October 14, 2023

C964: Computer Science Capstone Template

**Note:** This is the latest version of the Task 2 template. Following this template meets all the documentation requirements for C964 version SIM2 and SIM3. As it’s more succinct and clear, we recommend using this template for both SIM2 and SIM3. However, using the [previous template](https://westerngovernorsuniversity-my.sharepoint.com/:w:/g/personal/jim_ashe_wgu_edu/EcklZjLXTB5EpDS4BVYc8SEBhT3VHy3s_9lZSIZ5aH6Q5w?e=5tCTQb) is still acceptable.

Task 2 parts A, B, C and D

[Part A: Letter of Transmittal 1](#_Toc98598250)

[Letter of Transmittal Requirements 2](#_Toc1738085866)

[Letter Template 2](#_Toc1133266111)

[Part B: Project Proposal Plan 3](#_Toc1370766476)

[Project Summary 4](#_Toc904507251)

[Data Summary 4](#_Toc1736393957)

[Implementation 4](#_Toc1241988654)

[Timeline 4](#_Toc1357178365)

[Evaluation Plan 5](#_Toc623361460)

[Resources and Costs 5](#_Toc1538507987)

[Part C: Application 5](#_Toc1471073175)

[Part D: Post-implementation Report 6](#_Toc651895932)

[Solution Summary 7](#_Toc1134136520)

[Data Summary 7](#_Toc182221765)

[Machine Learning 7](#_Toc1505466430)

[Validation 7](#_Toc391434166)

[Visualizations 7](#_Toc201059345)

[User Guide 7](#_Toc1365484010)

[Reference Page 8](#_Toc1702353417)

# Part A: Letter of Transmittal

## Letter of Transmittal Requirements

The *Letter of Transmittal* should convince senior leadership to approve your project. Write a brief cover letter (suggested length 1-2 pages) describing the problem, how the application (part C) applies to the problem, the practical benefits to the organization, and a brief implementation plan. Include all artifacts typical of a professional (business) letter, e.g., subject line, date, greeting, signature, etc.

The letter should be concise and target a non-technical audience. Include the following:

* A summary of the problem.
* A proposed solution centering around your application.
* How the proposed solution benefits the organization.
* A summary of the costs, timeline, data, and any ethical concerns (if relevant).
* Your relevant expertise.

## Letter Template

03/23/2024

Daniel Newman

The World Health Lifeline

2025 Sunnyvale Lane, Cincinnati, OH,29323

Dear Mr. Newman,

Currently the rise of suicides in the world has been a huge concern for us over at The National Department of Health and with it there are a few problems that our newest application will address. The issue is that currently we have no way to report the correlation between population size and suicide rates which the government is required to provide to all health-related organizations. The government is also required to provide predictions on what the suicide numbers will be given a country’s population size. Sure, it’s easy to guess that a higher population will result in a more suicides , but we want to take it one step further and factor in other categories which will provide us with more accurate results. These categories include country, age, year of birth etc. After averaging all these categories, we will produce a scatterplot which will show the true effect of population and suicide rates. This means that in the future if a country’s population changes, we will be able to make an accurate assessment of what the suicide numbers could look like. This will allow the government to increase funding for specific countries in its efforts to reduce suicides. More mental health facilities will be made for countries with a higher population and more access to health care will be provided. Another problem is that we don’t currently have data about which gender is most likely to commit suicide given a specific country, age, and year of the suicide. This application is a great tool for you to use so that you have accurate data to provide to the health organizations which rely on you for its data.

As I stated in the previous paragraph the main benefit that this application has for your organization is that it will allow you to provide other health related organizations with the data they require in order to see the correlation between population size and suicide rates. Not only that but this application contains a UI to see the number of suicides that have occurred between men and women in a specific country, a specific year , and in a specific age range. This will allow you to distribute this application on a global scale so that each country can calculate its own suicide numbers between men and women within its own country. There are some country’s that can’t use this feature because their population is too low and the number of suicides that occur are negligible. This means that even if additional health funding was to be provided in that country, it’s very unlikely to make a difference in suicide numbers and that government funding would be better used elsewhere. With this application it will never be a surprise to us that we can expect regarding suicide numbers for a country that has a specific population. We will be able to take action and provide funding for large country’s which will help to protect the lives of its citizens.

To implement this application, we will need a csv file that contains data regarding suicides under specific conditions. We will need to see the correlation between gender, age, year of suicide and population size to the occurrence of a suicide. This application will use a scatterplot to graph the correlation between the number of suicides and population. We will also use machine learning to see what our model predicts the number of suicides to be as well as a line of best fit to see what we can expect from our predictions. These predictions are necessary when entering a new year when we have no data regarding what we can expect the number of suicides to look like. This will give your organization what it needs to send out to other health organizations that depend on you for its data and predictions regarding that data. As I previously mentioned we will also implement a UI that will represent the number of suicides when the user selects a country, year of suicide, and age range. We will use a pie chart to demonstrate the number of suicides per gender based on the previously mentioned parameters. This data will come from the csv file that I mentioned earlier, it has everything we need to report previous suicides for years from 1987 to 2015.

A close up of a paper

Description automatically generatedIn conclusion this application will be a great tool for you to use to distribute to other health organizations. Those health organizations will then also be able to use the UI provided to access their own country’s past suicide numbers. You will also be able to use the data the scatterplot predicts within your own records which is a known requirement for government health related organizations. We predict that this application will take around a month to produce as well as about 1000$ to create. I as well as my co-workers are knowledgeable in regard to python so that’s the programming language we’ve decided to use for this application. Some of our budget will be going towards training myself and a few others in data science since we are not familiar with building machine learning models. We hope that you will consider our services for this product since we believe it will be not only a great service to you, but to the entire world.

Sincerely,

Joseph Bonhomme, CEO of The World Health Lifeline

# Part B: Project Proposal Plan

The project proposal should target your client’s middle management. This audience may be IT professionals but have limited computer science expertise. Use appropriate industry jargon and sufficient technical details to describe the proposed project and its application. Remember, you’re establishing the technical context for your project and how it will be implemented for the client. **Write everything in the future tense.**

## Project Summary

* 1.Describe the problem.
* 2.Summarize the client and their needs as related to the problem.
* 3.Provide descriptions of all deliverables. For example, the finished application and a user guide.
* 4.Provide a summary justifying how the application will benefit the client.
* 1. The problem is that accurate suicide data is not available to health organizations around the world. No country has been able to accurately predict how many suicides will occur each year. This means that countries are not able to take accurate steps in reducing the number of suicides that occur within their country’s resulting in a growing problem with no solution.
* 2.The NDH(national department of health) has been issued a requirement to provide data regarding the number of suicides that are likely to occur given a population size. They also need to report suicide data and distribute it to health organizations around the world. Additionally, they need to report the number of suicides between men and women for the years 1987 to 2015.
* 3.The application will contain a data frame which will contain visibly relevant data that the scatterplot will use. There will be a bar graph representing suicides for men in the US, the scatterplot that uses linear regression for suicide predictions and a UI to see the difference between suicides among men and women.
* 4. This will benefit the client because they are responsible for having accurate information regarding suicides. Not only that but they are responsible for providing accurate data to other health organizations so that they can include the data within their own records. They will be able to distribute this application so that other countries can use it as well.

## Data Summary

* 1.Provide the source of the raw data, how the data will be collected, or how it will be simulated.
* 2.Describe how data will be processed and managed throughout the application development life cycle: design, development, maintenance, or others.
* 3.Justify why the data meets the needs of the project. If relevant, describe how data anomalies, e.g., outliers, incomplete data, etc., will be handled.
* 4.Address any ethical or legal concerns regarding the data. If there are no concerns, explain why.

<https://datalore.jetbrains.com/notebook/b9zSvVrCVgWZgWK88bBOIp/CqdXtmFkUaT7F4Ak50ukOQ/>

* 1.The data is provided by Kaggle, a website that provides users with downloadable excel files pertaining to specific categories. The data will be read in a Jupyter notebook using the pandas library.
* 2.The data will be collected by creating a data frame within the application so that various models can use it. Since the data is being read from a csv, there’s no need to do any additional maintenance regarding the data.
* 3.This data meets my needs because it has population numbers for specific country’s as well as suicide numbers that have occurred in those countries for a specific year. It’s also divided between men and women so this is extremely useful in my UI implementation where the suicides between men and women are represented.
* 4.Currently there are no legal concern’s regarding the use of this data since it is shared to the public. Anyone can access and use the data provided by Kaggle if it’s not used maliciously , which I’m not doing in this project.

## Implementation

* 1.Describe an industry-standard methodology to be used.
* 2.An outline of the project’s implementation plan. This outline can focus on the project’s development as a whole; or it may focus on only the implementation of the machine learning solution.
* 1.I decided to use the waterfall methodology for this project since there will be no changes after it’s complete. The client won’t be able to change the requirement’s, they are set in stone.
* 2.To implement the machine learning linear regression, we will take a week of our estimated month on reading up on the documentation provided by the sklearn and matplotlib libraries.

## Timeline

* Provide a projected timeline, including projected start dates and end dates for each milestone (a table is not required but encouraged).

|  |  |  |  |
| --- | --- | --- | --- |
| Milestone or deliverable | Duration  (hours or days) | Projected start date | Anticipated end date |
| Learned the basics of jupyter notebook | 5-20 hours | 01/01/2024 | 01/08/2024 |
| Learned how to use matplotlib and sklearn through documentation | 10-20 hours | 01/09/2024 | 01/16/2024 |
| Code the project in jupyter notebook | 10-30 hours | 01/17/2024 | 01/25/2024 |
|  |  |  |  |

## Evaluation Plan

* 1.Describe the verification method(s) to be used at each stage of development.
* 2.Describe the validation method to be used upon completion of the project.
* 1.At each stage I will verify my knowledge attained by completing online quizzes provided by online resources regarding the topic.
* 2.I will validate the completion of the project by making sure my project contains each of the requirement’s mentioned in the rubric.

## Resources and Costs

* 1.Itemize hardware and software costs.
* 2.Itemize estimated labor time and costs.
* 3.Itemize estimated environment costs of the application, e.g., deployment, hosting, maintenance, etc.
* 1.There will be no hardware costs however there will be software costs of around 100$ to enroll our employees in our selected Udemy Data Science course so that they have additional resources to learn what is required.
* 2.This labor will take around 100-150 hours and will cost us around 1000$ to complete.
* 3.To host this application we are looking to spend anywhere from 100-150 dollars.

**Part D: Post-implementation Report**

Create a post-implementation as outlined below. Provide sufficient detail so that a reader knowledgeable in computer science but unfamiliar with your project can understand what you have accomplished. Using examples and visualizations (including screenshots) beyond the three required is recommended (but not required). **Write everything in the past tense.**

## Solution Summary

* 1. Summarize the problem and solution.
* 2. Describe how the application provides a solution to the problem from parts A and B.
* 1. Our team has successfully implemented the solution to the problems surrounding the client. One of those problems were that they didn’t have the ability to predict the amount suicides given a population of a country. The other is that they didn’t have a means of seeing how the suicides were distributed from men and women. If our client doesn’t have this information, then other health organizations won’t have this information either.
* 2. The application provides a solution to the first problem above by using a scatterplot in order to show the amount of suicides that occur for a specific population. It uses a linear regression algorithm to predict those suicide numbers while using a line to show what the averages of what those predictions would be. The other problem that we addressed was not being able to see the difference in suicides between men and women for a specific country. We implemented a UI where the user selects a country, age range, and year of suicide. Then a pie chart displays the number of suicides that have occurred between men and women.

## Data Summary

* 1. Provide the source of the raw data, how the data was collected, or how it was simulated.
* 2. Describe how data was processed and managed throughout the application development life cycle: design, development, maintenance, or others.
* 1. As previously mentioned, the data is obtained from Kaggle.com. No additional information regarding the source of the data was provided.
* 2. The data was processed by using the pandas library and no other process was used in order to maintain it. It was simply referenced when necessary after it was read.

## Machine Learning

For each employed method (at least one is required) provide the following:

* 1. Identify the method and what it does (the “what”).
* 2. Describe how the method was developed (the “how”).
* 3. Justify the selection and development of the method (the “why”).
* 1. The machine learning method I used was linear regression. It displays what it believes will be the average predictions of suicides.
* 2. It was developed by using the linear regression model provided by Sklearn and the scatterplot module provided by the NumPy library.
* 3. I used linear regression because since suicides in correlation to population is mostly linear it would be a great way to show what we can expect from our predictions.

## Validation

For each employed method described in the section above provide the following:

* 1.An appropriate validation method. An appropriate validaiton method.
* 2.Results of the validation method or a future plan to obtain those results.
* 1.I will validate this by confirming my linear regression line is displayed.
* 2.Since the linear regression line is displayed, I know the model is working as expected.

## Visualizations

Identify the location of at least three unique visualizations. They can additionally be included here.

They visualizations are included in the jupyter notebook script.

## User Guide

Include an enumerated (steps 1, 2, 3, etc.) guide to execute and use your application.

* 1.Include instructions for downloading and installing any necessary software or libraries.
* 2.Provide an example of how the client should use the application.
* 1.To use this application you must have a datalore account: and navigate to: <https://datalore.jetbrains.com/notebook/b9zSvVrCVgWZgWK88bBOIp/CqdXtmFkUaT7F4Ak50ukOQ/>
* 2. The client should run each cell one after the other to ensure all required libraries will be used so no errors are thrown. To use the UI, you must select an option from each dropdown. If you want to rerun the UI, you need to re-run each dropdown and select your desired option so that each dropdown is accounted for when retrieving the results.