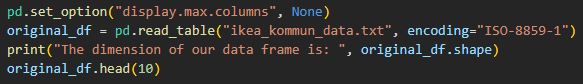
|  |
| --- |
| DALARNA UNIVERSITY |
| LAB # 3 |
| Submitted to: Hasan Felyeh, Saumya Gupta |

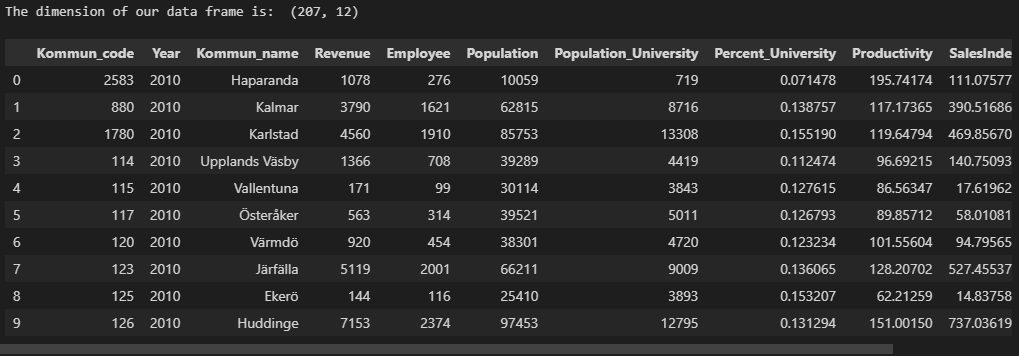
|  |
| --- |
| Sundas Khan  5-17-2022 |

**IKEA DATASET**

**Loading Of Original Dataset:**

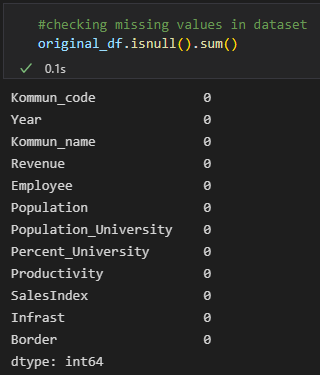


The dimensions of given data set along with top 10 rows are given below:

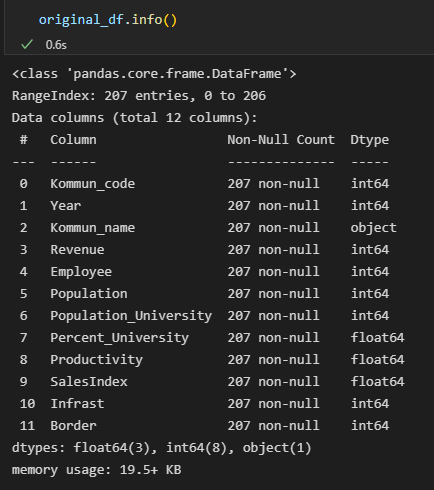


**Checking Missing Values:**

Checking of missing values is given dataset leads to following results:



**Checking Datatypes:**

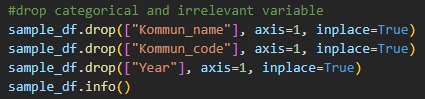


The checking of data types for given dataset reveal one non-numeric data type ***Kommun\_name***.

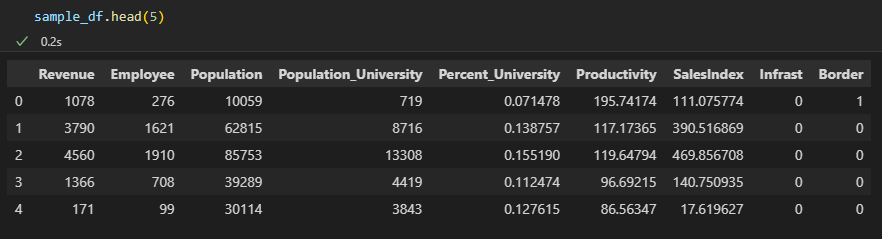
**Removal of Categorical & Irrelevant Variables:**

The original dataset is saved in sample dataset and irrelevant columns like Kommun\_code, Year and Kommun\_name is removed using drop function.





Now our new dataset holds following variables.



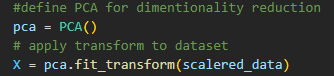
Now before fitting given dataset to Kmeans model, first we need to compute the optimal number of clusters using Elbow method, then we need to improve given dataset scaling and finally applying PCA for dimensionality reduction.

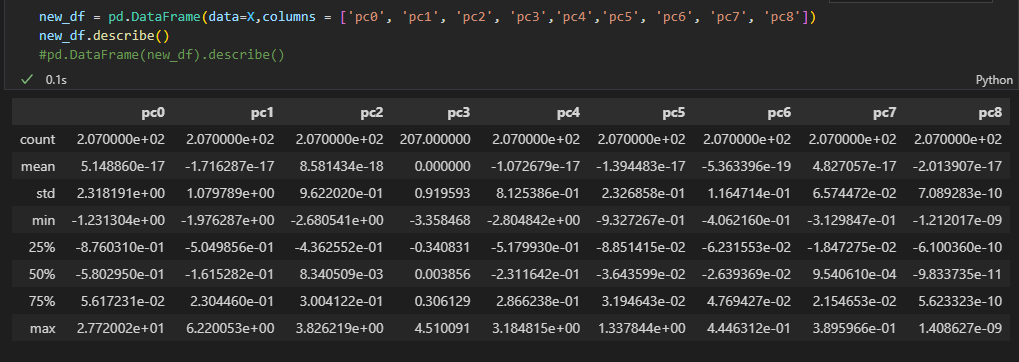
Chart, line chart

Description automatically generated

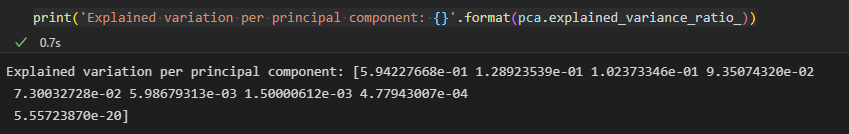
The Elbow method reveal number of clusters = 3





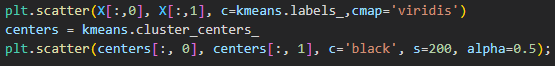


After applying PCA, the variation for each component is given by:



The output from PCA is now applied to Kmeans function

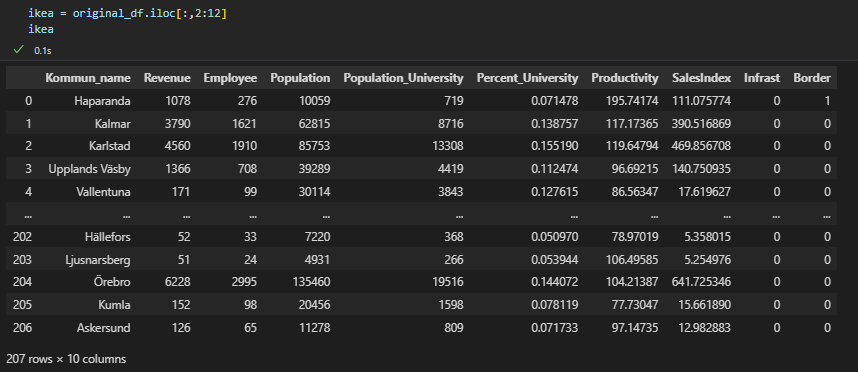




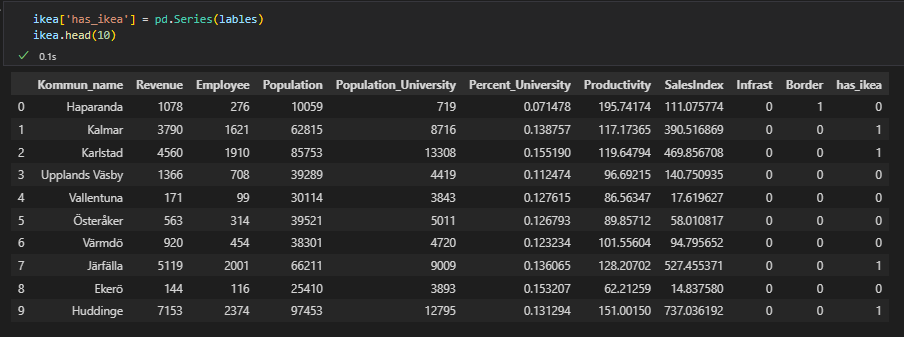
Chart, scatter chart

Description automatically generated

The original dataset has been saved to a new variable Ikea



The predicted labels from KMeans function have been appended to Ikea dataset with column name has\_ikea



**Clusters Formation**

Now we will assign each unique value of has\_ikea to a cluster variable



Below are the sizes of each cluster:



The maximum population for each cluster is given below:



The maximum sales index for each cluster is given below:



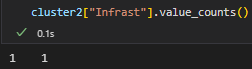
The maximum revenue for each cluster is given below:

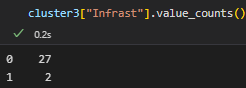


The examination of Infrast variable for each cluster is given below.

Text

Description automatically generated with medium confidence

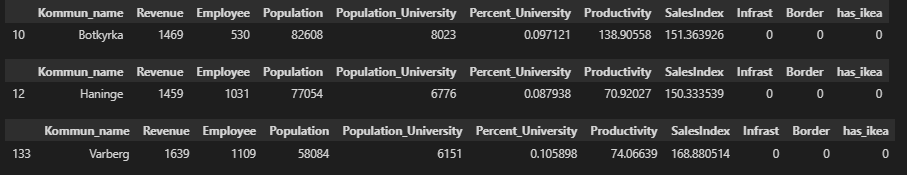




**Analysing Each Cluster based on Population**

Now if we examine Kommun for each cluster based on maximum population we found following results:

**For Cluster # 1**



For top 3 Kommun regions in Cluster 1, there is no Ikea infrastructure.

**For Cluster # 2**

There is Ikea infrastructure for Kommun in Cluster 2.

**For Cluster # 3**

A black screen with white text

Description automatically generated with low confidence

For Cluster 3, Malmo kommun already has Ikea infrastructure, while following Uppsala and Linkoping does not have Ikea.

**Observations:**

Top three Kommuns are:

|  |  |  |
| --- | --- | --- |
|  | **Kommun** | **Population** |
|  | Uppsala | 197787 |
|  | Linkoping | 146416 |
|  | Botkyrka | 82608 |

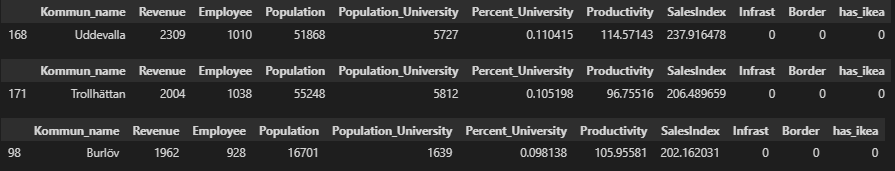
**Justification of Selection:**

The above selection is made purely on population basis however, above selection can also be justified based on physical distance between Kommuns. From Cluster 1, the distance between **Botkyrka** and Hainge is almost 27km and Botkyrka and Varberg is 9.3 km, so if Ikea open a store at Botkyrka, it will attract consumers from other near Kommuns as well. Also it can serve consumers from Stockholm as the distance between two kommuns is just 18 km. For Cluster 3, since Malmo already have Ikea and other two Kommuns with most population does not have Ikea infrastructure and the distance between Uppsala and Linkoping is 247 km, so we suggest having Ikea both in **Uppsala** & **Linkoping** since the distance between them is quite far also they have good population/ consumer size to cater.

**Analysing Each Cluster based on SalesIndex**

Now if we analyse each Kommun based on highest value of SalesIndex, we found following results.

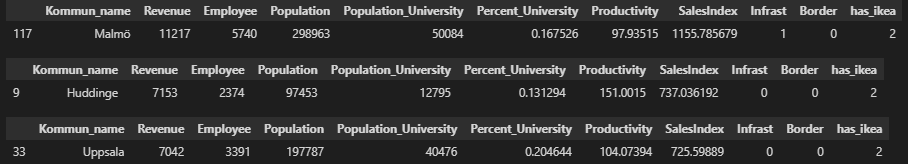
**For Cluster # 1**



**For Cluster # 2**



**For Cluster # 3**



**Observations:**

Top three Kommuns are:

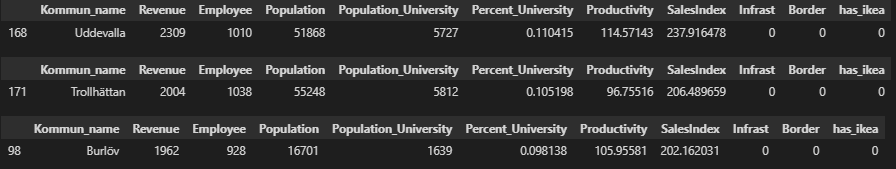
|  |  |  |
| --- | --- | --- |
|  | **Kommun** | **SalesIndex** |
|  | Huddinge | 737.036 |
|  | Uppsala | 725.598 |
|  | Uddevalla | 237.916 |

**Justification of Selection:**

The selection is made primarily on highest value of SalesIndex among all clusters. Though Stockholm has highest value for SalesIndex but it already has Ikea. **Huddinge** and **Uppsala** have highest value of SalesIndex, after Stockholm and the distance between them is 83km, so we suggest Ikea should open its new branches here. Our last choice is, Uddevalla with 51868 population, and SalesIndex 237.916. Also, Uddevalla is almost above 400km away in distance from Stockholm, Huddinge and Uppsala.

**Analysing Each Cluster based on Revenue**

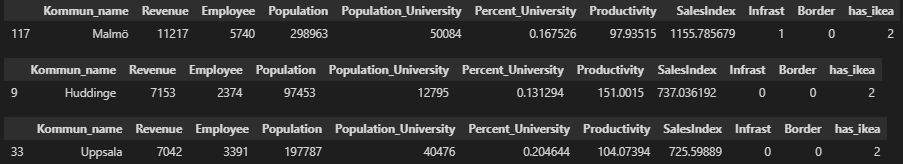
**Cluster # 1**

****

**Cluster # 2**

****

**Cluster # 3**

****

**Observations:**

Top three Kommuns are:

|  |  |  |
| --- | --- | --- |
|  | **Kommun** | **Revenue** |
|  | Huddinge | 7153 |
|  | Uppsala | 7042 |
|  | Uddevalla | 2309 |

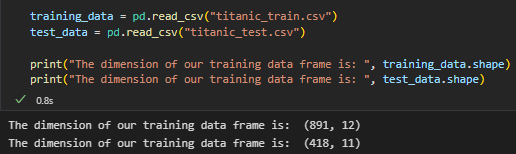
**Justification of Selection:**

The selection is made based on highest Revenue generated by the Kommun. Although Stockholm and Malmo come in first two positions in this regard, however they already have Ikea. From rest of the Kommuns, Huddinge, Uppsala and Uddevalla generate most of the revenue and they do not have Ikea, so If new Ikea location is suggested based on Revenue these three Kommuns come in top three positions.

**TITANIC DATASET**

**Loading Of Original Dataset:**

Both training and test datasets are loaded, and their shapes are printed.

****

**Checking & Imputating NULL values for both training & test data**

Initial null values both for training and test datasets are computed and then replaced with mean values.

|  |  |
| --- | --- |
| **Before Imputating** | **After Imputating** |
| **Training Dataset** | |
|  |  |
| **Test Dataset** | |
|  |  |

**Computing Correlation both for Training & Test Data**

The correlation computation shows Fare variable strong correlation with Pclass variable. So, idea is to create new variable using Fare to explore if we can surpass this correlation

**Training Dataset: Test Dataset:**

Graphical user interface, application

Description automatically generatedGraphical user interface, chart, application

Description automatically generated

**Recoding of Categorical Variables both for Training & Test Data**

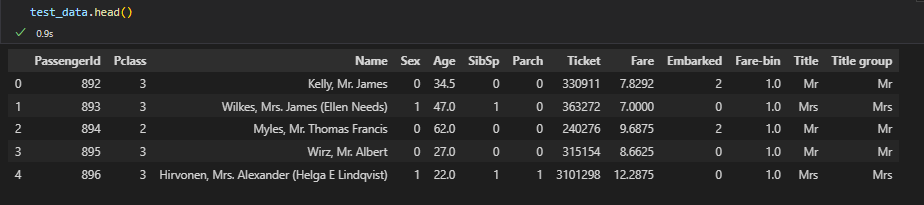
****

**Creating New Variables**

* Fare-bin variable is computed from Fare both for Training & Test Data
* Title variable is created using Name variable from given datasets
  + Title group is created using Titles
* Creating Travel Size using PassangerID and Ticket

**Training Dataset**

****

**Test Dataset**

**Computing Correlation**

Even after creating new feature Fare bin, Fare & Fare bin holds strong correlation with P-class so we are going to remove them from dataset

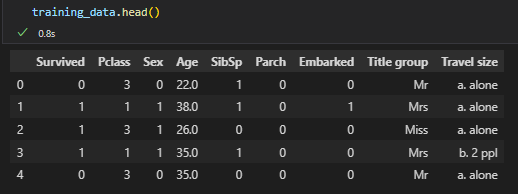
Chart, treemap chart

Description automatically generated

**Removing irrelevant variables from training and test dataset**

****

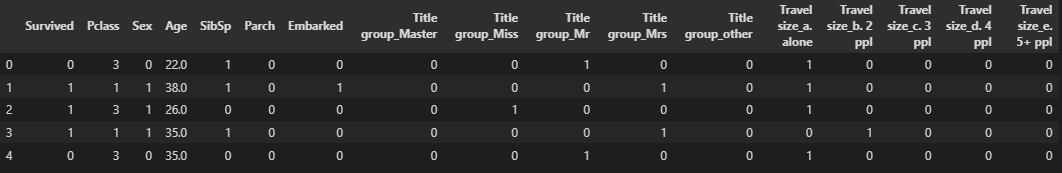
**Training Dataset**

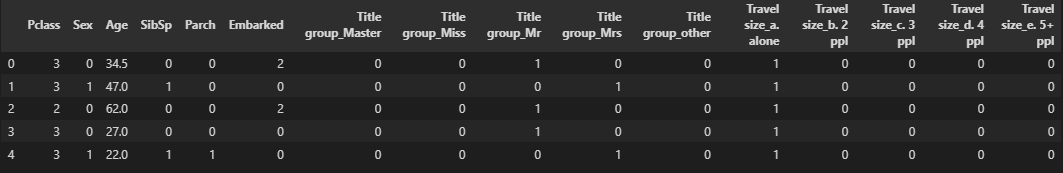
****

**Test Dataset**

****

Since our dataset have non-numeric variables so idea is to create dummy variables both for training and test datasets

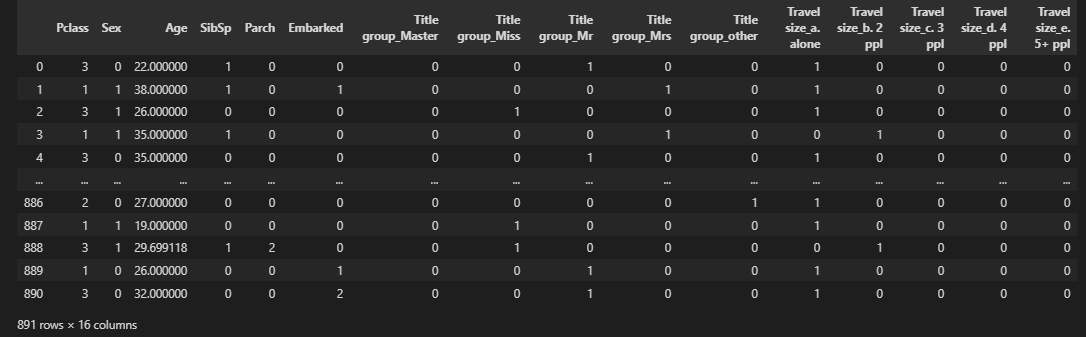
****

****

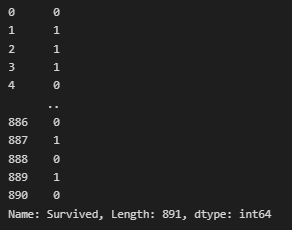
**Preparing Data for Modelling**

Defining predictor and response variables for training data. In training dataset, ***Survived*** is our response variable and all other variable are predictors.

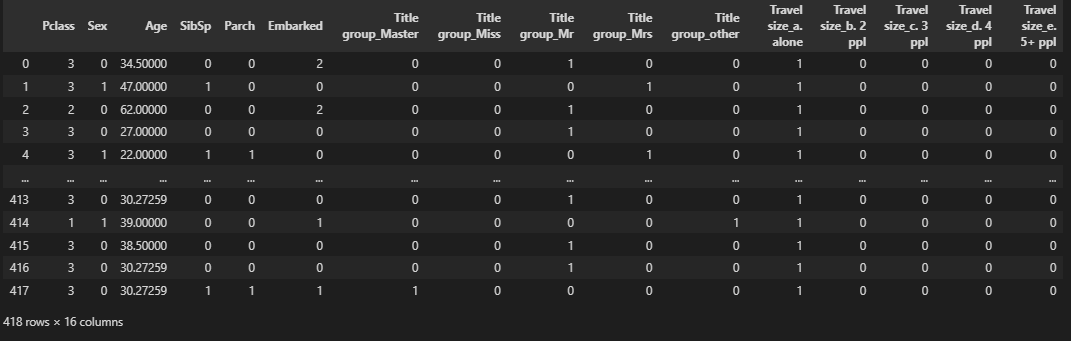
**X-train:**



**Y-train:**



**X-test:**

****

**Now our data is ready for use in SVM, Decision Trees and Logistic Regression Models**

The model’s accuracy scores are computed using k-fold cross validation

|  |  |
| --- | --- |
| **Models Accuracy for Training Dataset** | **Models Accuracy for Test Data** |
|  |  |

**Analysis:**

For training dataset, Logistic regression model performed best with highest score of 82.3%, while SVM scored 65.4%. For test dataset, in predicting Survived variable, Logistic regression stood first while SVM also performed better than its performance for training dataset. Decision trees score also improved for test dataset from 78.3% to 88.2%.