# Market Segmentation

Summary of fundamentals of Market Segmentation
and Replication of Case Study

By

**Team-C**
Thejaswin S
Ramniwash Kumar
Sameer Sekhar Bashistha

## Ten Steps of Market Segmentation Analysis

## Step 1: Deciding (not) to Segment

### Implications of Committing to Market Segmentation

Because of the significant implications of such a long-term organizational commitment, the decision to investigate the potential of a market segmentation strategy must be made at the highest executive level. It must be systematically and continuously communicated and reinforced at all organizational levels and across all administrative units. Cahill recommends not to segment unless the expected increase in sales is sufficient to justify implementing a segmentation strategy, stating that one of the truisms of segmentation strategy is that using the scheme has to be more profitable than marketing without it, net of the expense of developing and using the system itself. Before investing time and resources in a market segmentation analysis, it is essential to understand the implications of pursuing a market segmentation strategy.

Although market segmentation has developed to be an essential marketing strategy applied in many organizations, it is not always the best decision to pursue it. Croft (1994) recommends that – to maximize the benefits of market segmentation – organizations need to organize around market segments rather than to organize around products. Potentially required changes include the development of new products, the modification of existing products, changes in pricing and distribution channels used to sell the product, and all communications with the market. The commitment to market segmentation goes hand in hand with the willingness and ability of the organization to make substantial changes and investments.

### Implementation Barriers

Senior management can also prevent market segmentation from being successfully implemented by not making enough resources available, either for the initial market segmentation analysis itself or for the long-term implementation of a market segmentation strategy.

Lack of market or consumer orientation, resistance to change and new ideas, lack of creative thinking, poor communication and lack of sharing of information and insights across organizational units, short-term thinking, unwillingness to make changes, and office politics have been identified as preventing the successful implementation of market segmentation.

If senior management and the team tasked with segmentation do not understand the very foundations of market segmentation, or if they are unaware of the consequences of pursuing such a strategy, the attempt of introducing market segmentation is likely to fail.

Process-related barriers include not having clarified the objectives of the market segmentation exercise, lack of planning or bad planning, a lack of structured processes to guide the team through all steps of the market segmentation process, a lack of allocation of responsibilities, and time pressure that stands in the way of trying to find the best possible segmentation outcome.

## Step 2: Specifying the Ideal Target Segment

### Segment Evaluation Criteria

- It is essential to understand that – for a market segmentation analysis to produce results that are useful to an organization – user input cannot be limited to either a briefing at the start of the development of a marketing mix at the end.
- Instead, the user needs to be involved in most stages, literally wrapping around the technical aspects of market segmentation analysis.
- After committing to investigating the value of a segmentation strategy inStep 1, the organization has to make a significant contribution to market segmentation analysis in Step 2.
- In Step 2, the organization must determine two sets of segment evaluation criteria.
- These criteria are used to evaluate the relative attractiveness of the remaining market segments – those in compliance with the knock-out criteria.
- The second, much longer and more diverse set of attractiveness criteria represents a shopping list for the segmentation team.
- Members of the segmentation team need to select which of these criteria they want to use to determine how attractive potential target segments are.
- The segmentation team also needs to assess the relative importance of each attractiveness criterion to the organization.

### Knock-Out Criteria

The segment must match the organization's strengths; the organization must have the capability to satisfy segment members' needs. The part must be distinct; members of the part must be distinctly different from members of other features. The part must be reachable; there has to be a way to get in touch with members of the part in order to make the customized marketing mix accessible to them.

Knock-out criteria are used to determine if market segments resulting from the market segmentation analysis qualify to be assessed using segment attractiveness criteria. Kotler himself and several other authors have since recommended additional measures that fall into the knock-out criterion category:

- The segment must be large enough; the part must contain enough consumers to make it worthwhile to spend extra money on customizing the marketing mix for them.

- Members of the segment must be identifiable; it must be possible to spot them in the marketplace. The part must be homogeneous; members of the part must be similar.

### Attractiveness Criteria

The attractiveness across all criteria determines whether a market segment is selected as a target segment in Step 8 of market segmentation analysis. Instead, each market segment is rated; it can be more or less attractive concerning a specific criterion.

### Implementing a Structured Process

The most popular structured approach for evaluating market segments given selecting them as target markets are using a segment evaluation plot showing segment attractiveness along one axis and organizational competitiveness on the other axis. But there is a massive benefit in selecting the attractiveness criteria for market segments at this early stage: knowing precisely what it is about market segments that matter to the organization ensures that all of this information is captured when collecting data (Step 3). Suppose a core team of two to three people is primarily in charge of market segmentation analysis. In that case, this team could propose an initial solution and report their choices to the advisory committee – which consists of representatives of all organizational units – for discussion and possible modification. Optimally, approval by the advisory committee should be sought because the advisory committee contains representatives from multiple administrative units bringing a range of different perspectives to the challenge of specifying segment attractiveness criteria. At the end of this step, the market segmentation team should have a list of approximately six-segment attractiveness criteria. The segmentation team determines the segment attractiveness and organizational competitiveness values. The segment evaluation plot cannot be completed in Step 2 of the market segmentation analysis because – at his point – no segments are available to assess yet.

## Step 3: Collecting Data

### Segmentation Variables

The difference between commonsense and data-driven market segmentation is that data-driven market segmentation is based not on one, but multiple segmentation variables. We use the term segmentation variable to refer to the variable in the empirical data used in commonsense segmentation to split the sample into market segments. Empirical data is used to identify or create market segments and – later in the process – describe these segments in detail. Market segments are created by splitting the sample using this variable into a part of women and a segment of men. Empirical data forms the basis of both commonsense and data-driven market segmentation.

In the data-driven case, we may, for example, want to extract market segments of tourists who do not necessarily have gender in common but rather share a standard set of benefits they seek when going on vacation. These two simple examples illustrate how critical the quality of empirical data is for developing a correct segmentation solution. Empirical data for segmentation studies can come from a range of sources: from survey studies; from observations such as scanner data where purchases are recorded and, frequently, are linked to an individual customer`s long-term purchase history via loyalty programs; or from experimental studies. Optimally, data used in segmentation studies should reflect consumer behavior. Although it arguably represents the most common source of data for market segmentation studies, survey data can be unreliable in reflecting behavior, especially when the behavior of interest is socially desirable, such as donating money to a charity or behaving in an environmentally friendly way. Therefore, surveys should not be seen as the default source of data for market segmentation studies.

**Segmentation Criteria**

- The term segmentation criterion relates to the nature of the information used for market segmentation.
- The decision on which segmentation criterion to use cannot easily be outsourced to either a consultant or a data analyst because it requires prior knowledge about the market.
- The most common segmentation criteria are geographic, socio-demographic, psychographic, and behavioral.
- Bock and Uncles (2002) argue that the following differences between consumers are the most relevant in market segmentation: profitability, bargaining power, preferences for benefits or products, barriers to choice, and consumer interaction effects.
- If demographic segmentation works for your product or service, then use demographic segmentation.
- If geographic segmentation will work because your product will only appeal to people in a particular region, use it.
- Just because psychographic segmentation is sexier and more sophisticated than demographic or geographic segmentation, it is not better.
- Better is what works for your product or service at the least possible cost.

**Geographic Segmentation**

- Typically – when geographic segmentation is used – the consumer`s location of residence serves as the only criterion to form market segments.

- IKEA offers a similar product range worldwide, yet slight differences in offers, pricing, and the option to purchase online exist independent of the customer`s geographic location.
- While, for example, people residing in luxury suburbs may all be a good target market for luxury cars, location is rarely the reason for differences in product preference.
- The typical case is best illustrated using tourism: people from the same country of origin are likely to have a wide range of different ideal holidays, depending on whether they are single or travel as a family, whether they are into sports or culture.

**Socio-Demographic Segmentation**

- In some instances, the socio-demographic criterion may also explain specific product preferences (having children, for example, is the actual reason that families choose a family vacation village where previously, as a couple, their vacation choice may have been entirely different).
- But in many instances, the socio-demographic criterion is not the cause for product preferences, thus not providing sufficient market insight for optimal segmentation decisions.
- Yankelovich and Meer (2006) argue that socio-demographics do not represent a solid basis to market segmentation, suggesting that values, tastes, and preferences are more valuable because they are more influential in consumers` buying decisions.

**Psychographic Segmentation**

- When people are grouped according to psychological criteria, such as their beliefs, interests, preferences, aspirations, or benefits sought when purchasing a product, psychographic segmentation is used.
- Benefit segmentation, which Haley (1968) is credited for, is arguably the most popular kind of psychographic segmentation.
- Lifestyle segmentation is another popular psychographic segmentation approach (Cahill 2006) based on people's activities, opinions, and interests.
- Consequently, most psychographic segmentation studies use several segmentation variables, for example, several different travel motives and a number of perceived risks when going on vacation.
- The disadvantage of the psychographic approach is the increased complexity of determining segment memberships for consumers.

**Behavioral Segmentation**

- Another approach to segment extraction is to search directly for similarities in behavior or reported behavior.

- As such, behavioral expenses of consumers as segmentation variables, and Heilman and Bowman (2002) use actual purchase data across product categories.
- But behavioral data is not always readily available, especially if the aim is to include potential customers who have not previously purchased the product in the segmentation analysis, rather than limiting oneself to the study of existing customers of the organization.

**Data from Survey Studies**

The majority of market segmentation studies rely on survey data. Survey data is inexpensive and straightforward to obtain, making it a viable option for any business. However, as opposed to data gathered from observation of actual activity, survey data can be tainted by a variety of biases. As a result of these biases, the quality of solutions obtained from market segmentation analysis might suffer. A few important factors to consider while using survey data are outlined below.

**Choice of Variables**

Carefully selecting the variables included in commonsense segmentation and data-driven segmentation is critical to the quality of the market segmentation solution. Including unnecessary variables can make questionnaires long and tedious for respondents, which, in turn, causes respondent fatigue. Fatigued respondents tend to respond to lower quality. Noisy variables can result from not carefully developing survey questions or not selecting segmentation variables. The recommendation is to ask all necessary and unique questions while resisting the temptation to include unnecessary or redundant items. Redundant items are particularly problematic in the context of market segmentation analysis.

**Response Options**

- Answer options provided to respondents in surveys determine the scale of the data available for subsequent analyses.
- Because many data analytic techniques are based on distance measures, not all survey response options are equally suitable for segmentation analysis.
- Options allowing respondents to answer in only one of two ways, generate binary or dichotomous data.
- The distance between 0 and 1 is clearly defined and, as such, poses no difficulties for subsequent segmentation analysis.
- Options allow respondents to select an answer from a range of unordered categories to correspond to nominal variables.

- Metric data allow any statistical procedure to be performed (including the measurement of distance), and are therefore well suited for segmentation analysis.
- The most commonly used response option in survey research, however, is a limited number of ordered answer options larger than two.
- Preferably, therefore, either metric or binary response options should be provided to respondents if those options are meaningful with respect to the question asked.
- Using binary or metric response options prevents subsequent complications relating to the distance measure in the process of data-driven segmentation analysis.
- Although ordinal scales dominate both market research and academic survey research, using binary or metric response options instead is usually not a compromise.
- The visual analog scale allows respondents to indicate a position along a continuous line between two end-points and leads to data that can be assumed to be metric.

**Response Styles**

- A response bias is a systematic tendency to respond to a range of questionnaire items on some basis other than the specific item content (i.e., what the items were designed to measure).
- If a bias is displayed by a respondent consistently over time, and independently of the survey questions asked to represent a response style.
- A wide range of response styles manifest in survey answers, including respondents` tendencies to use extreme answer options (STRONGLY AGREE, STRONGLY DISAGREE), to use the midpoint (NEITHER AGREE NOR DISAGREE), and to agree with all statements.
- For example, some respondents displaying an acquiescence bias (a tendency to agree with all questions) could result in one market segment having a much higher than average agreement with all answers.
- It is critical, therefore, to minimize the risk of capturing response styles when data is collected for the purpose of market segmentation.
- Alternatively, respondents affected by such a response style must be removed before choosing to target such a market segment.

**Sample Size**

- Viennese psychologist Formann (1984) recommends that the sample size should be at least 2p (better five times 2p), where p is the number of segmentation variables.
- Qiu and Joe (2015) developed a sample size recommendation for constructing artificial data sets for studying the performance of clustering algorithms.
- According to Qiu and Joe (2015), the sample size should – in the simple case of equal cluster sizes – be at least ten times the number of segmentation variables times the number of segments in the data (10 · p · k where p represents the number of segmentation variables and k represents the number of segments).
- If segments are unequally sized, the smallest segment should contain a sample of at least 10 · p.
- Knowing the true structure of the data sets, they tested the sample size requirement for algorithms to correctly identify the true segments.
- The adjusted Rand index serves as the measure of correctness of segment recovery.
- The adjusted Rand index assesses the congruence between two segmentation solutions.
- To assess segment recovery, the adjusted Rand index is calculated for the true segment solution and the extracted one.
- (2016) extended this line of research to account for key features of typical survey data sets, making it more difficult for segmentation algorithms to identify correct segmentation solutions.
- Specifically, they investigated the effect on sample size requirements resulting from market characteristics not under the control of the data analyst and, data characteristics – at least to some degree under the control of the data analyst.
- Market characteristics studied included: the number of market segments present in the data, whether those market segments are equal or unequal in size, and the extent to which market segments overlap.
- De Craen et al.(2006) show that the presence of unequally sized segments makes it more difficult for an algorithm to extract the correct market segments.

**Data from Internal Sources**

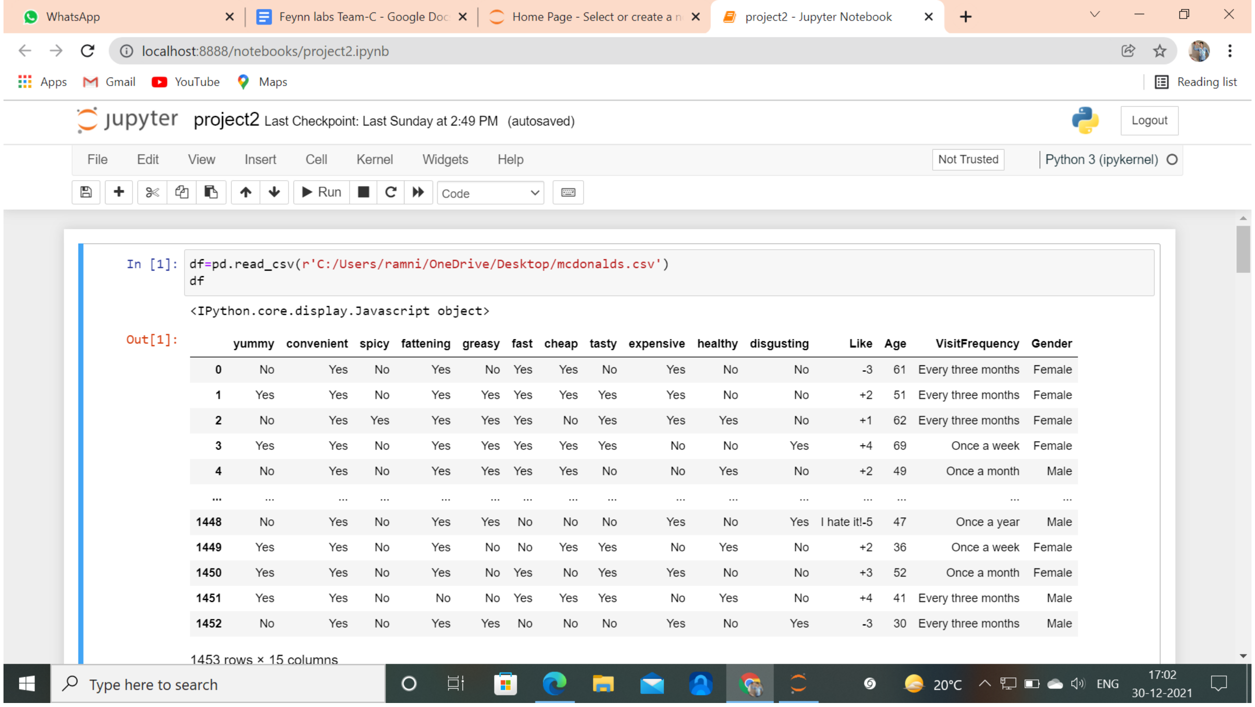
Organizations are increasingly having access to large volumes of internal data that may be mined for market segmentation studies. Scanning data available to grocery shops, booking data available through airline loyalty programs, and internet transaction data are just a few examples. The strength of such data comes from the fact that it represents actual consumer behavior rather than statements about consumer behavior or intentions, which are known to be influenced by imperfect memory (Niemi 1993) and a variety of response biases, such as social desirability bias or other response styles. Another advantage is that such data is typically created automatically, and no additional work is necessary to gather data provided organizations are capable of storing data in a way that makes it easy to retrieve. Internal data has the risk of being systematically skewed by over-representing current clients.

### Data from Experimental Studies

Experimental data is another alternative source of information for market segmentation studies. Experiments in the field or in the laboratory can yield data. They might, for example, be the product of studies on how individuals react to various commercials. The advertisement's response might then be utilized as a factor for segmentation. Choice experiments and conjoint analyses can also provide experimental data. The goal of such research is to provide customers with precisely crafted stimuli that include specified degrees of various product qualities. Consumers then choose which of the items – each of which is defined by a distinct set of attribute levels – they prefer. Conjoint research and choice experiments provide information on how each characteristic and attribute level influences decisions. This data can also be utilized as a criterion for segmentation.

## Step 4: Exploring Data

A First Glimpse at the Data-


WhatsApp
Feynn labs Team-C - Google Doc
Home Page - Select or create a n
project2 - Jupyter Notebook
localhost:8888/notebooks/project2.ipynb
Apps
Gmail
YouTube
Maps
Reading list
jupyter
project2 Last Checkpoint: Last Sunday at 2:49 PM (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3 (ipykernel)
Run
Code
Type here to search
20°C
ENG
17:02
30-12-2021

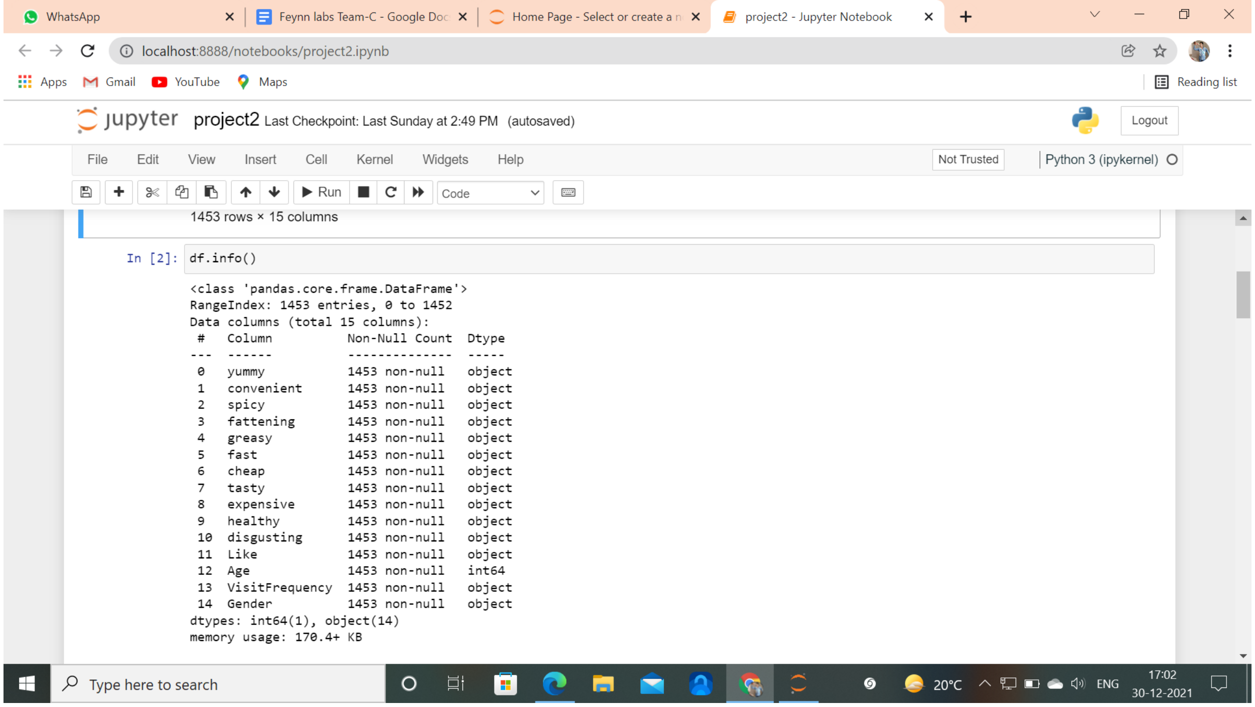
```
In [1]: df=pd.read_csv(r'C:/Users/ramni/OneDrive/Desktop/mcdonalds.csv')
        df

<IPython.core.display.Javascript object>
```

Out[1]:

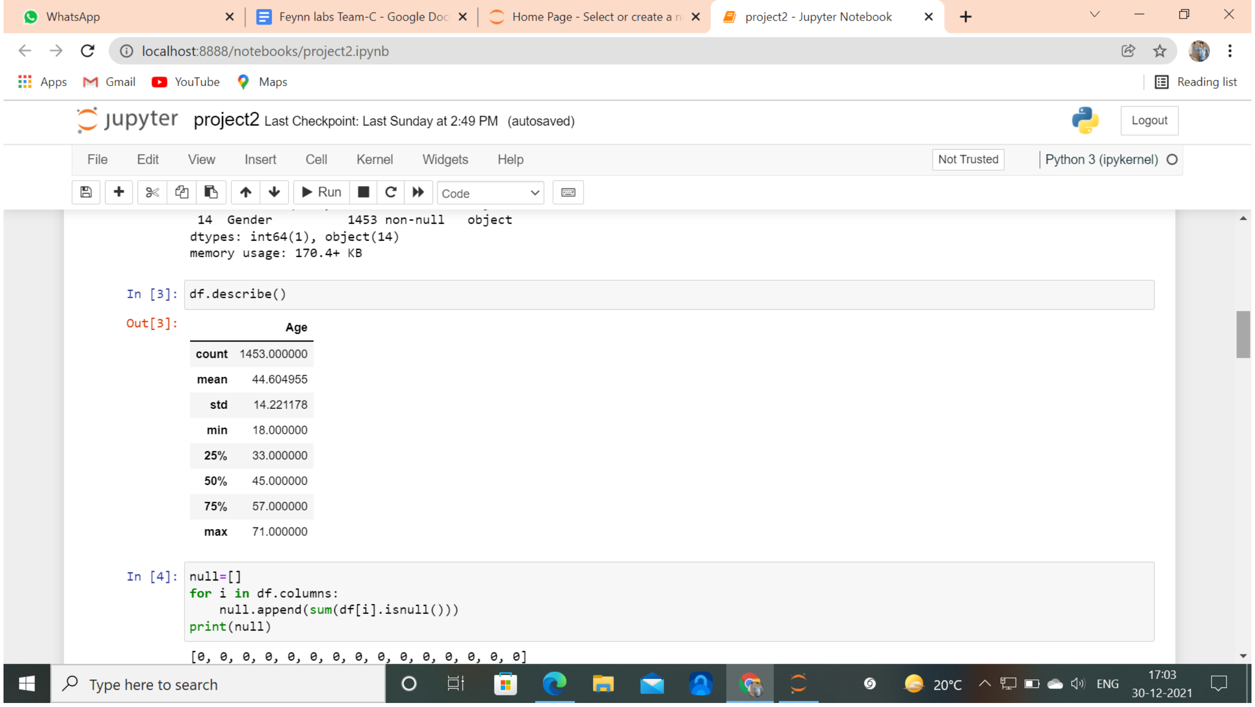
| | yummy | convenient | spicy | fattening | greasy | fast | cheap | tasty | expensive | healthy | disgusting | Like | Age | VisitFrequency | Gender |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | No | Yes | No | Yes | No | Yes | Yes | No | Yes | No | No | -3 | 61 | Every three months | Female |
| **1** | Yes | Yes | No | Yes | Yes | Yes | Yes | Yes | Yes | No | No | +2 | 51 | Every three months | Female |
| **2** | No | Yes | Yes | Yes | Yes | Yes | No | Yes | Yes | Yes | No | +1 | 62 | Every three months | Female |
| **3** | Yes | Yes | No | Yes | Yes | Yes | Yes | Yes | No | No | Yes | +4 | 69 | Once a week | Female |
| **4** | No | Yes | No | Yes | Yes | Yes | Yes | No | No | Yes | No | +2 | 49 | Once a month | Male |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **1448** | No | Yes | No | Yes | Yes | No | No | No | Yes | No | Yes | I hate it!-5 | 47 | Once a year | Male |
| **1449** | Yes | Yes | No | Yes | No | No | Yes | Yes | No | Yes | No | +2 | 36 | Once a week | Female |
| **1450** | Yes | Yes | No | Yes | No | Yes | No | Yes | Yes | No | No | +3 | 52 | Once a month | Female |
| **1451** | Yes | Yes | No | No | No | Yes | Yes | Yes | No | Yes | No | +4 | 41 | Every three months | Male |
| **1452** | No | Yes | No | Yes | Yes | No | No | No | Yes | No | Yes | -3 | 30 | Every three months | Male |

1453 rows × 15 columns

# Descriptive Analysis-


WhatsApp
Feynn labs Team-C - Google Doc
Home Page - Select or create a n
project2 - Jupyter Notebook
localhost:8888/notebooks/project2.ipynb
Apps
Gmail
YouTube
Maps
Reading list
jupyter project2 Last Checkpoint: Last Sunday at 2:49 PM (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
Type here to search
20°C
ENG
17:02
30-12-2021

```
1453 rows × 15 columns

In [2]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1453 entries, 0 to 1452
Data columns (total 15 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   yummy           1453 non-null   object
 1   convenient      1453 non-null   object
 2   spicy           1453 non-null   object
 3   fattening       1453 non-null   object
 4   greasy          1453 non-null   object
 5   fast            1453 non-null   object
 6   cheap           1453 non-null   object
 7   tasty           1453 non-null   object
 8   expensive       1453 non-null   object
 9   healthy         1453 non-null   object
 10  disgusting      1453 non-null   object
 11  Like            1453 non-null   object
 12  Age             1453 non-null   int64
 13  VisitFrequency  1453 non-null   object
 14  Gender          1453 non-null   object
dtypes: int64(1), object(14)
memory usage: 170.4+ KB
```


WhatsApp
Feynn labs Team-C - Google Doc
Home Page - Select or create a n
project2 - Jupyter Notebook
localhost:8888/notebooks/project2.ipynb
Apps
Gmail
YouTube
Maps
Reading list
jupyter project2 Last Checkpoint: Last Sunday at 2:49 PM (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
Type here to search
20°C
ENG
17:03
30-12-2021

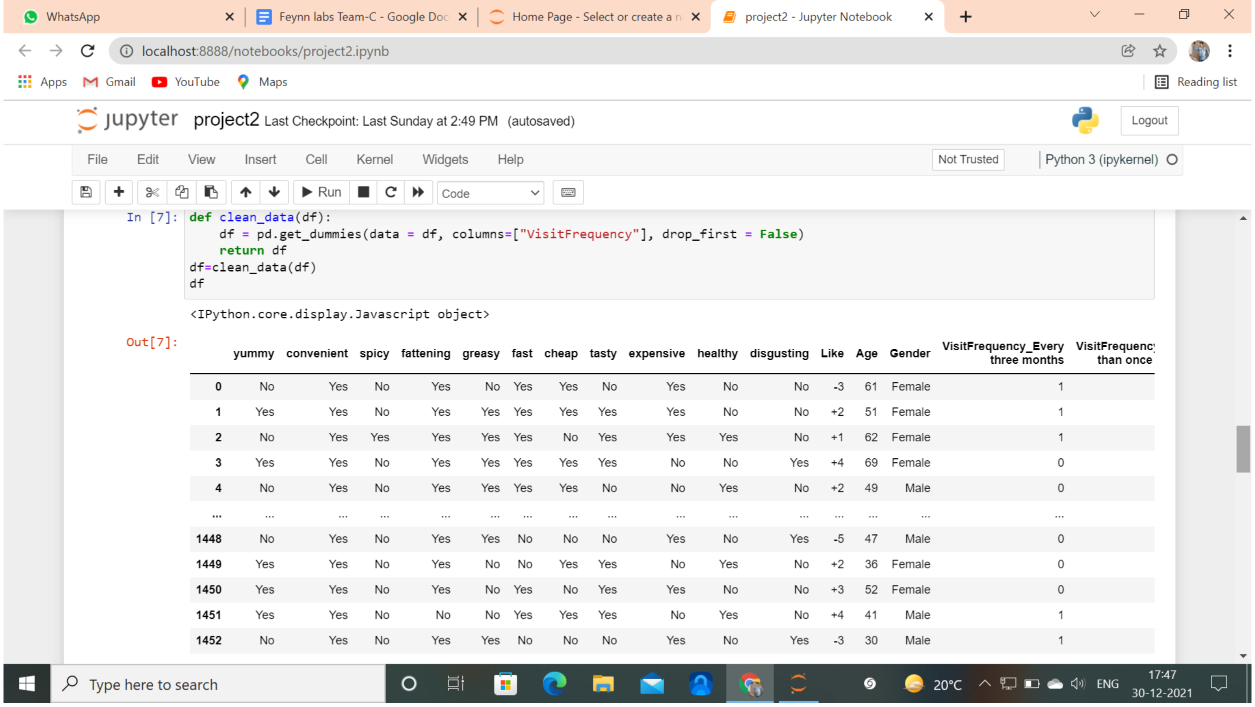
```
 14  Gender          1453 non-null   object
dtypes: int64(1), object(14)
memory usage: 170.4+ KB

In [3]: df.describe()
```

Out[3]:

|  | Age |
|---|---|
| count | 1453.000000 |
| mean | 44.604955 |
| std | 14.221178 |
| min | 18.000000 |
| 25% | 33.000000 |
| 50% | 45.000000 |
| 75% | 57.000000 |
| max | 71.000000 |

```
In [4]: null=[]
        for i in df.columns:
            null.append(sum(df[i].isnull()))
        print(null)

[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
```

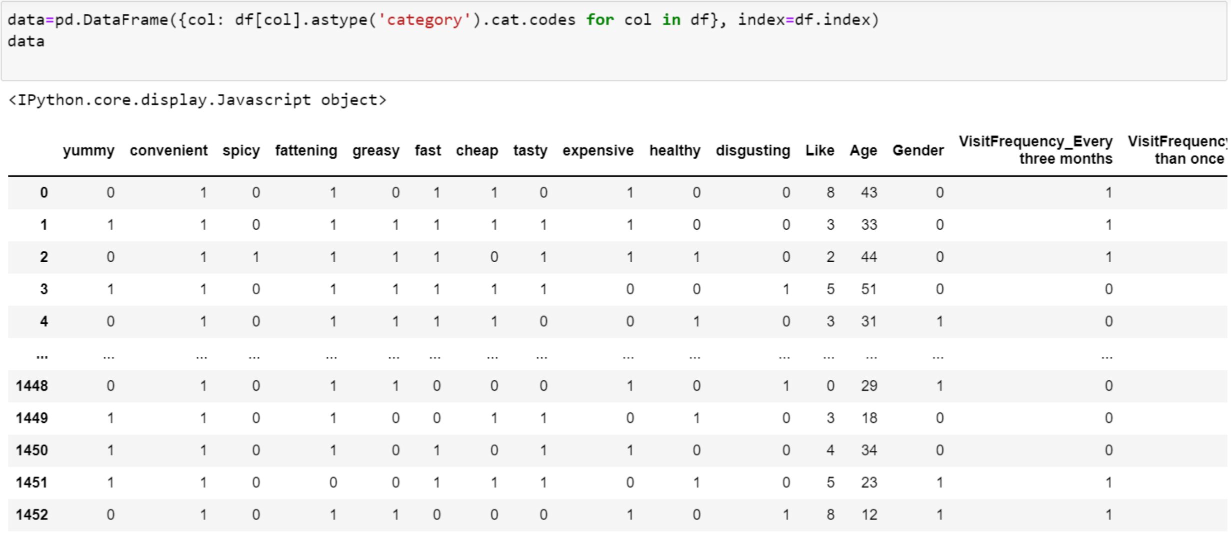
# Pre-processing(on Categorical and numerical variables):



WhatsApp
Feynn labs Team-C - Google Doc
Home Page - Select or create a n
project2 - Jupyter Notebook
localhost:8888/notebooks/project2.ipynb
Apps
Gmail
YouTube
Maps
Reading list
jupyter
project2 Last Checkpoint: Last Sunday at 2:49 PM (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3 (ipykernel)
Run
Code
Type here to search
20°C
ENG
17:47
30-12-2021

```
In [7]: def clean_data(df):
            df = pd.get_dummies(data = df, columns=["VisitFrequency"], drop_first = False)
            return df
        df=clean_data(df)
        df

        <IPython.core.display.Javascript object>

Out[7]:
```

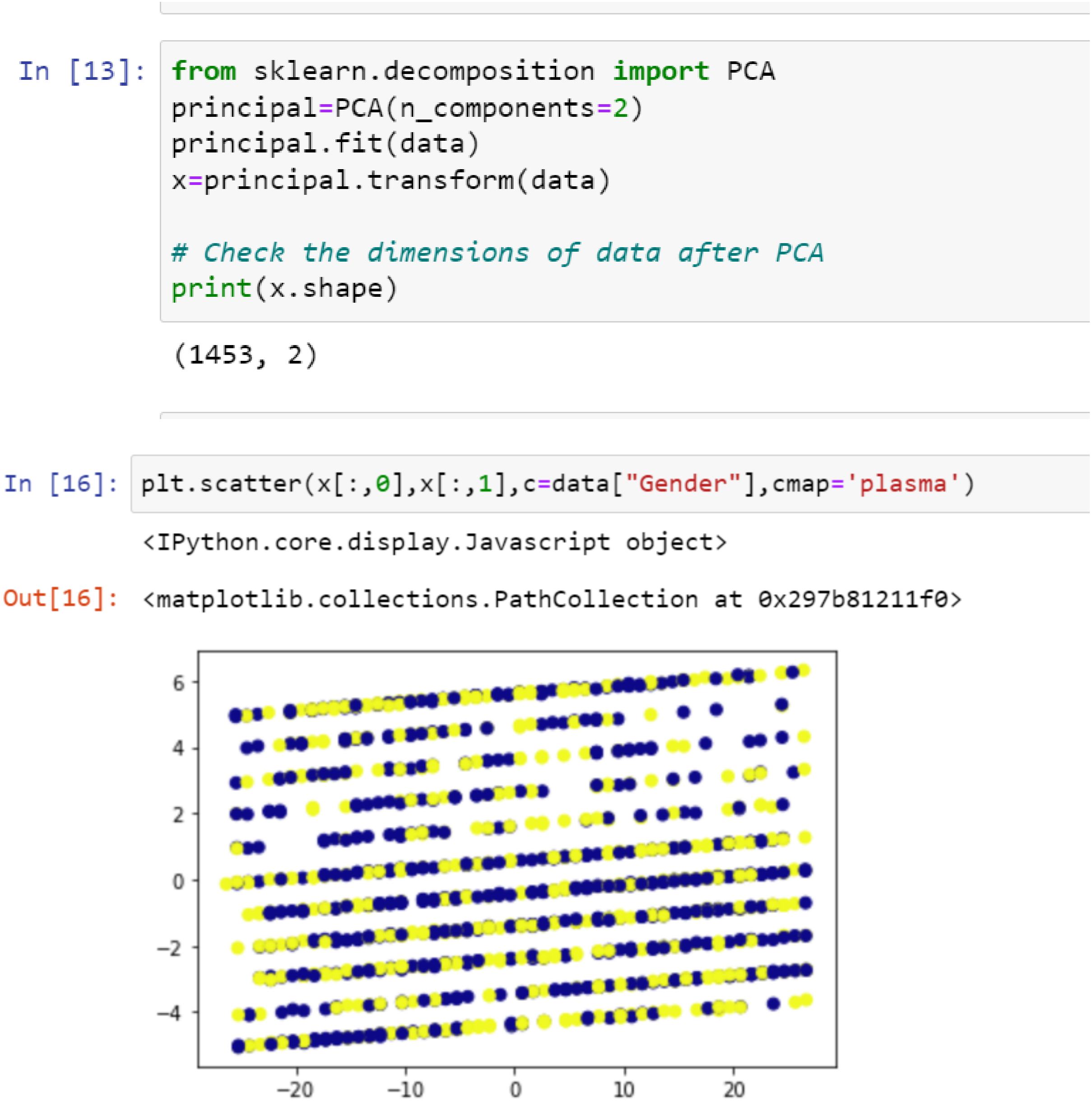
| | yummy | convenient | spicy | fattening | greasy | fast | cheap | tasty | expensive | healthy | disgusting | Like | Age | Gender | VisitFrequency_Every three months | VisitFrequenc than once |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | No | Yes | No | Yes | No | Yes | Yes | No | Yes | No | No | -3 | 61 | Female | 1 | |
| **1** | Yes | Yes | No | Yes | Yes | Yes | Yes | Yes | Yes | No | No | +2 | 51 | Female | 1 | |
| **2** | No | Yes | Yes | Yes | Yes | Yes | No | Yes | Yes | Yes | No | +1 | 62 | Female | 1 | |
| **3** | Yes | Yes | No | Yes | Yes | Yes | Yes | Yes | No | No | Yes | +4 | 69 | Female | 0 | |
| **4** | No | Yes | No | Yes | Yes | Yes | Yes | No | No | Yes | No | +2 | 49 | Male | 0 | |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **1448** | No | Yes | No | Yes | Yes | No | No | No | Yes | No | Yes | -5 | 47 | Male | 0 | |
| **1449** | Yes | Yes | No | Yes | No | No | Yes | Yes | No | Yes | No | +2 | 36 | Female | 0 | |
| **1450** | Yes | Yes | No | Yes | No | Yes | No | Yes | Yes | No | No | +3 | 52 | Female | 0 | |
| **1451** | Yes | Yes | No | No | No | Yes | Yes | Yes | No | Yes | No | +4 | 41 | Male | 1 | |
| **1452** | No | Yes | No | Yes | Yes | No | No | No | Yes | No | Yes | -3 | 30 | Male | 1 | |



```
data=pd.DataFrame({col: df[col].astype('category').cat.codes for col in df}, index=df.index)
data

<IPython.core.display.Javascript object>
```

| | yummy | convenient | spicy | fattening | greasy | fast | cheap | tasty | expensive | healthy | disgusting | Like | Age | Gender | VisitFrequency_Every three months | VisitFrequenc than once |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 8 | 43 | 0 | 1 | |
| **1** | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 3 | 33 | 0 | 1 | |
| **2** | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 2 | 44 | 0 | 1 | |
| **3** | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 5 | 51 | 0 | 0 | |
| **4** | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 3 | 31 | 1 | 0 | |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **1448** | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 29 | 1 | 0 | |
| **1449** | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 3 | 18 | 0 | 0 | |
| **1450** | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 4 | 34 | 0 | 0 | |
| **1451** | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 5 | 23 | 1 | 1 | |
| **1452** | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 8 | 12 | 1 | 1 | |

Principal Components Analysis-



```
In [13]: from sklearn.decomposition import PCA
         principal=PCA(n_components=2)
         principal.fit(data)
         x=principal.transform(data)

         # Check the dimensions of data after PCA
         print(x.shape)

         (1453, 2)
```

```
In [16]: plt.scatter(x[:,0],x[:,1],c=data["Gender"],cmap='plasma')

         <IPython.core.display.Javascript object>

Out[16]: <matplotlib.collections.PathCollection at 0x297b81211f0>
```

## Step 5: Extracting Segments

Two extraction methods that are commonly used in the market are-

(a) Distance-based method

(b) Model-based method

**Distance-based methods**

The most common distance measuring technique is-

(a) Euclidean distance

(b) Manhattan or absolute distance

(c) Asymmetric binary distance (applies only to binary vectors)

**1. Hierarchical methods**-

Hierarchical clustering methods are the most intuitive way of grouping data because they mimic how a human would approach the task of dividing a set of n observations (consumers) into k groups (segments). There are two extreme scenarios in hierarchical methods-

- **Divisive hierarchical clustering methods** start with the complete data set X and splits it into two market segments in a first step. Then, each of the segments is again split into two segments. This process continues until each consumer has their own market segment.

- **Agglomerative hierarchical clustering** approaches the task from the other end. The starting point is each consumer representing their own market segment (n singleton clusters). Step-by-step, the two market segments closest to one another are merged until the complete data set forms one large market segment.

**Dendrogram**

The result of hierarchical clustering is typically presented as a dendrogram. A dendrogram is a tree diagram. The root of the tree represents the one-cluster solution where one market segment contains all consumers. The leaves of the tree are the single observations (consumers), and branches in-between correspond to the hierarchy of market segments formed at each step of the procedure. The height of the branches corresponds to the distance between the clusters. Higher branches point to more distinct market segments.

**2. Partitioning methods-**

**(a) k-means & k-centroid clustering**

The algorithm:

1. K points(segments) are randomly placed into the object data space representing the initial group of centroids.
2. Each object or data point is assigned to the closest centroid.
3. After all, objects are assigned, the positions of the k centroids are recalculated by holding cluster membership fixed and minimizing the distance from each data to the corresponding cluster centroid.
4. Steps 2 and 3 are repeated until the positions of the centroids no longer move.

**Problem with k-means clustering method.**

Using randomly drawn data is suboptimal because it may result in some of those randomly drawn data being located very close to one another and thus not being representative of the data space. Using starting points that are not representative of the data space increases the likelihood of the k-means algorithm getting stuck in what is referred to as a local optimum.

**(b) Improved k-means**

Many attempts have been made to refine and improve the k-means clustering algorithm. The simplest improvement is to initialize k-means using “smart” starting values, rather than randomly drawing k consumers from the data set and using them as starting points. The best approach is to randomly draw many starting points and select the best set. The best starting points are those that best represent the data. Good representatives are close to their segment members; the total distance of all segment members to their representatives is small. Bad representatives are far away from their segment members; the total distance of all segment members to their representatives is high.

**(c) Hard Competitive learning**

Hard competitive learning like k-means also minimizes the sum of the distance from each data contained in the dataset to their closest representative (centroid), but the process by which this is achieved is slightly different. Hard competitive learning randomly picks one consumer and moves this consumer’s closest segment representative a small step into the direction of the randomly chosen consumer. As a consequence of this procedural difference, different segmentation solutions can emerge, even if the same starting points are used to initialize the algorithm.

### (d) Self-organising maps

The self-organizing map algorithm is similar to hard competitive learning: a single random consumer is selected from the data set, and the closest representative for this random consumer moves a small step in their direction. In addition, representatives who are direct grid neighbors of the closest representative move in the direction of the selected random consumer. The process is repeated many times; each consumer in the data set is randomly chosen multiple times and used to adjust the location of the centroids in the Kohonen map. What changes over the many repetitions, however, is the extent to which the representatives are allowed to change. The adjustments get smaller and smaller until a final solution is reached. The advantage of self-organizing maps over other clustering algorithms is that the numbering of market segments is not random.

## 3. Hybrid Approaches-

The basic idea behind hybrid segmentation approaches is to first run a partitioning algorithm because it can handle data sets of any size. But the partitioning algorithm used initially does not generate the number of segments sought. Rather, a much larger number of segments is extracted. Then, the original data is discarded, and only the centers of the resulting segments (centroids, representatives of each market segment) and segment sizes are retained and used as input for the hierarchical cluster analysis. At this point, the data set is small enough for hierarchical algorithms, and the dendrogram can inform the decision on how many segments to extract.

### (a) Two-step clustering

The two steps consist of running a partitioning procedure followed by a hierarchical procedure. The procedure has been used in a wide variety of application areas, including internet access types of mobile phone users, segmenting potential nature-based tourists based on temporal factors, identifying and characterizing potential electric vehicle adopters, and segmenting travel-related risks.

### (b) Bagged clustering

Bagged clustering, also combines hierarchical clustering algorithms and partitioning clustering algorithms, but adds bootstrapping. That means that the process of extracting segments is repeated many times with randomly drawn (bootstrapped) samples of the data. Bootstrapping has the advantage of making the final segmentation solution less dependent on the exact people contained in consumer data.

**Model-based methods-**

**1. Finite Mixtures of Distributions-**

**(a) Normal Distribution**

A mixture of normal distributions can be used for market segmentation when the segmentation variables are metric, for example, money spent on different consumption categories, time spent engaging in different vacation activities, or body measurements for the segments of different clothing sizes.

**Uncertainty plot -** The uncertainty plot illustrates the ambiguity of segment assignment. A data which cannot be clearly assigned to one of the market segments is considered uncertain**.**

**(b) Binary Distributions**

For binary data, finite mixtures of binary distributions, sometimes also referred to as latent class models or latent class analysis are popular. In this case, the p segmentation variables in the vector y are not metric, but binary (meaning that all p elements of y are either 0 or 1). Example- The elements of y, the segmentation variables, could be vacation activities where a value of 1 indicates that a tourist undertakes this activity, and a value of 0 indicates that they do not.

**2. Finite Mixtures of Regressions-**

Finite mixtures of distributions are similar to distance-based clustering methods and – in many cases – result in similar solutions. Compared to hierarchical or partitioning clustering methods, mixture models sometimes produce more useful, and sometimes less useful solutions.

**3. Algorithms with Integrated Variable Selection**

When the segmentation variables are binary, and redundant or noisy variables cannot be identified and removed during data pre-processing in Step 4, suitable segmentation variables need to be identified during segment extraction. A number of algorithms extract segments while – simultaneously – selecting suitable segmentation variables. Below are two such algorithms for binary segmentation variables:

**(a) Biclustering Algorithms**

Biclustering simultaneously clusters both consumers and variables. Biclustering algorithms exist for any kind of data, including metric and binary.

The biclustering algorithm which extracts these biclusters follows a sequence of steps. The starting point is a data matrix where each row represents one consumer and each column represents a binary segmentation variable:

Step 1- First, rearrange rows (consumers) and columns (segmentation variables) of the data matrix in a way to create a rectangle with identical entries of 1s at the top left of the data matrix. The aim is for this rectangle to be as large as possible.

Step 2 Second, assign the observations (consumers) falling into this rectangle to one bicluster.

Step 3 Remove from the data matrix the rows containing the consumers who have been assigned to the first bicluster. Once removed, repeat the procedure from step 1 until no more biclusters of sufficient size can be located.

**Advantages of Biclustering:**

No data transformation: Typically, situations, where the number of variables is too high, are addressed by pre-processing data. Pre-processing approaches such as principal components analysis reduce the number of segmentation variables by transforming the data. Any data transformation changes the information in the segmentation variables, thus risking that segmentation results are biased because they are not based on the original data. Biclustering does not transform data. Instead, original variables which do not display any systematic patterns relevant for grouping consumers are ignored.

Ability to capture niche markets: Because biclustering searches for identical patterns displayed by groups of consumers with respect to groups of variables, it is well suited for identifying niche markets. Biclustering methods, however, do not group all consumers. Rather, they select groups of similar consumers and leave ungrouped consumers who do not fit into any of the groups.

**(b) Variable Selection Procedure for Clustering Binary Data (VSBD)**

VSBD method is based on the k-means algorithm as clustering method the method assumes the presence of masking variables that needs to be identified and removed from the set of segmentation variables. Removing irrelevant variables helps to identify the correct segment structure, and eases interpretation.

Using the variable selection procedure generates a solution that is easy to interpret because only a small set of variables serve as segmentation variables, but each of them differentiates well between segments.

### (c) Variable Reduction: Factor-Cluster Analysis

The term factor-cluster analysis refers to a two-step procedure of data-driven market segmentation analysis. In the first step, segmentation variables are factor analysed. The raw data, the original segmentation variables, are then discarded. In the second step, the factor scores resulting from the factor analysis are used to extract market segments.

## Data Structure Analysis

Ideally, validation means calculating different segmentation solutions, choosing different segments, targeting them, and then comparing which leads to the most profit, or most success in mission achievement. This is clearly not possible in reality because one organization cannot run multiple segmentation strategies simultaneously just for the sake of determining which performs best. As a consequence, the term validation in the context of market segmentation is typically used in the sense of assessing reliability or stability of solutions across repeated after slightly modifying the data, or the algorithm. This approach is fundamentally different from validation using an external validation criterion which is referred to as stability-based data structure analysis.

### 1. Cluster Indices

Cluster indices represent the most common approach to making some of the most critical decisions, such as selecting the number of market segments to extract. Generally, two groups of cluster indices are distinguished: internal cluster indices and external cluster indices.

### (a) Internal Cluster Indices

Internal cluster indices use a single segmentation solution as a starting point. Solutions could result from hierarchical, partitioning, or model-based clustering methods. Internal clusters consider their combination as how compact is each of the market segments and how well separated are different market segments. Internal cluster indices fail to provide the best number of segments to extract.

**(b) External Cluster Indices**

External cluster indices evaluate a market segmentation solution using additional external information; they cannot be calculated using only the information contained in one market segmentation solution. Selecting any two consumers, the following four situations can occur when comparing two market segmentation solutions P1 and P2:

- A: Both consumers are assigned to the same segment twice.
- B: The two consumers are in the same segment in P1, but not in P2.
- C: The two consumers are in the same segment inP2, but not inP1.
- D: The two consumers are assigned to different market segments twice.

**2. Gorge Plots**

Gorge plots are used to visualize the similarity values. If natural, well-separated market segments are present in the data, we expect the gorge plot to contain many very low and many very high values. This is why this plot is referred to as the gorge plot. Optimally, it takes the shape of a gorge with a peak to the left and a peak to the right. For a real market segmentation analysis, gorge plots have to be generated and inspected for every number of segments. Producing and inspecting a large number of gorge plots is a tedious process, and has the disadvantage of not accounting for randomness in the sample used. These disadvantages are overcome by stability analysis, which can be conducted at the global or segment level.

**3. Global Stability Analysis**

To assess the global stability of any given segmentation solution, several new data sets are generated using resampling methods, and a number of segmentation solutions are extracted. Then the stability of the segmentation solutions across repeated calculations is compared. The solution which can best be replicated is chosen.

**4. Segment Level Stability Analysis**

Relying on global stability analysis could lead to selecting a segmentation solution with suitable global stability, but without a single highly stable segment. So, it is best to assess the segment level stability of market segments contained in those solutions to protect against discarding solutions containing interesting individual segments from being prematurely discarded.

### (a) Segment Level Stability Within Solutions (SLSW)

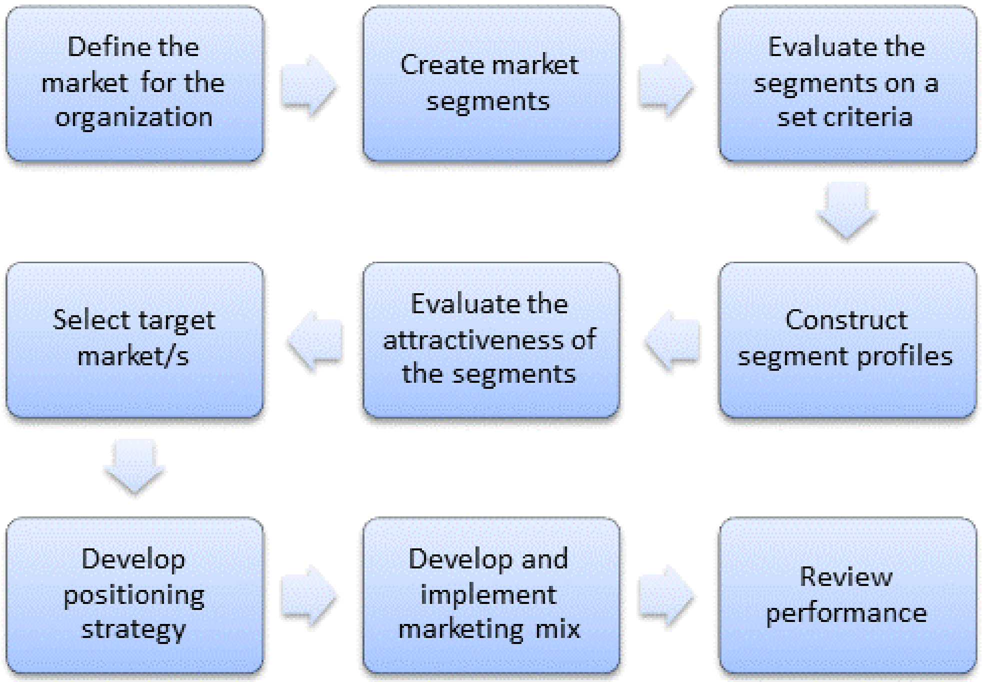
The criterion of segment level stability within solutions (SLSW) is similar to the concept of global stability. The difference is that stability is computed at the segment level, allowing the detection of one highly stable segment (for example a potentially attractive niche market) in a segmentation solution where several or even all other segments are unstable. Segment level stability within solutions (SLSW) measures how often a market segment with the same characteristics is identified across a number of repeated calculations of segmentation solutions with the same number of segments. It is calculated by drawing several bootstrap samples, calculating segmentation solutions independently for each of those bootstrap samples, and then determining the maximum agreement across all repeated calculations using the method which was proposed by Hennig.

### (b) Segment Level Stability Across Solutions (SLSA)

The purpose of this criterion is to determine the occurrence of a market segment across market segmentation solutions containing different numbers of segments. High values of segment level stability across solutions (SLSA) serve as indicators of market segments occurring naturally in the data, rather than being artificially created. Natural segments are more attractive to organizations because they actually exist, and no managerial judgment is needed in the artificial construction of segments. Segment level stability across solutions (SLSA), can be calculated in combination with an algorithm which extracts segments. However, for hierarchical clustering, segment level stability across solutions will reflect the fact that a sequence of nested partitions is created. If partitioning methods (k-means, k-medians, neural gas, ...) or finite mixture models are used, segmentation solutions are determined separately for each number of segments k.

## Step 7: Describing Segments:

- Developing a Complete Picture of Market Segments



Define the market for the organization

Create market segments

Evaluate the segments on a set criteria

Construct segment profiles

Evaluate the attractiveness of the segments

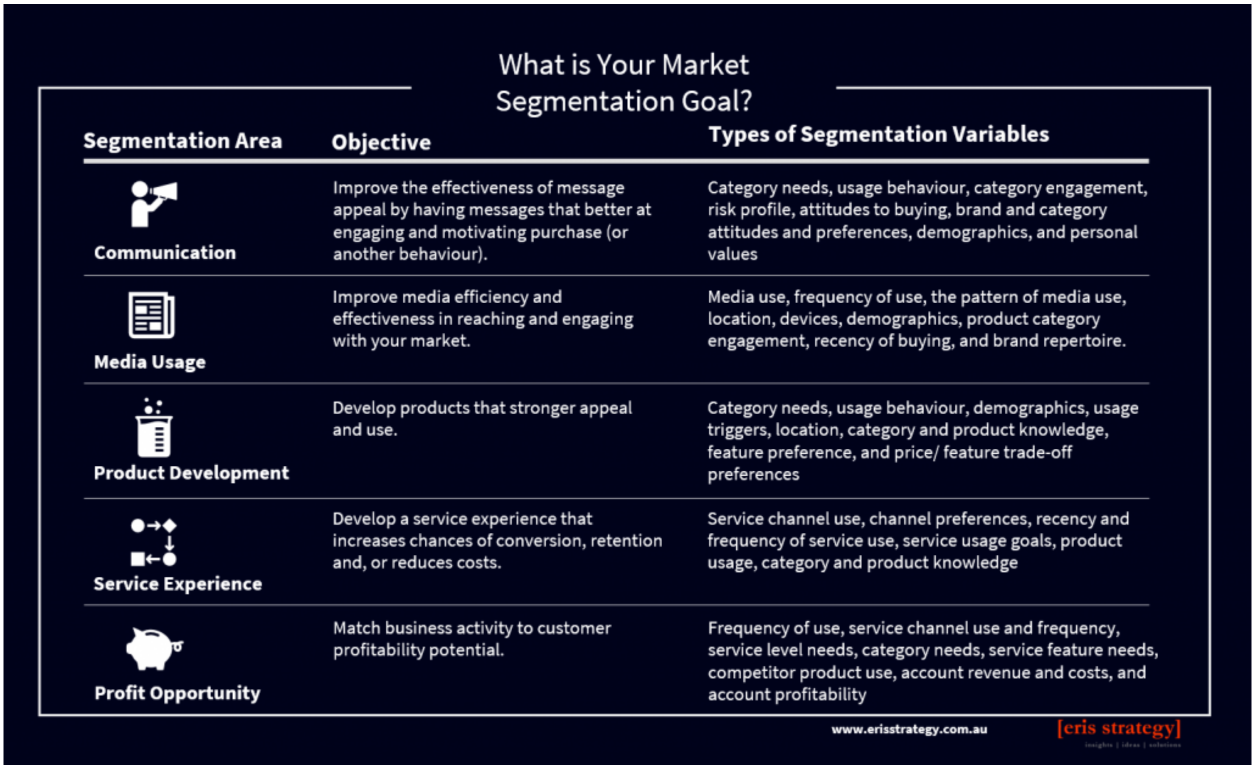
Select target market/s

Develop positioning strategy

Develop and implement marketing mix

Review performance

The full market segmentation process *Source: www.marketingstudyguide.com*

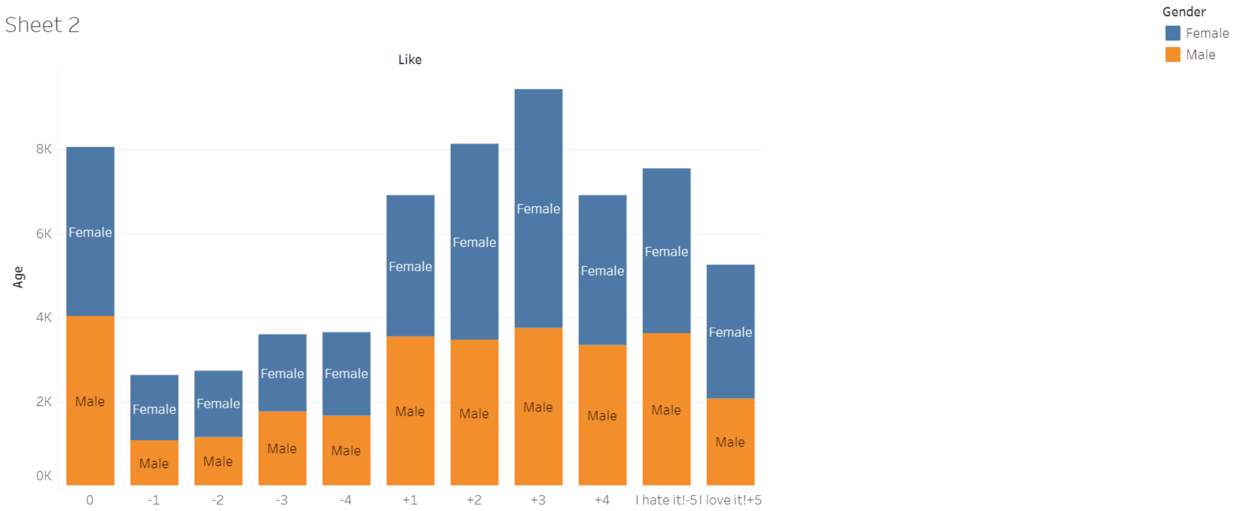


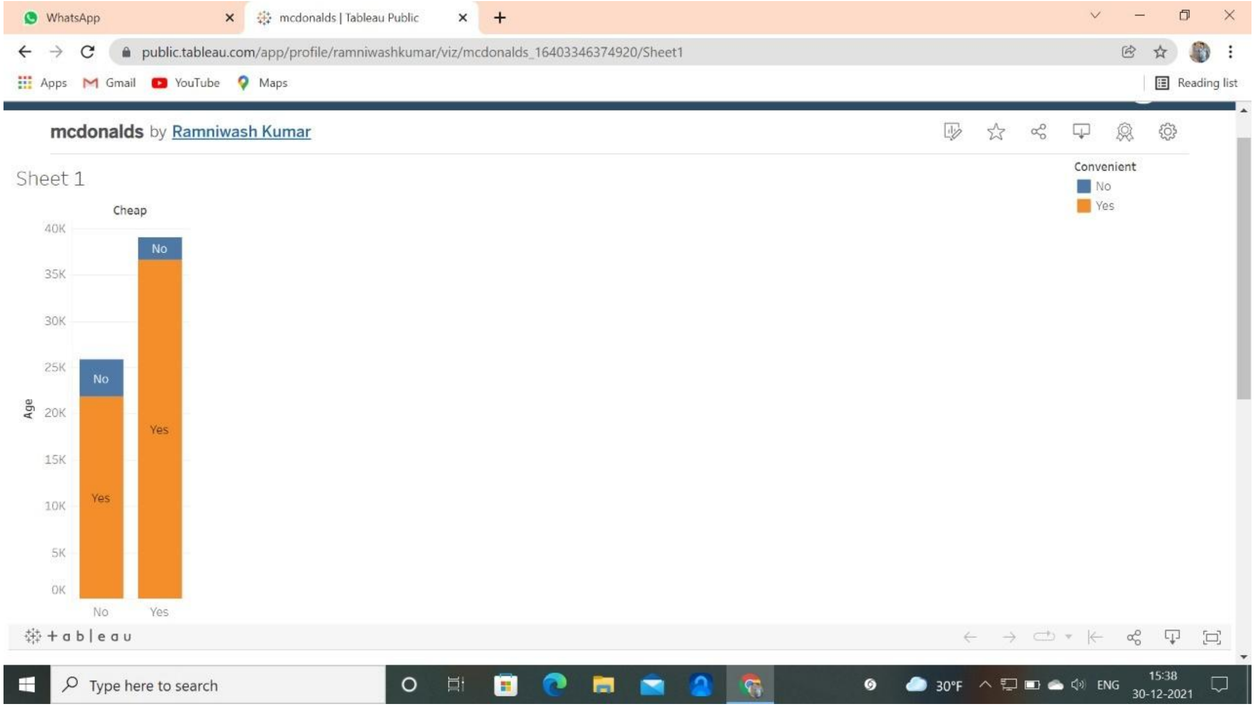
What is Your Market Segmentation Goal?

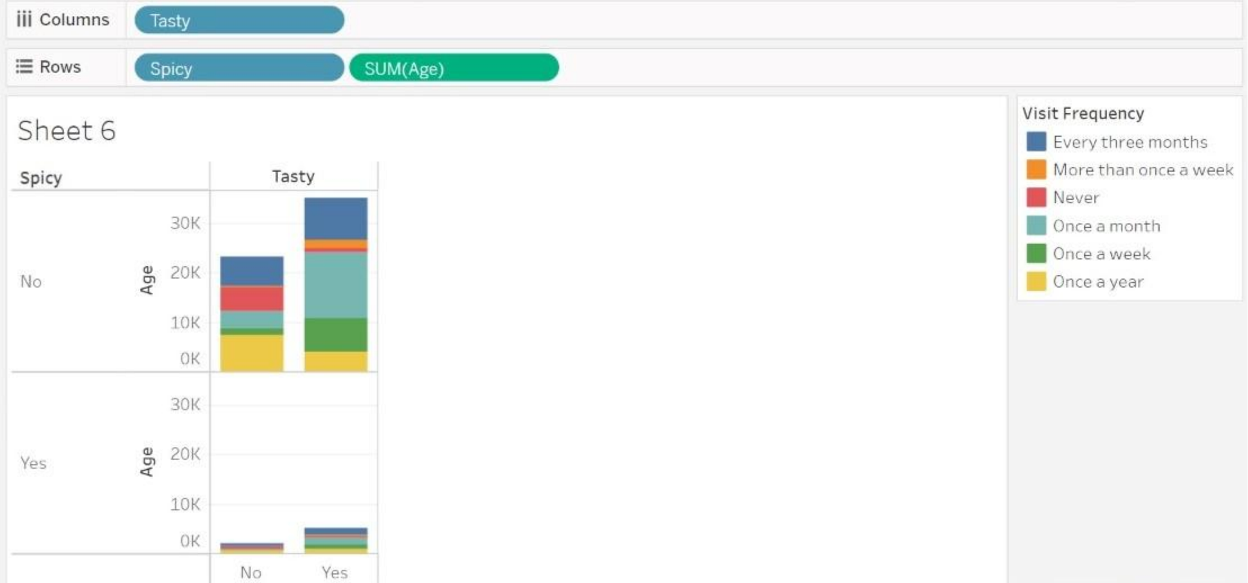
| Segmentation Area | Objective | Types of Segmentation Variables |
|---|---|---|
| Communication | Improve the effectiveness of message appeal by having messages that better at engaging and motivating purchase (or another behaviour). | Category needs, usage behaviour, category engagement, risk profile, attitudes to buying, brand and category attitudes and preferences, demographics, and personal values |
| Media Usage | Improve media efficiency and effectiveness in reaching and engaging with your market. | Media use, frequency of use, the pattern of media use, location, devices, demographics, product category engagement, recency of buying, and brand repertoire. |
| Product Development | Develop products that stronger appeal and use. | Category needs, usage behaviour, demographics, usage triggers, location, category and product knowledge, feature preference, and price/ feature trade-off preferences |
| Service Experience | Develop a service experience that increases chances of conversion, retention and, or reduces costs. | Service channel use, channel preferences, recency and frequency of service use, service usage goals, product usage, category and product knowledge |
| Profit Opportunity | Match business activity to customer profitability potential. | Frequency of use, service channel use and frequency, service level needs, category needs, service feature needs, competitor product use, account revenue and costs, and account profitability |

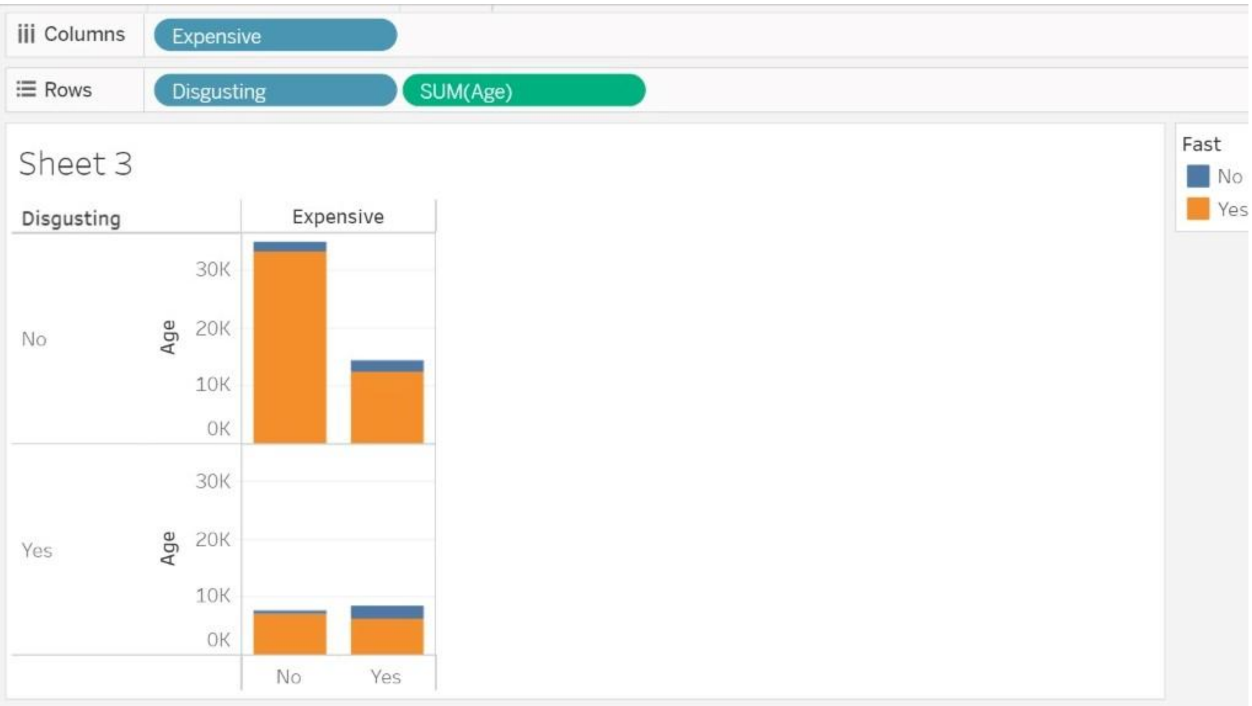
www.erisstrategy.com.au [eris strategy]

- Using Visualisations to Describe Market Segments









- Nominal Variables
  Gender is an example of a **nominal measurement** in which a number (e.g., 1) is used to label one gender, such as males, and a different number (e.g., 2) is used for the other gender, females. Numbers do not mean that one gender is better or worse than the other; they simply are used to classify persons.
  Similarly, an example of nominal in our data is VisitFrequency.

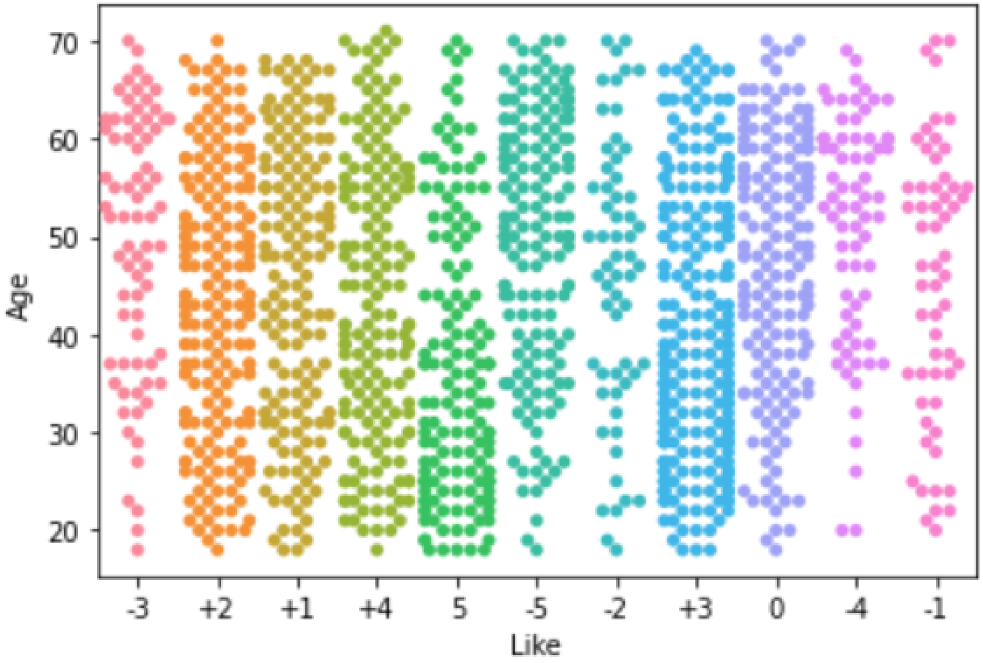
- Ordinal Descriptor Variables

  Ordinal data is **a statistical type of quantitative data in which variables exist in naturally occurring ordered categories.**

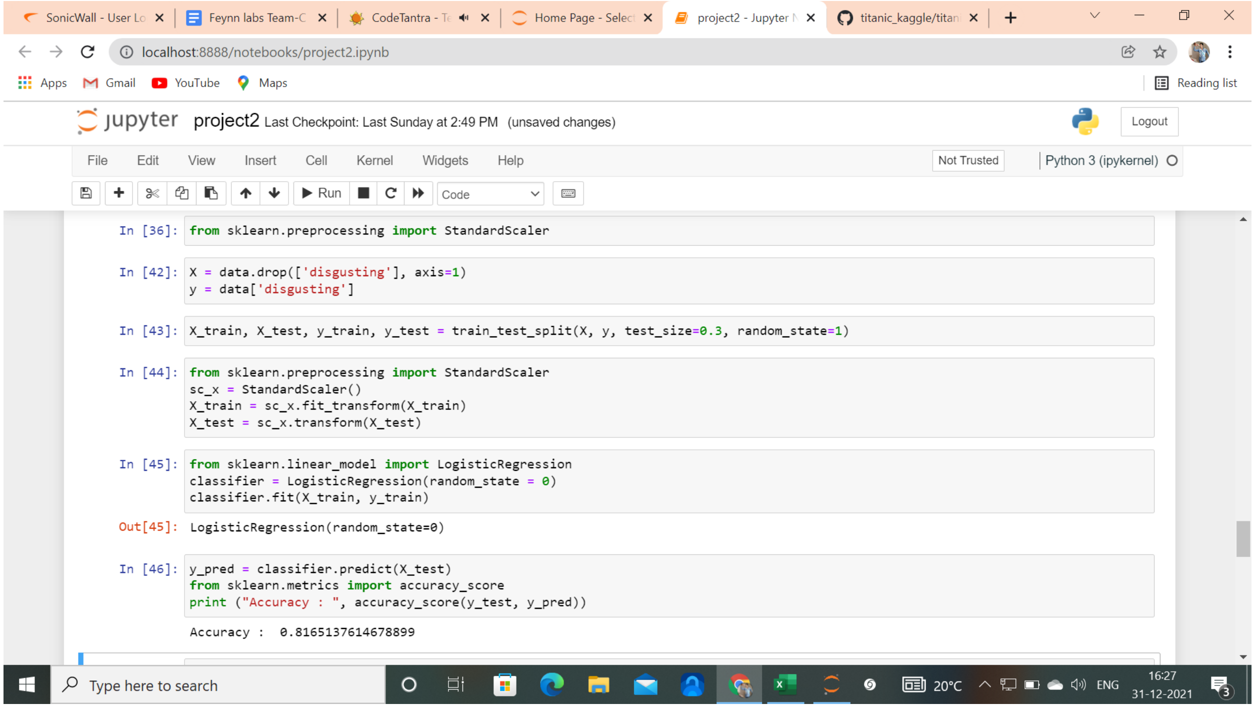
  Examples of ordinals in our data are yummy, convenient, spicy, Like, fattening, greasy, fast, cheap, tasty, expensive, healthy, disgusting.

- Metric Descriptor Variables-

  Simple statistical tests can be used to formally test for differences in descriptor variables across market segments. The simplest way to test for differences is to run a series of independent tests for each variable of interest. The outcome of the segment extraction step is segment membership, the assignment of each consumer to one market segment. Segment membership can be treated like any other nominal variable. It represents a nominal summary statistic of the segmentation variables. Therefore, any test for association between a nominal variable and another variable is suitable. The association between the nominal segment membership variable and another nominal or ordinal variable (such as Like)is visualized

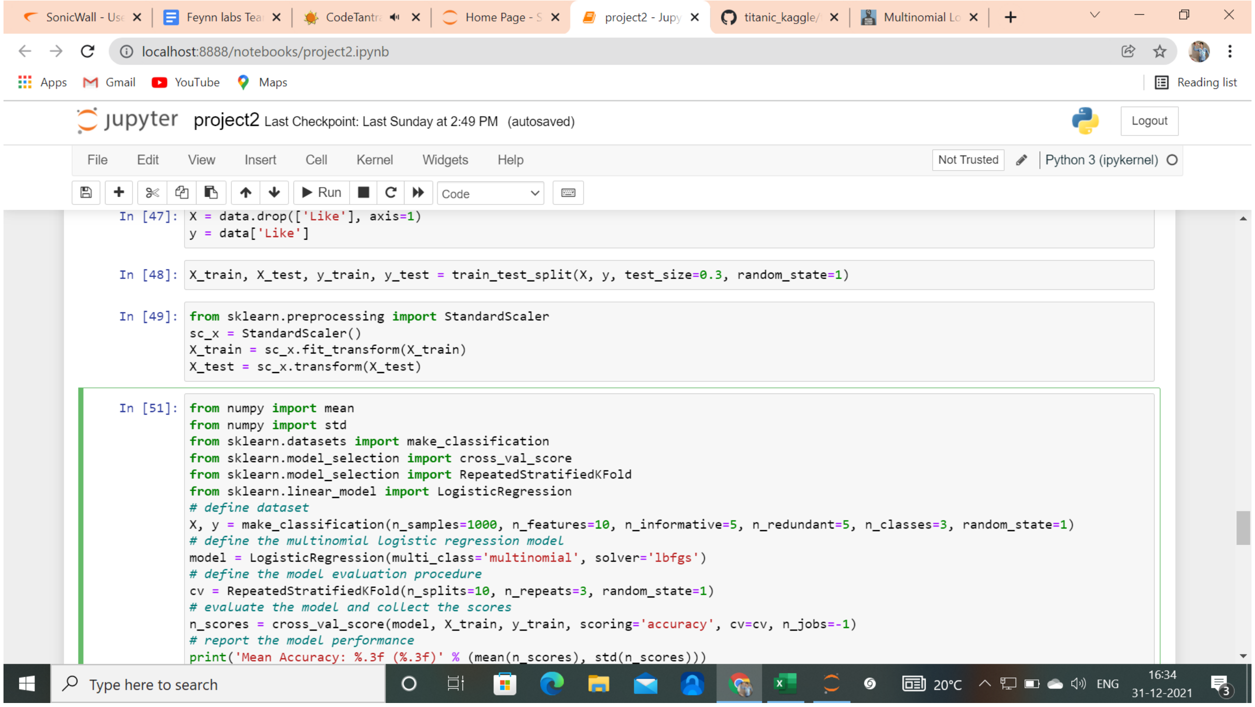


- Binary Logistic Regression


localhost:8888/notebooks/project2.ipynb
jupyter project2 Last Checkpoint: Last Sunday at 2:49 PM (unsaved changes)
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
Type here to search
20°C
ENG
16:27
31-12-2021

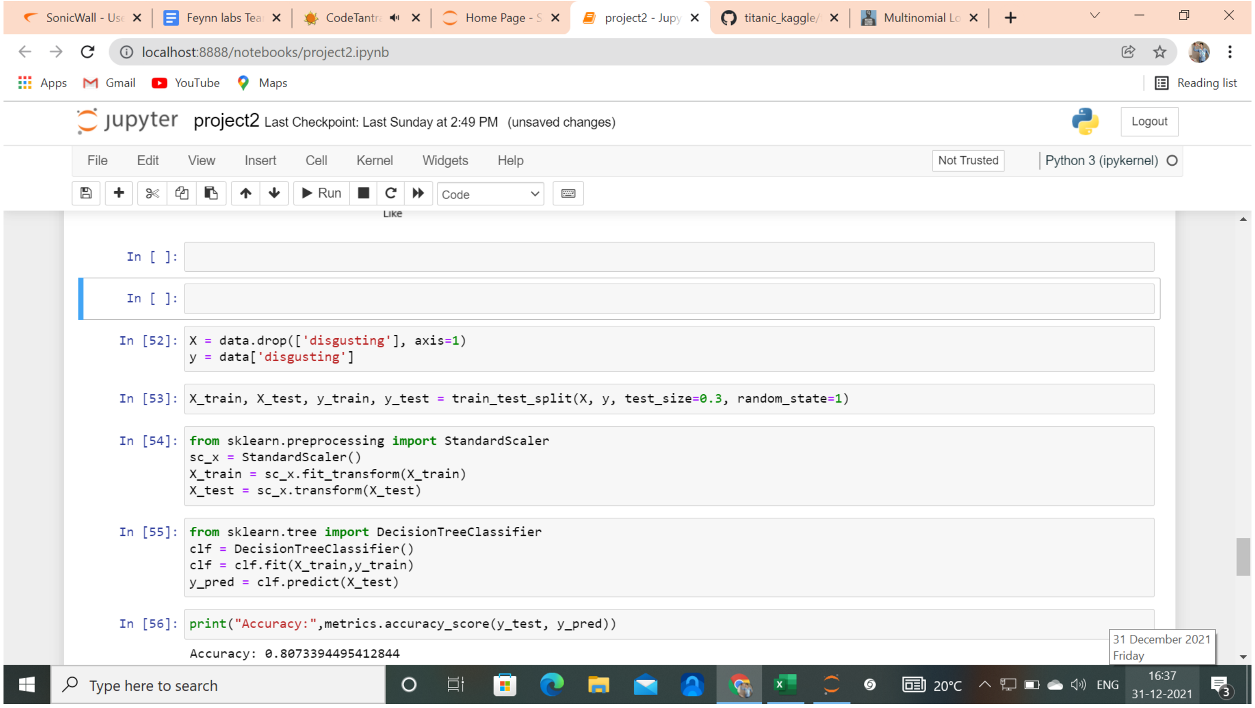
```
In [36]: from sklearn.preprocessing import StandardScaler

In [42]: X = data.drop(['disgusting'], axis=1)
         y = data['disgusting']

In [43]: X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=1)

In [44]: from sklearn.preprocessing import StandardScaler
         sc_x = StandardScaler()
         X_train = sc_x.fit_transform(X_train)
         X_test = sc_x.transform(X_test)

In [45]: from sklearn.linear_model import LogisticRegression
         classifier = LogisticRegression(random_state = 0)
         classifier.fit(X_train, y_train)

Out[45]: LogisticRegression(random_state=0)

In [46]: y_pred = classifier.predict(X_test)
         from sklearn.metrics import accuracy_score
         print ("Accuracy : ", accuracy_score(y_test, y_pred))

         Accuracy :  0.8165137614678899
```

- Multinomial Logistic Regression


localhost:8888/notebooks/project2.ipynb
jupyter project2 Last Checkpoint: Last Sunday at 2:49 PM (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
Type here to search
20°C
ENG
16:34
31-12-2021

```
In [47]: X = data.drop(['Like'], axis=1)
         y = data['Like']

In [48]: X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=1)

In [49]: from sklearn.preprocessing import StandardScaler
         sc_x = StandardScaler()
         X_train = sc_x.fit_transform(X_train)
         X_test = sc_x.transform(X_test)

In [51]: from numpy import mean
         from numpy import std
         from sklearn.datasets import make_classification
         from sklearn.model_selection import cross_val_score
         from sklearn.model_selection import RepeatedStratifiedKFold
         from sklearn.linear_model import LogisticRegression
         # define dataset
         X, y = make_classification(n_samples=1000, n_features=10, n_informative=5, n_redundant=5, n_classes=3, random_state=1)
         # define the multinomial logistic regression model
         model = LogisticRegression(multi_class='multinomial', solver='lbfgs')
         # define the model evaluation procedure
         cv = RepeatedStratifiedKFold(n_splits=10, n_repeats=3, random_state=1)
         # evaluate the model and collect the scores
         n_scores = cross_val_score(model, X_train, y_train, scoring='accuracy', cv=cv, n_jobs=-1)
         # report the model performance
         print('Mean Accuracy: %.3f (%.3f)' % (mean(n_scores), std(n_scores)))
```

- Tree-Based Methods



```
In [ ]:

In [ ]:

In [52]: X = data.drop(['disgusting'], axis=1)
         y = data['disgusting']

In [53]: X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=1)

In [54]: from sklearn.preprocessing import StandardScaler
         sc_x = StandardScaler()
         X_train = sc_x.fit_transform(X_train)
         X_test = sc_x.transform(X_test)

In [55]: from sklearn.tree import DecisionTreeClassifier
         clf = DecisionTreeClassifier()
         clf = clf.fit(X_train,y_train)
         y_pred = clf.predict(X_test)

In [56]: print("Accuracy:",metrics.accuracy_score(y_test, y_pred))

         Accuracy: 0.8073394495412844
```

## Step 8: Selecting the Target Segment(s)

### Targeting Decision

The first task here is to ensure that all the market segments that are still under consideration to be selected as target markets have well and truly passed the knock-out criteria test. Once this is done, the attractiveness of the remaining segments and the relative organisational competitiveness for these segments needs to be evaluated. In other words, the segmentation team has to ask a number of questions which fall into two broad categories:

1. Which of the market segments would the organisation most like to target? Which segment would the organisation like to commit to?

2. Which of the organisations offering the same product would each of the segments most like to buy from? How likely is it that our organisation would be chosen? How likely is it that each segment would commit to us?

### Market Segment Evaluation

Most books that discuss target market selection recommend the use of a decision matrix to visualize relative segment attractiveness and relative organizational competitiveness for each market segment. Some examples are the Boston matrix, General Electric / McKinsey matrix, directional policy matrix, and market attractiveness business strength matrix. The aim of all these decision matrices along with their visualizations is to make it easier for the organization to evaluate alternative market segments, and select one or a small number for targeting.