July 23, 2024

C964: Computer Science Capstone

**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 2](#_Toc172625144)

[Part B: Project Proposal Plan 4](#_Toc172625145)

[Project Summary 4](#_Toc172625146)

[Data Summary 4](#_Toc172625147)

[Implementation 4](#_Toc172625148)

[Timeline 4](#_Toc172625149)

[Evaluation Plan 5](#_Toc172625150)

[Resources and Costs 5](#_Toc172625151)

[Part C: Application 6](#_Toc172625152)

[Part D: Post-implementation Report 7](#_Toc172625153)

[Solution Summary 7](#_Toc172625154)

[Data Summary 7](#_Toc172625155)

[Machine Learning 7](#_Toc172625156)

[Validation 7](#_Toc172625157)

[Visualizations 7](#_Toc172625158)

[User Guide 7](#_Toc172625159)

[Reference Page 9](#_Toc172625160)

# Part A: Letter of Transmittal

July 22nd 2024

John Dollip, CEO

Dollip Foods

424 Merry Circle, Brooklyn NY

Dear John,

At Dollip Foods, we provide the highest quality food to our customers and communities. I’m writing this letter to first and foremost inform you of the most amazing opportunity to find places to expand our business to help grow Dollip Foods into the most successful food empire. Our success is not only a measurement of sales but also defined by the people we feed. As the population continues to grow in the US, new markets are emerging as potential sites for our grocery stores, food warehouses, and other commercial or industrial building sites. Of course, finding a suitable location for any of these sites requires a good understanding of the market and the overall population growth of the US.

I’m proposing a new solution to find these new emerging markets by developing a machine learning application that will accurately predict the population for these cities in 2024 for all 300 cities in the US. The application will also be able to predict any future cities population at a future year(input from the user). We plan to include graphs of the fastest growing cities as well as the slowest, as well as those that are most and least dense. With this information, Dollip Foods can make a highly informed decision as to which cities and or states we wish to expand our business into. Without this information, there is the possibility of making unfortunate market moves which may cost millions of dollars, and thousands of hours of wasted work hours planning for the construction of buildings, services, and employees that may cost us more than the potential incoming revenue.

As with any project, here is an overall summary of the costs, timeline, data, and concerns of the overall machine learning project.

Costs:

The overall estimated cost of the project will be $20,000. Here’s the breakdown:

* $20,000 - The biggest cost for the project is *two weeks' pay for 2 contractors*, with a possibility for added overtime depending on the project schedule.
* $0.00 - All the *software and libraries* are open source, and free under their respective licenses for commercial use as well as Jupyter Notebook.
* $0.00 for the *office space/internet connection/hardware* as we can use current facilities.
* $0.00 for the *dataset* since it is public domain and *free on Kaggle.com.*
* $0.00 for *ongoing maintenance* since the project will be hosted locally and currently paid employee salaries will cover any future features that might be added.

Timeline:

The estimated time to complete this project is 2 weeks. Here is a sample timeline.

* August 5th, 2024 – Start of Project, Data import and exploration
* August 9th, 2024 – Remove unnecessary features / empty data.
* August 12, 2024 – Create charts/visualizations for review
* August 14, 2024 – Choose a machine learning algorithm and start separating data into X and y.
* August 16, 2024 – Train model.
* August 19, 2024 – Review results and determine accuracy.

Data: We will be using public data given by the US census in 2020 as well as publicly available government information about 2024 city population. This information has already been gathered and compiled into one source on Kaggle.com that is free to use for commercial or academic purposes.

Concerns: Since the data is free and publicly available, there should be no legal concerns to note. There is also no ethical concerns with the project because although our business does contain food, the project itself is not about food safety or food-safe practices.

Additionally, there may be some roadblocks that come up one being that with any machine learning model. Predictions are just that, predictions. To counteract this, internal data should be used in conjunction with these predictions from our model. A second concern would be competition from our competitors who are looking to expand as well. With the tenancy of our plan and the brightest minds at Dollip Foods, we plan to be the first to break into these new markets, rest assured.

With over 30 years of experience in operations, this path has led me to new ways to operate and grow our company using the most up-to-date advances in technology to further our business endeavors. I believe now is the time to use AI to help us predict our future. As AI continues to be used in everything from cooking, working out, finance, to Tik-Tok. The future is now. Please accept my propsal to start our journey into the future of AI.

Sincerely,

Jeremy Pacailler

Jeremy Pacailler, Operations Director

**Part B: Project Proposal Plan**

## Project Summary

At Dollip Foods, we provide the highest quality food to our customers and communities. I’m writing this letter to first and foremost inform you of the most amazing opportunity to find places to expand our business to help grow Dollip Foods into the most successful food empire. Our success is not only a measurement of sales but also defined by the people we feed. As the population continues to grow in the US, new markets are emerging as potential sites for our grocery stores, food warehouses, and other commercial or industrial building sites. Of course, finding a suitable location for any of these sites requires a good understanding of the market and the overall population growth of the US.

I’m proposing a new solution to find these new emerging markets by developing a machine learning application that will accurately predict the population growth of all 300 cities in the US, by using a future year as input, the application will predict the annual growth rate for all 300 cities, show the current and future population information, as well as show a graph of top growth rates by state. With this information, Dollip Foods can make a highly informed decision as to which cities and or states we wish to expand our business into. Without this information, there is the possibility of making unfortunate market moves which may cost millions of dollars, and thousands of hours of wasted work hours planning for the construction of buildings, services, and employees that may cost us more than the potential incoming revenue.

Provide descriptions of all deliverables. For example, the finished application and a user guide.

In conclusion, I strongly believe this project will help Dollip Foods choose a premiere location for future facilities because of the low risk, cost, and time it will take to complete the project as well as the valuable information that will come with it.

## Data Summary

The source of the data comes from Kaggle.com under a free commercial license. The data was aquired from the 2020 census and is publicly available city data from government sites but was collected and arranged into one cvs file with 300 cities information across the US. The data contains 8 fields: rank (population highest to lowest), us city, us state, population 2024, population 2020, annual change (growth rate annually as a decimal), density (in miles squared), and area (miles squared).

Describe how data will be processed and managed throughout the application development life cycle: design, development, maintenance, or others.

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.

Address any ethical or legal concerns regarding the data. If there are no concerns, explain why.

## Implementation

* Describe an industry-standard methodology to be used.
* 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.

## 

## Timeline

|  |  |  |  |
| --- | --- | --- | --- |
| Milestone or deliverable | Duration  (hours or days) | Projected start date | Anticipated end date |
| Data Import and Exploration | 3 days | August 5th, 2024 | August 8th, 2024 |
| Data Trim | 1 day | August 9th, 2024 | August 9th, 2024 |
| Final Visualizations | 2 days | August 12th, 2024 | August 13th, 2024 |
| Model Chosen | 2 days | August 14th, 2024 | August 15th, 2024 |
| Model Trained | 1 day | August 16th, 2024 | August 16th, 2024 |
| Project Success Criteria Measured | 2 days | August 19th, 2024 | August 20th, 2024 |

## Evaluation Plan

* Describe the verification method(s) to be used at each stage of development.
* Describe the validation method to be used upon completion of the project.

## Resources and Costs

* $20,000 - The biggest cost for the project is *two weeks' pay for 2 contractors*, with a possibility for added overtime depending on the project schedule.
* $0.00 - All the *software and libraries* are open source, and free under their respective licenses for commercial use as well as Jupyter Notebook.
* $0.00 for the *office space/internet connection/hardware* as we can use current facilities.
* $0.00 for the *dataset* since it is public domain and *free on Kaggle.com.*
* $0.00 for *ongoing maintenance* since the project will be hosted locally and currently paid employee salaries will cover any future features that might be added.

# Part C: Application

Part C is your submitted application. This part of the document can be left blank or used to include a list of any submitted files or links.

The minimal requirements of the submitted *application* are as follows:

1. **The application functions as described.** Following the ‘User Guide’ in part D, the evaluator must be able to successfully review your application on a Windows 10 machine.
2. **A mathematical algorithm applied to data,** e.g., supervised, unsupervised, or reinforced machine learning method.
3. **A “user interface.”** Following the ‘User Guide’ in part D, the client must be able to use the application towards solving the proposed problem (as described in parts A, B, and D). For example, the client can input variables and the application outputs a prediction.
4. **Three visualizations.** The visualizations can be included separately when including them in the application is not ideal or possible, e.g., the visualizations describe proprietary data but the application is customer-facing.
5. **Submitted files and links are static and accessible.** All data, source code, and links must be accessible to evaluators on a Windows 10 machine. If parts of the project are able to be modified after submission, then matching source files must be submitted. For example, if the application is a website or hosted notebook, the .html or .ipynb files must be submitted directly to assessments.

Ideally, submitted applications should be reviewable using either Windows or Mac OS, e.g., Jupyter notebooks, webpages, Python projects, etc. If the source files exceed the 200 MB limit, consider providing screenshots or a Panopto video of the functioning application and contact your course instructor.

# Part D: Post-implementation Report

## Solution Summary

Describe how the application provides a solution to the problem from parts A and B.

## Data Summary

The data was a free dataset from Kaggle.com that contains 300 cities data about the 2020 and 2024 population, density, and area. It also includes the city name and state. It’s free for (license)

Describe how data was processed and managed throughout the application development life cycle: design, development, maintenance, or others.

## Machine Learning

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

KNeighbors Regression Model.

* Identify the method and what it does (the “what”).
* Describe how the method was developed (the “how”).
* Justify the selection and development of the method (the “why”).

## Validation

**Holdout Validation**: This is the simplest form of validation where the dataset is split into a training set and a validation set (often in an 80-20 or 70-30 ratio). The model is trained on the training set and evaluated on the validation set using regression metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), or R2R^2R2 score.An appropriate validation method. An appropriate validation method.

Results of the validation method or a future plan to obtain those results.

## 

## Visualizations

## A screenshot of a graph Description automatically generatedA graph with blue dots Description automatically generated

## User Guide

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

* Include instructions for downloading and installing any necessary software or libraries.
* Provide an example of how the client should use the application.

# Reference Page

Hussain, I. (2024, July 4). *Population of all US cities 2024*. Kaggle. https://www.kaggle.com/datasets/dataanalyst001/population-of-all-us-cities-2024

.