CUSTOMER SEGMENTATION

Using Unsupervised Learning

Name: R. Abishek Bhat [Email:abishekbhat31@gmail.com](mailto:abishekbhat31@gmail.com) Linkedin:https[://w](http://www.linkedin.com/in/abishekbhat31/)ww[.li](http://www.linkedin.com/in/abishekbhat31/)n[kedin.com/in/abishekbhat31/](http://www.linkedin.com/in/abishekbhat31/) Github:https://github.com/abishek-bhat

# ABSTRACT

Companies and organizations use customer segmentation analysis reports to optimize the corporate decisions that is derived from deep customer analysis. When used as part of good business practices, a company can improve its competitiveness as well as reduce the chances that important decisions are not grounded in a deep understanding of existing customer segmentation.

## Data set:

https://github.com/abishek-bhat/Customer-Segregation-using- UnsupervisedLearning/blob/main/Mall\_Customers.csv

Source Code:

https://github.com/abishek-bhat/Customer-Segregation-using-Unsupervised-

Learning/blob/main/Customer%20Segregation%20using%20Unsupervised%20Learning.ipynb

# ACKNOWLEDGE

I am using this opportunity to express my gratitude to everyone who supported me throughout the course of my capstone project. I am thankful for their aspiring guidance, invaluably constructive criticism and friendly advice during the project work. I am sincerely grateful to them for sharing their truthful and illuminating views on a number of issues related to the project.

Further, I have fortunate to have Mr. ANBU JOEL as my mentor. He has readily shared his immense knowledge in data analytics and guide me in a manner that the outcome resulted in enhancing my data skills.

I certify that the work done by me for conceptualizing and completing this project is original and authentic.

Date: 1/7/2022 Name: Abishek Bhat. R

# CERTIFICATION OF COMPLETION

I certify that the project titled "Customer Segmentation Using Unsupervised Learning" was undertaken and completed.

(1st August 2022).

Mentor Mr. ANBU JOEL

Date: 1st August, 2022

Place: Trichy.

# Table Of Content

|  |  |  |
| --- | --- | --- |
| Chapter Number | Title | Page Number |
| 1. | Introduction |  |
| 2. | Objective |  |
| 3. | Types of Customer Segmentation |  |
| 4. | Benefits of Customer Segmentation |  |
| 5. | Customer Segmentation Using Cluster Analysis |  |
| 6. | Conclusion |  |

INTRODUCTION

Customer segmentation reports are considered essential tactical analysis tools and are often used by managers to optimize products, services, sales and marketing strategies. Some of the key functionality in this type of dashboard report is that it simplifies analysis by combining charts with metrics. The top of the report shows customer count by industry both as figures and as a ranked chart. The second visualization shows a ranked list of countries where customers are located.

At its most basic, customer segmentation (also known as market segmentation) is the division of potential customers in a given market into discrete groups. That division is based on variables and descriptors of those customers having similar enough:

1. Needs, i.e., so that a single whole product can satisfy them.
2. Buying characteristics, i.e., responses to messaging, marketing channels, and sales channels, that a single go-to-market approach can be used to sell to them competitively and economically.

More details are mentioned below:

|  |  |  |
| --- | --- | --- |
| Segmentation | Consumer | Industrial |
| Segmentation Bases | Needs, wants benefits, solutions to problems, usage situation, usage rate. | Needs, wants benefits, solutions to problems, usage situation, usage  rate, size\*, industrial\*. |
| Descriptors Demographics | Age, income, marital status, family type & size,  gender, social class, etc. | Industry, size, location, current supplier(s),  technology utilization, etc. |
| Psychographics | Lifestyle, values, &  personality characteristics. | Personality characteristics of decision makers. |
| Behavior | Use occasions, usage level, complementary & substitute products us brand loyalty, etc. | Use occasions, usage level, complementary & substitute products used, brand loyalty, order size,  applications, etc. |
| Decision Making | Individual or group (family) choice, low or high involvement purchase, attitudes and knowledge about product class, price sensitivity, etc | Formalization of purchasing procedures, size & characteristics of decision making group, use of outside consultants, purchasing antena, (de) centralizing buying, price sensitivity,  switching costs, elc |
| Media Patterns | Level of use, types of media used, times of use, | Level of use, types of media used time of use, patronage at trade shows, receptivity of sales people, etc. |

# Objective

Customer segmentation analysis is the process performed when looking to discover insights that define specific segments of customers. Marketers and brands leverage this process to determine what campaigns, offers, or products to leverage when communicating with specific segments.

For example, a retail brand looking to determine how to reactivate lapsed customers might create a segment of customers who purchased in the past and haven’t purchased or browsed the eCommerce store in the past 30 days. It might then analyze that segment to understand what type of products these customers have purchased in the past, what is their discount affinity and more. Using this information, the marketing team can determine the best campaign to create in order to reactivate these lapsed customers.

Similarly, a company can use customer segmentation analysis to determine the value of certain segments by analyzing a segments predicted Future Value, average order value, loyalty tier distribution, and more.

Customer segmentation has the potential to allow marketers to address each customer in the most effective way. Using the large amount of data available on customers (and potential customers), a customer segmentation analysis allows marketers to identify discrete groups of customers with a high degree of accuracy based on demographic, behavioral and other indicators.

Since the marketer’s goal is usually to maximize the value (revenue and/or profit) from each customer, it is critical to know in advance how any particular marketing

action will influence the customer. Ideally, such “action-centric” customer segmentation will not focus on the short-term value of a marketing action, but rather the long-term customer lifetime value (CLV) impact that such a marketing

action will have. Thus, it is necessary to group, or segment, customers according to their CLV.

An additional approach to customer segmentation is leveraging machine learning algorithms to discover new segments. Different to marketer-designed segmentation models, as the ones described above, machine learning customer segmentation allows advanced algorithms to surface insights and groupings that marketers might find difficulty discovering on their own.

Furthermore, marketers that create a feedback loop between the segmentation model and campaign results will have ever improving customer segments. In these cases, the machine learning model will be not only able to refine its definition of segments, but also be able to identify if a specific subset of the segment is outperforming the rest, optimizing marketing performance.

# Types of Customer Segmentation

There are three main approaches to market segmentation or Customer Segmentation:

* A priori segmentation, the simplest approach, uses a classification scheme based on publicly available characteristics — such as industry and company size — to create distinct groups of customers within a market. However, a priori market segmentation may not always be valid, since companies in the same industry and of the same size may have very different needs.

## Needs-based segmentation is based on differentiated, validated drivers (needs) that customers express for a specific product or service being offered. The needs are discovered and verified through primary market research, and segments are demarcated based on those different needs rather than characteristics such as industry or company size.

* Value-based segmentation differentiates customers by their economic value, grouping customers with the same value level into individual segments that can be distinctly targeted.

# Benefits of Customer Segmentation

At the expansion stage, executing a marketing strategy without any knowledge of how your target market is segmented is akin to firing shots at a target 100 feet away — while blindfolded. The likelihood of hitting the target is a matter of luck more than anything else. Without a deep understanding of how a company’s best

current customers are segmented, a business often lacks the market focus needed to allocate and spend its precious human and capital resources efficiently.

Furthermore, a lack of best current customer segment focus can cause diffused go- tomarket and product development strategies that hamper a company’s ability to fully engage with its target segments. Together, all of those factors can ultimately impede a company’s growth.

If best current customer segmentation is done right, however, the business benefits are numerous. For example, a best current customer segmentation exercise can tangibly impact your operating results by:

* 1. Improving your whole product: Having a clear idea of who wants to buy your product and what they need it for will help you differentiate your company as the best solution for their individual needs. The result will be increased satisfaction and better performance against competitors. The benefits also extend beyond your core product offering, since any insights into your best customers will allow your organization to offer better customer support, professional services, and any other offerings that make up their whole product experience.
  2. Focusing your marketing message: In parallel with improvements to the product, conducting a customer segmentation project can help you develop more

focused marketing messages that are customized to each of your best segments, resulting in higher quality inbound interest in your product.

* 1. Allowing your sales organization to pursue higher percentage opportunities: By spending less time on less lucrative opportunities and more on your most successful segments, your sales team will be able to increase its win rate, cover more ground, and ultimately increase revenues.
  2. Getting higher quality revenues: Not all revenue dollars are created equal.

Sales into the wrong segment can be more expensive to sell and maintain, and may have a higher churn rate or lower upsell potential after the initial purchase has been made. Staying away from these types of customers and focusing on better ones will increase your margins and promote the stability of your customer base.

Conducting best current customer segmentation research can have numerous other ancillary benefits, of course, but this guide will focus primarily on how it can impact the four cited above. The bottom line is that if you are able to sell more of your product to your most profitable customers, then you will be able to scale the business more efficiently and ensure that everything you do — from lead generation to new product development — revolves around the right things.

# Customer Segmentation Using Cluster Analysis

The partional algorithms which divides the dataset into several clusters. In case of Kmeans where we need to know the number of clusters and in case of DBSCAN we dont. But the catch with DBSCAN is that we have two hyperparameters epsilon and min-samples that we will see below in the notebook.

Defining the number of clusters before hand requires domain knowledge which can be challenging. Various methods like the elbow chart, silhoute distance method, gap statistics are developed to bridge the shortcommings.

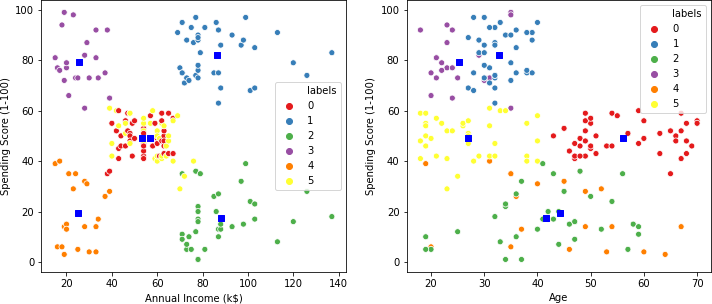
In brief, cluster analysis uses a mathematical model to discover groups of similar customers based on finding the smallest variations among customers within each group. The process is not based on any predetermined thresholds or rules (as are most simple segmentation methods), but rather the data itself generates the customer prototypes that inherently exist within the population of customers. The two main advantages of cluster analysis over simple threshold/rulebased segmentation are –

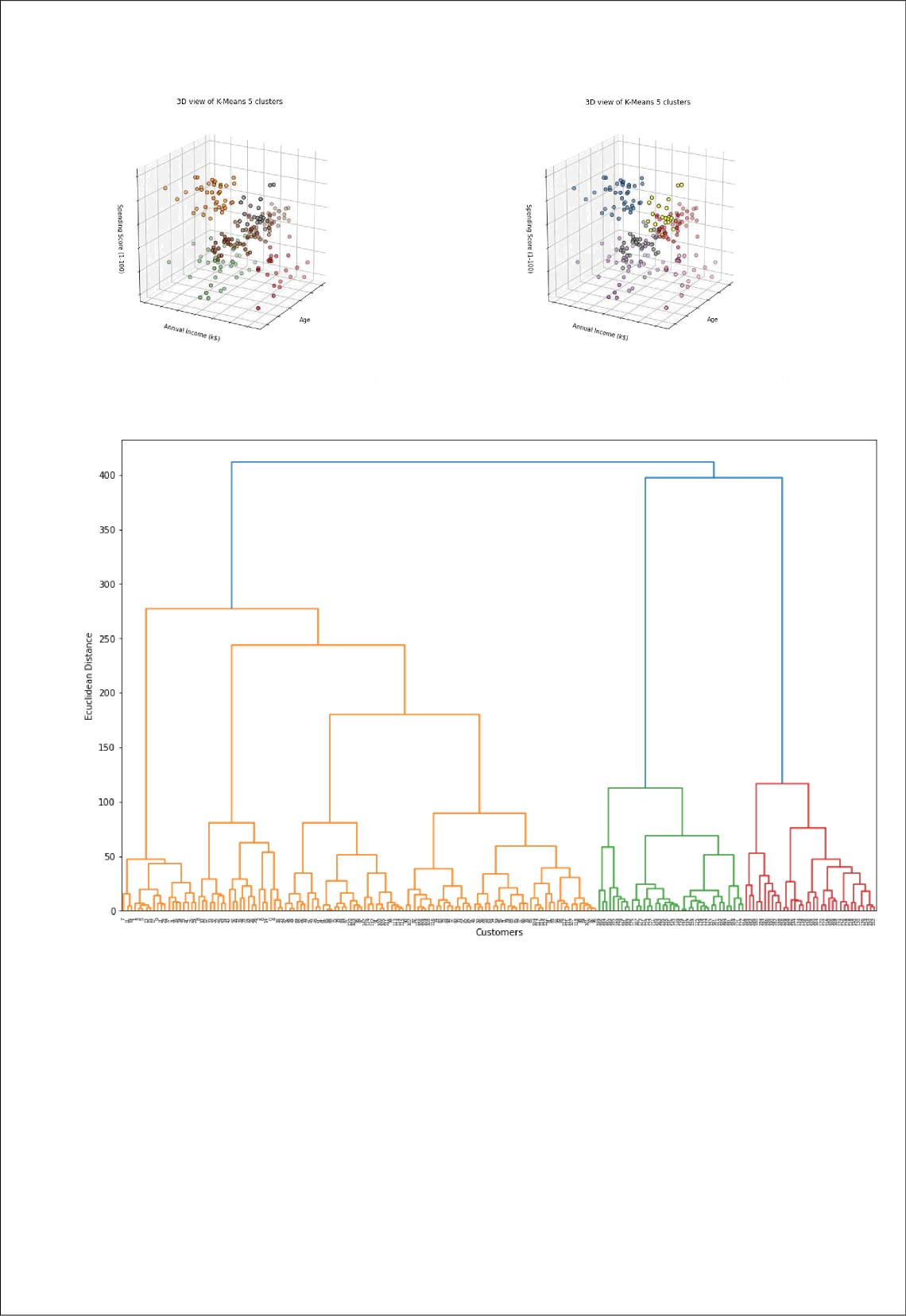
* Practicality – it would be practically impossible to use predetermined rules to segment customers over many dimensions, and
* Homogeneity – variances within each resulting group are very small in cluster analysis, whereas rule-based segmentation typically groups customers who are actually very different from one another. The customer segmentation process can be

performed with various clustering algorithms. We focused on k-means clustering in

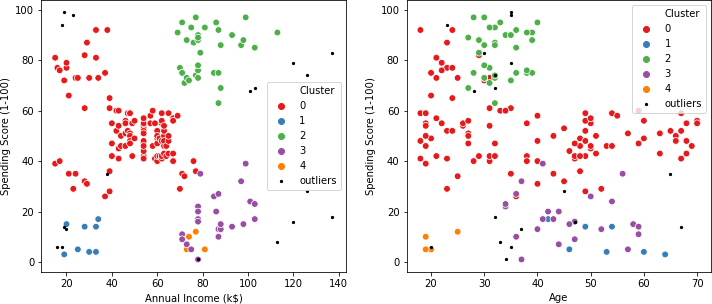
R. While the algorithm is quite simple to implement, half the battle is getting the data into the correct format and interpreting the results. We went over formatting the order data, running the kmeans(), Affinity propagation and DBSCAN function to cluster the data with several hypothetical kk clusters, using silhouette() from the cluster package to determine the optimal number of kk clusters, and interpreting the results by inspection of the k-means centroids.

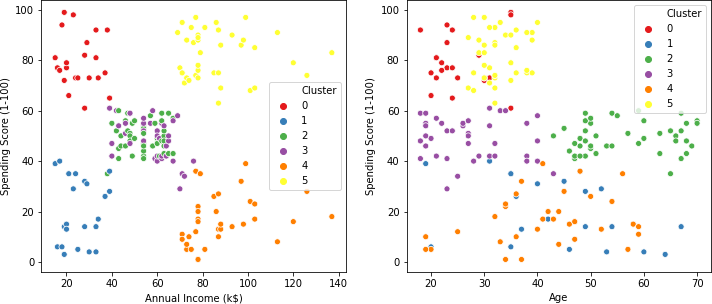
Some Sample Graphs Used:





Dend rogam





# Conclusion

# While this guide provides a step-by-step process for identifying, prioritizing, and targeting your best current customer segments, simply following it does not guarantee success. To be effective, you must prepare and plan for the various challenges and hurdles that each step may present, and always make sure to adapt your process to any new information or feedback that might change its output.

Additionally, you cannot force feed this process on your business. If the key stakeholders that will be impacted by the best current customers segmentation process do not fully buy-in, then the outputs produced from it will be relatively meaningless. If you properly manage the best current customer segmentation process, however, the impact it can have on every part of your organization — sales, marketing, product development, customer service, etc. — is immense. Your business will possess stronger customer focus and market clarity, allowing it to scale in a far more predictable and efficient manner. Ultimately, that means no longer needing to take on every customer that is willing to pay for your product or service, which will allow you to instead hone in on a specific subset of customers that present the most profitable opportunities and efficient use of resources. That is critical for every business, of course, but at the expansion stage, it can often be the difference between incredible success and certain failure.

# REFERENCES

* https://labs.openviewpartners.com/customer-segmentation/
* Source data related to our analysis has been collected from

## https://[www.kaggle.com/](http://www.kaggle.com/) and other you tube channels

CODE SAMPLE

In [235…

**import** pandas **as** pd **import** numpy **as** np **import** seaborn **as** sns

**import** matplotlib.pyplot **as** plt

**from** scipy **import** stats

**from** sklearn.cluster **import** KMeans, AgglomerativeClustering, DBSCAN, AffinityPropagation

**from** sklearn.metrics **import** silhouette\_score

**import** plotly **as** py

**import** plotly.graph\_objs **as** go

**import** scipy.cluster.hierarchy **as** sch

**from** itertools **import** product

**from** scipy.stats **import** pearsonr

**from** mpl\_toolkits.mplot3d **import** Axes3D **from** mpl\_toolkits.mplot3d **import** Axes3D **import** plotly **as** py

**import** plotly.graph\_objs **as** go

In [6]:

da\_ta **=** pd**.**read\_csv('Mall\_Customers.csv')*#importing the da\_ta set*

da\_ta**.**head()

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Out[6]: |  | **CustomerID** | **Gender** | **Age** | **Annual Income (k$)** | **Spending Score (1-100)** |
|  | **0** | 1 | Male | 19 | 15 | 39 |
|  | **1** | 2 | Male | 21 | 15 | 81 |
|  | **2** | 3 | Female | 20 | 16 | 6 |
|  | **3** | 4 | Female | 23 | 16 | 77 |
|  | **4** | 5 | Female | 31 | 17 | 40 |

In [7]:

print(da\_ta**.**shape)

(200, 5)

In [4]:

da\_ta**.**describe()

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Out[4]: | **CustomerID** | **Age** | **Annual Income (k$)** | **Spending Score (1-100)** |
|  | **count** 200.000000 | 200.000000 | 200.000000 | 200.000000 |
|  | **mean** 100.500000 | 38.850000 | 60.560000 | 50.200000 |
|  | **std** 57.879185 | 13.969007 | 26.264721 | 25.823522 |
|  | **min** 1.000000 | 18.000000 | 15.000000 | 1.000000 |
|  | **25%** 50.750000 | 28.750000 | 41.500000 | 34.750000 |
|  | **50%** 100.500000 | 36.000000 | 61.500000 | 50.000000 |
|  | **75%** 150.250000 | 49.000000 | 78.000000 | 73.000000 |
|  | **max** 200.000000 | 70.000000 | 137.000000 | 99.000000 |

In [5]:

Out[5]:da\_ta**.**isnull()**.**sum()

CustomerID 0

Gender 0

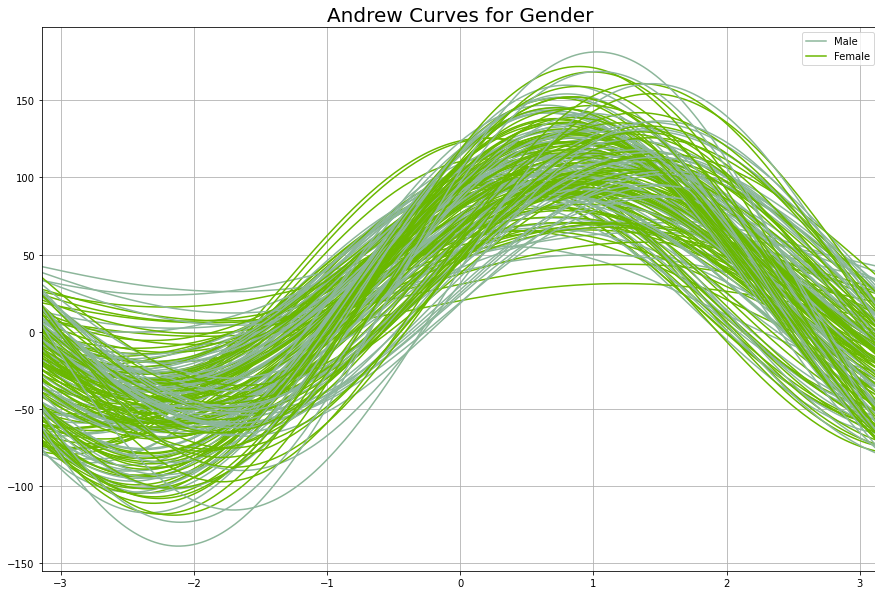
Age 0

Annual Income (k$) 0

Spending Score (1-100) 0

dtype: int64

In [218…



In [9]:

plt**.**rcParams['figure.figsize'] **=** (15, 10) pd**.**plotting**.**andrews\_curves(da\_ta**.**drop("CustomerID", axis**=**1), "Gender") plt**.**title('Andrew Curves for Gender', fontsize **=** 20)

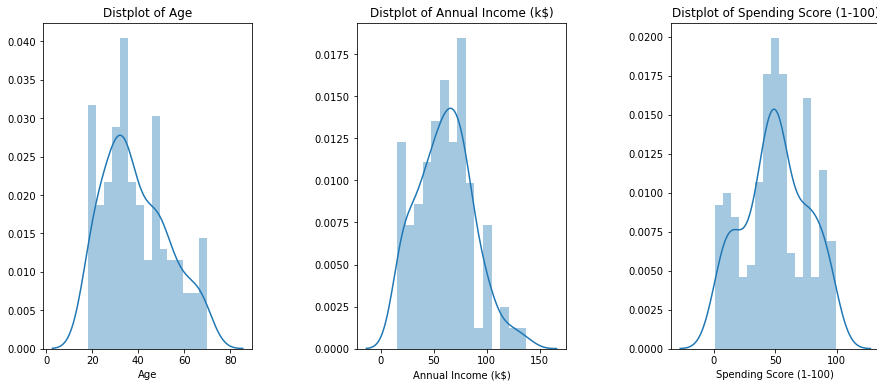
plt**.**show()

plt**.**figure(1 , figsize **=** (15 , 6)) number **=** 0

**for** x **in** number ['Age' , 'Anumbernumberual Inumbercome (k$)' , 'Spenumberdinumberg Score number **+=** 1

plt**.**subplot(1 , 3 , number) plt**.**subplots\_adjust(hspace **=** 0.5 , wspace **=** 0.5) snumbers**.**distplot(da\_ta[x] , binumbers **=** 15) plt**.**title('Distplot of {}'**.**format(x))

plt**.**show()



In [ ]:

In [10]:

*#Subsetting the dataframes for different gender groups* ma\_da\_ta\_le **=** data[data['Gender'] **==** 'Male'] fe\_da\_ta\_ma\_le **=** data[data['Gender'] **==** 'Female']

*#Explore the Age , Income and Spending score for Males*

plt**.**figure(1 , figsize **=** (15 , 6)) n **=** 0

**for** x **in** ['Age' , 'Annual Income (k$)' , 'Spending Score (1-100)']: n **+=** 1

plt**.**subplot(1 , 3 , n) plt**.**subplots\_adjust(hspace **=** 0.5 , wspace **=** 0.5) sns**.**distplot(ma\_da\_ta\_le[x] , bins **=** 20) plt**.**title('Distplot of male {}'**.**format(x))

plt**.**show()

*#Explore the Age, Income and Spending score of Females*

plt**.**figure(1 , figsize **=** (15 , 6)) n **=** 0

**for** x **in** ['Age' , 'Annual Income (k$)' , 'Spending Score (1-100)']:

Out[10]:

In [13]:

sns**.**countplot(y **=** 'Gender' , da\_ta **=** da\_ta)

<matplotlib.axes.\_subplots.AxesSubplot at 0x23f7f4d8c88>

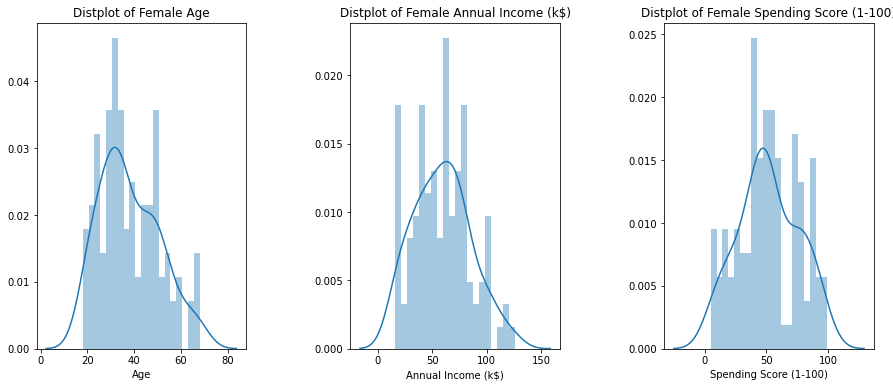
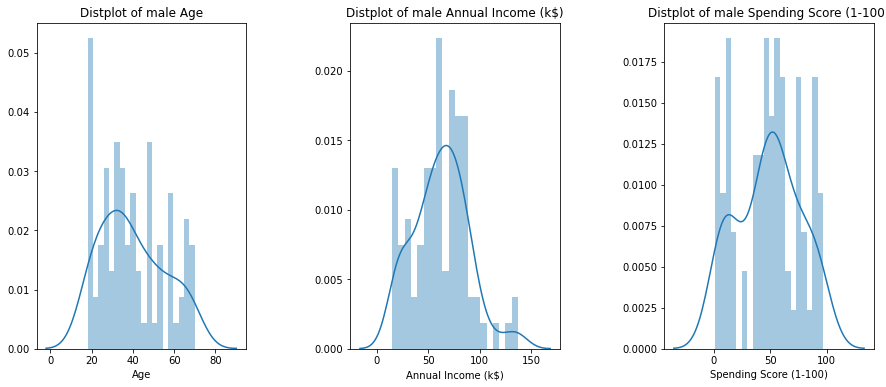


n **+=** 1

plt**.**subplot(1 , 3 , n)

plt**.**subplots\_adjust(hspace **=** 0.5 , wspace **=** 0.5) sns**.**distplot(fe\_da\_ta\_ma\_le[x] , bins **=** 20) plt**.**title('Distplot of Female {}'**.**format(x))

plt**.**show()



In [18]:

*#Investigate different age groups for males and females*

age\_da\_male\_ta **=** da\_ta[da\_ta['Gender']**==**'Male']['Age'] *# subset with males age*

fe\_age\_da\_male\_ta **=** da\_ta[da\_ta['Gender']**==**'Female']['Age'] *# subset with females age*

age\_bins **=** range(15,75,5)

*# males histogram*

fig2, (ax1, ax2) **=** plt**.**subplots(1, 2, figsize**=**(12,5), sharey**=True**) sns**.**distplot(males\_age, bins**=**age\_bins, kde**=False**, color**=**'#0066ff', ax**=**ax1, hist\_kws**=**dict( ax1**.**set\_xticks(age\_bins)

ax1**.**set\_ylim(top**=**25) ax1**.**set\_title('Males') ax1**.**set\_ylabel('Count')

ax1**.**text(45,23, "TOTAL count: {}"**.**format(age\_da\_male\_ta**.**count())) ax1**.**text(45,22, "Mean age: {:.1f}"**.**format(age\_da\_male\_ta**.**mean()))

In [24]:

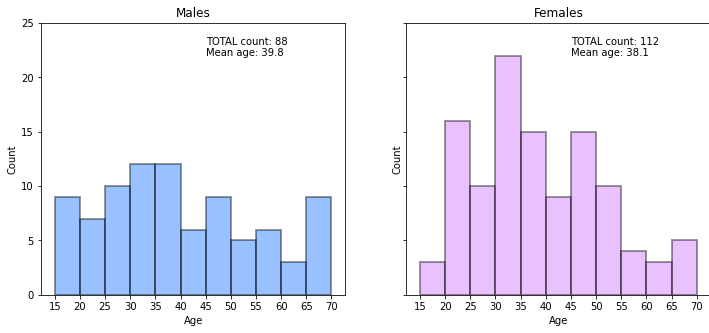
*# females histogram*

sns**.**distplot(females\_age, bins**=**age\_bins, kde**=False**, color**=**'#cc66ff', ax**=**ax2, hist\_kws**=**dic ax2**.**set\_xticks(age\_bins)

ax2**.**set\_ylim(top**=**25) ax2**.**set\_title('Females') ax2**.**set\_ylabel('Count')

ax2**.**text(45,23, "TOTAL count: {}"**.**format(fe\_age\_da\_male\_ta**.**count())) ax2**.**text(45,22, "Mean age: {:.1f}"**.**format(fe\_age\_da\_male\_ta**.**mean()))

plt**.**show()



In [29]:

*#finding the maximum Expendicture with respect to GENDER and AGE*

spend\_by\_male **=** da\_ta[da\_ta['Gender']**==**'Male']['Spending Score (1-100)'] *# subset with ma*

fe\_spend\_by\_male **=** da\_ta[da\_ta['Gender']**==**'Female']['Spending Score (1-100)'] *# subset wi*

spending\_bin **=** range(0,105,5)

*# males histogram*

fig2, (ax1, ax2) **=** plt**.**subplots(1, 2, figsize**=**(18,5), sharey**=True**) sns**.**distplot(spend\_by\_male, bins**=**spending\_bin, kde**=False**, color**=**'#0066ff', ax**=**ax1, hist\_k ax1**.**set\_xticks(spending\_bin)

ax1**.**set\_ylim(top**=**25) ax1**.**set\_title('Males') ax1**.**set\_ylabel('Count')

ax1**.**text(50,15, "Mean spending score: {:.1f}"**.**format(spend\_by\_male**.**mean())) ax1**.**text(50,14, "Median spending score: {:.1f}"**.**format(spend\_by\_male**.**median())) ax1**.**text(50,13, "Std. deviation score: {:.1f}"**.**format(spend\_by\_male**.**std()))

*# females histogram*

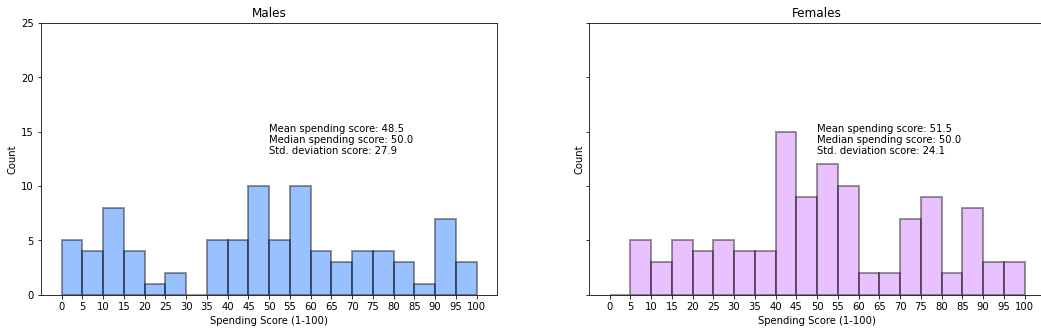
sns**.**distplot(fe\_spend\_by\_male, bins**=**spending\_bin, kde**=False**, color**=**'#cc66ff', ax**=**ax2, his ax2**.**set\_xticks(spending\_bin)

ax2**.**set\_ylim(top**=**25) ax2**.**set\_title('Females') ax2**.**set\_ylabel('Count')

ax2**.**text(50,15, "Mean spending score: {:.1f}"**.**format(fe\_spend\_by\_male**.**mean())) ax2**.**text(50,14, "Median spending score: {:.1f}"**.**format(fe\_spend\_by\_male**.**median())) ax2**.**text(50,13, "Std. deviation score: {:.1f}"**.**format(fe\_spend\_by\_male**.**std()))

Out[29]:

Text(50, 13, 'Std. deviation score: 24.1')



In [40]:

Age\_median **=** da\_ta**.**groupby(["Gender",pd**.**cut(da\_ta['Age'], age\_bins)])**.**median() Age\_median**.**index **=** Age\_median**.**index**.**set\_names(['Gender', 'Age\_group']) Age\_median**.**reset\_index(inplace**=True**)

In [41]:

Age\_median**.**head(10)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Out[41]: | **Gender** | **Age\_group** | **CustomerID** | **Age** | **Annual Income (k$)** | **Spending Score (1-100)** |
|  | **0** Female | (15, 20] | 112.0 | 19.0 | 63.0 | 50.0 |
|  | **1** Female | (20, 25] | 46.0 | 23.0 | 39.0 | 65.0 |
|  | **2** Female | (25, 30] | 139.5 | 29.0 | 74.5 | 73.0 |
|  | **3** Female | (30, 35] | 126.0 | 32.0 | 70.0 | 69.0 |
|  | **4** Female | (35, 40] | 138.5 | 38.0 | 73.5 | 41.0 |
|  | **5** Female | (40, 45] | 137.0 | 44.0 | 73.0 | 28.0 |
|  | **6** Female | (45, 50] | 82.0 | 49.0 | 54.0 | 44.5 |
|  | **7** Female | (50, 55] | 87.0 | 54.0 | 57.0 | 43.0 |
|  | **8** Female | (55, 60] | 74.0 | 58.0 | 50.0 | 35.0 |
|  | **9** Female | (60, 65] | 79.0 | 64.0 | 51.5 | 39.0 |

In [42]:

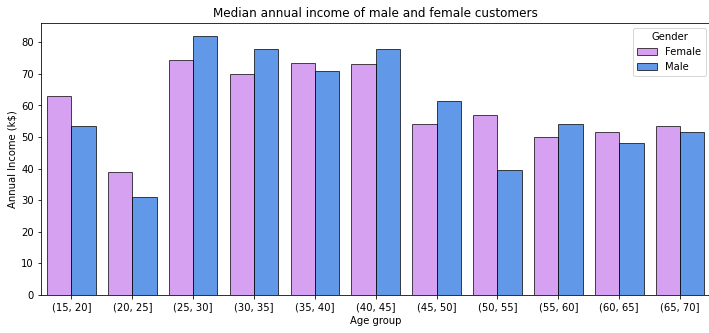
fig, ax **=** plt**.**subplots(figsize**=**(12,5))

sns**.**barplot(x**=**'Age\_group', y**=**'Annual Income (k$)', hue**=**'Gender', data**=**Age\_median, palette**=**['#cc66ff','#0066ff'],

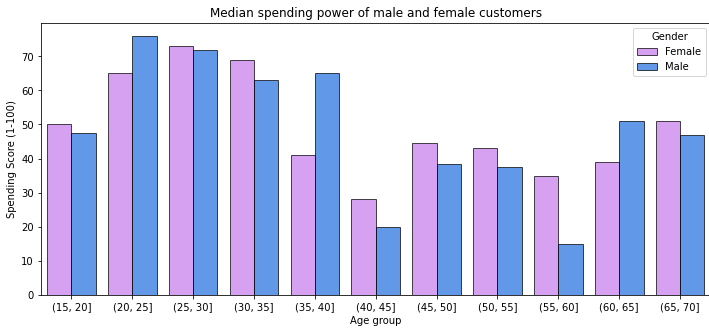
alpha**=**0.7,edgecolor**=**'k', ax**=**ax)

ax**.**set\_title('Median annual income of male and female customers') ax**.**set\_xlabel('Age group')

plt**.**show()



In [43]:



In [58]:

fig, ax **=** plt**.**subplots(figsize**=**(12,5))

sns**.**barplot(x**=**'Age\_group', y**=**'Spending Score (1-100)', hue**=**'Gender', data**=**Age\_median, palette**=**['#cc66ff','#0066ff'],

alpha**=**0.7,edgecolor**=**'k', ax**=**ax)

ax**.**set\_title('Median spending power of male and female customers') ax**.**set\_xlabel('Age group')

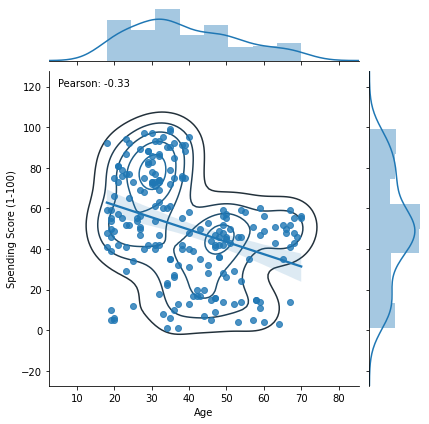
plt**.**show()

*# calculating Pearson's correlation*

corr, \_ **=** pearsonr(da\_ta['Age'], da\_ta['Spending Score (1-100)'])

joint\_ploting **=** (sns**.**jointplot('Age', 'Spending Score (1-100)', data**=**da\_ta, kind**=**'reg'))**.**plot\_joint(sns**.**kdeplot, zorder**=**0, n\_levels**=**6)

plt**.**text(5,120, 'Pearson: {:.2f}'**.**format(corr)) plt**.**show()



In [60]:

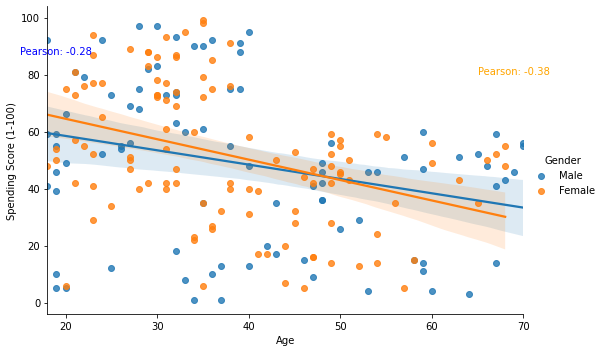
*# calculating Pearson's correlation betweem ,age groups and spending power*

corr1, \_ **=** pearsonr(age\_da\_male\_ta**.**values, spend\_by\_male**.**values) corr2, \_ **=** pearsonr(fe\_age\_da\_male\_ta**.**values, fe\_spend\_by\_male**.**values)

sns**.**lmplot('Age', 'Spending Score (1-100)' , data**=**da\_ta, hue**=**'Gender', aspect**=**1.5)

plt**.**text(15,87, 'Pearson: {:.2f}'**.**format(corr1), color**=**'blue') plt**.**text(65,80, 'Pearson: {:.2f}'**.**format(corr2), color**=**'orange')

plt**.**show()



### K-MEANS

In [209…

*#Trying K-Means Clustering for AGE AND SPENDING SCORE TO SEE THE VARIATION..*

X1 **=** data[['Age' , 'Spending Score (1-100)']]**.**iloc[: , :]**.**values inter **=** []

point\_s **=** []

**for** n **in** range(2 , 11):

algorithm **=** (KMeans(n\_clusters **=** n ,init**=**'k-means++', n\_init **=** 10 ,max\_iter**=**300, tol**=**0.0001, random\_state**=** 111 , algorithm**=**'elkan') )

algorithm**.**fit(X1) inter**.**append(algorithm**.**inter\_)

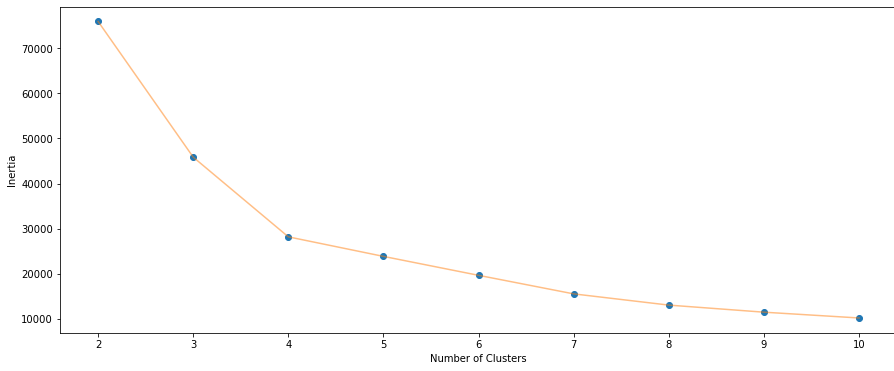
silhouette\_avg **=** silhouette\_score(X1, algorithm**.**labels\_) point\_s**.**append(silhouette\_avg) *# data for the silhouette score method*

In [211…

*#The Elbow Chart*

plt**.**figure(1 , figsize **=** (15 ,6)) plt**.**plot(np**.**arange(2 , 11) , inter , 'o')

plt**.**plot(np**.**arange(2 , 11) , inter , '-' , alpha **=** 0.5) plt**.**xlabel('Number of Clusters') , plt**.**ylabel('inter') plt**.**show()



### Similarly we can go for the combinations of other variables

In [79]:

X **=** da\_ta[['Age' ,'Annual Income (k$)' ,'Spending Score (1-100)']] inter\_2 **=** []

point\_s\_2 **=** []

**for** n **in** range(2 , 11):

algorithm **=** KMeans(n\_clusters **=** n ,init**=**'k-means++', n\_init **=** 10 ,max\_iter**=**300, tol**=**0.0001, random\_state**=** 111 , algorithm**=**'elkan')**.**fit(X)

inter\_2**.**append(algorithm**.**inter\_2\_)

silhouette\_avg **=** silhouette\_score(X, algorithm**.**labels\_) point\_s\_2**.**append(silhouette\_avg) *# data for the silhouette score method*

In [81]:

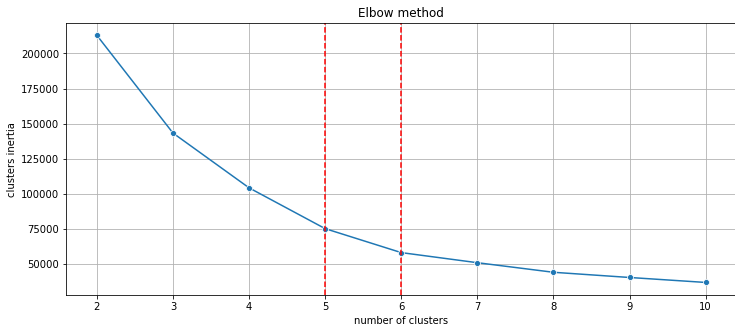
fig, ax **=** plt**.**subplots(figsize**=**(12,5))

ax **=** sns**.**lineplot(np**.**arange(2 , 11), inertia, marker**=**'o', ax**=**ax) ax**.**set\_title("Elbow method")

ax**.**set\_xlabel("number of clusters") ax**.**set\_ylabel("clusters inertia") ax**.**axvline(5, ls**=**"--", c**=**"red")

ax**.**axvline(6, ls**=**"--", c**=**"red") plt**.**grid()

plt**.**show()



### A choice of 5 or 6 clusters seems to be fair. Let's see the silhouette score.

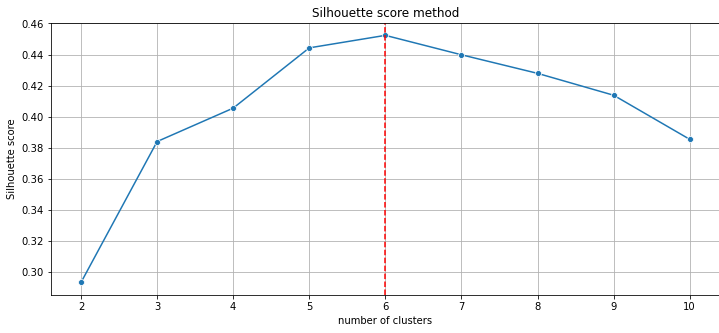
In [82]:

fig, ax **=** plt**.**subplots(figsize**=**(12,5))

ax **=** sns**.**lineplot(np**.**arange(2 , 11), s\_scores, marker**=**'o', ax**=**ax) ax**.**set\_title("Silhouette score method")

ax**.**set\_xlabel("number of clusters") ax**.**set\_ylabel("Silhouette score") ax**.**axvline(6, ls**=**"--", c**=**"red") plt**.**grid()

plt**.**show()



### Silhouette score method indicates the best options would be respectively 6 or 5 clusters. Let's compare both.

In [91]:

*#For clusters of K=6*

K\_means\_6 **=** (KMeans(n\_clusters **=** 6 ,init**=**'k-means++', n\_init **=** 10 ,max\_iter**=**300,

tol**=**0.0001, random\_state**=** 111 , algorithm**=**'elkan') )

K\_Means\_6**.**fit(X)

labels\_6 **=** K\_Means\_6**.**labels\_

center\_6 **=** K\_Means\_6**.**cluster\_centers\_ KM6\_df **=** da\_ta**.**copy() KM6\_df['labels'] **=** labels6

In [94]:

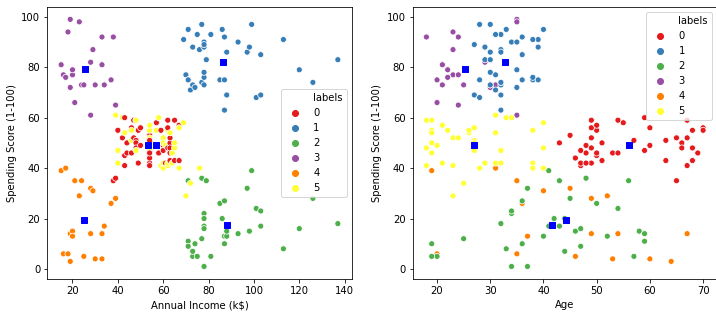
fig1, (axes) **=** plt**.**subplots(1,2,figsize**=**(12,5))

scat\_1 **=** sns**.**scatterplot('Annual Income (k$)', 'Spending Score (1-100)', data**=**KM6\_df, hue**=**'labels', ax**=**axes[0], palette**=**'Set1', legend**=**'full')

sns**.**scatterplot('Age', 'Spending Score (1-100)', data**=**K\_Means\_6\_df, hue**=**'labels', palette**=**'Set1', ax**=**axes[1], legend**=**'full')

axes[0]**.**scatter(center\_6[:,1],center\_6[:,2], marker**=**'s', s**=**40, c**=**"blue")

axes[1]**.**scatter(center\_6[:,0],center\_6[:,2], marker**=**'s', s**=**40, c**=**"blue") plt**.**show()



### K-Means algorithm generated the following 6 clusters:

younger clients with **medium** annual and **medium** spending score clients with **high** annual income and **low** spending score

younger clients with **medium** annual and **medium** spending score clients with **high** annual income and **high** spending score

### clients with **low** annual income and **low** spending score clients with **low** annual income and **high** spending score

In [100…

K\_Means\_clustering\_sizes **=** K\_Means\_6\_df**.**groupby('labels')**.**size()**.**to\_frame() K\_Means\_clustering\_sizes**.**columns **=** ["KM\_size"]

K\_Means\_clustering\_sizes

|  |  |  |
| --- | --- | --- |
| Out[100… | **labels** | **KM\_size** |
|  | **0** | 45 |
|  | **1** | 39 |
|  | **2** | 35 |
|  | **3** | 22 |
|  | **4** | 21 |
|  | **5** | 38 |

In [104…

In [110…

fig **=** plt**.**figure(figsize**=**(7, 7))

ax **=** Axes3D(fig, rect**=**[0, 0, .99, 1], elev**=**20, azim**=**210) ax**.**scatter(K\_Means\_6\_df['Age'],

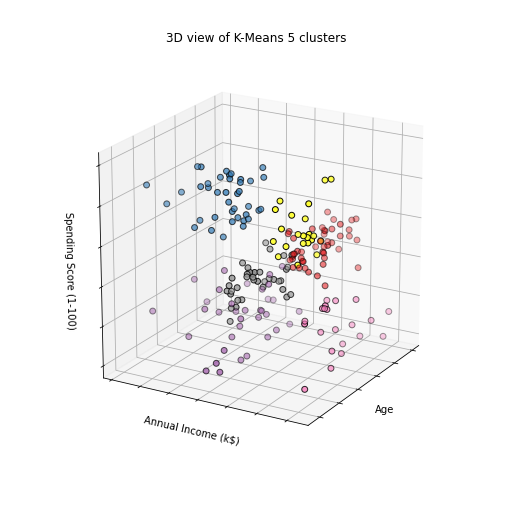
K\_Means\_6\_df['Annual Income (k$)'], K\_Means\_6\_df['Spending Score (1-100)'], c**=**K\_Means\_6\_df['labels'],

s**=**35, edgecolor**=**'k', cmap**=**plt**.**cm**.**Set1)

ax**.**w\_xaxis**.**set\_ticklabels([]) ax**.**w\_yaxis**.**set\_ticklabels([]) ax**.**w\_zaxis**.**set\_ticklabels([]) ax**.**set\_xlabel('Age') ax**.**set\_ylabel('Annual Income (k$)') ax**.**set\_zlabel('Spending Score (1-100)')

ax**.**set\_title('3D view of K-Means 5 clusters') ax**.**dist **=** 12

plt**.**show()



*#For clusters of K=5*

K\_Means\_5 **=** (KMeans(n\_clusters **=** 5 ,init**=**'k-means++', n\_init **=** 10 ,max\_iter**=**300,

tol**=**0.0001, random\_state**=** 111 , algorithm**=**'elkan') )

K\_Means\_5**.**fit(X)

fig1, (axes) **=** plt**.**subplots(1,2,figsize**=**(12,5))

labels\_5 **=** K\_Means\_5**.**labels\_

center\_5 **=** K\_Means\_5**.**cluster\_centers\_ K\_Means\_5\_df **=** da\_ta**.**copy() K\_Means\_5\_df['labels'] **=** labels\_5

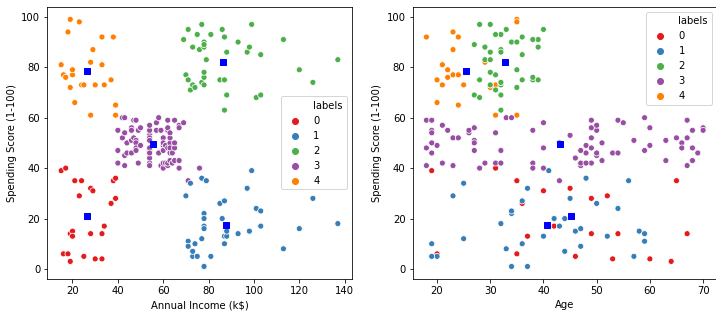
In [111…

scat\_1 **=** sns**.**scatterplot('Annual Income (k$)', 'Spending Score (1-100)', data**=**KM5\_df, hue**=**'labels', ax**=**axes[0], palette**=**'Set1', legend**=**'full')

sns**.**scatterplot('Age', 'Spending Score (1-100)', data**=**K\_Means\_5\_df, hue**=**'labels', palette**=**'Set1', ax**=**axes[1], legend**=**'full')

axes[0]**.**scatter(center\_5[:,1],center\_5[:,2], marker**=**'s', s**=**40, c**=**"blue")

axes[1]**.**scatter(center\_5[:,0],center\_5[:,2], marker**=**'s', s**=**40, c**=**"blue") plt**.**show()



### K-Means algorithm generated the following 5 clusters: clients with **low** annual income and **high** spending score

clients with **medium** annual income and **medium** spending score clients with **high** annual income and **low** spending score

### clients with **high** annual income and **high** spending score clients with **low** annual income and **low** spending score

There are no distinct groups is terms of customers age.

In [112…

K\_Means\_clust\_sizes5 **=** KM5\_df**.**groupby('labels')**.**size()**.**to\_frame() K\_Means\_clust\_sizes5**.**columns **=** ["KM\_size"]

K\_Means\_clust\_sizes5

fig **=** plt**.**figure(figsize**=**(7, 7))

ax **=** Axes3D(fig, rect**=**[0, 0, .99, 1], elev**=**20, azim**=**210) ax**.**scatter(KM5\_df['Age'],

|  |  |  |
| --- | --- | --- |
| Out[112… | **labels** | **KM\_size** |
|  | **0** | 23 |
|  | **1** | 36 |
|  | **2** | 39 |
|  | **3** | 79 |
|  | **4** | 23 |

In [113…

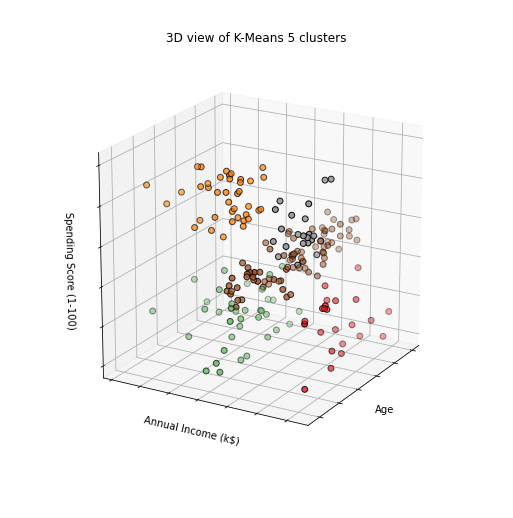
KM5\_df['Annual Income (k$)'], KM5\_df['Spending Score (1-100)'], c**=**KM5\_df['labels'],

s**=**35, edgecolor**=**'k', cmap**=**plt**.**cm**.**Set1)

ax**.**w\_xaxis**.**set\_ticklabels([]) ax**.**w\_yaxis**.**set\_ticklabels([]) ax**.**w\_zaxis**.**set\_ticklabels([]) ax**.**set\_xlabel('Age') ax**.**set\_ylabel('Annual Income (k$)') ax**.**set\_zlabel('Spending Score (1-100)')

ax**.**set\_title('3D view of K-Means 5 clusters') ax**.**dist **=** 12

plt**.**show()

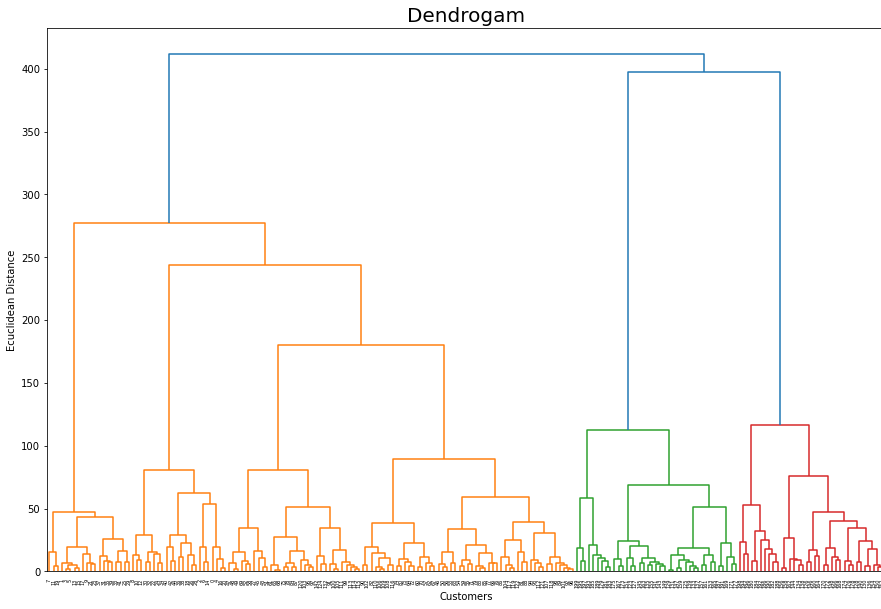


In [221…

*#Using Dendrograms to find out optimal clusters*

dendrogram **=** sch**.**dendrogram(sch**.**linkage(X, method **=** 'ward')) plt**.**title('Dendrogam', fontsize **=** 20) plt**.**xlabel('Customers')

plt**.**ylabel('Ecuclidean Distance') plt**.**show()



In [226…

clustering\_agglomerative **=** AgglomerativeClustering(n\_clusters **=** 5, affinity **=** 'euclidean'

clustering\_agglomerative**.**fit(X)

labels\_clustering\_agglomerative **=** clustering\_agglomerative**.**labels\_ clustering\_agglomerative\_df **=** data**.**copy() clustering\_agglomerative\_df['labels'] **=** labels\_clustering\_agglomerative

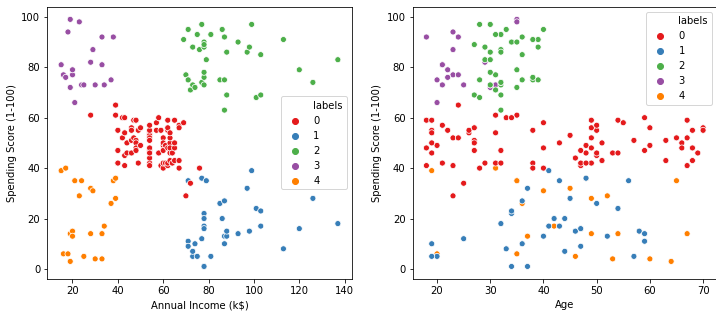
In [234…

fig1, (axes) **=** plt**.**subplots(1,2,figsize**=**(12,5))

sns**.**scatterplot('Annual Income (k$)', 'Spending Score (1-100)', data**=**clustering\_agglomera hue**=**'labels', ax**=**axes[0], palette**=**'Set1', legend**=**'full')

sns**.**scatterplot('Age', 'Spending Score (1-100)', data**=**clustering\_agglomerative\_df, hue**=**'labels', palette**=**'Set1', ax**=**axes[1], legend**=**'full')

plt**.**show()



**DBSCAN**

In [121… eps\_values **=** np**.**arange(8,12.75,0.25) *# eps values to be investigated*

min\_samples **=** np**.**arange(3,10) *# min\_samples values to be investigated*

DBSCAN\_params **=** list(product(eps\_values, min\_samples))*# creates a mxn combinations of eps*

In [123…

no\_of\_clusters **=** [] sil\_score **=** []

**for** p **in** DBSCAN\_params:

DBSCN\_clustering **=** DBSCAN(eps**=**p[0], min\_samples**=**p[1])**.**fit(X) no\_of\_clusters**.**append(len(np**.**unique(DBSCN\_clustering**.**labels\_))) sil\_score**.**append(silhouette\_score(X, DBSCN\_clustering**.**labels\_))

In [124…

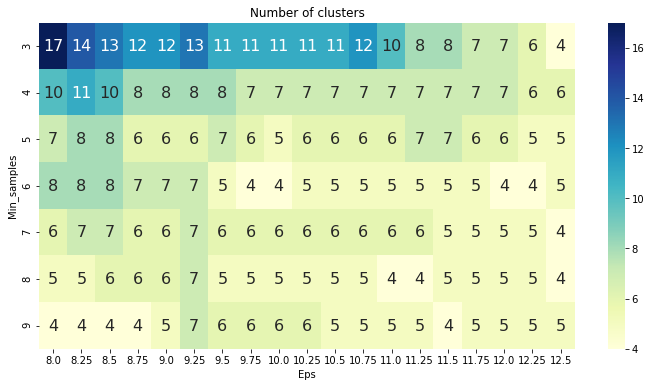
tmp **=** pd**.**DataFrame**.**from\_records(DBSCAN\_params, columns **=**['Eps', 'Min\_samples']) tmp['No\_of\_clusters'] **=** no\_of\_clusters

pivot\_1 **=** pd**.**pivot\_table(tmp, values**=**'No\_of\_clusters', index**=**'Min\_samples', columns**=**'Eps'

fig, ax **=** plt**.**subplots(figsize**=**(12,6))

sns**.**heatmap(pivot\_1, annot**=True**,annot\_kws**=**{"size": 16}, cmap**=**"YlGnBu", ax**=**ax) ax**.**set\_title('Number of clusters')

plt**.**show()



### Range of clusters is between 17 to 4. Now we see which cluster has the maximum value from the heat map and choose the corresponding min and eps values

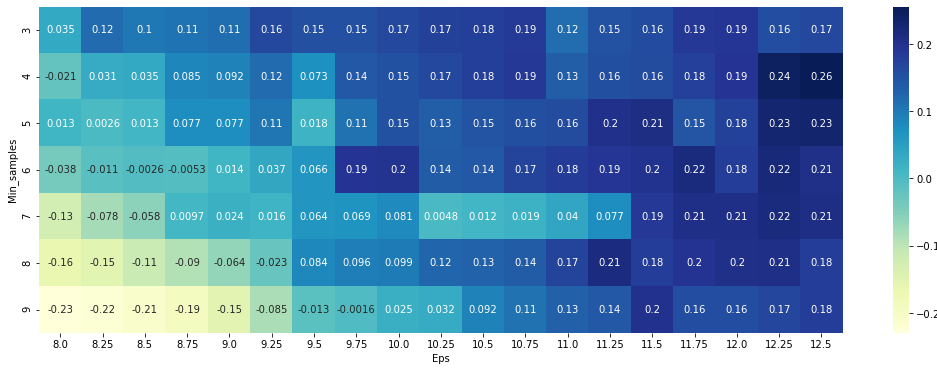
In [127…

tmp **=** pd**.**DataFrame**.**from\_records(DBSCAN\_params, columns **=**['Eps', 'Min\_samples']) tmp['Sil\_score'] **=** sil\_score

pivot\_1 **=** pd**.**pivot\_table(tmp, values**=**'Sil\_score', index**=**'Min\_samples', columns**=**'Eps') fig, ax **=** plt**.**subplots(figsize**=**(18,6))

sns**.**heatmap(pivot\_1, annot**=True**, annot\_kws**=**{"size": 10}, cmap**=**"YlGnBu", ax**=**ax)

plt**.**show()



### The highest value of 0.26 is for 12.5 and 4

In [129…

DBSCN\_clustering **=** DBSCAN(eps**=**12.5, min\_samples**=**4)**.**fit(X) DBSCAN\_cluster **=** X**.**copy()

DBSCAN\_cluster**.**loc[:,'Cluster'] **=** DBSCN\_clustering**.**labels\_ *# append labels to points*

DBSCAN\_clustered\_sizes **=** DBSCAN\_cluster**.**groupby('Cluster')**.**size()**.**to\_frame()

In [130…

DBSCAN\_clustered\_sizes**.**columns **=** ["DBSCAN\_size"] DBSCAN\_clustered\_sizes

|  |  |  |
| --- | --- | --- |
| Out[130… | **Cluster** | **DBSCAN\_size** |
|  | **-1** | 18 |
|  | **0** | 112 |
|  | **1** | 8 |
|  | **2** | 34 |
|  | **3** | 24 |
|  | **4** | 4 |

### We have

Outliers = 18 denoted by cluster -1

### Core clusters = 0-4 where 0 has around 112 data points

In [131…

outliers **=** DBSCAN\_cluster[DBSCAN\_cluster['Cluster']**==-**1] fig2, (axes) **=** plt**.**subplots(1,2,figsize**=**(12,5))

sns**.**scatterplot('Annual Income (k$)', 'Spending Score (1-100)', data**=**DBSCAN\_clustered[DBSCAN\_clustered['Cluster']**!=-**1], hue**=**'Cluster', ax**=**axes[0], palette**=**'Set1', legend**=**'full', s**=**45)

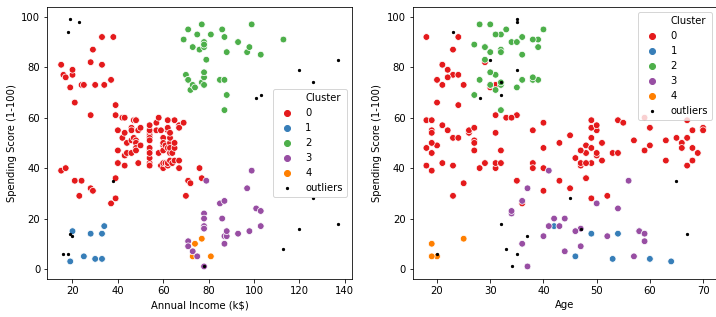
sns**.**scatterplot('Age', 'Spending Score (1-100)',

data**=**DBSCAN\_cluster[DBSCAN\_cluster['Cluster']**!=-**1], hue**=**'Cluster', palette**=**'Set1', ax**=**axes[1], legend**=**'full', s**=**45)

axes[0]**.**scatter(outliers['Annual Income (k$)'], outliers['Spending Score (1-100)'], s**=**5, axes[1]**.**scatter(outliers['Age'], outliers['Spending Score (1-100)'], s**=**5, label**=**'outliers axes[0]**.**legend()

axes[1]**.**legend() plt**.**setp(axes[0]**.**get\_legend()**.**get\_texts(), fontsize**=**'10') plt**.**setp(axes[1]**.**get\_legend()**.**get\_texts(), fontsize**=**'10')

plt**.**show()



In [245…

no\_of\_clusters **=** []

preferences **=** range(**-**20000,**-**500,200) af\_sil\_score **=** [] *# silouette scores*

**for** p **in** preferences:

AF **=** AffinityPropagation(preference**=**p, max\_iter**=**200)**.**fit(X) no\_of\_clusters**.**append((len(np**.**unique(AF**.**labels\_)))) af\_sil\_score**.**append(silhouette\_score(X, AF**.**labels\_))

af\_results **=** pd**.**DataFrame([preferences, no\_of\_clusters, af\_sil\_score], index**=**['preference af\_results**.**sort\_values(by**=**'sil\_score', ascending**=False**)**.**head()

|  |  |  |  |
| --- | --- | --- | --- |
| Out[245… | **preference** | **clusters** | **sil\_score** |
|  | **41** -11800.0 | 6.0 | 0.451649 |
|  | **14** -17200.0 | 6.0 | 0.451440 |
|  | **55** -9000.0 | 6.0 | 0.451440 |
|  | **26** -14800.0 | 6.0 | 0.451440 |
|  | **32** -13600.0 | 6.0 | 0.451440 |

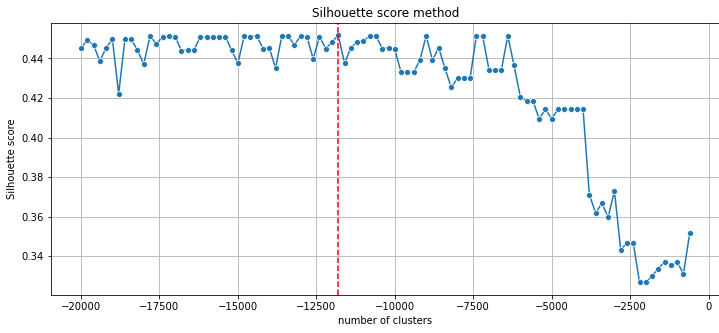
In [246…

fig, ax **=** plt**.**subplots(figsize**=**(12,5))

ax **=** sns**.**lineplot(preferences, af\_sil\_score, marker**=**'o', ax**=**ax) ax**.**set\_title("Silhouette score method")

ax**.**set\_xlabel("number of clusters") ax**.**set\_ylabel("Silhouette score") ax**.**axvline(**-**11800, ls**=**"--", c**=**"red") plt**.**grid()

plt**.**show()



In [248…

AF **=** AffinityPropagation(preference**=-**11800)**.**fit(X)

In [250…

AF\_clustered **=** X**.**copy()

AF\_clustered**.**loc[:,'Cluster'] **=** AF**.**labels\_ *# append labels to points*

In [251…

AF\_clust\_sizes **=** AF\_clustered**.**groupby('Cluster')**.**size()**.**to\_frame() AF\_clust\_sizes**.**columns **=** ["AF\_size"]

AF\_clust\_sizes

|  |  |  |
| --- | --- | --- |
| Out[251… | **Cluster** | **AF\_size** |
|  | **0** | 22 |
|  | **1** | 22 |
|  | **2** | 44 |
|  | **3** | 39 |
|  | **4** | 34 |
|  | **5** | 39 |

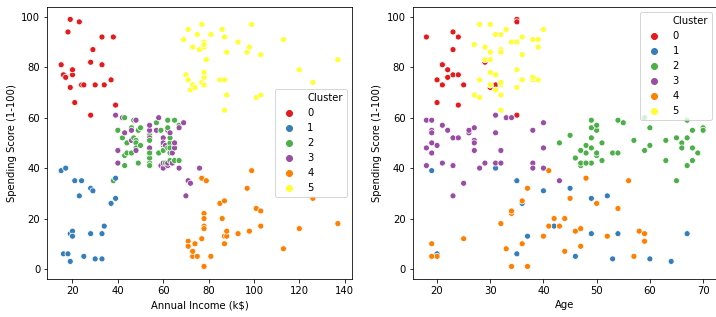
In [252…

fig3, (ax\_af) **=** plt**.**subplots(1,2,figsize**=**(12,5))

scat\_1 **=** sns**.**scatterplot('Annual Income (k$)', 'Spending Score (1-100)', data**=**AF\_clustere hue**=**'Cluster', ax**=**ax\_af[0], palette**=**'Set1', legend**=**'full')

sns**.**scatterplot('Age', 'Spending Score (1-100)', data**=**AF\_clustered, hue**=**'Cluster', palette**=**'Set1', ax**=**ax\_af[1], legend**=**'full')

plt**.**setp(ax\_af[0]**.**get\_legend()**.**get\_texts(), fontsize**=**'10') plt**.**setp(ax\_af[1]**.**get\_legend()**.**get\_texts(), fontsize**=**'10') plt**.**show()



In [254…

clusters **=** pd**.**concat([KM\_clust\_sizes,KM\_clust\_sizes5, DBSCAN\_clust\_sizes, AF\_clust\_sizes]

clusters

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Out[254… |  | **KM\_size** | **KM\_size** | **DBSCAN\_size** | **AF\_size** |
|  | **-1** | NaN | NaN | 18.0 | NaN |
|  | **0** | 45.0 | 23.0 | 112.0 | 22.0 |
|  | **1** | 39.0 | 36.0 | 8.0 | 22.0 |
|  | **2** | 35.0 | 39.0 | 34.0 | 44.0 |
|  | **3** | 22.0 | 79.0 | 24.0 | 39.0 |
|  | **4** | 21.0 | 23.0 | 4.0 | 34.0 |
|  | **5** | 38.0 | NaN | NaN | 39.0 |

### From the above comparisons, it is clear that DBSCAN failed to generate reasonable clusters. It is due to its problems in recognising clusters of various densities (which are present in this case).

In turn, K-Means and Affinity Propagation algorithms created reasonable 6 clusters.