QUESTIONS TO ASK THIRD-PARTY DATA PROVIDERS

CATEGORY	QUESTION	WHY IS IT IMPORTANT?
Data Sources	What are the sources of data the provider uses for race, ethnicity, language, gender identity, age, etc.?	• It is important to know how the underlying raw data was obtained. U.S. Census records, probability panels/ surveys, public records, transactions, searches, social activity, physical location visits, cookie data, mobile event data, proprietary algorithms – any or all can be the source of a particular segment.
Data Sources	Does the provider collect data themselves or aggregate it from other sources? Has it been sent through a chain of different providers? Does the provider build its own identity graph, or simply license other providers' data?	• When third parties aggregate data from multiple sources, there is often limited insight and input into the quality or accuracy of the data. • If the provider licenses data from another provider, then they are likely to have limited insight into how the data comes together.
Coverage	What is your audience? What is the provider's coverage for that audience?	• It is unrealistic to expect providers to have a 100 percent coverage rate of the population or specific segments. Some have higher coverage of homeowners, credit card holders, or smaller-sized households. This introduces bias when targeting multicultural consumers due to low coverage in the underlying data.
Definition	How does the provider define race/ethnicity, language dominance, and gender identity?	• There are different ways providers define race/ethnicity or language such as first name, surname, country of origin, English proficiency, U.S. Census definitions, etc. There is no standard (Census-based) definitions for sexual or gender identity, acculturation level, Spanish language dominance, etc. So, it is important to understand how these segments are defined by the provider.

CATEGORY	QUESTION	WHY IS IT IMPORTANT?
Definition: Race and Ethnicity	How does the provider account for people of multiple races/ ethnicities?	• Some definitions do not account for mixed-race populations.
Accuracy	Are there certain demographics that the vendor is more or less confident in? Is it possible to understand why and which ones?	• This is the key to understand the limitations/strengths the dataset may have in regard to your specific audience.
Mapping	How are identities being mapped? Is the process deterministic or probabilistic?	• A deterministic method uses collected known personally identifiable information (PII), while a probabilistic method relies on anonymized data to create likelihoods for a potential match (i.e. demographic assignment). • Due to this difference in how people are matched, the probabilistic approach would tend to be less accurate. It is important to understand the details of the matching process and how it works.
Mapping	What are the parameters (match keys) used to perform the mapping of people to data (PII, IP address, etc.)?	• Many consumers do not visit or register on in-language or cultural sites, using IP address to do mapping could have some limitations. • Cookie value to household mapping could introduce a significant amount of error (i.e. the wrong demographic could be mapped to the household/ person).
Mapping	How does the provider assign race/ethnicity/ language to persons or households (e.g. geography, zip code, Census block group penetrations, etc.)?	• An increasing number of multicultural consumers do not live in ethnically concentrated areas. Demographic assignment models that use only “geography” may provide less accurate results.

CATEGORY	QUESTION	WHY IS IT IMPORTANT?
Validation	How does the provider know that as a specific demographic group has been identified correctly? How does the provider define "accuracy" in the demographic assignment process? How are accuracy and coverage related for different demographic buckets?	• Providers may see lower accuracy for certain demos that have higher coverage. This is common because it means that a lot of people could be assigned to a demographic bucket without any certainty that the person/ household belongs in that bucket.
Validation	How does the provider correct for biases in the data?	• Some third-party data providers have higher coverage of homeowners, credit card holders, smaller-sized households, etc. thus leading to low coverage for multicultural consumers. • Ask to see the results of prior matches that have been done in which representative and reliable datasets were used for evaluation. It would be reasonable to expect the majority of the identities to be correct.
Validation	Is the data validated/ corrected against a "truth" dataset or validation source? "Truth set"= directly collected information from a representative survey or panel.	• To reduce bias and fix coverage gaps, big datasets must be cleaned up and calibrated using a representative panel/ surveys to ensure proper racial/ethnic representation.
Quality	How often is the dataset refreshed/updated? Are processes in place that would trigger a refresh? What is that process?	• Even though someone's race /ethnic identity is unlikely to change over time, multicultural consumers tend to move at higher rates than the general market. It is important to account for mobility (changes in address, zip code, etc.). • Other variables like age and size of the home can also change over time (e.g. children born, grandparents move in, etc.).