k$)
##  Min.   :  1.00   Length:200         Min.   :18.00   Min.   : 15.00    
##  1st Qu.: 50.75   Class :character   1st Qu.:28.75   1st Qu.: 41.50    
##  Median :100.50   Mode  :character   Median :36.00   Median : 61.50    
##  Mean   :100.50                      Mean   :38.85   Mean   : 60.56    
##  3rd Qu.:150.25                      3rd Qu.:49.00   3rd Qu.: 78.00    
##  Max.   :200.00                      Max.   :70.00   Max.   :137.00    
##  Spending Score (1-100)
##  Min.   : 1.00         
##  1st Qu.:34.75         
##  Median :50.00         
##  Mean   :50.20         
##  3rd Qu.:73.00         
##  Max.   :99.00
```

```
colnames(data) <- c("id","sex","age","annual_score","spending_score")
data <- data %>% select(-1)
```

```
data %>% count(sex) %>% ggplot(aes(n,sex,fill=sex)) + geom_col() + scale_fill_grey() + theme_light()
```




```
for i in data.columns[1:] :
  sns.displot(x=i,data = data)
  plt.title(i)
  plt.show();
```



Clustering

Recipe

```
t_sne_rec <- recipe(~.,data = data) %>%
  step_normalize(all_numeric()) %>%
  step_dummy(all_nominal())

df <- t_sne_rec %>% prep(data) %>% juice()
```

Cross-Validation

```
cv <- vfold_cv(data,v = 5)
```

Tuning grid

```
grid <- tibble(num_clusters = seq(1,15))
```

K-means workflow

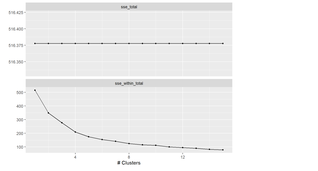
```
k_means_spec <- k_means(num_clusters = tune()) %>% set_engine("stats")

kmeans_workflow <- workflow() %>%
add_model(k_means_spec) %>% add_recipe(t_sne_rec)
```

Tuning

```
tune_res <- tune_cluster(object = kmeans_workflow,resamples = cv,grid = grid)


tune_res %>% autoplot()
```



Elbow method (5 clusters)

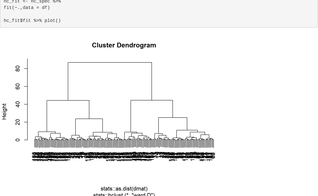
```
final_kmeans <- kmeans_workflow %>%
update_model(k_means_spec %>% set_args(num_clusters = 5)) %>%
fit(data)
```

3D representation

```
augment(final_kmeans, new_data = data) %>%
plot_ly(x = ~spending_score , y = ~annual_score, z = ~age , color = ~.pred_cluster)
```

```
## No trace type specified:
##   Based on info supplied, a 'scatter3d' trace seems appropriate.
##   Read more about this trace type -> https://plotly.com/r/reference/#scatter3d
```

```
## No scatter3d mode specified:
##   Setting the mode to markers
##   Read more about this attribute -> https://plotly.com/r/reference/#scatter-mode
```

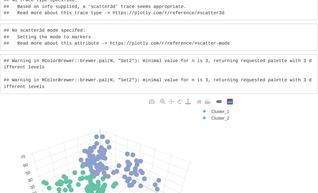


Hierarchical clustering

```
hc_spec <- hier_clust(linkage_method = "ward.D")

hc_fit <- hc_spec %>%
fit(~.,data = df)

hc_fit$fit %>% plot()
```




```
final_hir <- kmeans_workflow %>%
update_model(k_means_spec %>% set_args(num_clusters = 2)) %>%
fit(data)

augment(final_hir, new_data = data) %>%
plot_ly(x = ~spending_score , y = ~annual_score, z = ~age , color = ~.pred_cluster)
```

```
## No trace type specified:
##   Based on info supplied, a 'scatter3d' trace seems appropriate.
##   Read more about this trace type -> https://plotly.com/r/reference/#scatter3d
```

```
## No scatter3d mode specified:
##   Setting the mode to markers
##   Read more about this attribute -> https://plotly.com/r/reference/#scatter-mode
```

```
## Warning in RColorBrewer::brewer.pal(N, "Set2"): minimal value for n is 3, returning requested palette with 3 d
ifferent levels

## Warning in RColorBrewer::brewer.pal(N, "Set2"): minimal value for n is 3, returning requested palette with 3 d
ifferent levels
```