Car Insurance Customer Segmentation Analysis

1. Problem Statement

This project aims to perform the K-means clustering and define optimal number of clusters of health insurance customers who are also interested in car insurance. First, I will create so-called 'business cluster' which will cover the trends from four business related variables: Annual Premium, Policy Sales Channel, Vintage and Region Code. Then, I will create 'demographic cluster' which will group customers based on common demographic characteristics such as Age, Gender, Region Code, Driving License and Vehicle Age. The project also aims to derive insights from the data that can be used to inform business decisions.

2. Data Import and Check

Libraries needed

In [1]:
```
# libraries for visualizations
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib
import matplotlib.pyplot as plt

%matplotlib inline
```

In [2]:
```
# libraries for clustering and PCA
from sklearn import preprocessing
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
from sklearn.metrics import silhouette_score
```

Data Import

In [3]:
```
train_df = pd.read_csv("car_train_df.csv")
train_df.shape
```

Out[3]:
```
(370789, 12)
```

In [4]:
```
train_df.head()
```

Out[4]:

	Gender	Age	Driving_License	Region_Code	Previously_Insured	Vehicle_Damage	Annual_Premium	Policy_Sales_Chan
0	1.0	44	1.0	28.0	0.0	1.0	40454	2
1	1.0	76	1.0	3.0	0.0	0.0	33536	2
2	1.0	47	1.0	28.0	0.0	1.0	38294	2
3	1.0	21	1.0	11.0	1.0	0.0	28619	14
4	0.0	29	1.0	41.0	1.0	0.0	27496	14

In [5]:
```
train_df.describe()
```

Out[5]:

	Gender	Age	Driving_License	Region_Code	Previously_Insured	Vehicle_Damage	Annual_Premi
count	370789.000000	370789.000000	370789.000000	370789.000000	370789.000000	370789.000000	370789.0000
mean	0.540251	38.670344	0.997942	26.437281	0.459666	0.503146	29264.643
std	0.498378	15.440166	0.045316	13.310391	0.498371	0.499991	14743.056
min	0.000000	20.000000	0.000000	0.000000	0.000000	0.000000	2630.0000
25%	0.000000	25.000000	1.000000	15.000000	0.000000	0.000000	24102.0000
50%	1.000000	36.000000	1.000000	28.000000	0.000000	1.000000	31319.0000
75%	1.000000	49.000000	1.000000	36.000000	1.000000	1.000000	38596.0000
max	1.000000	85.000000	1.000000	52.000000	1.000000	1.000000	61892.0000

3. Data Pre-processing

The data manipulation and cleaning process has been done in my previous project, Car Insurance Sales Prediction.

Subset the Interested Customers

In [6]:
```
# subset data set of customers interested
customer_df = train_df[train_df['Response']==1].copy()
customer_df = customer_df.drop(columns = ['Response'])
```

Standardization

In [7]:
```
# Standardization
sc = preprocessing.StandardScaler()
df_std = sc.fit_transform(customer_df)
df_std = pd.DataFrame(data = df_std, columns = customer_df.columns)
```

4. Business Cluster Analysis

Subset Business Data

In [8]:
```
business = ['Region_Code', 'Annual_Premium', 'Policy_Sales_Channel', 'Vintage']
business_df = customer_df[business]
business_std = df_std[business]
business_std.head()
```

Out[8]:

	Region_Code	Annual_Premium	Policy_Sales_Channel	Vintage
0	0.102859	0.671918	-1.193396	0.752312
1	0.102859	0.533091	-1.193396	-1.515807
2	0.102859	0.130556	-1.193396	-0.978621
3	0.691056	1.129662	0.568628	-1.288995
4	-1.745763	-1.759100	1.125057	-0.083311

Principal Component Analysis

I will choose reasonable components in order to obtain a better clustering solution than with the standard K-Means so that we can see a nice and clear plot for our segmented groups.

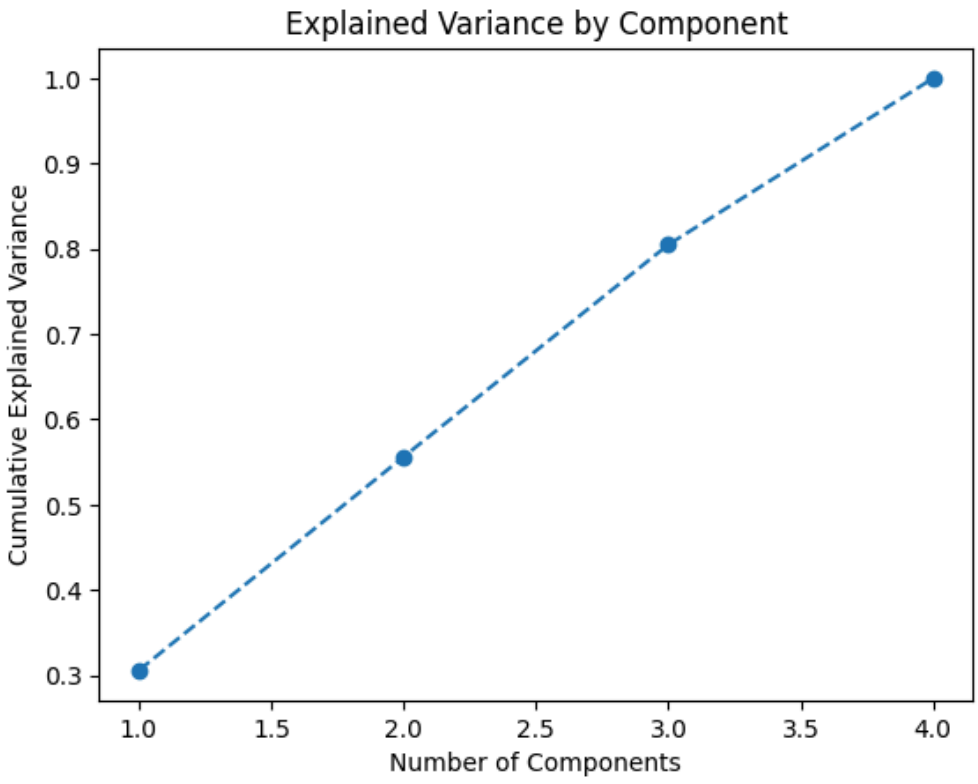
In [9]:
```
# principal component analysis
pca = PCA()
pca.fit(business_std)

# explained variance of each component
pca.explained_variance_ratio_
```

Out[9]:
```
array([0.30551573, 0.25045271, 0.24885113, 0.19518043])
```

We observe that the first component explains around 31% of the variability of the data. The second one explains 25% and so on.

In [10]:
```
# explained variance by principal component
plt.plot(range(1, 5), pca.explained_variance_ratio_.cumsum(), marker = 'o', linestyle = '--')
plt.title('Explained Variance by Component')
plt.xlabel('Number of Components')
plt.ylabel('Cumulative Explained Variance')
plt.show()
```



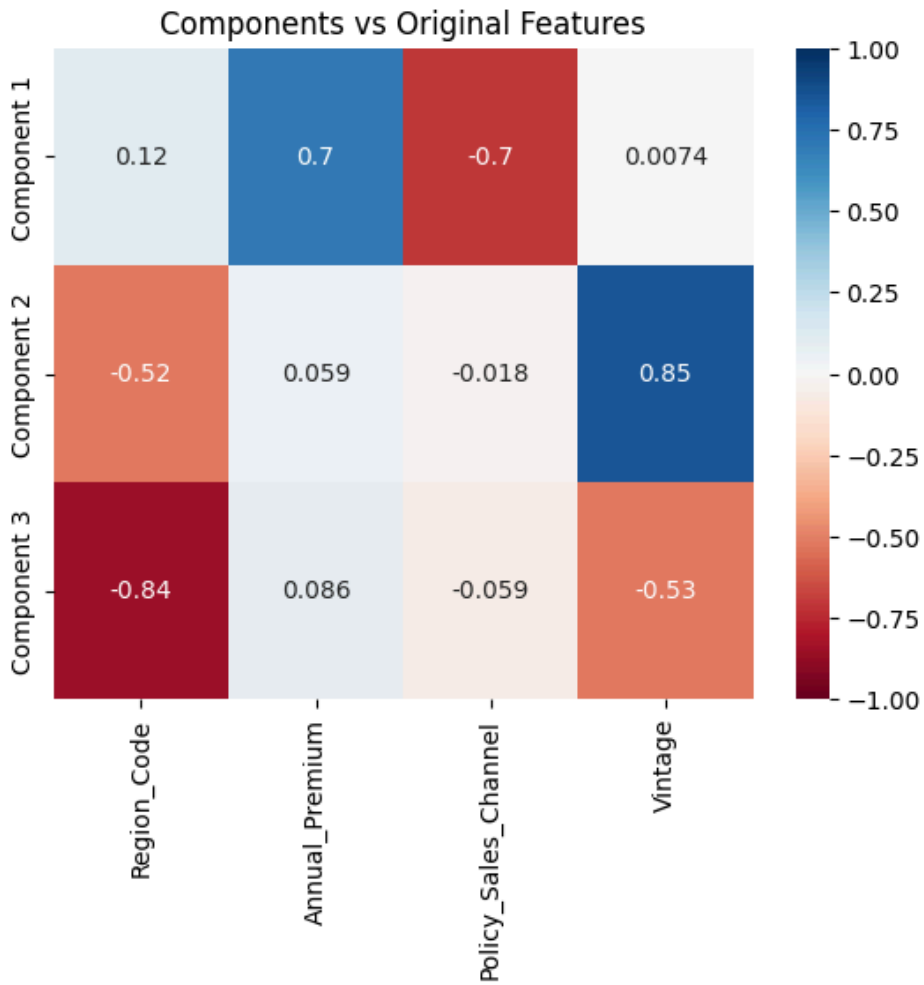
80% of the variance of the data is explained by 3 components. Let's keep the first 3 components for our further analysis.

In [11]:
```
pca = PCA(n_components = 3)
pca.fit(business_std)
```

Out[11]:
```
PCA
PCA(n_components=3)
```

Correlation Heatmap of Components vs Original Features

In [12]:
```
df_pca_comp = pd.DataFrame(data = pca.components_,
                           columns = business_std.columns,
                           index = ['Component 1', 'Component 2', 'Component 3'])
sns.heatmap(df_pca_comp,
            vmin = -1,
            vmax= 1,
            cmap = 'RdBu',
            annot = True)
plt.title('Components vs Original Features')
plt.show()
```



We reduced our features to three components from the original four values that explain the shape the values themselves show the so-called loadings. Loadings are correlations between an original variable and the component.

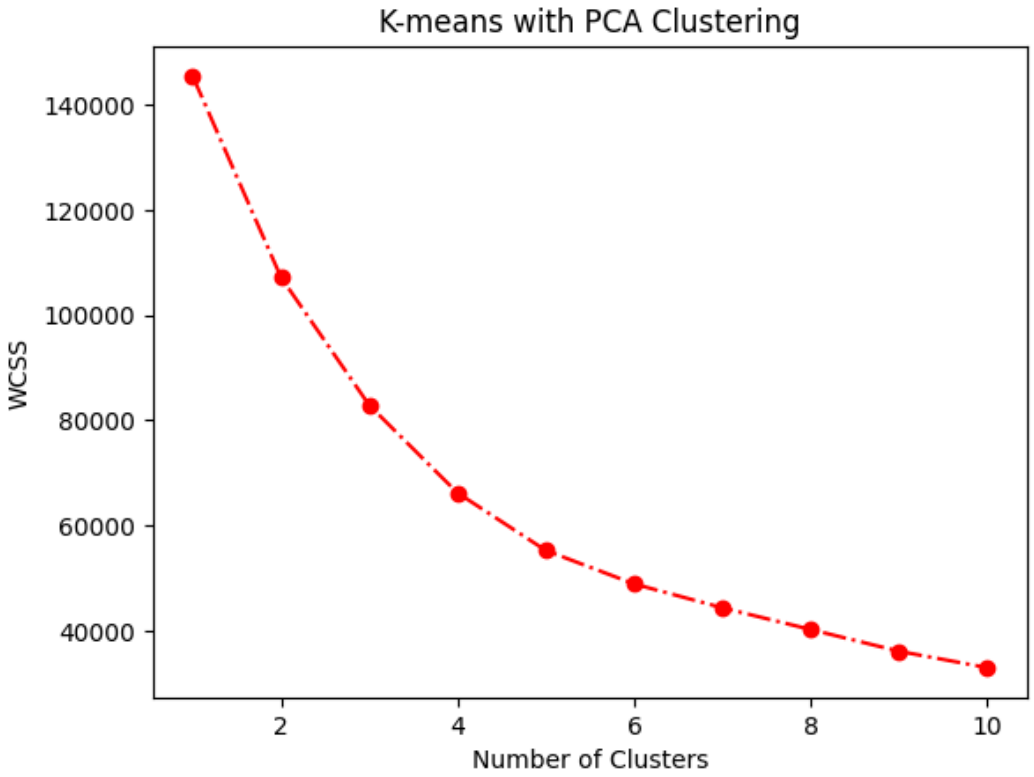
- We see that the first component has high positive correlation with Annual Premium and negative correlation with Policy Sales Channel. This component shows a pattern where customers who pay higher annual premiums tend to prefer certain channels for policy sales. Higher value of the first component represents higher annual premium and preference for lower Policy Sales Channel code.
- For the second component, Vintage is the most important feature. We see moderate negative correlation with Region Code as well. This component captures the relationship between customer longevity with the company and their association with specific geographic regions. It indicates that long-standing customers tend to be concentrated in particular regions.
- For the final component, we see strong negative correlation with Region Code. The final component could represent the relationship between customers from specific geographic regions (lower region codes) and those who have been associated with the company for a shorter period of time.

In [13]:
```
scores_pca = pca.transform(business_std)
```

Within Clusters Sum of Squares

In [14]:
```
wcss = []
for i in range(1, 11):
    kmeans_pca = KMeans(n_clusters = i, init = 'k-means++', random_state = 0, n_init = 10)
    kmeans_pca.fit(scores_pca)
    wcss.append(kmeans_pca.inertia_)
```

In [15]:
```
plt.plot(range(1, 11), wcss, marker = 'o', linestyle = '-.', color = 'red')
plt.xlabel('Number of Clusters')
plt.ylabel('WCSS')
plt.title('K-means with PCA Clustering')
plt.show()
```



Looking at the plot of within sum of squares, we see that optimal cluster number by within sum of square is 4.

K-Means Clustering with 4 Clusters

In [16]:
```
kmeans_pca = KMeans(n_clusters = 4, init = 'k-means++', random_state = 0, n_init = 10).fit(scores_pca)
label_pca = kmeans_pca.predict(scores_pca)
```

In [17]:
```
df_pca_kmeans = pd.concat([business_df.reset_index(drop=True), pd.DataFrame(scores_pca)], axis = 1)
df_pca_kmeans.columns.values[-3: ] = ['Component_1', 'Component_2', 'Component_3']
df_pca_kmeans['Segment_Kmeans_PCA'] = kmeans_pca.labels_
```

Segment Profiling

In [18]:
```
pca_kmeans_analysis = df_pca_kmeans.groupby(['Segment_Kmeans_PCA']).mean()
pca_kmeans_analysis['Total_Revenue'] = df_pca_kmeans[['Annual_Premium', 'Segment_Kmeans_PCA']].groupby(
pca_kmeans_analysis['No. Obs'] = df_pca_kmeans[['Segment_Kmeans_PCA', 'Region_Code']].groupby(['Segment
pca_kmeans_analysis['Prop Obs'] = pca_kmeans_analysis['No. Obs']/pca_kmeans_analysis['No. Obs'].sum()
pca_kmeans_analysis
```

Out[18]:

Segment_Kmeans_PCA	Region_Code	Annual_Premium	Policy_Sales_Channel	Vintage	Component_1	Component_2	C
0	9.488866	22335.320018	117.956612	153.942264	-0.900502	0.719843	
1	28.199768	38522.833195	56.878280	76.531136	0.801693	-0.805387	
2	36.121287	18587.199210	125.179873	149.650045	-0.904095	-0.510527	
3	28.399396	38395.489548	61.661277	234.471419	0.749392	0.786643	

Based on the values of variables and component scores, I will label each of the four segments as 'Middle-range insurer', 'Premium short-term insurer', 'Thrifty insurer' and 'Premium long-term insurer'.

0. Middle-range insurer

- This segement of customers pays the middle range price of average annual premium (22335 dollars) compared to other clusters. We see that this segment is from lower region code (9) compared to other groups on average.

1. Premium short-term insurer

- This is the segment that pays the highest price of average annual premium (38522 dollars), but have been associated with the company for the shortest period of time (77 days on average).
- They prefer the lowest Policy Sales Channel code of 56 on average.

2. Thrifty insurer

- This is a segment of customers that pay low range price of average annual premium (18587 dollars).
- They prefer the highest Policy Sales Channel code of 125 on average.

3. Premium long-term insurer

- This is the segment that pays the high range price of average annual premium (38395 dollars), and have been associated with the company for the longest period of time (234 days on average).
- They prefer the low Policy Sales Channel code of 61 on average.

Segment Analysis

In [19]:
```
pca_kmeans_analysis.rename({0: 'Middle-range insurer',
                            1: 'Premium short-term insurer',
                            2: 'Thrifty insurer',
                            3: 'Premium long-term insurer'})
```

Out[19]:

Segment_Kmeans_PCA	Region_Code	Annual_Premium	Policy_Sales_Channel	Vintage	Component_1	Component_2	C
Middle-range insurer	9.488866	22335.320018	117.956612	153.942264	-0.900502	0.719843	
Premium short-term insurer	28.199768	38522.833195	56.878280	76.531136	0.801693	-0.805387	
Thrifty insurer	36.121287	18587.199210	125.179873	149.650045	-0.904095	-0.510527	
Premium long-term insurer	28.399396	38395.489548	61.661277	234.471419	0.749392	0.786643	

We will analyze these segments by looking at some business index to evaluate in terms of revenue potential, loyalty or growth opportunity.

1. Total Revenue

- Premium long-term insurer is the segment with the highest revenue potential (470,191,165 dollars).
- The segment with the second highest revenue potential is Premium short-term insurer (463,969,003 dollars).

2. Customer Life Time Value

- Premium long-term insurer is the segment with the highest CLV in terms of the average annual premium spend per customer (38,395 dollars).
- The segment with the second highest CLV is Premium short-term insurer (38,522 dollars).

3. Market Size

- Premium long-term insurer is the segment with the largest market size (27.1%).
- The segment with the second largest market size is Thrifty insurer (26.9%).

4. Customer retention

- Premium long-term insurer is the segment with the longest customer retention. Each customer has been associated with the company for 234 days on average.
- The segment with the shortest customer retention is Premium short-term insurer (76 days).

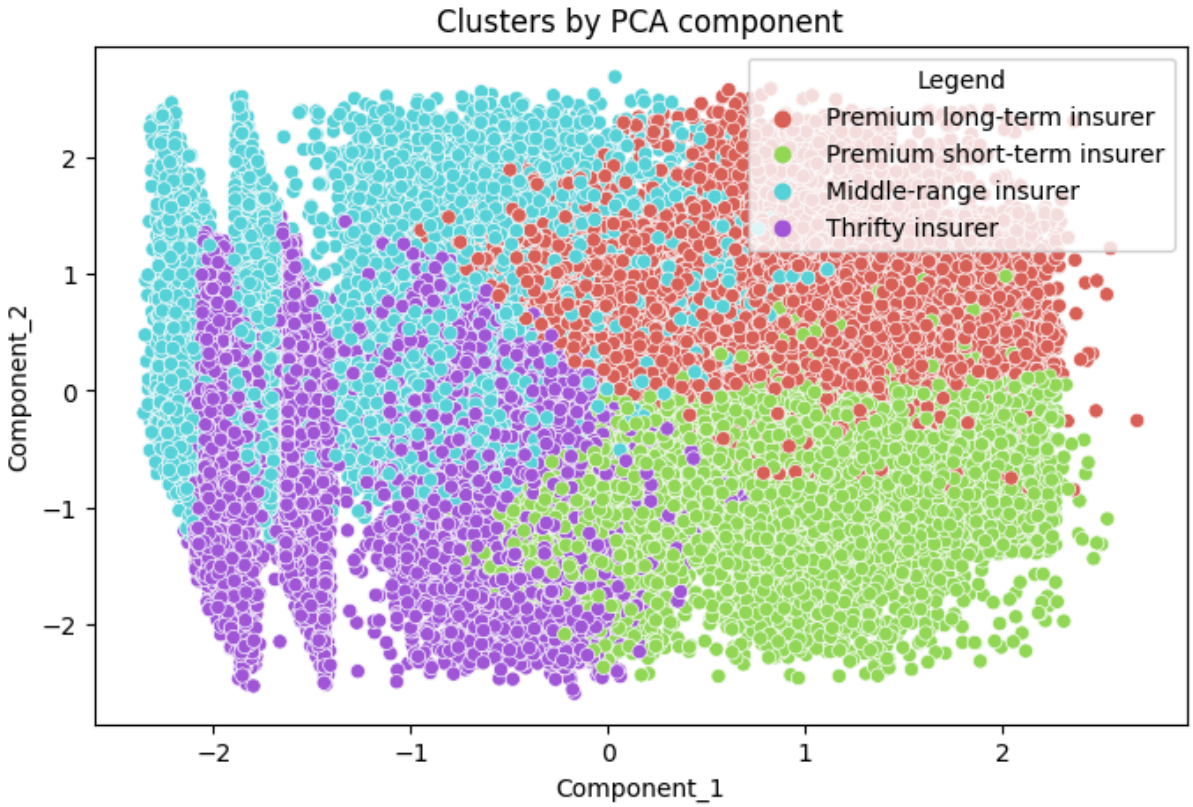
5. Policy Sales Channel preference

- Insurers paying higher annual premium tend to prefer to be outreached by lower Policy Sales Channel code.
- Premium long-term and short-term insurer prefer 62 and 56 on average, while middle-range and thrifty insurer prefer 117 and 125 on average.

Visualization of Business Clusters by PCA Component

In [20]:
```
df_pca_kmeans['Legend'] = df_pca_kmeans['Segment_Kmeans_PCA'].map({0: 'Middle-range insurer',
                            1: 'Premium short-term insurer',
                            2: 'Thrifty insurer',
```

```
                            3: 'Premium long-term insurer'})
plt.figure(figsize=(8, 5))
sns.scatterplot(df_pca_kmeans, x = 'Component_1', y = 'Component_2', hue = 'Legend', palette = 'hls')
plt.title('Clusters by PCA component')
plt.show()
```



- Looking at the plot, we realize that the first component separates the customers based on annual premium price and their preference on Policy Sales Channel. Higher score means that a customer pays high annual premium and prefers to be outreached by lower Public Policy Channel code.
- The second component separates the customers based on the duration of their association with the company. Higher score means that a customer has been associated with the company for a longer period of time.

5. Demographic Cluster Analysis

Subset demographic data

In [21]:
```
demographic = ['Gender', 'Region_Code', 'Age', 'Driving_License', 'Vehicle_Age_num', 'Vehicle_Damage']
demo_df = customer_df[demographic]
demo_std = df_std[demographic]
demo_std
```

Out[21]:

	Gender	Region_Code	Age	Driving_License	Vehicle_Age_num	Vehicle_Damage
0	0.799164	0.102859	0.060231	0.029776	2.123928	0.147932
1	0.799164	0.102859	0.307817	0.029776	2.123928	0.147932
2	-1.251308	0.102859	1.050574	0.029776	0.120252	0.147932
3	-1.251308	0.691056	0.307817	0.029776	0.120252	0.147932
4	0.799164	-1.745763	-0.517469	0.029776	0.120252	0.147932
...	...	...	...	...	...	...
45150	0.799164	1.615367	-0.269883	0.029776	0.120252	0.147932
45151	-1.251308	0.102859	-0.187355	0.029776	0.120252	0.147932
45152	-1.251308	0.102859	0.225288	0.029776	0.120252	0.147932
45153	-1.251308	0.102859	1.463217	0.029776	0.120252	0.147932
45154	-1.251308	0.102859	-0.434940	0.029776	0.120252	0.147932

45155 rows × 6 columns

Principal Component Analysis

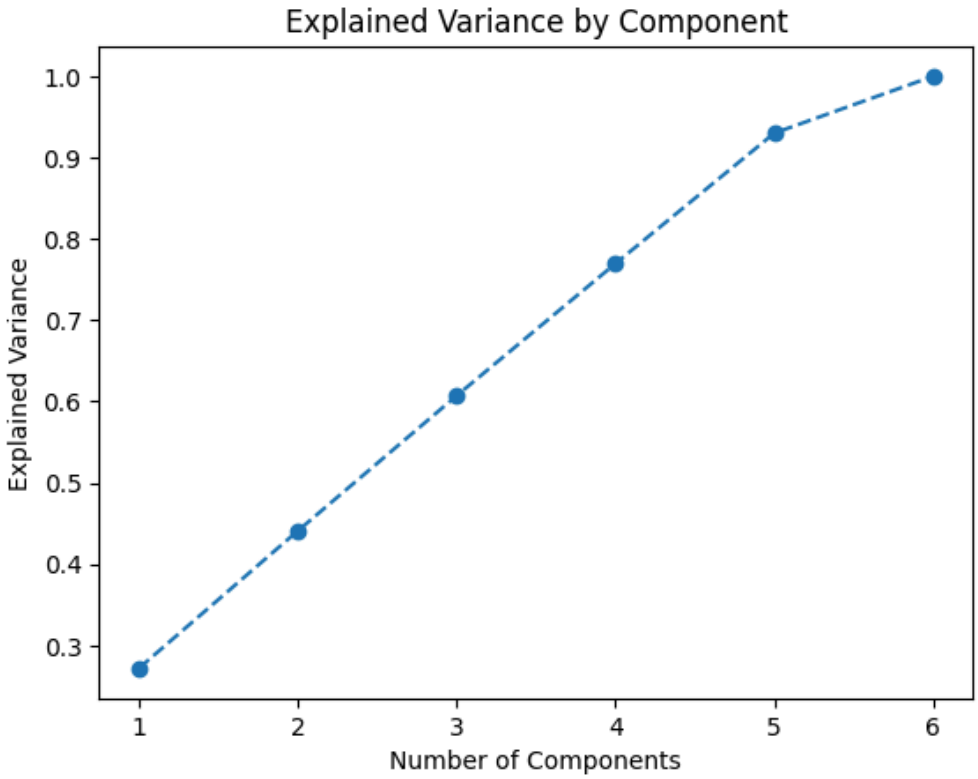
In [22]:
```
pca = PCA()
pca.fit(demo_std)

#cumulative sum of total variance
pca.explained_variance_ratio_.cumsum()
```

Out[22]:
```
array([0.27173146, 0.44069828, 0.60698634, 0.76954837, 0.93000219,
       1.        ])
```

The first four components can explain about 77% of the total variance.

In [23]:
```
# explained variance by principal component
plt.plot(range(1, 7), pca.explained_variance_ratio_.cumsum(), marker = 'o', linestyle = '--')
plt.title('Explained Variance by Component')
plt.ylabel('Explained Variance')
plt.xlabel('Number of Components')
plt.show()
```



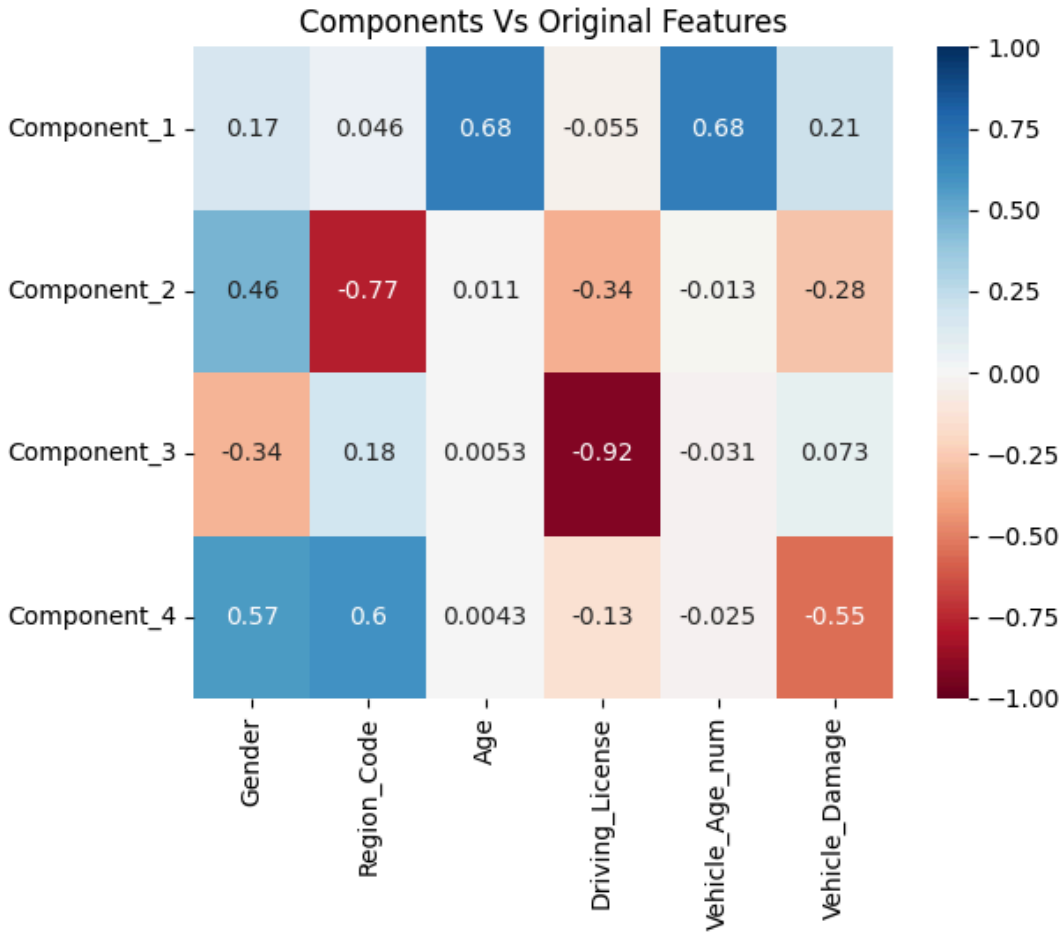
77% of the variance of the data is explained by 4 components. Let's keep the first 4 components for our analysis.

In [24]:
```
pca = PCA(n_components = 4)
pca.fit(demo_std)
```

Out[24]:
```
PCA(n_components=4)
```

Correlation Heatmap of Components Vs Original Features

In [25]:
```
demo_pca_comp = pd.DataFrame(data = pca.components_,
                             columns = demo_std.columns,
                             index = ['Component_1', 'Component_2', 'Component_3', 'Component_4'])
sns.heatmap(demo_pca_comp,
            vmin = -1,
            vmax = 1,
            cmap = 'RdBu',
            annot = True)
plt.title('Components Vs Original Features')
plt.show()
```



- We see that there is a strong positive correlation between the first component and Age and Vehicle Age. It seems like the first component is a representation of a pattern where older age groups tend to own older vehicles.
- For the second component, we see strong negative correlation with Region Code and moderate positive correlation with Gender.
- For the third component, we see strong negative correlation with Driving License. The third component seems to represent customers who do not have driving license.
- For the final component, we realize that Gender and Region Code are the most important features. We observed that Vehicle Damage load negatively but are still important. It seems like the final component is a representation of a pattern where customers in particular gender group and region are less likely to have damage on their vehicles.

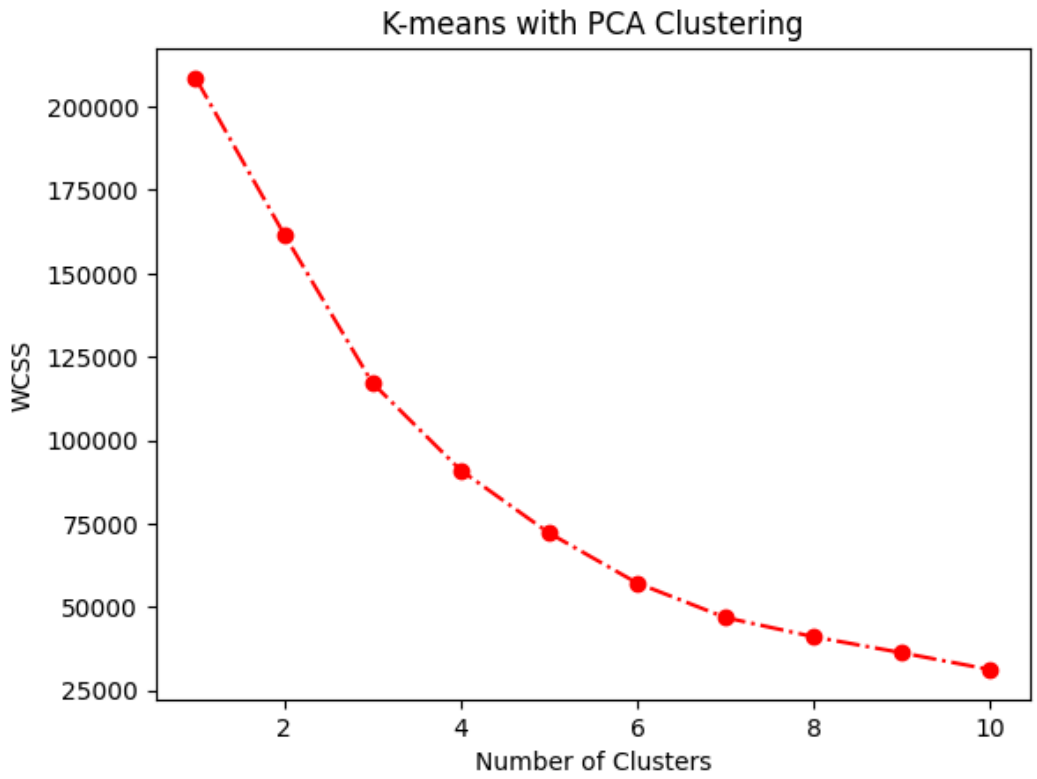
In [26]:
```
demo_scores_pca = pca.transform(demo_std)
```

Within Clusters Sum of Squares

In [27]:
```
wcss = []
for i in range(1, 11):
    kmeans_pca = KMeans(n_clusters = i, init = 'k-means++', random_state = 0, n_init = 10)
    kmeans_pca.fit(demo_scores_pca)
    wcss.append(kmeans_pca.inertia_)
```

In [28]:
```
plt.plot(range(1, 11), wcss, marker = 'o', linestyle = '-.', color = 'red')
plt.title('K-means with PCA Clustering')
plt.xlabel('Number of Clusters')
```

```
plt.ylabel('WCSS')
plt.show()
```



According to the elbow rule, the optimal number of clusters seems to be 7.

Silhouette Score

In [ ]:
```
#for i in range(2,10):
#    kmeans_pca_sil = KMeans(n_clusters = i, init = 'k-means++', random_state = 0, n_init = 10).fit(dem
#    preds = kmeans_pca_sil.predict(demo_scores_pca)
#    silhouette_avg = silhouette_score(demo_scores_pca, preds)
#
#    print('Silhouette Score for %i Clusters: %0.4f' % (i, silhouette_avg))
```

K-Means Clustering with 7 Clusters

In [29]:
```
demo_kmeans_pca = KMeans(n_clusters = 7, init = 'k-means++', random_state = 0, n_init = 10).fit(demo_s
demo_label_pca = demo_kmeans_pca.predict(demo_scores_pca)
```

In [30]:
```
df_pca_kmeans = pd.concat([customer_df[demographic + ['Annual_Premium'] + ['Policy_Sales_Channel']].re
df_pca_kmeans.columns.values[-4:] = ['Component_1', 'Component_2', 'Component_3', 'Component_4']
df_pca_kmeans['Segment_KMeans_PCA'] = demo_label_pca
```

Segment Profiling

In [31]:
```
demo_pca_kmeans_analysis = df_pca_kmeans.groupby(['Segment_KMeans_PCA']).mean()
demo_pca_kmeans_analysis['Total_Revenue'] = df_pca_kmeans[['Annual_Premium', 'Segment_KMeans_PCA']].gro
demo_pca_kmeans_analysis['No. Obs'] = df_pca_kmeans[['Segment_KMeans_PCA', 'Gender']].groupby(['Segment
demo_pca_kmeans_analysis['Prop Obs'] = demo_pca_kmeans_analysis['No. Obs']/demo_pca_kmeans_analysis['No
demo_pca_kmeans_analysis
```

Out[31]:

Segment_KMeans_PCA	Gender	Region_Code	Age	Driving_License	Vehicle_Age_num	Vehicle_Damage	Annual_Pr
0	0.537118	26.369824	26.433670	1.0	0.012498	1.000000	29073.
1	0.000000	31.614896	44.143668	1.0	1.035989	1.000000	30786.
2	0.999928	32.771720	44.083989	1.0	0.999350	0.999928	30757.
3	0.682865	8.905087	44.658694	1.0	1.037965	1.000000	25790.
4	0.571429	25.687371	35.824017	1.0	0.601449	0.000000	23195.
5	0.675000	24.750000	58.575000	0.0	1.125000	1.000000	32062.
6	0.776598	28.962176	59.494131	1.0	1.663872	1.000000	35073.

Based on the values of variables and component scores, I will label each of the seven segments as 'Young New Vehicle Owner', 'Mature Female Vehicle Owner', 'Mature Male Vehicle Owner', 'Lower Regionals', 'Non-Damaged Vehicle Owner', 'Elderly Vehicle Owner without Licenses' and 'Elderly Vehicle Owner'.

0. Young New Vehicle Owner

- This is the youngest segment with an average age of 27.
- They own the vehicles that are relatively the newest and have damage.

1. Mature Female Vehicle Owner

- This segment is comprised entirely of women with an average age of 44.
- They pay high average annual premium price of 30787 dollars.

2. Mature Male Vehicle Owner

- This segment is comprised almost entirely of men with an average age of 44.
- They pay high average annual premium price of 30757 dollars.

3. Lower Regionals

- This is the segment from region code 9 on average.

4. Non-Damaged Vehicle Owner

- This is a segment with an average age of 36 who does not have a damage on their vehicles.
- It seems to be a segment of customers who have not yet experienced vehicle damage.
- They pay the smallest average annual premium price of 23195 dollars.

5. Elderly Vehicle Owner without Licenses

- This is an oldest segment with an average age of 59 who does not have driving license.
- They own older vehicles compared to other clusters.

6. Elderly Vehicle Owner

- This is an oldest segment with an average age of 59 who possesses driving license.
- They own the oldest vehicles among other clusters.
- They pay the highest average annual premium price of 35073 dollars.

Segment Analysis

In [32]:
```
demo_pca_kmeans_analysis.rename({0:'Young New Vehicle Owner',
                                 1: 'Mature Female Vehicle Owner',
                                 2: 'Mature Male Vehicle Owner',
                                 3: 'Lower Regionals',
                                 4: 'Non-Damaged Vehicle Owner',
                                 5: 'Elderly Vehicle Owner without Licenses',
                                 6: 'Elderly Vehicle Owner'})
```

Out[32]:

Segment_KMeans_PCA	Gender	Region_Code	Age	Driving_License	Vehicle_Age_num	Vehicle_Damage	Annual_Pr
Young New Vehicle Owner	0.537118	26.369824	26.433670	1.0	0.012498	1.000000	29073.
Mature Female Vehicle Owner	0.000000	31.614896	44.143668	1.0	1.035989	1.000000	30786.
Mature Male Vehicle Owner	0.999928	32.771720	44.083989	1.0	0.999350	0.999928	30757.
Lower Regionals	0.682865	8.905087	44.658694	1.0	1.037965	1.000000	25790.
Non-Damaged Vehicle Owner	0.571429	25.687371	35.824017	1.0	0.601449	0.000000	23195.
Elderly Vehicle Owner without Licenses	0.675000	24.750000	58.575000	0.0	1.125000	1.000000	32062.
Elderly Vehicle Owner	0.776598	28.962176	59.494131	1.0	1.663872	1.000000	35073.

Here are several factors and metrics to evaluate each customer segment in terms of revenue potential, loyalty and growth opportunity.

1. Total Revenue

- Mature Male Vehicle Owner is the segment with the highest revenue potential (425,896,572 dollars).
- The segment with the second highest revenue potential is Mature Female Vehicle Owner (319,937,722 dollars).

2. Customer Life Time Value

- Elderly Vehicle Owner is the segment with the highest CLV in terms of the average annual premium spend per customer (35,073 dollars).

3. Market Size

- Mature Male Vehicle Owner is the segment with the largest market size (30.7%).
- The segment with the second largest market size is Mature Female Vehicle Owner (23%).

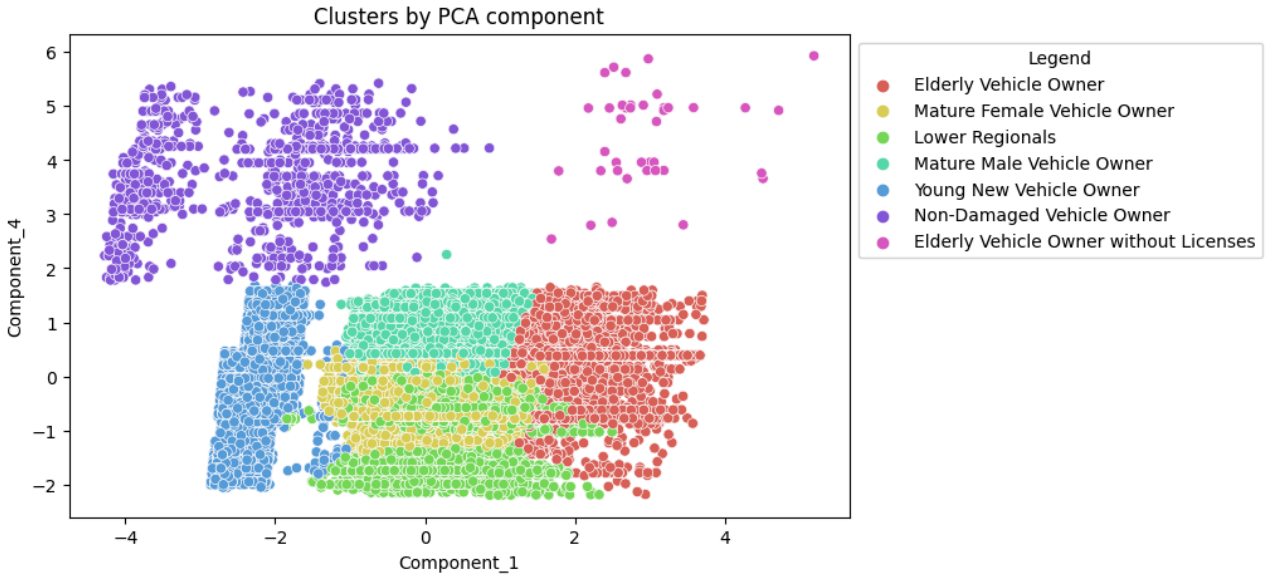
5. Policy Sales Channel preference

- Insurers in older groups tend to prefer to be outreached by lower Policy Sales Channel code.
- Elderly Vehicle Owner and Elderly Vehicle Owner without License prefer 60 and 67 on average, while Young New Vehicle Owner, the youngest segment, prefer the code 127 on average.

Visualization of Demographic Clusters by PCA Components

In [33]:
```
df_pca_kmeans['Legend'] = df_pca_kmeans['Segment_KMeans_PCA'].map({0:'Young New Vehicle Owner',
                                    1: 'Mature Female Vehicle Owner',
                                    2: 'Mature Male Vehicle Owner',
                                    3: 'Lower Regionals',
                                    4: 'Non-Damaged Vehicle Owner',
                                    5: 'Elderly Vehicle Owner without Licenses',
                                    6: 'Elderly Vehicle Owner'})

plt.figure(figsize=(8, 5))
ax = sns.scatterplot(df_pca_kmeans, x = 'Component_1', y = 'Component_4', hue = 'Legend', palette = 'h
plt.title('Clusters by PCA component')
sns.move_legend(ax, "upper left", bbox_to_anchor=(1, 1))
```



- Looking at the plot, we realize that the first component separates the customers well into age groups. High scores mean that they are in older age group and own older vehicle.
- The fourth component separates the customers into two big groups. 'Non-Damaged Vehicle Owner' and 'Elderly Vehicle Owner without Licenses' both have higher score compared to the rest and seem to be the group that is less likely to have damage in their vehicles.

6. Business Questions

Q1. Can we identify distinct customer segments based on particular variables? What are the key characteristics of each customer segment?

For the business cluster, we identified 4 customer segments based on Annual Premium, Policy Sales Channel and Vintage:

- Thrifty insurer: low price of average annual premium
- Middle-range insurer: medium price of average annual premium
- Premium short-term insurer: high price of average annual premium, short duration of customer association
- Premium long-term insurer: high price of average annual premium, long duration of customer association

For the demographic cluster, we identified 7 customer segments based on Age, Gender, Region Code, Driving License and Vehicle Age:

- Young New Vehicle Owner: the youngest segment, relatively the newest vehicle, low price of average annual premium
- Mature Female Vehicle Owner: females in 40-50 age range, high price of average annual premium
- Mature Male Vehicle Owner: males in 40-50 age range, high price of average annual premium
- Lower Regionals: segment from lower region code, average age of 44
- Non-Damaged Vehicle Owner: segment with non-damaged vehicles, average age of 35, low price of average annual premium
- Elderly Vehicle Owner without Licenses: the olderst segment without driving license, highest price of average annual premium
- Elderly Vehicle Owner: the olderst segment, highest price of average annual premium

Q2. Which segments are most valuable to the company in terms of revenue potential, loyalty, or growth opportunities?

Understanding the most valuable segment can help in allocating resources and designing targeted marketing strategies.

- From the business cluster, 'Premium long-term insurer' seems to be the most valuable segment to the company in terms of revenue potential and loyalty. 'Premium long-term insurer' is a segment with the largest total revenue of 470,191,165 dollars and the longest average duration of 234 days, suggesting a higher level of loyalty or retention among these customers.

- From the demographic cluster, 'Mature Male Vehicle Owner' seems to be the most valuable segment to the company. 'Mature Male Vehicle Owner' emerges as the segment with the highest revenue potential of 425,896,572 dollars and has the largest market size which present greater growth opportunities due to their size and potential for expansion

Q3. Do different segments prefer different communication channels?

Understanding preferred communication channels can optimize marketing spend and improve campaign effectiveness.

- There is a tendency that insurers who pay higher annual premium and are in older age groups prefer to be outreached by lower Policy Sales Channel code.

- 'Premium long-term insurer' and 'Premium short-term insurer' from business cluster and 'Elderly Vehicle Owner' and 'Elderly Vehicle Owner without License' from demographic cluster prefer the channel code around between 60 and 70 on average.

- 'Thrifty insurer' from business cluster and 'Young New Vehicle Owner', the youngest segment from the demographic cluster, prefer the code greater than 120 on average.

Q4. Are there differences in duration of a customer's association with the company across segments?

Yes, there are differences in the duration of a customer's association with the company across segments.

- 'Premium short-term insurer' has the shortest duration of association with the company, with an average of 77 days. Despite paying the highest price of average annual premium, customers in this segment are relatively new to the company, indicating a recent acquisition of customers who opt for higher-priced insurance plans.

- In contrast, 'Premium long-term insurer' has the longest duration of association with the company, with an average of 234 days. Despite also paying a high price of average annual premium, customers in this segment have been with the company for a significantly longer period, suggesting a higher level of loyalty or retention among these customers.

- Overall, the differences in the duration of association with the company across segments suggest varying levels of customer loyalty. Understanding these differences can inform customer relationship management strategies, retention initiatives, and targeted marketing efforts aimed at maximizing customer lifetime value and fostering long-term relationships with the company.

Q5. What are the age distribution? Are there the age-related trends and preferences?

The analysis highlights several age-related trends and preferences among the identified customer segments.

- Customers in older age groups pay a high average annual premium price, suggesting a potential preference for comprehensive coverage and higher levels of insurance protection.
- The largest age group is 40-50 age range (70% in total), contributing the most to total revenue potential (949,627,836 dollars in total).
- Younger age groups in their 20s and 30s tend to have relatively newer vehicles and pay a lower average annual premium price than older age groups.
- Elderly segments, both with and without licenses, spend the highest average annual premium per customer.

7. Conclusion

In conclusion, the customer segmentation analysis has provided valuable insights into the characteristics, behaviors, and preferences of health insurance customers interested in vehicle insurance. By leveraging PCA and k-means clustering algorithms, we identified distinct customer segments based on both business and demographic variables.

From the analysis, we have identified key customer segments such as 'Premium long-term insurer' from the business cluster and 'Mature Male Vehicle Owner' from the demographic cluster, which demonstrate high revenue potential, market size, and growth opportunities. Understanding the Policy Sales Channel preferences of each segment can optimize marketing strategies and improve campaign effectiveness.

Furthermore, differences in customer retention across segments highlight varying levels of loyalty and engagement. Strategies aimed at customer retention and relationship management should be tailored to the specific needs and preferences of each segment.

Overall, the age-related trends and preferences identified underscore the importance of understanding demographic dynamics in shaping customer behavior and preferences. By leveraging these insights, businesses can tailor their products, services, and marketing strategies to better meet the needs of different customer segments, ultimately driving customer satisfaction, loyalty, and revenue growth.