****

**Milestone 4 - The Basics of Data Prep & Cleaning**

**~5 hours**



Your colleague, Mateo, thinks you’re ready to begin learning about data analysis. Mateo is a Data Analyst and he’s scheduled a meeting with you to review the dataset you’ll be working with.

****

[**Video Link**](https://youtu.be/ohRiL_Kecng)

Mateo shares two files with you for Milestone 4, linked below. Milestone 4 Data contains a main tab called “Data” and four other tabs which provide examples for other tasks that we’ll tackle later. The Data Dictionary file should have just one tab. ***Note that the links are to “View Only” Google Drive files. You will need to download the files or create a copy in your own Google Drive to work with the files.*** You can use either Excel or Google Docs to work with these files (the data may load more slowly in Google).

* [Milestone 4 Data](https://docs.google.com/spreadsheets/d/1Yq9I9JYKDMqGx0wSfATiNugZaBN-Mfvx/edit?usp=sharing&ouid=118334720634078226839&rtpof=true&sd=true)
* [Milestone 4 Data Dictionary](https://docs.google.com/spreadsheets/d/1zGl0yrPVIJhxhP8pRT2j3ri1bktfqlYl/edit?usp=sharing&ouid=118334720634078226839&rtpof=true&sd=true)

Task 6: Getting Familiar with the Data Dictionary

Take a moment to review these documents, beginning with the Milestone 4 Data Dictionary. Mateo has taken the raw data from the Census Bureau and [transformed](https://www.ibm.com/think/topics/data-transformation) it into an understandable, workable document for your analysis. [A Data Dictionary](https://library.ucmerced.edu/data-dictionaries) is a list of the names, definitions, attributes, and questions associated with each survey entry.

There are several things that Mateo has done to make this dataset easier for you to work with. Read through the following bullet points carefully.

* Mateo has color coded the data so that the columns are grouped by subject. For example, you’ll see that Columns B - Q (highlighted in orange) all relate to demographics. Columns R - W relate to employment (yellow), and so on. The questions which correspond with each of the columns are also color coordinated in your Data Dictionary.
* There is a connection between the two spreadsheets. For instance, you’ll see that the “EST\_ST” Column in your Data Sheet (Column B) corresponds to Question #1 (Row 5) in the Dictionary. The possible values for Question #1 are listed in Column D in the Dictionary. We know, for instance, that residents who live in New Mexico will have the value “35” in Column B in the Data Sheet. 
  + The reason that you see the value “35” rather than the words “New Mexico” in Column B is because the data sheet is “coded.” [Learn more about data entry here](https://home.csulb.edu/~msaintg/ppa696/696codes.htm). We will learn the basics of data coding in a moment.
* You’ll notice that Mateo’s dataset is a shortened version of the full data survey survey from Milestone 3. This is to make the analysis process easier for you. Mateo has removed survey data not related to this project (e.g., veteran status, natural disaster preparedness questions) to simplify your analysis.
* Mateo has also removed entries where the participants declined to provide responses to one or more of the questions in the survey. This was done to make it easier for you to analyze the dataset as a whole. (This is why Khaled asked you to learn about bias in Milestone 3 - your team is already making choices that will impact the outcomes of your research!)
* Column D is blank - we’ll get to that in a bit!

For the moment, explore the Data Dictionary and Dataset to get familiar with the types of data available to you from the broader Census Bureau survey.

Using Google Sheets / Excel for Data Preprocessing

We want to be careful here! It’s easy to make mistakes while cleaning. **Save an “original” copy of your work before making any major changes to your work.**Having an “original” or “clean” copy of your data ensures that you can always start over if you make a mistake later in your analysis. Don’t skip this step!!

At this point, it will be helpful for you to know a bit more about Google Sheets / Excel, and how it can help you perform basic data analysis. Begin by reviewing these key concepts:

* You can quickly navigate your workspace using shortcuts for [Google Sheets](https://support.google.com/docs/answer/181110?hl=en&co=GENIE.Platform%3DDesktop) or for [Excel](https://support.microsoft.com/en-us/office/keyboard-shortcuts-in-excel-1798d9d5-842a-42b8-9c99-9b7213f0040f)
* Understand the [search and replace function](https://support.google.com/docs/answer/62754?hl=en&co=GENIE.Platform%3DDesktop)
* Spend time [learning about filters and sorting](https://www.w3schools.com/googlesheets/google_sheets_sort.php)
* [Conditional formatting](https://www.w3schools.com/googlesheets/google_sheets_conditional_formatting.php) is another tool that can help you with data cleaning

Key Formulas

There are many essential formulas for you to know about in Google Sheets / Excel. **Look at Tab 2 of your Milestone 4 Data (‘Key Formulas’) to see demos of these formulas.**

* [Sum](https://support.google.com/docs/answer/3093669?hl=en&sjid=2644863441701132713-NA): returns the sum of a series of numbers and/or cells
* [Average](https://support.google.com/docs/answer/3093615?hl=en&sjid=2644863441701132713-NA): returns the numerical average value in a dataset, ignoring text
* [Count](https://support.google.com/docs/answer/3093620?hl=en&sjid=2644863441701132713-NA): returns the number of numeric values in a dataset
* [Concatenation](https://support.google.com/docs/answer/3094123?hl=en&sjid=2644863441701132713-NA): ties strings to one another
* [Average](https://support.google.com/docs/answer/3093615?hl=en&sjid=2644863441701132713-NA): returns the numerical average value in a dataset, ignoring text
* [TRIM](https://support.google.com/docs/answer/3094140?hl=en&sjid=13000517817542846310-NA): (you can also trim whitespace by going to Data > Data Cleanup > Trim Whitespace).

Data Cleaning

Next, let’s explore basic data cleaning principles. We clean data to remove the “bad” data we talked about in Milestone 2. Get acquainted with the basics of data cleaning with [this video from the NSDC Video Library](https://youtu.be/t8WkoGLkdTk) (Note: This video includes some tips on data cleaning with Pandas - we will not be using Pandas in this project).

Task 7: Deduplicating the Data

We’ll begin our review of the data in the Milestone 4 Data sheet by confirming that there are no duplicate participant entries in the sheet. To do this, we’ll take a look at Column A, which contains the ID Number (‘SCRAM’) of each participant who completed the survey. *We are only deduplicating Column A. Save a copy of your work before you do anything.*

Option 1:

* Use the Conditional Formatting tool to identify duplicates using [color coding](https://www.coursera.org/articles/tutorial-highlight-duplicates-google-sheets) (see above).
* Remove the duplicate entries by deleting those rows.

Option 2:

* Go to Data and Click Remove Duplicates (for Google Sheets, Data > Data Cleanup > Remove Duplicates). Make sure that you select *all* columns for analysis, or the program will just delete the duplicates in that row, leaving you with messy data.

Task 8: Preparing the Data

Mateo started the data preparation process for you. Earlier today, he began to replace values for Column C in the Datasheet (EST\_MA, or Estimated Metro area). Pick up where he left off and do a search / replace for each of the values in this Column. This is called “data coding.”

Refer to the Data Dictionary’s Row 8 to find the text string that should replace the numbers in Column C.

You’ll need to perform a search and replace tool - [instructions on how to restrict search and replace to a single column can be found here](https://superuser.com/questions/1500058/find-and-replace-in-excel-can-one-restrict-to-just-one-column). When you’ve replaced all the values for Column C, continue this process for Columns G, and H.

Task 9: Inconsistent Data

You’ll notice in Column E (TBIRTH) that there is some inconsistency in the data. Most of the entries provide only the year of birth for the respondent. A handful include the month of birth as well (e.g., ‘1961-4’). This birth month data is not helpful and is the result of collection errors.

Find all inconsistent entries and remove the birth months. You can do this by filtering the column by Largest to Smallest value. At the top, you’ll see the entries you need to edit.

Outliers

There are a handful of spaces where we can look for [Outliers](https://www.freecodecamp.org/news/what-is-an-outlier-definition-and-how-to-find-outliers-in-statistics/) in this data. Outliers extremely high or extremely low data point relative to the median. With outliers, we run the risk of having a skewed analysis.

Take a look at Column L, which represents the total size of the respondent’s household (THHLD\_NUMPER). If we sort this column by size, we see that there are a small number of households with 8 or more residents (50 of 10,136 respondents’ households, or 0.4%). We don’t need to do anything with those data points at this time, but it’s important to know that they are there. Depending on our research question, those outliers may skew our data.

For example, if we were to compare the average responses of this subgroup vs. the larger dataset in Column BR (EXPNS\_DIF) responses to the question “In the last 7 days, how difficult has it been for your household to pay for usual household expenses?”, we see that families with 8 or more members reported having more difficulty affording household expenses (an average score of 2.5 out of 4 relative to the overall average score of 1.9. You can see a visual display of the respondents’ answers in the ‘Outliers’ tab of your spreadsheet.

And this makes sense! In an economy where food prices are rising, it would be more expensive to keep a large family healthy and fed. Again, it’s important to keep Outliers in mind when we’re working with data. Note that some data analysts will use a process called [smoothing](https://medium.com/@j749023/introduction-to-data-smoothing-98100cdfd361) to remove outliers and make patterns more visible. Just remember to check for outliers when reviewing your data.

Task 10: Supplementing the Data

[](https://youtu.be/ohRiL_Kecng?feature=shared&t=1818)

[**Video Link**](https://youtu.be/ohRiL_Kecng?feature=shared&t=1818)

Next, Mateo asks you to find ways to add new data to this datasheet. In particular, you can fill out Column D with information about the region each respondent lives in based on what U.S. state is indicated in Column B. Each U.S. state belongs to one of four regions (Northeast, South, Midwest, and West). A key which tells you which of the 50 states (plus Washington, D.C.) belong to each region can be found in your Data Dictionary (Rows 9-11).

We can supplement our data by using the IF function.

* [IF Function](https://support.microsoft.com/en-us/office/if-function-69aed7c9-4e8a-4755-a9bc-aa8bbff73be2): compares a value to an expected value and returns a result based on whether the comparison is true or false
* [Nested IFs](https://exceljet.net/formulas/nested-if-function-example): used when you want your formula to show a particular value after taking several conditions into account
* Hint: use the below formula to get started

=IF(OR(B4=9, B4=23, B4=25, B4=33, B4=34, B4=36, B4=42, B4=44, B4=50), "Northeast", IF(OR( …………

Add a region code for each participant entry in Column D. We will learn other formulas which can help support our data analysis in the next milestone.

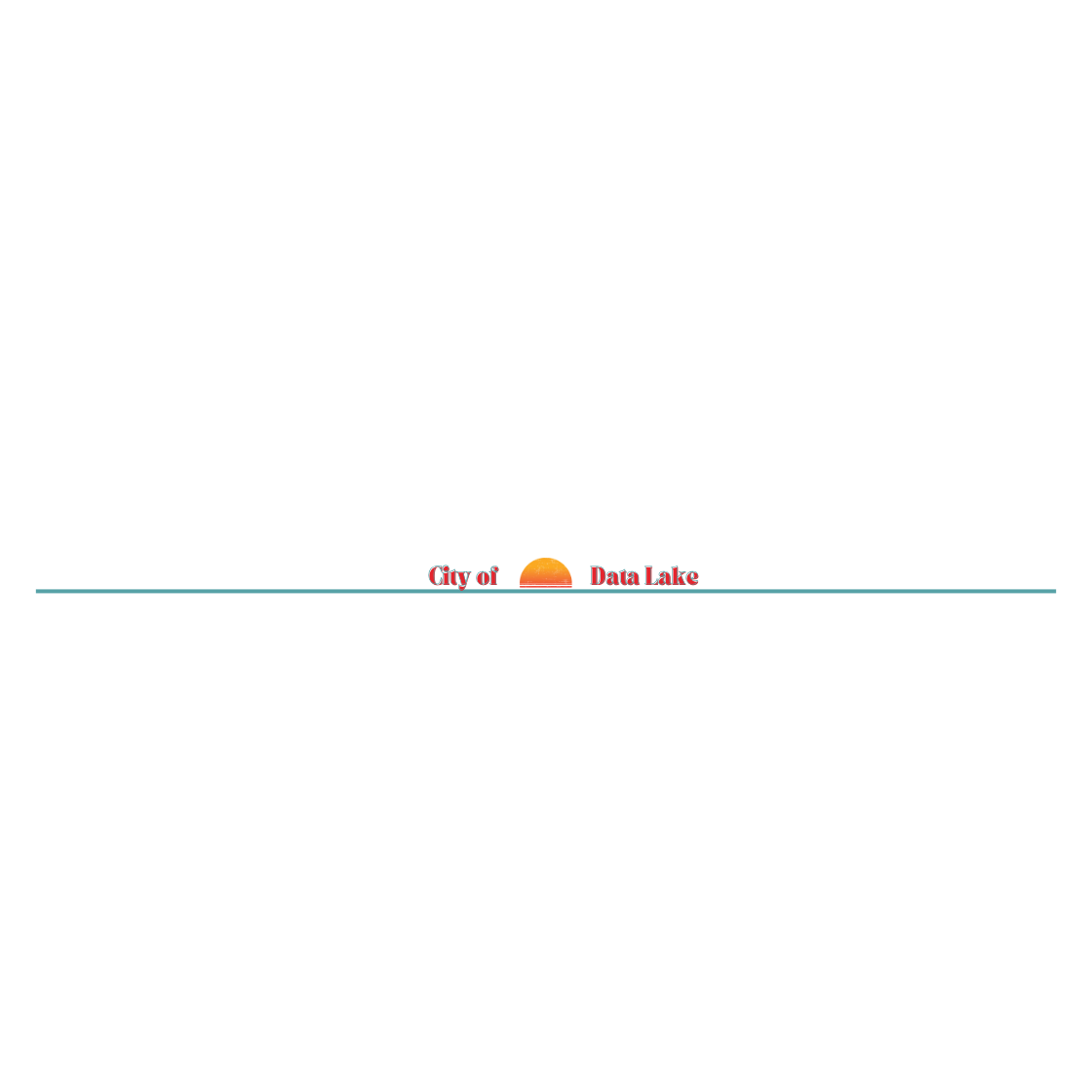
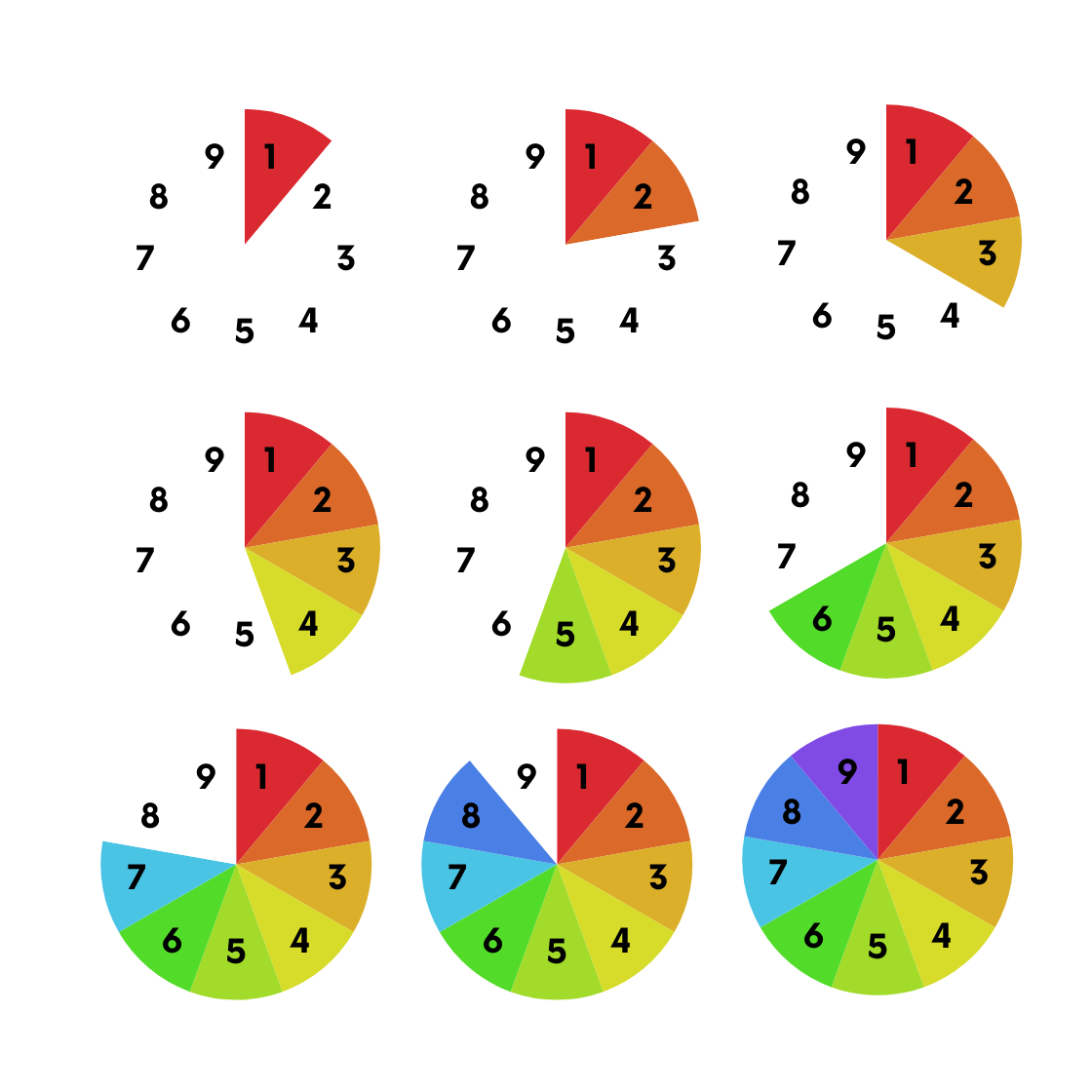
Open-Ended Data

This dataset doesn’t include participant responses to [open-ended questions](https://www.qualtrics.com/experience-management/research/open-ended-questions/) (e.g., “What is your favorite thing about remote work?”, which would be answered in an open text box). However, it’s important to understand how you might use those kinds of responses in a data analysis.

It is possible to manually code qualitative data responses using a variety of tools such as sentiment analysis or data visualization. The National Student Data Corps (NSDC) has a separate [data science project on movie reviews and sentiment analysis](https://nebigdatahub.org/nsdc-data-science-project-sentiment-analysis/) which might interest you. Bookmark this for your next project!

Data Cleaning Ethics

As noted above, there are several ethical considerations at hand when we clean data. Although we haven’t begun our analysis yet, we have already begun to make choices about what data will be included in our final report. It’s worth thinking about how reducing “noisy” data or removing null values may skew our analysis. If you’re interested in learning more about data ethics and the “data science pipeline,” we encourage you to review this [NSDC Data Science Flashcard video series](https://www.youtube.com/playlist?list=PLNs9ZO9jGtUDZONu0E2zVHOvtjiuqwtDe).



**You’ve completed the 4th Milestone!**

Great job! This Milestone was a big one and we covered a lot of new information. You should now feel a bit more confident in your ability to understand and interpret a raw dataset. You now have a baseline understanding of how the variables of your data are represented and related to one another.

**Next Steps:**

In Milestone 5: Exploratory Data Analysis, we’ll start thinking about key formulas we can use to identify specific patterns in our data.

**Milestone 5 - Exploratory Data Analysis**

**~5 hours** 

In this Milestone, we’ll move beyond simple data cleaning and further explore the formulas which will help you perform a deeper analysis of remote work trends in the U.S.

[****](https://youtu.be/8mb6uoDWVfM)

[**Video Link**](https://youtu.be/8mb6uoDWVfM)

Exploratory Data Analysis (EDA)

You have now arrived at the EDA stage of your research! This is where most data scientists spend their time in a given project. Take a moment to [learn about EDA](https://www.geeksforgeeks.org/what-is-exploratory-data-analysis/). As you can see, we have already begun to do some EDA, including data cleaning and a review of the data structures.

More Excel / Google Sheets Functions:

[](https://youtu.be/8mb6uoDWVfM?feature=shared&t=171)

[**Video Link**](https://youtu.be/8mb6uoDWVfM?feature=shared&t=171)

In Milestone 4, you learned about some useful Excel / Google Sheets formulas. To really dive into an analysis on remote work, we’ll need to refine our knowledge and practice with additional formulas.

* [SUMIF](https://support.google.com/docs/answer/3093583?hl=en&sjid=13000517817542846310-NA): useful for adding together values in a range of cells that meet multiple criteria
* [COUNTIF](https://support.google.com/docs/answer/3093480?hl=en&ref_topic=9054531&sjid=13000517817542846310-NA;): helpful when you’re trying to identify the number of values within a certain range that match a particular set of criteria
* [COUNTUNIQUE](https://support.google.com/docs/answer/3093405?hl=en&sjid=2644863441701132713-NA): a useful formula that counts the number of unique values in a list of specified values and ranges

Having trouble? These are some [common Google sheets errors and how to fix them](https://skills.ai/blog/master-formulas-in-google-sheets-the-ultimate-guide/#:~:text=Common%20Formula%20Errors%20and%20Their%20Fixes).

Want to learn more functions? [We recommend this list](https://www.w3schools.com/googlesheets/google_sheets_and.php). This is [another great overview](https://www.goskills.com/Excel/Resources/Most-useful-Excel-functions-for-data-analysis).

Task 11: Experiment with Formulas

Let’s practice with our formulas. Create a new tab in your data spreadsheet (as always, save your work first!). Then:

* Use the COUNTIF formula to identify how many participants self-identified as White, Black, Asian, and Multi-racial (Column G).
* Use the COUNTIF formula to identify how many participants have a total household income of $100,000 or more (Column Q).
* Use the COUNTIF (or SUMIF) formula to calculate how many respondents indicated that their household recently experienced a job less (Column R).
* Use the COUNTUNIQUE formula to see if all 50 U.S. States and Washington D.C. are represented in Column B. (For those using Excel: the UNIQUE or COUNTUNIQUE formulas may not be available in your version of Excel. Try the formula COUNTA(UNIQUE(B3:B10131)) if needed. Also feel free to skip this bullet if you cannot use this formula.)

Come back here when you’re done.

Task 12: Build a Pivot Table

[](https://youtu.be/8mb6uoDWVfM?feature=shared&t=658)

[**Video Link**](https://youtu.be/8mb6uoDWVfM?feature=shared&t=658)

Pivot Tables are an enormously powerful tool that can allow you to analyze multiple data variables at a time. Get started by reading [this Google tutorial on pivot tables](https://support.google.com/docs/answer/1272900). This is also [a helpful overview](https://youtu.be/0bojnxjNMTM?feature=shared&t=31).

Using the tutorial shared above, practice building at least one pivot table in a new tab in your spreadsheet. This is a great space for you to start identifying the variables that you’d like to compare in your final analysis. What relationships seem to exist between certain columns?

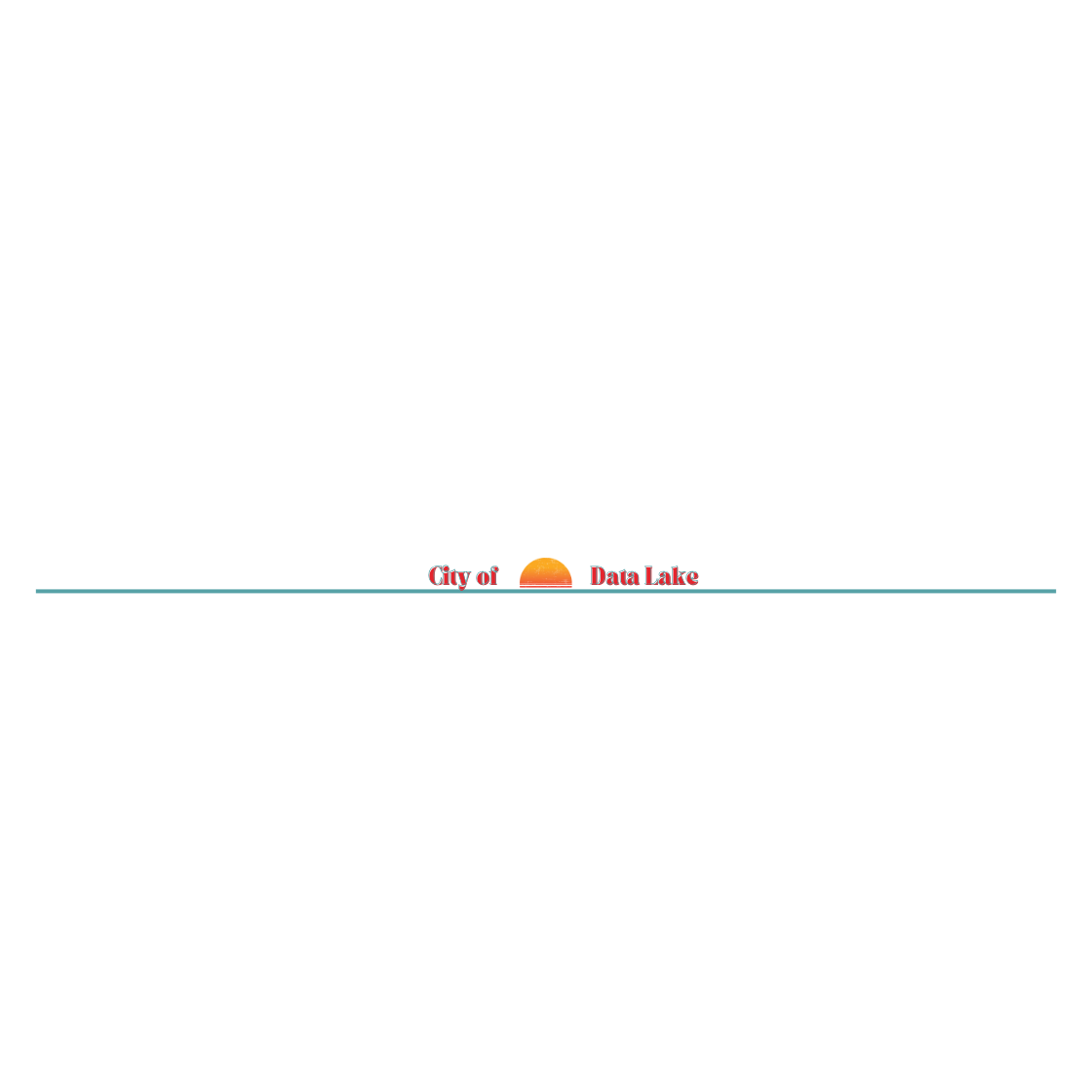
Hint: You can also use [Slicers](https://support.microsoft.com/en-us/office/use-slicers-to-filter-data-249f966b-a9d5-4b0f-b31a-12651785d29d) with your pivot tables to compare and contrast new variables.

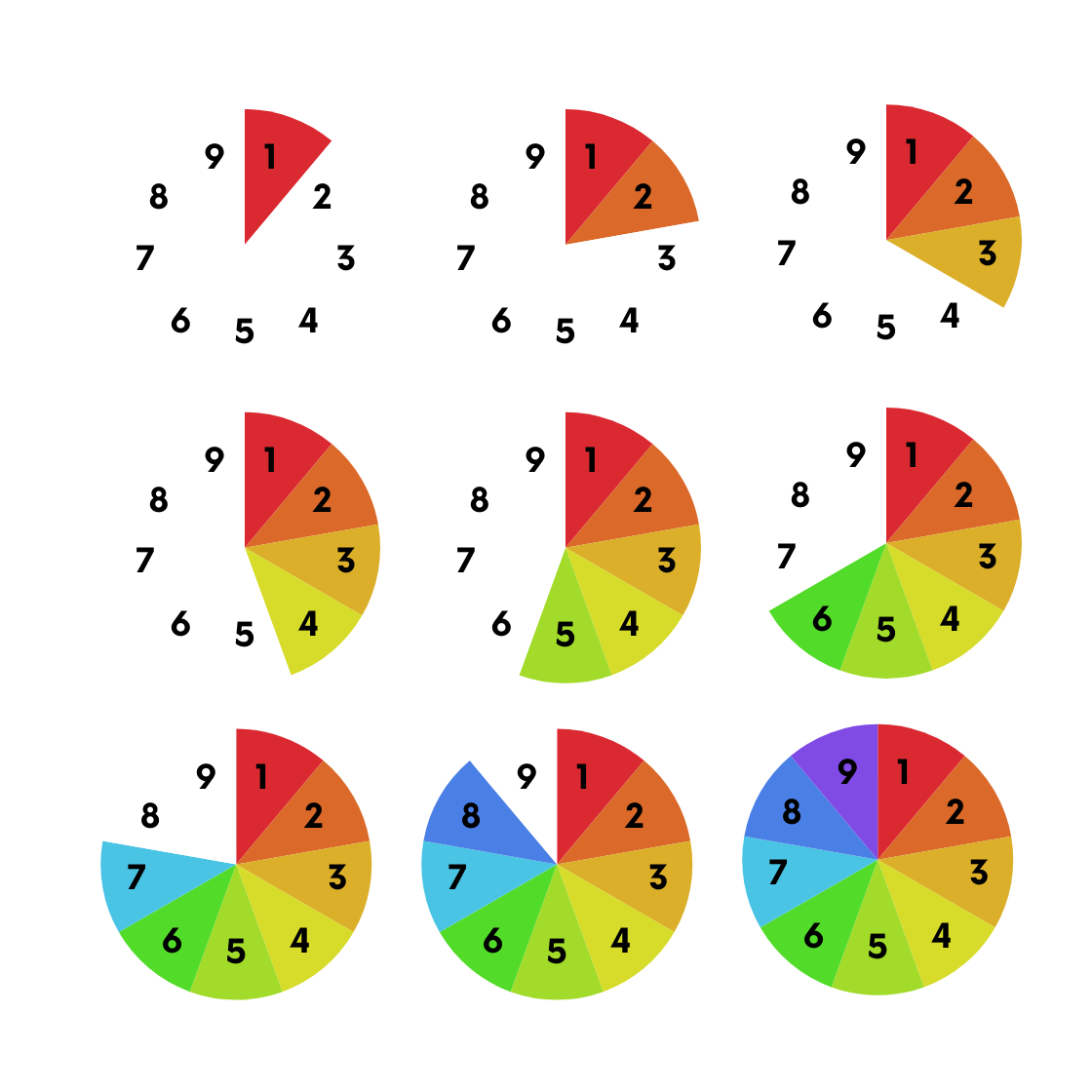
This Task that may take some time to complete! Don’t worry, keep experimenting - pivot tables are a sandbox for you to explore the data. Come back here when you’re done.

Other Data Science Concepts

Of course, this is just the tip of the iceberg and there are many other data and statistics concepts you’ll need to know to advance your career in data science. We won’t cover them all here, but here is a partial list of other techniques and terms you may want to know about.

* [Statistical Significance](https://www.qualtrics.com/experience-management/research/statistical-significance-calculator/)
* [Data Normalization](https://www.splunk.com/en_us/blog/learn/data-normalization.html)
* [Data Aggregation](https://www.ibm.com/docs/hu/tnpm/1.4.2?topic=data-aggregation)
* [Skew](https://faculty.washington.edu/swithers/seestats/SeeingStatisticsFiles/seeing/plotagain/skew.html)
* [Shape, Center, Spread of a Distribution](https://math.oxford.emory.edu/site/math117/shapeCenterAndSpread/)
  + [Central Tendency](https://statistics.laerd.com/statistical-guides/measures-central-tendency-mean-mode-median.php)

**You’ve completed the 5th Milestone!**



Excellent job! You should now be developing a sense of the types of research questions you want to ask in this analysis. The relationships between various data points should become clearer in Milestone 6. Keep going!

**Next Steps:**

We’ve reached the end of Document B. [Please take a moment to complete this very short, five question Google Survey on Milestones 4 - 5](https://forms.gle/ZA4wkzuWv8SiXQ639). Your input here will help our team understand what resources helped you learn (or not) and what lessons you found particularly helpful.

Our next step is to head over to Document C, which contains Milestones 6 & 7. Here, we’ll take the information we learned from this Document and begin applying it more rigorously to our research question. We’ll practice developing and testing a hypothesis.