Exercise 2

A hypotetical social study could be that of looking at tweets from Twitter, analyze the text and determine if they contain hate speech, with eventually the objective to automatically detect hate speech.

The data collection strategy could be that of retrieving the tweets using Twitter's API, and filtering the tweets by only selecting those that are in English, e.g. coming from the USA.

Some research questions could be to understand what percentage of tweets contain hate speech, and if there are some groups or individuals that are predominantly the offenders/receivers of hate speech.

a)

As general problems there could be sparsity, which makes it easier to analayse the more frequent elements and harder for the rare elements.

Furthermore the content could be incomplete, corrupted or containing errors, what is called noise. There are also some general data bias that exist in processing social data:

- Population biases: Systematic distortions in demographics or other user characteristics between a population of users
- Behavioral Biases: Systematic distortions in user behavior across platforms or contexts, or across users represented in different datasets
- Content Production Biases: Behavioral biases that are expressed as lexical, syntactic, semantic, and structural differences in the content generated by users
- Linking Biases: Behavioral biases that are expressed as differences in the attributes of networks obtained from user connections, interactions or activity
- Temporal Biases: Systematic distortions across user populations or behaviors over time

Another problem with datasets is redundancy which means that values can occur more than one times. In doing so these values can be identical or near identical.

In our scenario, there would be no real problem of sparsity, given the huge amount of tweets on Twitter. It hardly would be incomplete or corrupted, but it could very well contain errors, such as typos, and a potential solution would be to discard such tweets.
Some biases such as population bias would definitely affect the research, but by stating that the research is focused on a particular group or population, then it should be fine.
Temporal biases could be avoided by considering the tweets only in certain periods of times, or by comparing them on different time scales.
Other biases could also have an influence, and it would be hard to counter these biases given the text nature of the data.
Moreover, the Twitter's API shouldn't yield duplicates, but identical or near identical tweets may very well exist, but that wouldn't be actual redundancy to be eliminated.

b)

Again there are different types of biases:

- Functional biases: Biases that are a result of platform-specific mechanisms or affordances, that is, the possible actions within each system or environment.
- Normative biases Biases that are a result of written or unwritten norms and expectations of acceptable patterns of behavior on a given online platform or medium.
- External biases). Biases resulting from factors outside the social platform, including considerations of socioeconomic status, ideological/religious/political leaning, education, personality, culture, social pressure, privacy concerns, and external events.

Additionally, it is possible that accounts are not individuals, instead there are non-individual agents, e.g. organizations or bots.

Such types of biases would be hard to counter in our scenario, but could be of interest in the research outcome. For instance, external biases could be somehow taken into consideration by analyzing the profiles of the writers of the tweets, to predict/understand if for instance some groups or individuals with certain socioeconomic statuses, ideological/religious/political leanings, cultures, etc. are particular targets or offenders of hate speech. Moreover, a known issue on Twitter is that of bots, or accounts that are not actual individuals, but rather "robots", which should definitely be taken into consideration in the analysis of the tweets.

c)

Yes there are two additional biases which occur during collecting and processing:

- Data collection biases: Bias that arises from the selection of data sources or from the way in which data from those sources are collected and processed. The way in which the selection of particular data sources affects the observations and thus the research results can be referred to as source selection bias.

Data processing biases: Biases introduced by data processing operations such as cleaning, enrichment, and aggregation.

In our scenario, some biases are introduced by force, e.g. by deciding to only analyze tweets in a certain language or from a certain country, a data collection bias is introduced, which should be the basis of careful research findings, that should not extend outside of the context on which the data source is restricted to.

Data processing biases could also be introduced in cleaning the data and in a potential enrichment by also analyzing the profiles of the authors of tweets, because some profiles may be richer, more detailed etc. than others.

d)

- Qualitative analyses: they tend to be in-depth, open-ended, and exploratory, answering questions about the how, what, or why of a social phenomenon.
- Descriptive analyses: they are the basis of many studies, quantitatively depicting social data through numerical or graphical summaries of variables of interest. Such analyses capture the distribution, variability, or correlations among variables.
- Observational Studies

A combination of different types of analysis would definitely help mitigate the risk of a biased insight and generally give better results.

For instance, a quantitative analysis would greatly benefit from the amount of data available, and could be used to quantitatively depict behavior and populations on Twitter, specifically to get a general sense of the hate speech problem on Twitter, even by simply determining on a large scale the percentage of tweets affected by this problem.

This could be integrated and reinforced by a qualitative analysis, with which the research could go into more depth and answer more specific questions, but construct new hypotheses.

e)

How the evaluation is performed may lead to biased conclusions or outcomes, including due to metrics selection or results assessment and interpretation, which can both pose threats to construct validity. Moreover, our own biases, perspectives and experience may be reflected in the way in which our analysis' results are assessed and interpreted, and may also be dependent on the assumptions made about the data and the methods that were used.

An issue that may occur is that of determining the definition of hate speech, and to correctly apply that definition to correctly categorize the tweets. This issue could be mitigated by applying the definitions established by the laws in the country of reference, and to also apply the definitions provided by other prominent research papers on the matter, which would also counter the issue that the interpretation and assessment of results are too often done by data experts, not by domain experts.

Another issue is that of the context of the tweets, something that is hard to discern automatically at evaluation time.