# Worksheet for Yelp Business Data Analysis

# Section 1: Sharing and understanding data

1. How does a data dictionary help increase data understanding? What are the advantages to including a data dictionary with a dataset? Give an example of how data might be used incorrectly without a data dictionary.

* For more information about data dictionaries and metadata, check out this link: <http://library.ucmerced.edu/data-dictionaries>
* For some examples of more complex data dictionaries, check out this link: <https://data.nal.usda.gov/data-dictionary-examples>

# Section 2: Asking the right questions

One important way that businesses can take advantage of this data is by comparing themselves to similar businesses. This involves grouping businesses by certain categories.

1. What are some of the ways that similar businesses can be grouped, based on the columns in the data table above?
2. How can businesses take advantage of the data collected on Yelp users? Hint: Think about this question from a marketing perspective.
3. What other data would you wish to collect from a marketing perspective? What kinds of questions could you answer with this data?

# Section 3: Interpreting data visualizations

One way that businesses can be grouped is by state. This is shown in the bar chart below.

A screenshot of a cell phone

Description automatically generated

Plots like this one can also be useful to Yelp users. For example, if you were trying to decide where to move, you could examine which cities have the most of certain businesses you like to visit, such as restaurants or museums. Note that this is just a subset of data – states such as Florida (FL) certainly don’t contain 0 businesses with Yelp reviews!

1. If you wanted to move to a state that seems to have the most saturated market (i.e., the most businesses) based purely on the graph above, which state would you move to?

Another way to visualize this data is by location (specifically, coordinates), which can be seen below.

A screenshot of a cell phone

Description automatically generated

This visualization shows businesses grouped by city, and then plotted using the businesses’ latitudinal and longitudinal coordinates.

1. Which city on the heat map appears to have the highest concentration of businesses? Is this surprising given your answer to question 5?
2. Look back at question 4. If you had collected that kind of data, what kind of visual representation could you create to understand the results?
3. What might you infer if you mistook this data set to be representative of all businesses in the US?

# Section 4: Data ethics and privacy

The way Yelp has access to all of this data is by collecting user feedback. While it is not presented here, data is kept on specific users as well as businesses. Yelp is able to track the kinds of places you like to go, as well as your location.

1. Is Yelp’s data collection a violation of privacy, or is it necessary to provide the best possible service to its users?
2. What other types of applications are keeping track of your data?

* To continue this discussion of the ethics involved in Artificial Intelligence, check out these materials: <https://github.com/generationai/Johns-AI-Intro-2020>

# Challenge problem: Try it yourself

**Challenge Problem:** Using the Excel file *NC\_Yelp.xlsx* containing data Yelp data from North Carolina (NC), add filters to different columns and examine the different ways to filter the data based on your preferences. For example, try looking for restaurants in Charlotte, NC with a rating above 3 stars using Excel filters. How many businesses did you find?

\*\*Note: This data is a small subset of data from one single state. If you wanted to look at this entire data set, or a data set containing years of Yelp reviews across the United States, you would need to use more advanced tools to open and manipulate the larger data set. For instance, you could learn a statistical programming language such as R or Python, or you could create a *data base* that you can *query* (or ask questions to) in a language called SQL. In the year 2020, you may have a data scientist who does this for you. In this case, it’s important to be able to communicate effectively with the data scientist and have some understanding of how data can be used for analysis. In the future, it may be increasingly useful to be capable of manipulating large data sets for yourself regardless of your field or job title. To learn more about how to manipulate the entire dataset used here and see how these graphs were made, take a look at the code [here](https://mitreextorg.sharepoint.com/sites/GenAI/Shared%20Documents/Forms/AllItems.aspx?id=%2Fsites%2FGenAI%2FShared%20Documents%2FGeneral%2FPurdue%2FYelpBusinessesLesson%2FSolution%5FSet%5FYelp%2Eipynb&parent=%2Fsites%2FGenAI%2FShared%20Documents%2FGeneral%2FPurdue%2FYelpBusinessesLesson).