👉 **Want 40+ data science cases like these? Join datainterview.com 🚀**

Question Type: Product

Duration: 40 Minutes
Difficulty: Medium
Domains: Product

Problem

Suppose you have 10 million transactions of Apple customers. How would you find meaningful segmentation of customers?

DataInterview.co

👉 **Want 40+ data science cases like these? Join datainterview.com 🚀**

Solution

> *Suppose you have 10 million transactions of Apple customers. How would you find meaningful segmentation of customers?*

[Candidate] Thank you for the question. I want to first discuss the business problem with you. You mentioned “transactions” and “customers.” I understand that Apple isn’t just a company that sells iPhones but, various other services and products such as subscriptions (i.e. Apple Music, Books, News), payments (i.e. Apple Pay) and app stores. Which slice of transactions would you want me to take a look at?

[Interviewer] That’s a great question. How about just focusing on devices for now?

[Candidate] Okay understood, so devices could be iPhones, MacBooks, iPads and such. Is this correct?

[Interviewer] Yes.

[Candidate] So, I’m going to assume that if these are the transactions people focus on then, the type of customers we care about are store visitors whether brick & mortar or online.

[Interviewer] Sounds good. Anything else?

[Candidate] Just another business-related question. As you said, segmentations should have meanings but it kind of depends on the type of problem. Could I ask a bit more context on where this is plugged-into?

[Interviewer] Let’s suppose that the segmentation is used for a marketing purpose. How would you approach this problem then?

[Candidate] Understood. So, perhaps we are investigating trends and demography of users who would purchase certain device types.

[Interviewer] Sure that could work.

[Candidate] Okay great. I think I have enough information to start formulating an analysis approach.

[Interviewer] Great. Let’s hear it.

Datainterview.com

👉 **Want 40+ data science cases like these? Join datainterview.com** 🚀

[Candidate] There's a framework that can be readily applied when it comes to segmentation analysis, something called RFM, recency-frequency-monetary value. I know that it's usually employed in customer segmentation to understand who are top-tiered customers and who aren't. This framework can help categorize users.

[Interviewer] Interesting approach. What kind of data would you need for such analysis?

[Candidate] I will need a user's purchase history. Can I assume that Apple can track user's device and accessory purchases based on user email address?

[Interviewer] Yes.

[Candidate] Great. Then, here's the data source I'm thinking. The data could consists of the following variables - email_address, transaction_date and sales. For instance, I recently bought the MacBook Air 2020, so my email address XYZ@gmail.com, 2020-02-14 and $1,100, will appear in the transaction data.

[Interviewer] So, how would you use this data to create a segmentation?

[Candidate] I'd roll-up the data at the user level to generate the following - (1) Recency - days since the last transaction, (2) Frequency - count of transactions, (3) Monetary - total sales. I can use this information to apply segmentation on users.

[Interviewer] Okay, how would you segment the user on the variables?

[Candidate] I'm thinking I could apply quantiles. So, the top 33th percentiles on recency, frequency and monetary could represent a bucket of customers who are the most loyal customers of Apple. When a new product launches, they are the first ones to purchase them regardless of the price.

[Interviewer] Interesting approach. Can you think of another customer group?

[Candidate] Each dimension could be binned into three intervals based high (x <= 33th percentile), mid (33th percentile < x <= 66th percentile) and low (< 66th percentile), such that a total of 27 segments are generated. The low levels across all three buckets are users who are the opposite of loyal Apple customers. They are difficult to funnel into sales.

datainterview.co

👉 **Want 40+ data science cases like these? Join datainterview.com 🚀**

Interviewer Solution

To ace this question:

1. **Ask questions** - given that the question is open-ended, ask the interviewer questions to define a focus. For instance, you could ask about the business objective of the analysis and tailor a methodology based on the objective.
2. **Understand the problem first** - a rookie mistake is diving into methodology before fleshing out the business problem and data required for analysis. Flesh out key details on those topics first before presenting an approach to analysis.
3. **Explain business impact** - Suppose you conducted clustering and discovered that customers can be grouped into five groups. Explain how the result is significant to the business.

Discuss business problem

The interviewer will test your ability to analyze the problem within the business context. You need to demonstrate that you have business sense. First, discuss business applications of customer segmentation. Here are three potential ways customer segmentation could be helpful:

1. **Targeted Marketing** – Imagine a billboard at the Time Square in New York City. Apple is about to release a new product – the next generation of iPhone. They want to craft an advertisement board that will resonate with the passersby of the city. How should the marketing team define an effective message? The process begins with understanding their core customer using customer segmentation. Based on demographic data consisted of gender, age, location, occupation and income, an effective message can be tailored to inspire their core customer to purchase the new iPhone.
2. **Product Trends among Different Customer Groups** – What combination of products are popular among different customer types? Suppose that Apple wants to target college students, aged 18 to 24. Apple's data scientists could use customer segmentation to identify popular products among this group. Perhaps, they may uncover a pattern that, among college students, a purchase of MacBook Pro is strongly correlated with that of Dr. Dre headphones. Apple can leverage this insight to offer student discounts when the two products are purchased together during the back-to-school week.
3. **Identifying and Dissecting Loyal Customers** – Almost every company, including Apple, desires loyal customers that generate strong sales and profit. In other words, they are the ones that buy Apple products frequently and expensively. Who are such loyal customers for Apple? Knowing this information can be insightful for Apple. For instance, suppose that a small fraction of customers produces the highest sales and profit than every other types of customers. This insight is vital for fostering business intelligence and evaluating strategy to grow a robust fanbase across customers. Perhaps, Apple could investigate demographical information of loyal customers to understand why they purchase expensive goods often the first place. If could unravel why they are successful

👉 **Want 40+ data science cases like these? Join datainterview.com 🚀**

with those customers, then they could devise a business strategy to emulate the success across different customer groups.

CUSTOMER TABLE

Customer ID	First Seen Date	Age	Gender	Annual Income	Current Location
1	01/21/18	25	Male	$80,000	Fredmont, CA
2	01/22/18	18	Female		Chicago, IL
3	03/09/18	45	Male	NA	NYC, NY
4	12/12/08	33	Male	$130,000	Washington, D.C.

PRODUCT TABLE

Product ID	Release Year	Model	Family	Current Price
A	2014	Macbook Air 2014	MacBook Air	NA
B	2017	Fifth-Gen iPad	iPad	$300
C	2015	Apple Watch 1	Apple Watch	$100
D	2018	iPhone XS	iPhone	$700

TRANSACTION TABLE

Transaction ID	Timestamp	Customer ID	Product ID	Sale ($)	Channel
234-3413-1431	09-13-18 12:34:55	D	F	$700	Online
435-5435-6577	12-04-18 09:42:43	D	G	$300	Online
054-4324-4123	08-12-18 01:34:34	C	C	$100	Store

Discuss how you would process the data for analysis:

Using the data described above, discuss preprocessing and feature engineering to perform before clustering.

1. **Feature cleaning -** clean data for potential duplicate entries and missing values
2. **Feature engineering -** Conduct feature engineering to derive metrics on recency, frequency, and monetary value (RFM), a standard attribute set in customer segmentation.
3. **Scaling -** Normalize data to reduce noise that may affect clustering algorithms that use the mean to calculate the position of the centroid vector.

Discuss potential segmentation techniques that could be used:

Recency-Frequency-Monetary Value Segmentation (RFM)

👉 **Want 40+ data science cases like these? Join datainterview.com 🚀**

This is an industry standard approach for customer segmentation across high tech companies including Google. RFM stands for recency, frequency and monetary value. For each of the three variables, compute quantiles and create three intervals of quantiles, representing high, medium and low group. For instance, the high group in frequency could represent the top 33% customers who frequently purchase items from Apple. With each of the three variables binned into three groups, there are 27 different combinations in which customers could be identified. A customer who is in the top 33% (high) group across recency, frequency and monetary value could be viewed as most invaluable customer. Such individual would purchase often, keep up-to-date with the latest product offerings from Apple and generate a lot of sales from Apple.

K-Means Clustering

K-Means is also a decent initial approach to customer segmentation. If you already have business intuition or hypothesis on how many clusters may exist, then use the default K. Otherwise, external validation approaches, Elbow technique or Silhouette analysis, can help you identify optimal K. It is highly recommended that you conduct several iterations of clusters with each iteration having different initial centroid positions and features. For each iteration, conduct exploratory data analysis on the cluster results to ensure that the cluster convey practical meaning that could be applicable in business. For instance, if you identify that a cluster in an experiment with K=7 clusters and 12 features, contained a population with the majority, being 18-to-24 year olds, purchasing MacBook and Dr. Dre headphones together, what could this convey? It could mean that this population represents college students who are purchasing the goods during back-to-school weeks. Apple could leverage this information to provide discount sales with the purchase of two products.

Hierarchical Clustering

Another useful approach when K cluster size is unknown is hierarchical clustering. In general, there is bottom-up and top-down approach for hierarchical clustering. It does not matter which one you use as long as the final set of clusters are reasonable and meaningful.

👉 **Want 40+ data science cases like these? Join datainterview.com 🚀**

Interviewer Assessment

In the statistics section, a candidate is assessed based on correctness and soundness of statistical methodology, product sense and communication. For each dimension the candidate is rated in the following scale: (5) superior, (4) good, (3) adequate, (2) marginal, (1) not competent.

Assessments	Rating	Comments
Statistical Methodology	5	The methodology from the candidate sounds great. He understood how to apply the RFM framework. When asked about what kind of data source he might use, he didn't shy away from being able to respond clearly and specifically. He also understood how to find meanings in the segmentation applied which is a huge plus.
Product Sense	5	The candidate understands the business model of Apple. He knew how to turn a generic question on "transactions" and "customers" into specifics (i.e. device users, subscribers). This suggests that he researched things about Apple before the interview which is one of the qualities our team looks for as stakeholder engagements often start with a vague, open-ended business problem.
Communication	5	The candidate structured his response very well. He began with a business discussion, asking questions about the problem and clarifying assumptions about Apple. This shows that the candidate knows how to turn an open-ended question into a specific problem. Most other candidates will often jump into the solution first. But, this candidate stood out given that he made sure there's no room for vague interpretation. Then, he followed up with a solution laying things out in a step by step manner. He was coherent, clear and confident with his delivery. A+ for his communication performance.

DataInterview.com