**––**

Welcome to the Office of the Mayor! Let’s get started on our data analysis for the (fictional) City of Data Lake.

\*\*Note: in order to use this document, you will need to download a copy as a .doc (to be used in either Microsoft or your own Google Drive)

[](https://youtu.be/AlhcPOugRP4)

[**Video Link**](https://youtu.be/AlhcPOugRP4)

**About Data Lake, West Dakota**

Here are some things you should know about your hometown of Data Lake, West Dakota. These facts may inform your final recommendations to Mayor Ellation and her team (or may not, depending on the research question you pick).



* Data Lake’s population (314,159 residents) has grown by 20,471 (6.5%) since Spring 2020. Of that group, 59% are adults of working age (ages 18-67), 23% are children (ages 0-18), and 18% are adults who are retired (ages 67+).
  + Many of Data Lake’s remote workers say they relocated to this area to achieve a higher quality of life. They selected Data Lake specifically because of the city's close proximity to Random Forest National Park (~20 miles away).
  + The adult newcomers to Data Lake are 49% female and 48% male (3% identified as non-binary) and 63% are either currently employed or actively seeking employment.
  + While the population growth has given Data Lake a larger tax base to draw from, the city’s physical infrastructure (transportation, housing, schools, police department, and local health and business communities) are looking to the Mayor’s Office for guidance on how to appropriately scale up and meet the ever-increasing demand for public services.
* This influx of new community members reverses a long-time trend for Data Lake and many other medium-sized cities across the U.S. - population shrinkage. In previous years, Data Lake has lost an average of 2-3% of its population annually as citizens moved to larger cities in pursuit of greater educational and employment opportunities.
* Data Lake has a diverse economy - the largest employers in the area are regional hospitals, city schools, local banks, insurance companies, and a car manufacturing plant. Many of the remote employees that have newly relocated to Data Lake work in technology (31%) or in digital marketing and sales (27%).
* Data Lake’s new mayor, Tess Ellation, was elected on a platform which emphasized job growth and revitalization of the city’s business district. While campaigning, she stated that her team would be using data-driven insights to recommend adjustments to the city’s annual budget of $600 million. *Our goal for this project is to give her at least two data-driven insights about the lives, needs, and habits of remote workers that the Office of the Mayor can adopt for public policy.*

**Milestone 1 - Preliminary Research**

**~2 hours**

Before we do anything related to data, we first need to learn a bit about the modern work landscape. This will help us make specific recommendations to Mayor Ellation and her policy team. We’ll spend the first Milestone diving into the current state of modern work culture in the U.S. This “desk research” will help us in the next stages of the project, when we identify patterns in our dataset.

Task 1: Beginning Desk Research

Spend some time reading through the following news articles on the modern American workforce.

|  |  |
| --- | --- |
| **Article** | **Read?** |
| [Nearly Half Of Workers Will Consider A Job Change If Ordered Back To The Office, Survey Finds](https://www.forbes.com/sites/jackkelly/2025/01/17/nearly-half-of-workers-will-consider-a-job-change-if-ordered-back-to-the-office-survey-finds/) |  |
| [Will Americans Ever Lose Their Taste for Telework?](https://www.bloomberg.com/news/articles/2025-01-16/as-remote-work-lingers-in-us-the-future-of-commuting-is-murky) |  |
| [Traffic on the Rise as Employers Scale Back Remote Work](https://www.govtech.com/transportation/traffic-on-the-rise-as-employers-scale-back-remote-work) |  |
| [Remote work is eroding your colleagues’ social skills, survey finds](https://financialpost.com/fp-work/remote-work-eroding-social-skills) |  |
| [How Loneliness And Remote Work Are Shaping The Employee Experience](https://www.forbes.com/councils/forbesbusinesscouncil/2024/10/28/how-loneliness-and-remote-work-are-shaping-the-employee-experience/) |  |
| [The rise in remote work since the pandemic and its impact on productivity](https://www.bls.gov/opub/btn/volume-13/remote-work-productivity.htm) |  |
| [Study finds hybrid work benefits companies and employees](https://news.stanford.edu/stories/2024/06/hybrid-work-is-a-win-win-win-for-companies-workers) |  |
| [Tulsa's $10,000 remote worker incentive boosts city's economy and retention](https://kfoxtv.com/news/nation-world/tulsas-10000-remote-worker-incentive-boosts-citys-economy-and-retention-brain-drain-oklahoma-housing-cost-of-living-finances-income-tax-sales-community-slack-work-from-home-pandemic-economy) |  |
| [Seattle, the remote work capital of the U.S., is in denial about its effects](https://www.seattletimes.com/seattle-news/seattle-the-remote-work-capital-of-the-us-is-in-denial-about-its-effects/) |  |
| [The Small Cities Benefiting From Remote Workers](https://www.forbes.com/sites/lizfarmer/2021/09/15/the-small-cities-capturing-remote-workers/) |  |
| [Despite An Uptick In Remote Work, Parents Are Still Immensely Struggling To Balance It All](https://www.essence.com/news/money-career/parents-work-life-balance/) |  |

Hint: You may be tempted to use AI to summarize articles for you. Try not to do this - you’ll learn more and retain information better if you read it yourself. You’ll need detailed subject matter expertise in future Milestones.

Next, list three other news articles on remote work from reputable sources (major news outlets, trusted scientific organizations) that we want to recommend for Mayor Ellation and her team. We want to establish a well-rounded baseline for what remote work is and its impacts.

|  |  |
| --- | --- |
| Recommendation 1: | https://www.forbes.com/sites/julianhayesii/2025/04/25/is-remote-work-effective-google-big-tech-and-the-data-say-maybe-not/ |
| Recommendation 2: | https://www.khaleejtimes.com/jobs/uae-calls-for-more-remote-work-options-to-reduce-commuting-make-daily-jobs-easier |
| Recommendation 3: | https://www.cnbc.com/2025/03/23/5-years-into-the-remote-work-boom-the-return-to-office-push-is-stronger-than-everheres-why.html |



Hint: Try to find research that was published in the last 12 months - you want current data.

Task 2: Summarizing Desk Research

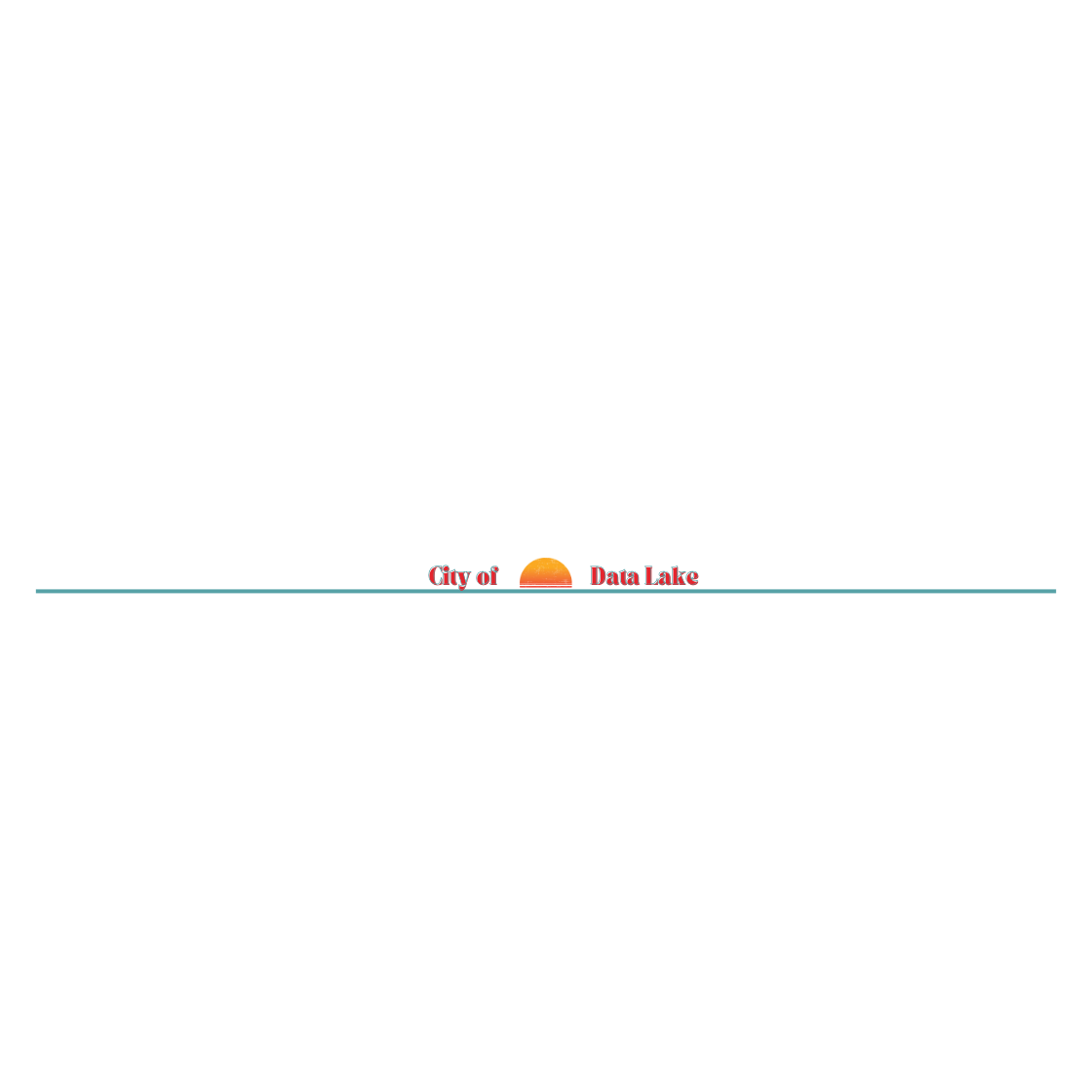
Soren, a Project Manager for the City of Data Lake, asks about your preliminary desk research. Write a short summary (6-7 sentences or bullet points) on remote work for him. Include: what questions researchers are asking, what kinds of work tends to hire remote employees (or not), what data has been collected, and what questions the research raises for you. This exercise is simply meant to help you collect your thoughts!

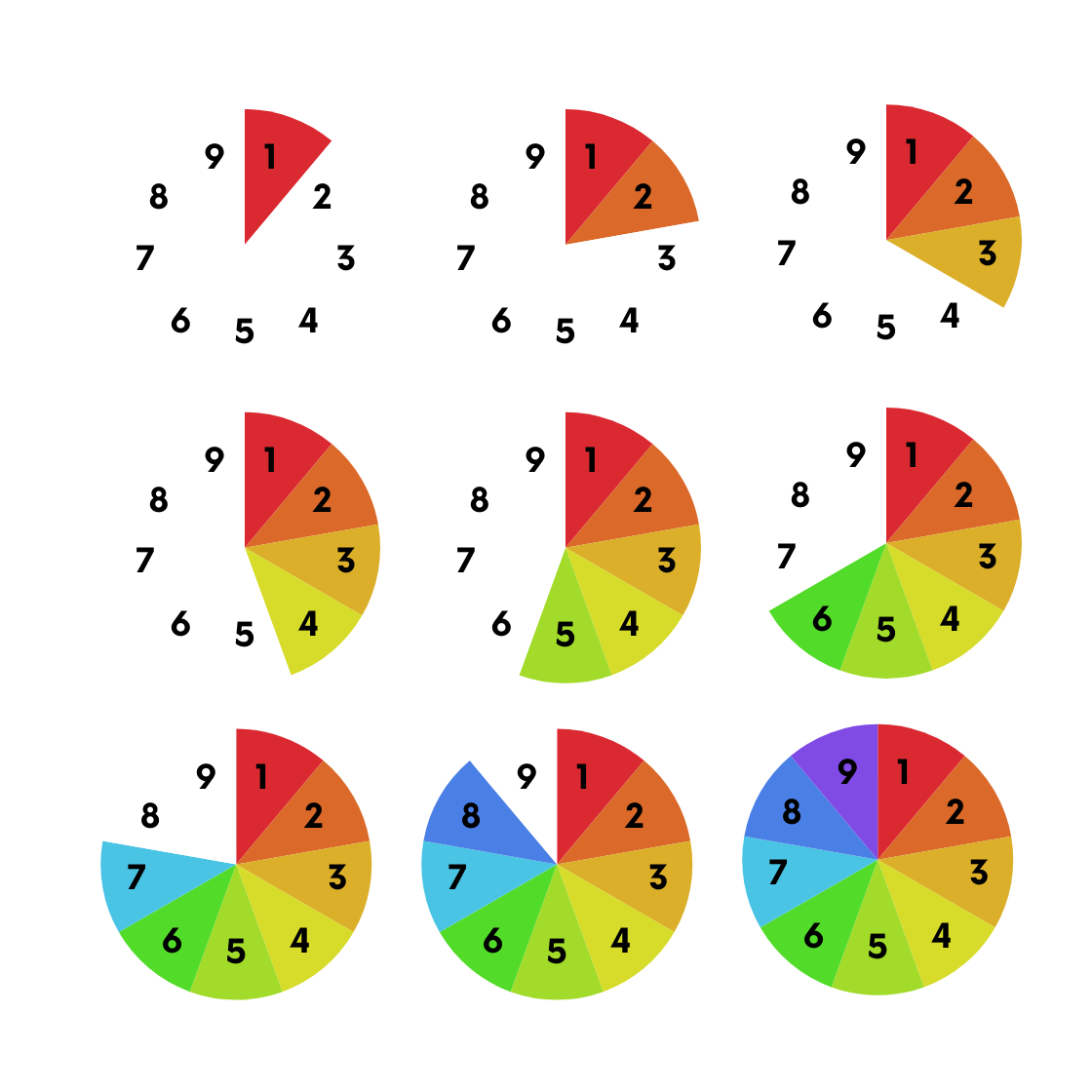
|  |
| --- |
| * Remote jobs increases productivity in sectors like finance and IT, especially in task when you have to focus in an individual task. * Make employees less prone to quit jobs, they can get mad if they are force to get back to full-time office and loose more productivity than in their in-office times. * It can downgrade their soft skills like verbal communication and their productivity in team creative tasks. * They spend less in commuting and make the enterprise spend less in their time in the office. * Is possible to have fake job applicants. * Some of them still struggle with their domestic/social life. |

Hint: As above, it will be tempting to use AI to help you write this summary. Try not to do this - you will get more out of this exercise if you do the writing yourself. Don’t worry about writing well, just pick out some key points and share your research findings here.

Since you and Soren work for a city government, you should also understand how remote work has impacted other cities. Include below 3-4 pieces of information (findings) about remote work that will be useful for your colleagues to understand (we provided one insight to help get you started).

|  |  |
| --- | --- |
| (Example) | Remote workers are choosing to live in small or mid-sized cities for a variety of reasons, including: greater work-life balance, better cost of living, access to outdoor spaces, and because of incentive programs ([source](https://www.forbes.com/sites/lizfarmer/2021/09/15/the-small-cities-capturing-remote-workers/)). New residents are improving these cities’ tax base and can help revitalize “brain drain” zones. |
| Finding 1 | Young people like remote jobs and remote/hybrid employee’s trend to quit if they are force to come back to the office at full time. |
| Finding 2 | Formation programs are still needed and their social skills need more training. |
| Finding 3 | Traffic haven’t reduce so much during these years, but morning peak hours aren’t so problematic as before. |



**You’ve completed the 1st Milestone!**

Great work! You established a baseline of knowledge about the topic of remote work policies. This “desk research” should help you better understand and recognize patterns in the dataset you’re about to analyze.

**Next Steps:**

On the next page, we’ll tackle Milestone 2: Finding & Sourcing Quality Data. In this Milestone, we’ll learn about how we can identify datasets that will help us accurately answer some of the questions raised by our desk research.

**Milestone 2 - Finding & Sourcing Quality Data**

**~3 hours**



Now that we’ve done our desk research, we can start looking for data that we can analyze ourselves. A good dataset is essential to any research project and it will help us achieve high accuracy in our report on remote work for Mayor Ellation.

Note: Your Data Lake team will provide you with a quality dataset later on in this project. However, you should know about other ways to find data for future research projects. Let’s learn about data sourcing!

[****](https://youtu.be/yOGxrNt55Io)

[**Video Link**](https://youtu.be/yOGxrNt55Io)

Data Sourcing:

We can find data in a few different ways. Two commons methods include:

|  |  |  |
| --- | --- | --- |
| **Method** | **Advantages / Disadvantages** | **Tips** |
| Internet Search | Fast, easy / Not always reliable or accurate | - Here are [some great sources for free datasets](https://www.atlassian.com/data/business-intelligence/free-datasets) easily accessible online.  - Here are some [other places where you can find reliable data](https://resources.nu.edu/researchprocess/datasets#:~:text=Most%20datasets%20can%20be%20located,a%20good%20place%20to%20look.). |
| Library Research | Under-rated, can answer very specific research questions, easily accessed through a University or local library | - For example, Columbia University has several [data consultation programs for current students](https://library.columbia.edu/services/research-data-services.html). Your school [may also give you access to external platforms like Statista with a login](https://library.columbia.edu/collections/eresources/databases.html).  - The Library of Congress also has [a useful set of links to various dataset repositories](https://guides.loc.gov/datasets/repositories#s-lib-ctab-22713457-0). |

You can also collect your own data - we’ll cover that later!

Data Vetting:

Not all data is created equal. First, we want to consider where we are getting our data from. Is this a legitimate and trustworthy source? Here is some [useful advice on how to find reliable data sources](https://www.datatopolicy.org/navigator/identify-data-sources-ensure-reliability).

There are some questions we should always ask ourselves when reviewing a dataset. Those include:

* Are the dataset’s sources clearly documented?
* Looking at the dataset’s collaborators, do we see any conflicts of interest?
* Who is funding this research (e.g. the government, companies who may have an interest in the research showing a specific result, academic organizations)?
* Where does the data itself come from (e.g., a survey, scientific measurements, algorithms)?
* Do other reputable sources cite this data?

Identifying “Bad Data”

We ask these questions because we want to know if there is “bad data” in our dataset. [Bad data can be data that is incomplete, outdated, inconsistent, or unstructured](https://medium.com/@shang.eng2018/how-to-spot-bad-data-in-your-datasets-9bc728d42e53).

[Sometimes bad data can be fixed](https://www.datacamp.com/blog/10-signs-bad-data-quality#:~:text=our%20online%20course.-,How%20to%20Manage%20Bad%20Data,-Hopefully%2C%20you%20are), and sometimes not! All in all, we want to make sure that we don’t use poor quality data because this can hurt the accuracy and validity of our analysis.

There are several things we want to review to determine if data is “bad.” We should [read the data documentation](https://data.library.arizona.edu/data-management/best-practices/data-documentation-readme-metadata) to review:

|  |  |
| --- | --- |
| [Sample Size](https://www.qualtrics.com/en-gb/experience-management/research/determine-sample-size/) and Population Sample | Is the [data set large enough](https://medium.com/@analyticsemergingindia/decoding-data-size-pros-and-cons-of-working-with-small-data-sets-bc1ea0792da6) to show statistical significance? Is the dataset representative of a given population?  Hint: Small sample sizes can lead to [overfitting in Machine Learning](https://www.kaggle.com/code/rafjaa/dealing-with-very-small-datasets) and may lead you to over-generalize in your analysis. |
| Survey Length | If our research topic is defined by long periods of time (e.g. economic recessions, children’s educational development), does the study cover change over time? How much time is sufficient for our research? |
| Survey Variables | What does the study claim to measure? Is that variable measurable, either directly or indirectly?  Example: The World Happiness Report does an annual review of global citizens’ “happiness” rates. Happiness, however, is a relative and subjective measurement! [Researchers break down the question of “happiness” into several metrics](https://worldhappiness.report/data/), including self-reported emotional satisfaction as well as six other variables more easily measured (e.g., GDP, administrative corruption, legal freedoms). The result is a composite measurement of happiness. |
| Scope | Sometimes, a dataset might be interesting, but it’s simply not relevant to our research.  Example: A well-structured dataset on remote workers’ favorite colors might be interesting and strictly related to our research for the city of Data Lake, but it won’t help us make useful recommendations for Mayor Ellation. |
| Data Documentation | Does the owner share clearly written data documentation? Is the source of the data set neatly outlined? Is the data itself well-structured and easy to work with? If not, that might set off alarm bells for you. |
| Date | Depending on what you’re researching, data that is more than a few years old may no longer be relevant. Generally, the more recent the data collection, the better. |
| Source | Who collected and published this data? Do we trust that they are capable of processing the data?  Example: Data sourced from the U.S. Census Bureau is probably quite reliable. We can imagine that data from “RemoteWorkIsTheBestWork.com” is probably biased and may not have been collected with rigorous and scientific methods. |

Task 3: Vetting Data Sources



Khaled, a Data Engineer for the City of Data Lake, gives you the below list of datasets. He asks you to consider one of the following for your research on remote workers. Using the information you learned on data vetting, review this list and identify the potential advantages and disadvantages of each dataset. For each dataset that you discard, note why you think it’s not a good dataset.

* [U.S. Survey of Working Arrangements and Attitudes (SWAA)](https://wfhresearch.com/data/)

|  |  |
| --- | --- |
| **Advantages** | **Disadvantages** |
| Reputable source, has great consistency over time, cited by other reputable sources ([link](https://www.stlouisfed.org/on-the-economy/2024/dec/trends-work-from-home-us-insights-six-datasets)) | Missing some important variables like family size, income, etc. that we’d want to know about |

* [Remote Work Productivity from Kaggle](https://www.kaggle.com/datasets/mrsimple07/remote-work-productivity)

|  |  |
| --- | --- |
| **Advantages** | **Disadvantages** |
| Consistent over time, easy to manipulate as an example dataset, clean and large enough. | Unknown sources, some only one measurement is objective and has a backup. Employee ID is not useful at all. |

* [NSW Remote Working Survey](https://data.nsw.gov.au/data/dataset/nsw-remote-working-survey)

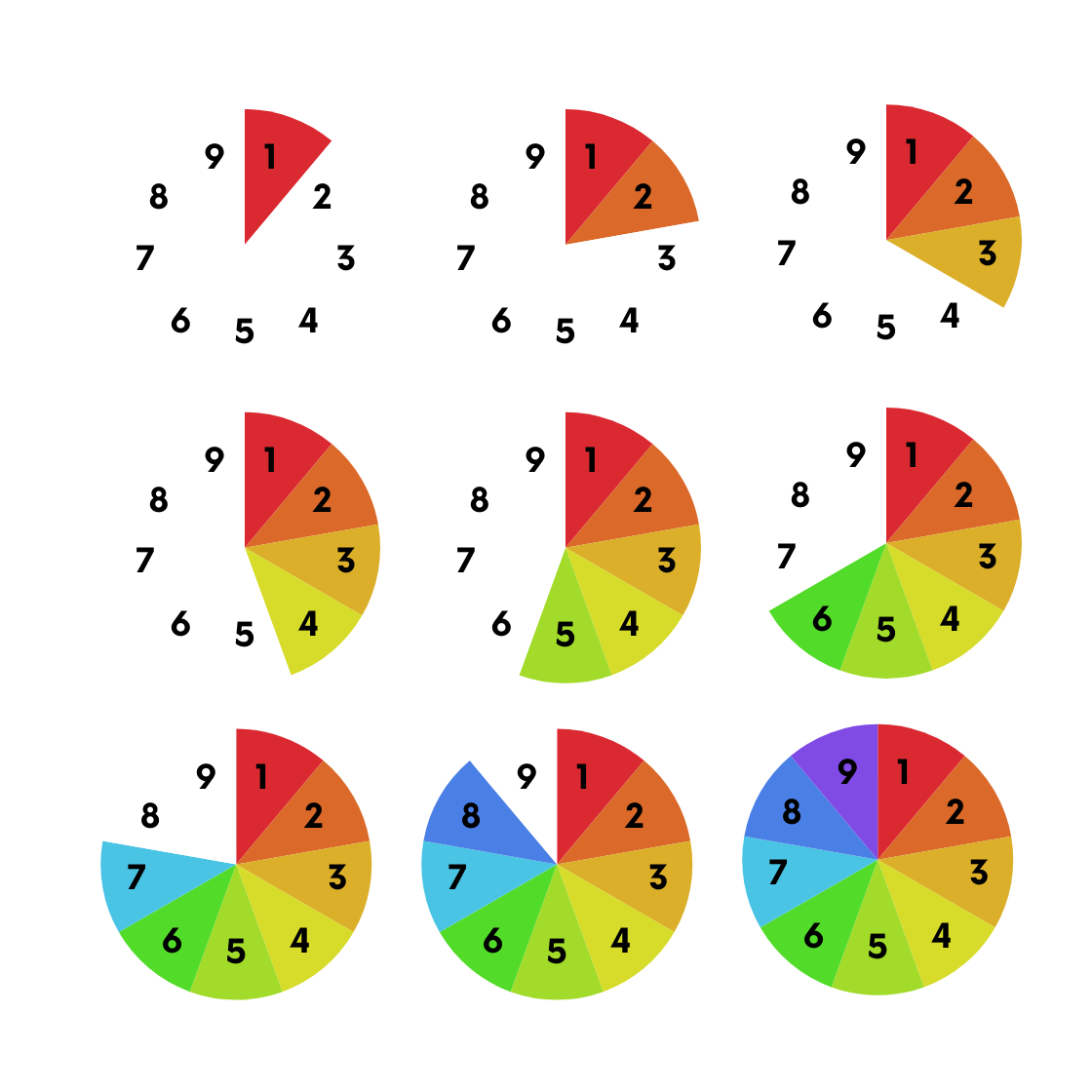
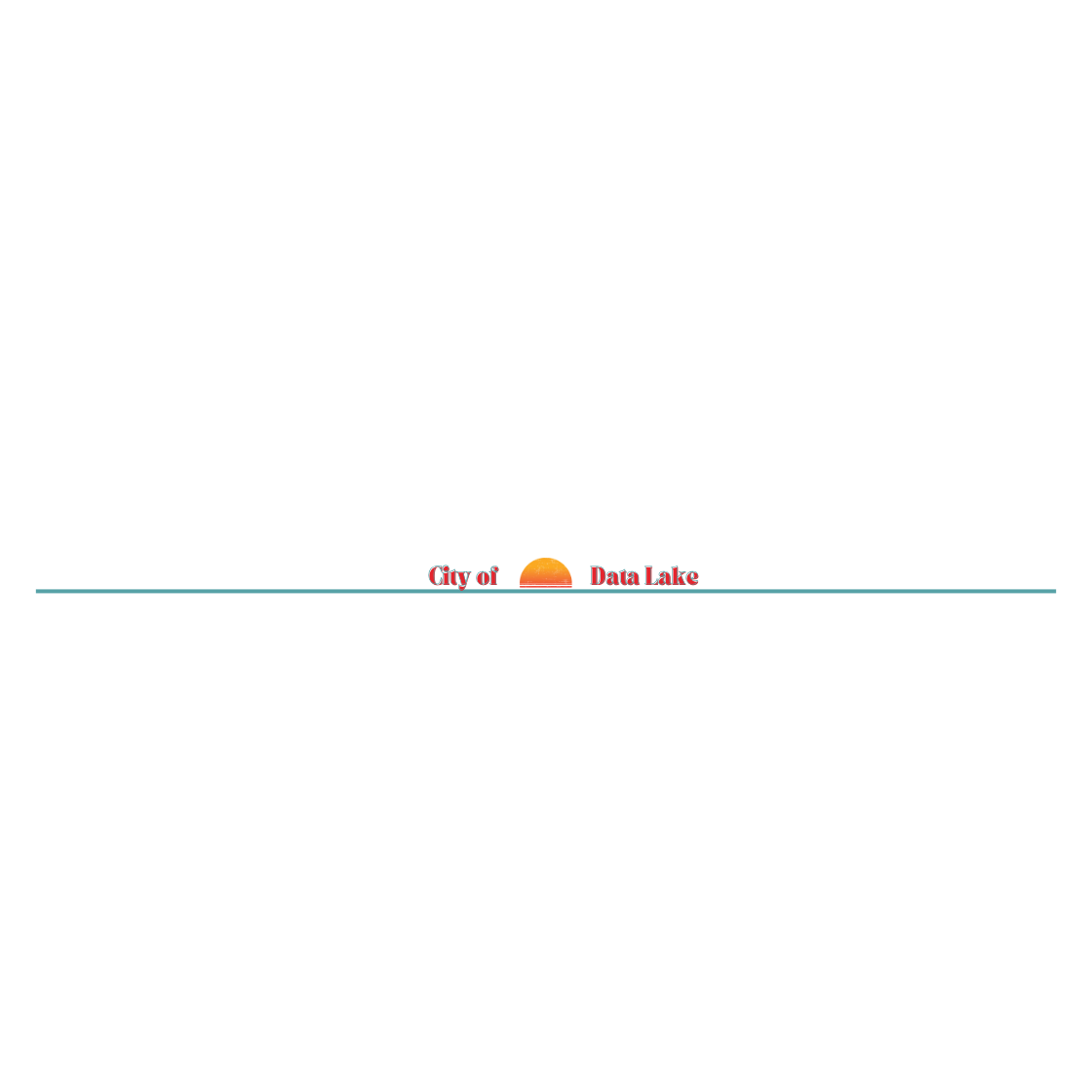
|  |  |
| --- | --- |
| **Advantages** | **Disadvantages** |
| Large amounts of data with a lot of categories. Trusted sources. Has info about the topic of my research | A lot of noise due to the excess of columns and their names. Some metrics need to be cleaned or have null values. The data is from 3-5 years ago. |

* [Household Pulse Survey: Measuring Emergent Social and Economic Matters Facing U.S. Households](https://www.census.gov/data/experimental-data-products/household-pulse-survey.html) (specifically, [this survey](https://www.census.gov/data/tables/2024/demo/hhp/cycle09.html))

|  |  |
| --- | --- |
| **Advantages** | **Disadvantages** |
| Large enough, easy to manipulate and filter. Explanatory documentation and trusted source. | The information is not useful enough for my research about remote jobs. |

* [Working From Home Caused a Spatial Shift in Daytime Population Away From Traditional Job Centers](https://www.census.gov/library/stories/2023/02/remote-work-during-pandemic-shifted-daytime-population.html) (Note: you will not be able to see the data directly from this article - that’s fine!)

|  |  |
| --- | --- |
| **Advantages** | **Disadvantages** |
| Meaningful, already cleaned and diverse data. Extremely documented and fresh. | Its dispersed all over the article, some datasets are too small or their fields are repeated. |



**You’ve completed the 2nd Milestone!**

Excellent work! We’ve learned how to identify the datasets and data sources that are most likely to help us with our research. We now know that every dataset has its potential risks and advantages - sometimes we will need to merge multiple datasets to answer our particular question.

**Next Steps:**

Up next is Milestone 3: Diving into the Documentation & Identifying Bias. In this Milestone, we’ll look at the specific dataset that Khaled has picked for us to work with.

**Milestone 3 - Diving into the Documentation & Identifying Bias**

**~2 hours** 

You and Khaled agree that you should work with the dataset he sourced from the Census Bureau website for this remote work analysis. You can now explore the survey’s documentation. This process will give you a sense of the type of analysis you will be able to perform and what types of questions you’ll be able to ask in your report.

Begin by reviewing the [U.S. Census Bureau’s documentation](https://www.census.gov/data/experimental-data-products/household-pulse-survey.html) on this survey. For this analysis, you will use the [data from Phase 4.2, Cycle 9](https://www.census.gov/data/tables/2024/demo/hhp/cycle09.html) (collected August 20 – September 16, published October 3, 2024).

[****](https://youtu.be/qgYkS3XSIoQ)

[**Video Link**](https://youtu.be/qgYkS3XSIoQ)

Task 4: Interpreting a Survey Questionnaire

Next, we’ll want to read the questionnaire that was used to collect our survey responses. This is important because we need to understand how our data was collected and what participants were thinking as they answered our questions.

Briefly read through the Phase 4.2 questionnaire (available in [English](https://www2.census.gov/programs-surveys/demo/technical-documentation/hhp/Phase_4-2_HPS_Questionnaire_ENGLISH.pdf) and [Spanish](https://www2.census.gov/programs-surveys/demo/technical-documentation/hhp/Phase_4-2_HPS_Questionnaire_SPANISH.pdf)). Make some notes as you read, writing down anything that sticks out to you.

|  |
| --- |
| * Questions should be short and clear * Separated sentences are more legible * Is better to give the context and the goals of the survey * Estimated time required should be provided among the number of questions * Questions should coverage all the data needed for the goals |

You may notice that this is a well-written questionnaire, suitable for digital responses. There are many reasons why this survey works well, including:

* Researchers clearly articulated their goals and objectives for the survey at the beginning of the document. These key principles ensure that no questions are extraneous or outside the scope of the study.
* The survey is available in multiple languages. This ensures that the survey is accessible to a diverse group of participants.
* The survey was designed with a specific amount of time in mind (20 minutes). In general, the shorter the survey, the more responses you get.
* The survey includes a clear disclaimer about privacy and identity at the top. Links to other online resources and contact information have been provided for respondents.
  + This survey was written with an IRB (Institutional Review Board)’s approval. [What is an IRB?](https://www.youtube.com/watch?v=U8fme1boEbE) (video)
* The survey asks many demographic questions, including information about participants’ age, race, gender, sexuality, and household size. Later, the survey asks for other insightful information about participants’ disability status and location.
* Because of the sensitive nature of some of the survey questions, we also see that the researchers included key resources at the bottom of the survey, including materials for mental health and unemployment support.

Survey Bias

Despite our best efforts, all datasets include [some kind of bias](https://www.youtube.com/playlist?list=PLNs9ZO9jGtUDZONu0E2zVHOvtjiuqwtDe). This is particularly true of datasets that collect some subjective data, like this dataset does. This data is both quantitative (numerical data, such as the number of residents in a given home) and qualitative fields in the questionnaire (subjective and non-numerical data, such as an open-ended question about why a family might choose to relocate).

There are several types [of Survey Bias](https://www.qualtrics.com/experience-management/research/survey-bias/) that we need to be aware of:.

* Selection Bias / Sampling Bias: When your data is not representative of a given population.
  + We want our data to reflect the average family as well as a subset of individuals who can show us what trends and interests are unique to the remote worker population.
* Interviewer Bias: Bias introduced by the actions of the interviewer.
  + We don’t want questions to suggest a preferred answer. For example, the question “What is your main *reason* for not working for pay or profit?” is more neutral than “What is your main *excuse* for not working for pay or profit?”. A respondent might not feel they could answer without judgement if the second question was asked instead. Subtle adjustments to language can help us reduce Interviewer Bias.
* Non-response Bias: For any number of reasons, respondents may not complete your survey. We want to eliminate barriers to people completing our questionnaire.
  + In this case, the survey was conducted electronically and was designed to take 20 minutes or less to complete. The research purpose of the survey was also clearly explained, increasing the likelihood that respondents would complete the survey.

If you’re interested in learning more about the impacts of bias on AI development, we encourage you to watch [this video built from IBM’s Open DS4All materials](https://www.youtube.com/watch?v=wL53xKmnruw).

Task 5: Crafting a Quantitative Questionnaire

Another Data Lake resident and colleague, Abeni, asks you for help with a quantitative survey she is designing. She is concerned that some of her draft questions for a survey on public health may prompt Bias in respondents. Prepare to help her with this project by [reading about best practices for writing survey questions](https://www.qualtrics.com/blog/writing-survey-questions/). You’ll learn about leading words, mutually exclusive language, and more. 

Next, take a look at the below survey questions. Identify the problems with the way these questions are phrased. Explain what would make them stronger survey questions and rewrite each question.

1. How much do you agree with the following statement?: “The government should force citizens to attend an annual wellness checkup.” (select one)

* Strongly agree
* Agree
* Neutral
* Disagree
* Strongly disagree

|  |  |
| --- | --- |
| Type of bias this question may prompt: | This could prompt interview bias because of the use of suggestive words like “force.” Someone might agree that certain healthcare mandates should be required to remain employed in particular civil service roles, for example, but “force” has a strong negative implication, perhaps prompting that person to select “strongly disagree.” |
| Recommended edits: | Rephrase the question as: “The federal government should mandate certain preventative healthcare measures (with appropriate medical exceptions).” |

1. What are your thoughts about your current healthcare provider? (open response)

|  |  |
| --- | --- |
| Type of bias this question may prompt: | The question is too ambiguous. The target person could think too much about what to answer. Is not provided if the question what its goal is. It can also lead to useless answers |
| Recommended edits: | Emphasize the question with the quality of your current healthcare provider services. Maybe do more questions if we want to get more information about people’s thoughts about it. |

1. Have you received any vaccines in the last 6 months? (select one)

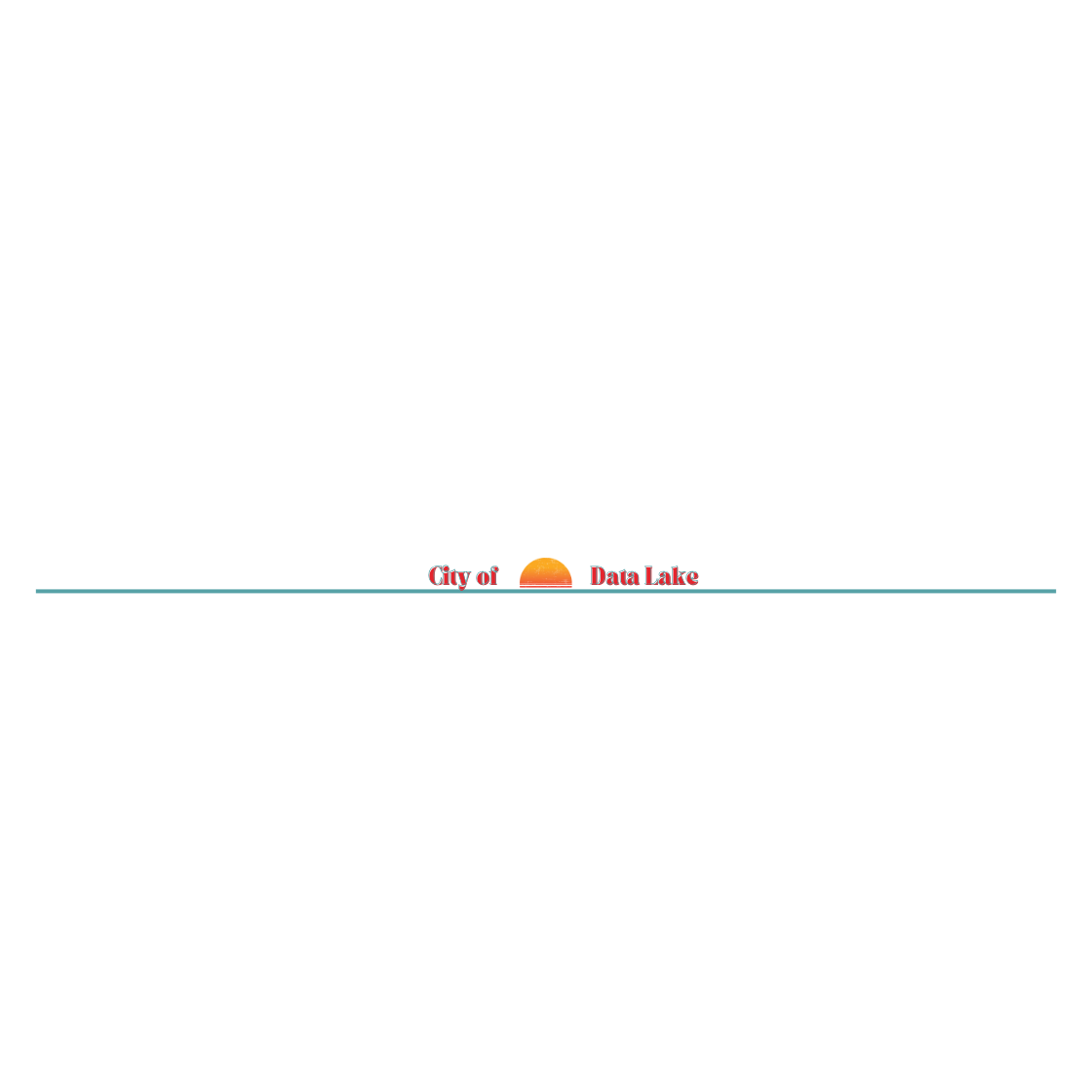
* Yes
* No

|  |  |
| --- | --- |
| Type of bias this question may prompt: | Is not normal to receive multiple vaccines in 6 months. Even one its rare in a year. Some people would prefer not to answer |
| Recommended edits: | Ask them if they received a vaccine in the last 6 months. And also add the “prefer not to answer” option. Or even maybe a list of ranges of the last time you received a vaccine. |

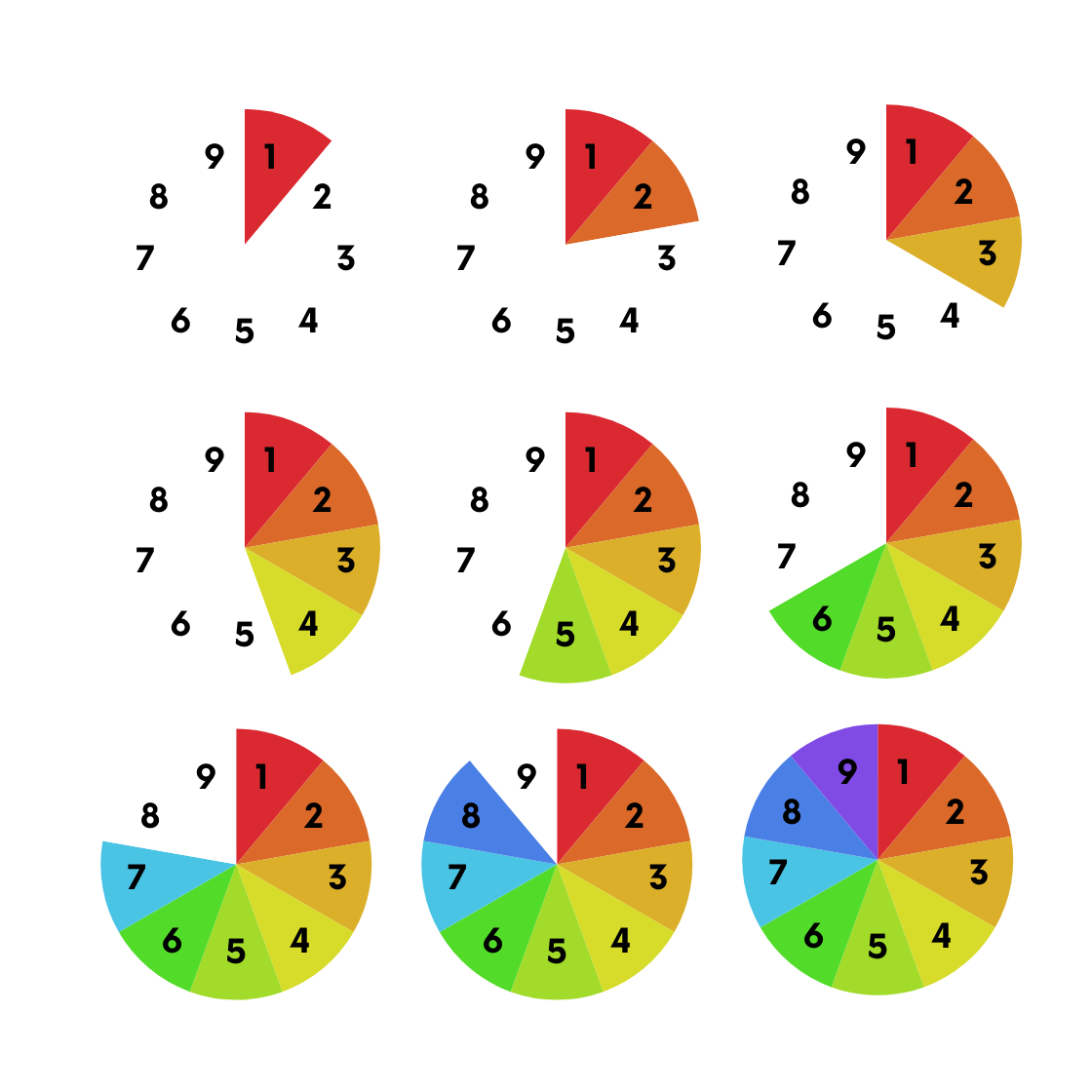
1. Approximately how much time do you spend watching or reading health-related news each day? (select one)

* Less than 1 hour
* 1-3 hours
* 3-5 hours
* More than 5 hours

|  |  |
| --- | --- |
| Type of bias this question may prompt: | Repeated mutually exclusive choices |
| Recommended edits: | Make the times smaller. The 2nd and 3rd options limits should not coincidere |



**You’ve completed the 3rd Milestone!**



Great work! We’ve done a bit more research into what makes for a good research survey and identified some of the potential pitfalls of bias. Knowing specifically how the Census Bureau’s dataset and questionnaire were structured has made us more prepared for our analysis.

**Next Steps:**

We’ve reached the end of Document A. [Please take a moment to complete this very short, five question Google Survey on Milestones 1 - 3](https://forms.gle/B13yzt62xdR6UhMu7). Your input here will help our team understand what resources helped you learn (or not) and what lessons you found particularly helpful.

Next, we’ll switch to Document B, which includes Milestones 4 & 5. We’ll get into the data itself and begin the data cleaning process.