**Unit 3.2.7 – Evaluating Data Sources Drill**

*In each of the scenarios, find possible shortcomings of the data sources to answer the given question. What could be done to either adjust the analysis or reframe the question so that you can answer it accurately?*

Data Source: Amsterdam availability data scraped from AirBnB on December 24th. Question: What are the popular neighborhoods in Amsterdam?

* The issue with this data is that it is only from AirBnB and not from other Web sites or services (including hotels) where residences or rooms can be booked for travel. AirBnB may not yet be fully used in Amsterdam and skew results only toward cities where it is heavily used or not prohibited by law. Also, different type of demographic might use AirBnB versus other services, thereby skewing results.
* Either reframe the question to make it specific to AirBnB or reconduct the analysis to include many other booking services (VRBO and hotels to name a few.
* From solution –
  + Need to also note that the AirBnB data is only scrapped on December 24th, Christmas eve, making for a skewed analysis not representative of the year.
  + Christmastime will bias availability in two ways- more AirBnBs may be taken due to vacationers, or fewer AirBnBs may be on offer because people don’t want to share their homes over the holidays. Choose a sampling period that doesn’t include a major holiday, or reframe the question to focus on popular neighborhoods during major holidays.

Data Source: Mental health services use on September 12, 2001 in San Francisco, CA and New York City, NY. Question: How do patterns of mental health service use vary between cities?

* The issue here stems from the immediate aftermath of the terrorist attacks on September 11, 2001 in NYC. Trying to assess the difference between mental health of people in San Francisco and NYC the day after such a tragic event would only begin to shed light on how deeply the attacks impacted people’s mood in NYC.
* The correct the I issue, either conduct the study at a later time—several years after the event—or try to analyze data in periods prior to the attacks—such as September 10, 2001.
* From solution –
  + Mental health services use on September 12, 2001 in San Francisco, CA and New York City, NY: Given the terrorist attack on NYC the day before, patterns of mental health services use are likely to be very different in the two cities-- even if they were pretty similar on September 10, 2001. Reframe the question to mental health use in the aftermath of a nearby vs distant crisis, perhaps.

Data Source: Armenian Pub Survey. Question: What are the most common reasons Armenians visit local pubs? <https://www.kaggle.com/erikhambardzumyan/pubs>

* The issue with the data is it appears to only take place over a week-long period, which is unlikely to be representative of why Armenians viksit pubs in general throughout the year. Also, the average appears to skew toward very young individuals (students), which would not properly represent why older people visit local pubs.
* The data should be recollected to include equal samples of all age demographics and be taken through all periods of the year. Also, the question could be altered to mention that young people are the primary target and that we are trying to figure out why people go to local pubs in late February/early March.
* From solution –
  + Armenian Pub Survey: Respondents are all university students. Reframe to ‘university students’ reasons to visit pubs.’