Data Analyst Test

#### Fill in your details here

| Name | **Alphaeus Mutemi** |
| --- | --- |
| Email address | **alphaeusmutemi@gmail.com** |
| Date completed |  |

#### Once you have completed the test, send an email to Mourine Musee [mmusee@mercycorps.org](mailto:mmusee@mercycorps.org) to confirm.

#### WRITTEN TEST

This test contains 4 tasks, measuring the skills and competencies of the Data Analyst candidates on their ability to work through data analysis on Excel, presentation in Word and PowerPoint, data manipulation & management, and statistics & data science.

Aim to complete the test in **2 to 3 hours**.

Consider the following when completing the tasks:

1. Show all of your workings. Workings are more important than the correct answer. **Answers without workings will not be considered**. Workings do not have to be shown in this document, as long as they are in the accompanying Excel file or in your written code.
2. Use formulas or scripts as much as possible. **Manual data manipulation will not be considered favorably.**
3. Upload **all files** you create into the folder that is relevant to the task. Only files in your Google Drive folder will be marked.
4. You do not have to work online. Please feel free to download all the files and work locally, then upload/replace the files on the drive at the end.

**Task 1 - Data analysis 1:**

1. Open the Excel file called **Task 1 & 2\_Survey Data.xlsx** in the **Task 1 & 2** folder.

*This data is from an endline survey at the end of the project. The objective is to understand the impact that the project has had on its female participants.*

1. In the space below, produce a maximum 2-page summary of the below **key insights** in this dataset. Perform any analysis in the **Analysis** sheet in the Excel file. The summary should contain written analysis and figures as needed. Highlight what you believe could be important information for decision making.

The **key insights** to include in your summary are:

* 1. Average time taken per interview. Are there any outliers?
  2. Statistics about, and comparison between, high and low season income.
  3. Comparison: increase in profits by location.
  4. Loan acquisition trends.
  5. Perceived freedom of decision making and agency.

**Task 2 - Data analysis 2:**

* 1. Copy the below data into a **new sheet** within the same Excel file**.** Examine the data which summarises responses from a survey question on income. The numbers in the columns represent the number of households for each response.

| Governorate | The household has diverse and reliable income sources.  (2 or more) | The household only has 1 income source | The household doesn’t have any reliable income source |
| --- | --- | --- | --- |
| Governorate 1 | 12 | 51 | 73 |
| Governorate 2 | 30 | 65 | 51 |
| Governorate 3 | 130 | 90 | 30 |
| Governorate 4 | 33 | 60 | 40 |
| Governorate 5 | 95 | 88 | 26 |
| Governorate 6 | 28 | 110 | 56 |
| Governorate 7 | 22 | 77 | 60 |
| Governorate 8 | 19 | 66 | 39 |

* 1. Add columns and rows as needed and use the appropriate formulas to calculate the following (add answers below each bullet point here):
     + - 1. Total assessed households?
* *There were a total of 1352 households assessed*.
  + - * 1. Total households assessed in each governorate?

| Governorate | Total assessed households |
| --- | --- |
| Governorate 1 | 136 |
| Governorate 2 | 146 |
| Governorate 3 | 250 |
| Governorate 4 | 133 |
| Governorate 5 | 209 |
| Governorate 6 | 194 |
| Governorate 7 | 159 |
| Governorate 8 | 124 |

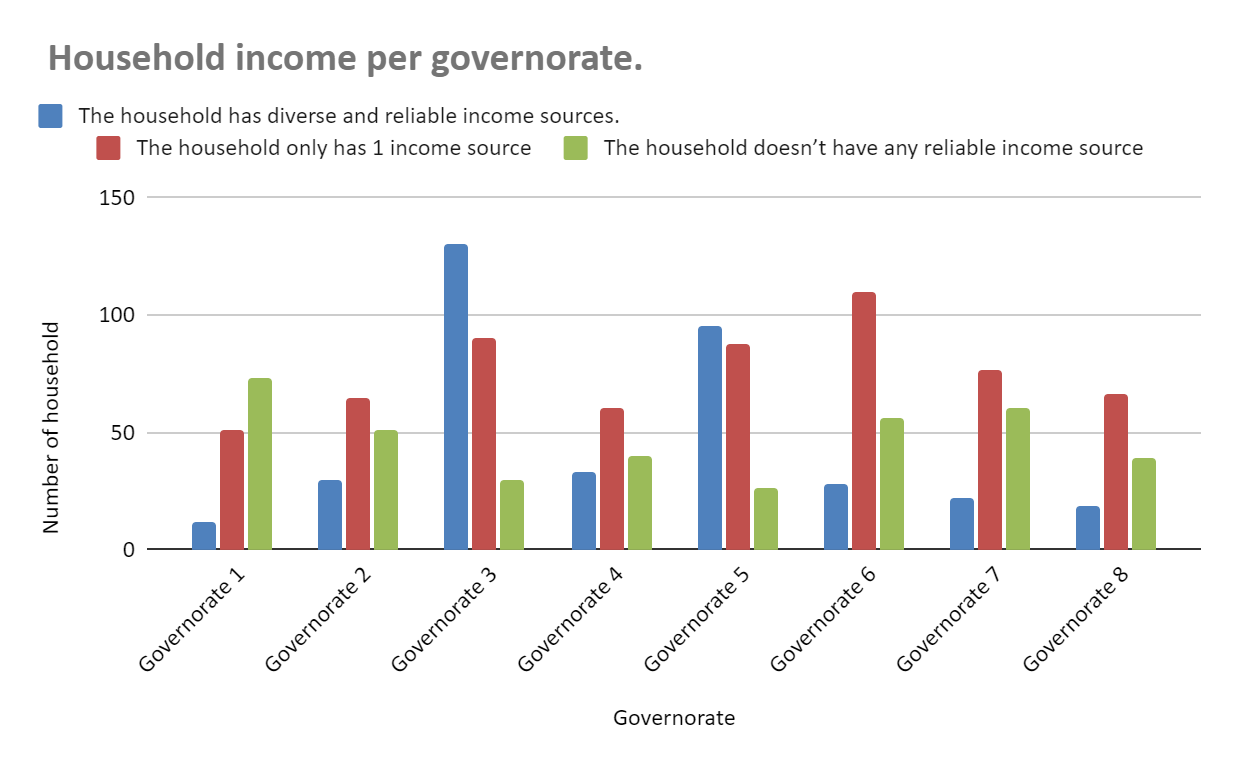
* + - * 1. Total households for each of the income categories?

|  | The household has diverse and reliable income sources.  (2 or more) | The household only has 1 income source | The household doesn’t have any reliable income source |
| --- | --- | --- | --- |
| Total households | 369 | 607 | 375 |

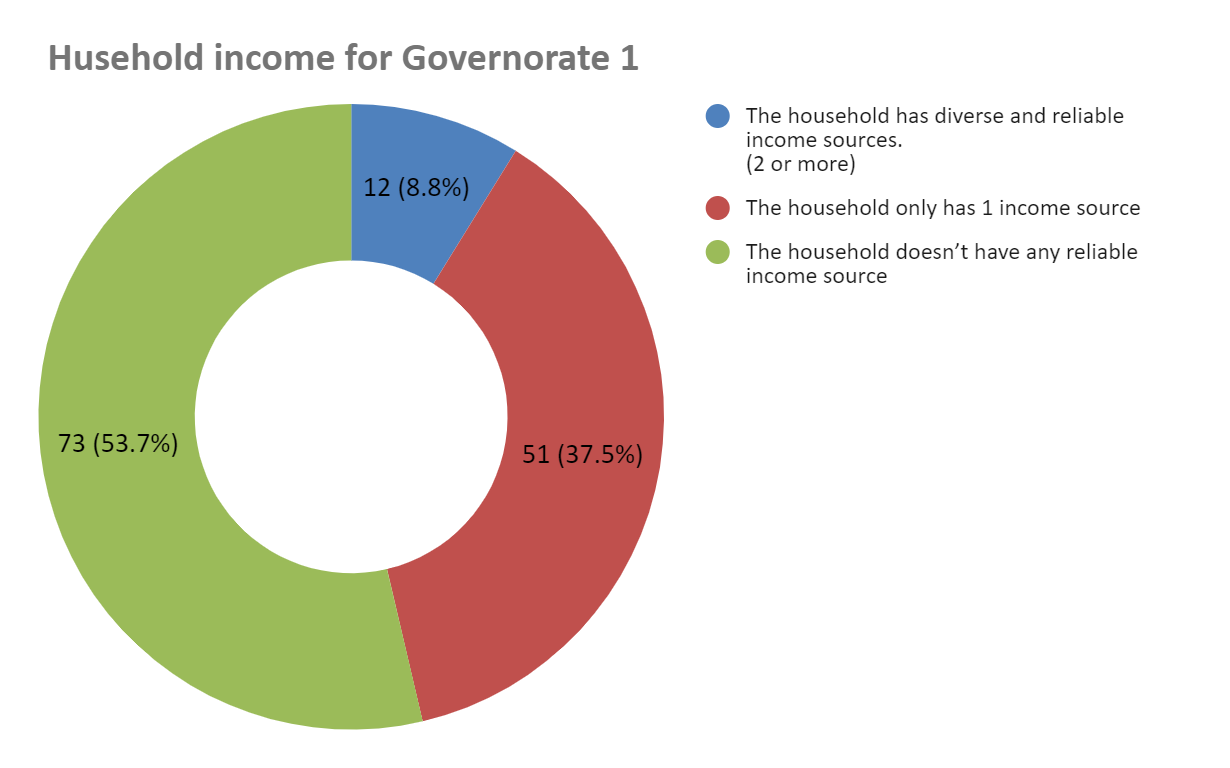
* + - * 1. Proportion of households in each income category, per governorate, relative to total number of assessed households?

|  | Proportion of households |
| --- | --- |
| Governorate 1 | 0.1006661732 |
| Governorate 2 | 0.1080680977 |
| Governorate 3 | 0.1850481125 |
| Governorate 4 | 0.09844559585 |
| Governorate 5 | 0.1547002221 |
| Governorate 6 | 0.1435973353 |
| Governorate 7 | 0.1176905996 |
| Governorate 8 | 0.0917838638 |

* + - * 1. What proportion of households in Governorate 4 have only one source of income?
* 0.4511278195
  + - * 1. What proportion of households that don’t have a reliable source of income are in Governorates 2 and 3?
* Governorates 2 = 0.3493150685
* Governorates 3 = 0.12
  1. Paste here an appropriate chart to depict household income data per governorate.



* 1. Paste here an appropriate chart to depict household income data for Governorate 1.



* 1. Assuming the programme wants to reach out to those households with the least income, prepare one slide in PowerPoint or Google Slides with your recommendations for this programme using the dataset above. Upload this the task’s folder on Google Drive.

**Task 3 - Data management:**

This task will involve writing code. Upload all the code you write into the **Task 3 & 4** folder, as well as all files that are created.

***NOTE: Questions will only be marked if the relevant code has been uploaded. Favourable marking will be given to submissions with clear & understandable naming of variables, and with appropriate comments.***

1. Locate the **.csv** files in the **Task 3 & 4** folder.
2. Using Python, or another scripting language if you don’t know Python, merge the two databases
3. **customer\_database.csv** and **orders\_database.csv** into one table with the following headings:
   1. Name
   2. Age
   3. Height
   4. Location
   5. Month
   6. Water bill (monthly)
   7. Income
   8. Number of orders

Notes for this question:

* **Sex** should not be included in the merged database
* The monthly values for **Income** and **Number of orders** appear in a single column, with the respective month in another column

1. Using **data\_dictionary.csv,** replace the human name column headings with the respective variable names.
2. Save the merged database as a **merged\_database.csv**. What is the size of this dataset (number of rows / columns / file size)?

**Task 4 - Statistics and data science:**

This task will involve writing code. Upload all the code you write into the **Task 3 & 4** folder, as well as all files that are created.

***NOTE: Questions will only be marked if the relevant code has been uploaded. Favourable marking will be given to submissions with clear & understandable naming of variables, and with appropriate comments.***

1. Using **merged\_database.csv** that was created in **Task 3**, produce a correlation coefficient matrix for the numerical variables in the database. Answer the following questions within this document:
   1. Paste the matrix as a table here - not as a picture:
   2. Which variables are most highly correlated?
   3. In the context of a machine learning problem how does this matrix help you?
2. Split the data into a training and test set, creating two .csv files: **training\_set.csv** and **test\_set.csv**. Split as follows:

| **Test set** | **Training set** |
| --- | --- |
| Peter | Faith |
| Stella | John |
|  | Juan |
|  | Julia |
|  | Lynn |
|  | Marcus |
|  | Betty |

1. Using a statistical or machine learning model, predict the number of orders that Peter and Stella make in May, June, and July. Paste below a table comparing the true number of orders and the predicted number of orders.
2. Is this an example of a supervised or unsupervised learning problem?
3. Comment on the accuracy of the predictions, providing relevant accuracy metrics to support your assessment.