Introduction to the Study Session

In this study, you will be tasked with answering a series of questions using the Jupyter Notebooks provided. Some are in the standard, 1D, top-to-bottom list of cells format you are used to. Others are in a multi-column, 2D format.

We are measuring both how long it takes you to complete each question AND your accuracy on each question. Thus, try to complete the questions **accurately** and **quickly**.

This session should take approximately 1 hour to complete.

We do not anticipate any risks from completing this study.

You can choose whether to be in this study or not. If you volunteer to be in this study, you may withdraw at any time without consequences of any kind. The investigator may withdraw you from this research if circumstances arise which warrant doing so, such as providing false information or being under the age of 18.

For completing this user study session, you will be compensated **$20** in the form of an **Amazon gift card.**

Please note that, for this study, you are NOT allowed to do the following:
* Move cells around.
* Delete cells.
* Add cells.

After entering your email, which is needed to send you the compensation, please press the START button to begin.

* Name (First and last name, e.g. John Doe)

* What is your email?

TASK: Finding & Comparing Results in 1D

Open the tab with the 1D COVID Analysis notebook but do NOT look over it yet.

You will be asked questions that require comparing the results starting in Section 4. Please make sure to read each question carefully.

When you are ready to begin, press the NEXT button.

You may look over the 1D COVID Analysis notebook now.

Which state's analysis is found between the analysis of New York data and the analysis of Mississippi data?

- Florida
- All States
- Nevada
- New Jersey
- Missouri

- ◯ Ohio
- ◯ Virginia

* Note that there are 3 bar charts in each section, starting with Section 4; for this question, only consider Sections 5-9. Look at the relevant bar charts to answer the following question:

  Out of those shown in the relevant bar charts, **which *county* in which state,** EXCLUDING the ALL STATES section, had the highest number for **deaths** of COVID-19? Example Answer: Blacksburg, Virginia

* Look at the scatterplots, which are in Sections 4-9, and each one's associated value for the coefficient of determination (how well the line of best fit fits the data) to answer the following question:

  Which section's scatterplot graph's line of best fit **least fits** the data (coefficient of determination **closest to 0**)?

- ◯ All States
- ◯ Florida
- ◯ Mississippi
- ◯ Ohio
- ◯ New York
- ◯ Missouri

TASK: Parameter Tuning in 1D

Open the tab with the 1D KNN Parameter Tuning notebook and briefly look over it.

Do NOT move any of the cells in this notebook.

You will be asked questions that require tuning the parameter "k" in Section 1 and choosing the distance metric in Section 4. Only run the necessary cells (the "k-value" cell in Section 1, and the cells in Section 4) to test each possible parameter set (k-value and distance metric).

You will be evaluating each parameter set (k-value and distance metric) based on the generated accuracy of the model on the test dataset. There is a cell near the end of the notebook which will generate the accuracy score as the following fraction: number correctly predicted / total number of test instances.

Please make sure to read each question carefully.

When you are ready to begin, press the NEXT button.

REMINDERS:
- The "k-value" cell is in Section 1.
- The distance metric cell is in Section 4.
- Only run the necessary cells (the "k-value" cell in Section 1, and the cells in Section 4) to test each parameter set (k-value and distance metric).
- The cell which outputs the accuracy is in Section 4.
- The accuracies are fairly close, so take notes on paper if necessary.

---

* Which of the following k-values produces the most accurate model with the given dataset for the **Euclidean** distance metric?

◯ 43

- ◯ 47
- ◯ 51
- ◯ 55
- ◯ 59

---

* Which of the following k-values produces the most accurate model with the given dataset for the **Manhattan** distance metric?

- ◯ 43
- ◯ 47
- ◯ 51
- ◯ 55
- ◯ 59

* Given each **distance metric with its optimal k-value**, which distance metric produces the **most accurate model** on the given dataset?

- ◯ Euclidean
- ◯ Manhattan

TASK: Code & Results Comparison in 1D

Open the tab with the 1D KNN Code & Result Comparison notebook but DO NOT look over it until you have read the question on the next page.

Do not move any of the cells in this notebook.

Please make sure to read the question carefully.

When you are ready to begin, press the NEXT button.

Compare the code from the two analyses in Sections 2 & 3, respectively, to answer the following question:

Which of the following items appear differently between the two analyses?

- [ ] The use of the head (data.head()) or tail (data.tail()) of the data
- [ ] The numbers assigned in the conversion of "stabf" class names from string to numeric
- [ ] The cutoff number for the training and testing splits (e.g. my_data.iloc[:555] means the cutoff number is 555)
- [ ] Whether the data is normalized or not
- [ ] Different distance metrics (Manhattan, Euclidean) used
- [ ] Same distance metric but different code for calculating it
- [ ] The variable name for the distance matrix
- [ ] The value of k (number of nearest neighbors)
- [ ] The text of the print message showing the accuracy of the model

* Post-1D Survey

Please state your level of agreement or disagreement for the following statements based on your experience in this study with the 1D notebooks.

	Strongly Agree	Agree	Agree a little	Neutral	Disagree a little	Disagree	Strongly Disagree

It was easy to navigate the 1D notebooks.	○	○	○	○	○	○	○
I could quickly find the relevant information in the 1D notebooks.	○	○	○	○	○	○	○
It was easy to make comparisons between visuals in the 1D notebooks.	○	○	○	○	○	○	○
It was easy to make comparisons between numerical results in the 1D notebooks.	○	○	○	○	○	○	○
It was easy to make comparisons between different sections of code and results in the 1D notebooks.	○	○	○	○	○	○	○

TASK: Finding & Comparing Results in 2D

Open the tab with the 2D COVID Analysis notebook but do NOT look over it yet.

You will be asked questions that require comparing the results starting in Section 4. Please make sure to read each question carefully.

When you are ready to begin, press the NEXT button.

You may look over the 2D COVID Analysis notebook now.

Which state's analysis is found between the analysis of Pennsylvania data and the analysis of Virginia data?

- ○ Texas
- ○ Maryland
- ○ Washington
- ○ Illinois
- ○ All States

- ◯ Maine
- ◯ Georgia

* Note that there are 3 bar charts in each section, starting with Section 4; for this task, only consider Sections 5-9. Look at the relevant bar charts to answer the following question:

  Out of those shown in the relevant bar charts, **which *county* in which state**, EXCLUDING the ALL STATES section, had the highest number for **deaths per case** of COVID-19? Example Answer: Blacksburg, Virginia

* Look at the scatterplots, which are in Sections 4-9, and each one's associated value for the coefficient of determination (how well the line of best fit fits the data) to answer the following question:

  Which section's scatterplot graph's line of best fit **best fits** the data (coefficient of determination **closest to 1**)?

- ◯ All States
- ◯ Texas
- ◯ Georgia
- ◯ Virginia
- ◯ Illinois
- ◯ Pennsylvania

TASK: Parameter Tuning in 2D

Open the tab with the 2D KNN Parameter Tuning notebook and briefly look over it.

Do NOT move any of the cells in this notebook.

You will be asked questions that require tuning the parameter "k" in Section 1 and choosing the distance metric in Section 4. Only run the necessary cells (the "k-value" cell in Section 1, and the cells in Section 4) to test each possible parameter set (k-value and distance metric).

You will be evaluating each parameter set (k-value and distance metric) based on the generated accuracy of the model on the test dataset. There is a cell near the end of the notebook which will generate the accuracy score as the following fraction: number correctly predicted / total number of test instances.

Please make sure to read each question carefully.

When you are ready to begin, press the NEXT button.

REMINDERS:
- The "k-value" cell is in Section 1.
- The distance metric cell is in Section 4.
- Only run the necessary cells (the "k-value" cell in Section 1, and the cells in Section 4) to test each parameter set (k-value and distance metric).
- The cell which outputs the accuracy is in Section 4.
- The accuracies are fairly close, so take notes on paper if necessary.

---

* Which of the following k-values produces the most accurate model with the given dataset for the **Euclidean** distance metric?

◯ 45

- ◯ 49
- ◯ 53
- ◯ 57
- ◯ 61

* Which of the following k-values produces the most accurate model with the given dataset for the **Manhattan** distance metric?

- ◯ 45
- ◯ 49
- ◯ 53
- ◯ 57
- ◯ 61

* Given each **distance metric with its optimal k-value**, which distance metric produces the **most accurate model** on the given dataset?

- ◯ Euclidean
- ◯ Manhattan

TASK: Code Comparison in 2D

Open the tab with the 2D Code Comparison notebook but DO NOT look over it.

Do not move any of the cells in this notebook.

Please make sure to read the question carefully.

When you are ready to begin, press the NEXT button.

Compare the code from the two analyses in Sections 2 & 3, respectively, to answer the following question:

Which of the following items appear differently between the two analyses?

- [ ] The use of the head (data.head()) or tail (data.tail()) of the data
- [ ] The numbers assigned in the conversion of "stabf" class names from string to numeric
- [ ] The cutoff number for the training and testing splits (e.g. my_data.iloc[:555] means the cutoff number is 555)
- [ ] Whether the data is normalized or not
- [ ] Different distance metrics (Manhattan, Euclidean) used
- [ ] Same distance metric but different code for calculating it
- [ ] The variable name for the distance matrix
- [ ] The value of k (number of nearest neighbors)
- [ ] The text of the print message showing the accuracy of the model

Post-2D Survey

Please state your level of agreement or disagreement with the following statements based on your experience in this study with the 2D notebooks.

	Strongly Agree	Agree	Agree a little	Neutral	Disagree a little	Disagree	Strongly Disagree
It was easy to navigate the 2D notebooks.	○	○	○	○	○	○	○

I could quickly find the relevant information in the 2D notebooks.	○	○	○	○	○	○	○
It was easy to make comparisons between visuals in the 2D notebooks.	○	○	○	○	○	○	○
It was easy to make comparisons between numerical results in the 2D notebooks.	○	○	○	○	○	○	○
It was easy to make comparisons between different sections of code and results in the 2D notebooks.	○	○	○	○	○	○	○

POST-TASKS SURVEY

Now that you have completed the tasks, the last part is to fill out a brief post-task survey on your experiences with 1D and 2D computational notebooks.

Please read each question carefully.

REMINDER:
1D Notebooks refers to the standard Jupyter Notebook format of a top-to-bottom list of cells.
2D Notebooks refers to the multi-column 2D format Jupyter Notebooks.

* Based on your experiences in this study, please state your level of agreement or disagreement related to the following question:

Compared to 1D Notebooks, I believe 2D Notebooks would be more useful for the following tasks:

	Strongly Agree	Agree	Agree a little	Neutral	Disagree a little	Disagree	Strongly Disagree
Navigating through the notebook	○	○	○	○	○	○	○
Locating items (code, results) in the notebook	○	○	○	○	○	○	○

Organizing/cleaning a notebook	○	○	○	○	○	○	○
Presenting a notebook	○	○	○	○	○	○	○
Data exploration and preparation	○	○	○	○	○	○	○
Performing analysis and development	○	○	○	○	○	○	○
Debugging	○	○	○	○	○	○	○
Comparing results	○	○	○	○	○	○	○
Collaborating on a shared notebook	○	○	○	○	○	○	○

* Based on your experiences in this study, please state your level of agreement or disagreement with the following statement:

	Strongly Agree	Agree	Agree a little	Neutral	Disagree a little	Disagree	Strongly Disagree
I feel that the spatial layout of the cells and sections in the 2D notebooks improved my performance on the tasks.	○	○	○	○	○	○	○
I feel that having more of the notebook cells on the screen in the 2D notebooks improved my performance on the tasks.	○	○	○	○	○	○	○
I feel that the 2D notebooks better utilized the screen space I had for this study.	○	○	○	○	○	○	○
If I had the choice, I would use 2D layouts (e.g. multi-column) instead of the 1D, 1-column layout.	○	○	○	○	○	○	○

If you would like to elaborate on your answers to the survey questions or add any final comments, please do so here.

Thank you for completing the study!