Predict Customer Personality to boost marketing campaign by using Machine Learning

Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com

Created by:
Fairuz Nur Indah Putri
fnputri31@gmail.com
https://www.linkedin.com/in/fairuz-nur-indah-p/

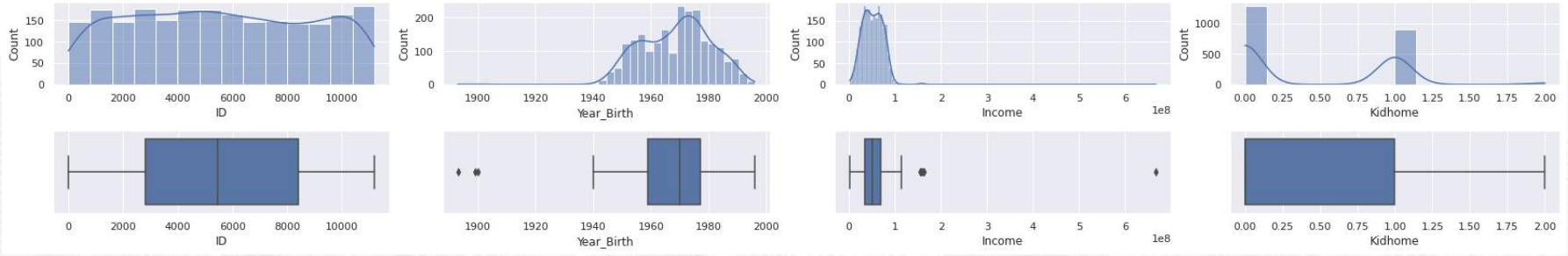
An Engineering Graduate that highly motivated about learning new things to improve my skills, especially in machine learning, quantitative analytics, and statistics. I am currently interested in working in the data field, especially being a data scientist. I have experience in using **Python**, **R**, and **SQL** and have several machine learning projects that I have worked on including **supervised**, **unsupervised**, **product recommendation**, **forecasting**, and **deep learning** with **tensorflow**.

Overview

“Sebuah perusahaan dapat berkembang dengan pesat saat mengetahui perilaku customer personality nya, sehingga dapat memberikan layanan serta manfaat lebih baik kepada customers yang berpotensi menjadi loyal customers. Dengan mengolah data historical marketing campaign guna menaikkan performa dan menyasar customers yang tepat agar dapat bertransaksi di platform perusahaan, dari insight data tersebut fokus kita adalah membuat sebuah model prediksi kluster sehingga memudahkan perusahaan dalam membuat keputusan ”

Conversion Rate Analysis Based on Income, Spending and Age

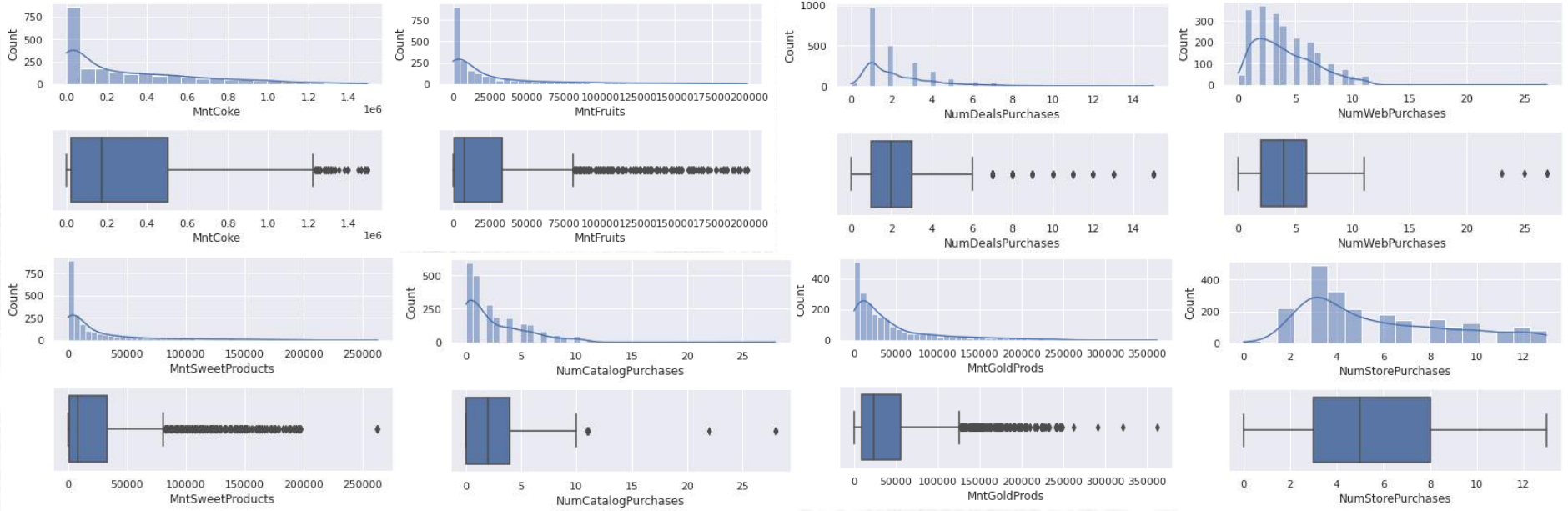
1. EDA (Numeric Columns)



Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Analysis Based on Income, Spending and Age

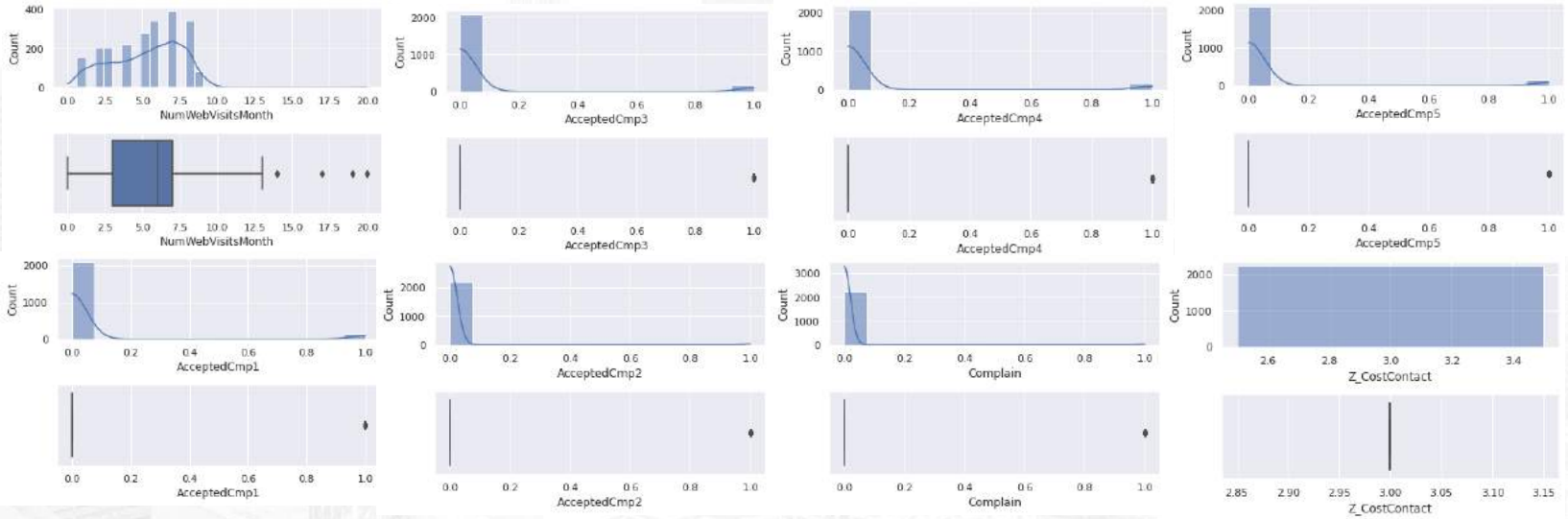
1. EDA (Numeric Columns)



Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Analysis Based on Income, Spending and Age

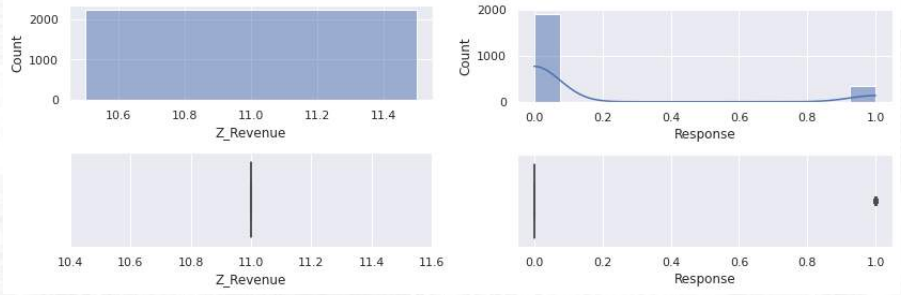
1. EDA (Numeric Columns)



Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Analysis Based on Income, Spending and Age

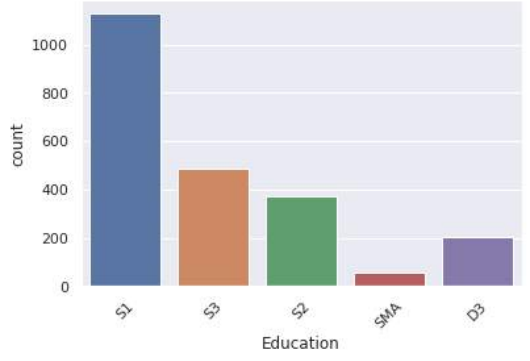
1. EDA (Numeric Columns)

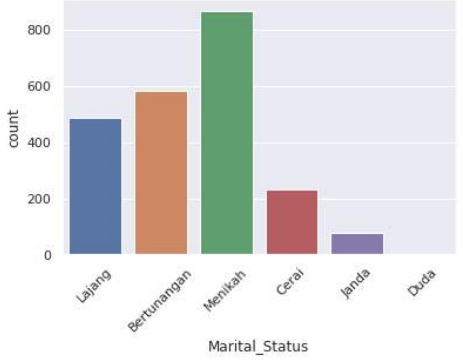


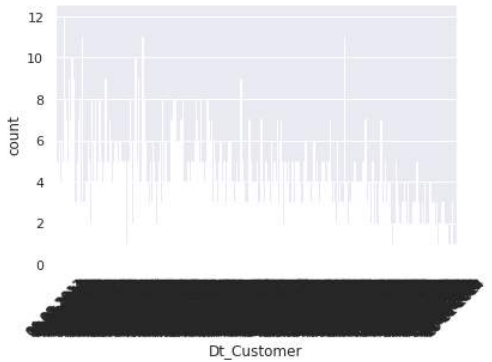
Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Analysis Based on Income, Spending and Age

1. EDA (Category Columns)







Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Analysis Based on Income, Spending and Age

1. EDA (Correlation)

```
threshold = 0.7
couple = []
for col in corr.columns:
  for idx in corr.index:
    if col != idx:
      if [col, idx] not in couple:
        couple.append([idx, col])
        if np.abs(corr.loc[idx, col]) > threshold:
          print("Correlation between {} and {} is {}".format(idx, col, corr.loc[idx, col]))
```

```
Correlation between MntCoke and Income is 0.830056050542334
Correlation between MntMeatProducts and Income is 0.8168150203811545
Correlation between NumCatalogPurchases and Income is 0.7918406763718138
Correlation between NumStorePurchases and Income is 0.7317522657608556
Correlation between MntMeatProducts and MntCoke is 0.8236521541141395
Correlation between NumWebPurchases and MntCoke is 0.74019511433521
Correlation between NumCatalogPurchases and MntCoke is 0.8234208892802969
Correlation between NumStorePurchases and MntCoke is 0.8069178612408089
Correlation between MntMeatProducts and MntFruits is 0.7131685597805494
Correlation between MntFishProducts and MntFruits is 0.7050161898585481
Correlation between MntFishProducts and MntMeatProducts is 0.7262415247350231
Correlation between NumCatalogPurchases and MntMeatProducts is 0.8516600511783947
Correlation between NumStorePurchases and MntMeatProducts is 0.7793355920302663
Correlation between MntSweetProducts and MntFishProducts is 0.7008709124970488
Correlation between NumStorePurchases and NumCatalogPurchases is 0.7086124704074878
```

Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Analysis Based on Income, Spending and Age

2. Data Preparation

Total_kids	CmpAt	spending	WEBConvertion_rate%	Age	Year_Join
0	0	1617000	114.29	65	10
2	0	27000	20.00	68	8
0	0	776000	200.00	57	9
1	0	53000	33.33	38	8
1	0	422000	100.00	41	8

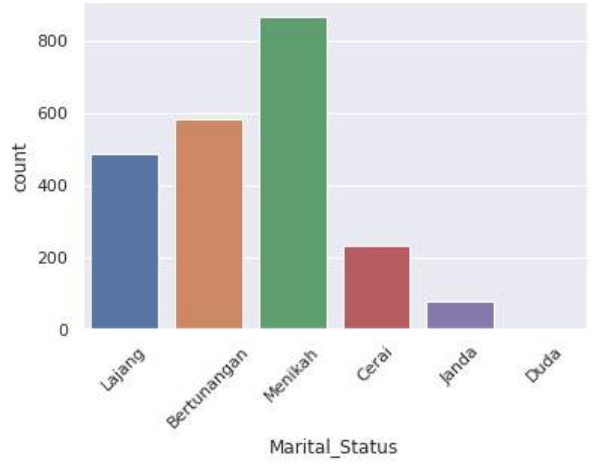
Several columns resulting from feature extraction : Total_kids, spending, WEBConvertion_rate%, Age, Year_Join

- Total_kids = Kidhome + Teenhome
- cmpAt = customer accepted at which campaign (scale 1 – 5, there are 5 campaign)
- spending = MntCoke + MntFruits + MntMeatProducts + MntFishProducts + MntSweetProducts + MntGoldProds
- WEBConvertion_rate% = (NumWebPurchases / NumWebVisitsMonth) * 100
- Age = date.today().year – Year_Birth
- Year_Join = date.today().year - Dt_Customer

Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Analysis Based on Income, Spending and Age

2. Data Preparation

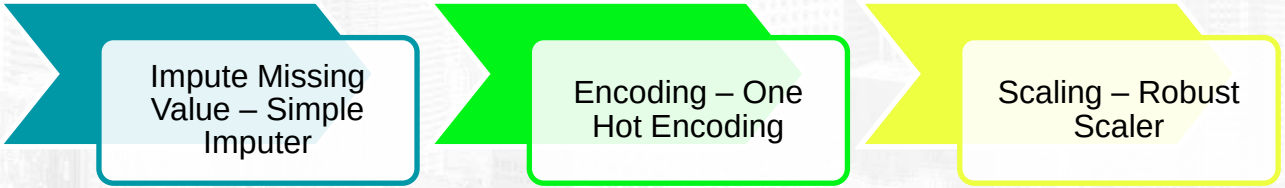


Change and merge Janda and Duda value with Cerai

Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Analysis Based on Income, Spending and Age

3. Feature Engineering

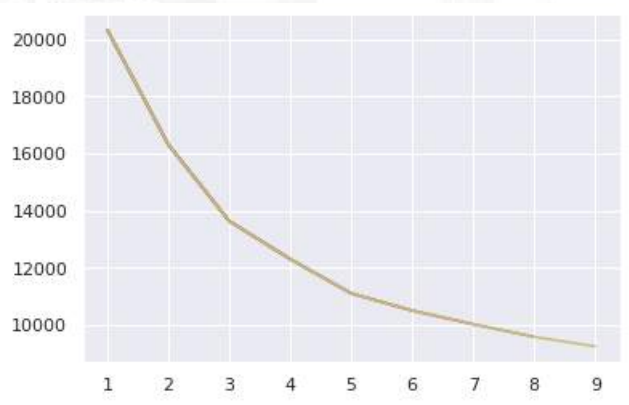


Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Analysis Based on Income, Spending and Age

4. Modeling

Elbow Method



Based on **elbow method** result, we can choose 3 or 5 for the k. I choose 3 for the k because after looking forward through the model **k = 3** gives good cluster.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Analysis Based on Income, Spending and Age

4. Modeling (With PCA)

	pc1	pc2	pc3
Income	0.320892	0.165327	0.153707
Recency	-0.003536	-0.005331	0.008796
NumDealsPurchases	-0.140378	-0.148551	0.854620
NumCatalogPurchases	0.317337	0.141822	0.181721
NumStorePurchases	0.237649	0.126658	0.211856
Complain	-0.002048	0.000159	0.000177
Total_kids	-0.265622	-0.162135	0.281297
CmpAt	0.523267	-0.841411	-0.072983
spending	0.309929	0.100428	0.131556
WEBConvertion_rate%	0.525454	0.410646	0.051570
Age	0.055519	0.018333	0.143354
Year_Join	-0.004253	-0.016011	0.200512
Education_D3	-0.008725	-0.003070	-0.012588
Education_S1	0.004847	0.023128	-0.004970
Education_S2	0.000767	-0.008711	0.010736
Education_S3	0.014768	-0.004729	0.024908
Education_SMA	-0.011656	-0.006617	-0.018085
Marital_Status_Bertunangan	-0.000927	-0.002903	0.006021
Marital_Status_Cerai	0.003954	-0.000661	0.016702
Marital_Status_Lajang	0.002290	0.010319	-0.033949
Marital_Status_Menikah	-0.005317	-0.006755	0.011225

After fit and transform the model with PCA using data after feature engineering. We can get the pca component with **pca.components_**, and according to the picture beside we can see all features that more dominant and representative to the PC (have same direction)

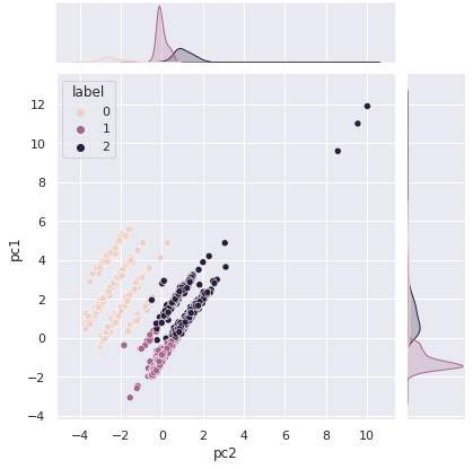
- CmpAt and WebConvertion_rate% represent pc1 with pca component score 0.523 and 0.524. Choose **WebConvertion_rate%** because have highest score than CmpAt
- CmpAt and WebConvertion_rate% represent pc2 with pca component score -0.841 and 0.416. Choose **CmpAt** because have highest score than WebConvertion_rate% (minus means different direction).
- **NumDealsPurchaces** represent pc3 with pca component score 0.854.

Untuk selengkapnya, dapat melihat jupyter notebook disini

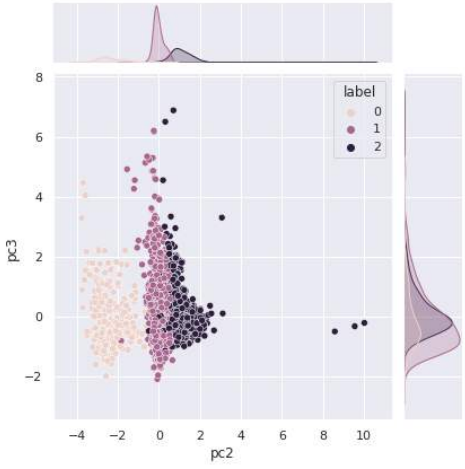
Conversion Rate Analysis Based on Income, Spending and Age

4. Modeling (With PCA)

Clustering result



- **pc1(WebConversion_rate%):** label 1 is distributed in the middle value, label 0 in the low value, and label 2 in the high value.
- **Pc2(CmpAt):**label 0 is distributed in the low value, label 1 is distributed in the high value and label 2 is distributed in the middle value.
- In addition, label 2 has a high standard deviation because the distribution is spread when viewed on labels pc1 and pc2.



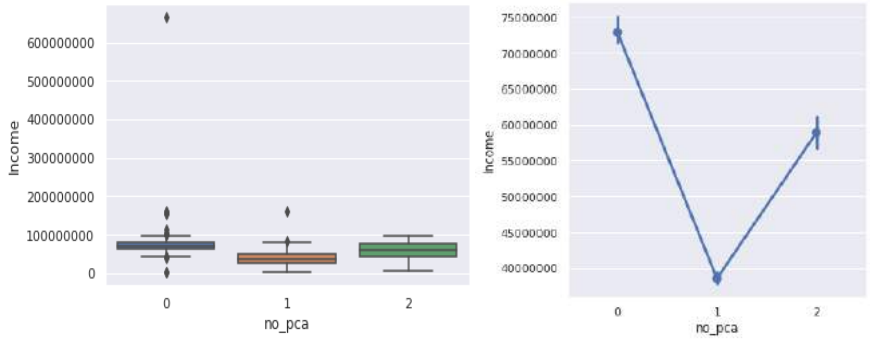
pc3 (**NumDealsPurchaces**), it is difficult to do clustering because the data distribution coincides.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Analysis Based on Income, Spending and Age

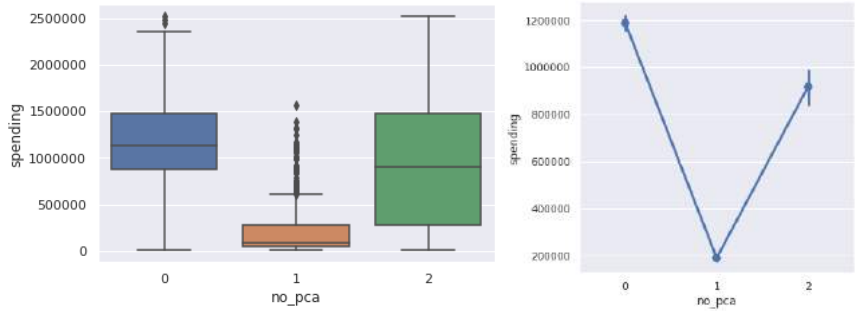
4. Modeling (Without PCA)

Clustering result based on Income, Spending, and Age



For feature **income**, it can be seen that label 0 has a higher average value between 70,000,000 million to 75,000,000 million. label 1 has the lowest average value between 0 – 40,000,000 million, and label 2 has an average value in the middle which is around 55,000,000 million to 60,000,000 million.

For feature **spending**, it can be seen that label 0 has a higher average value between 1,000,000 million to 1,200,000 million. Label 1 has the lowest average value between 0 - 200,000 thousand, and label 2 has an average value in the middle, which is around 800,000 thousand to 1,000,000 million.

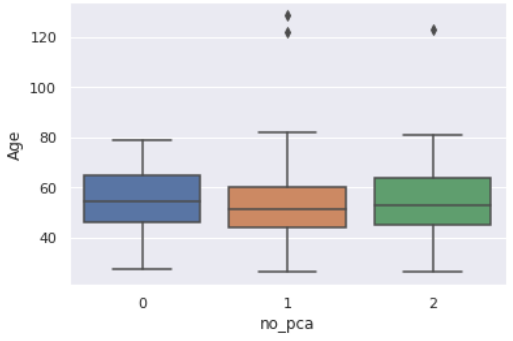


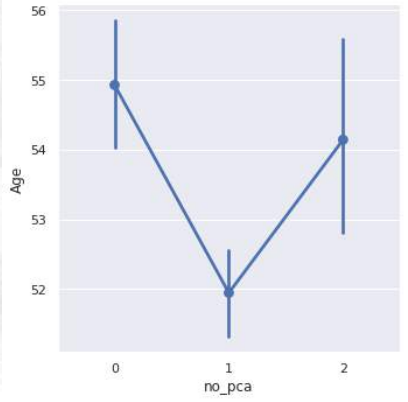
Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Analysis Based on Income, Spending and Age

4. Modeling (Without PCA)

Clustering result based on Income, Spending, and Age





For feature **Age**, it can be seen that label 0 has a higher average age that is 55. Label 1 has the lowest average age that is 52, and label 2 has an average age in the middle, that is 54.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Analysis Based on Income, Spending and Age

4. Bussines Recomendation

- For label 0, because it has a large average income and spending, luxury products can be offered along with discounts
- For labels because they have a small average income and spending, we can offer several products that have discounted prices
- For label 2 we can offer both those offered on customer label 1 and label 0.

Untuk selengkapnya, dapat melihat jupyter notebook disini