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

[Today’s date]

[Recipient’s name]

[Company name]

[Address]

Dear [Recipient’s name],

(Lorem ipsum dolor sit amet, consectetur adipiscing elit. Integer at sodales leo, nec fermentum tellus. Vivamus leo lorem, semper eget erat nec, ultricies vulputate sapien. Integer eget nisi at erat condimentum dictum. Nam sed ornare leo, non pretium nulla. Nam dolor arcu, condimentum et maximus quis, bibendum ut risus. Nam hendrerit ac erat sit amet luctus. Quisque lacinia sapien sed nisl porta, id rutrum odio tristique. Sed sodales nisi a condimentum fermentum. Suspendisse lobortis diam in orci consectetur congue. Sed ligula felis, accumsan eu venenatis ac, hendrerit ac nulla. In ultrices, sem a semper ultrices, orci ex dictum sapien, ac dapibus ligula lectus at est. Quisque at posuere purus, vitae ultricies risus. Donec tincidunt, ipsum eget euismod lobortis, tortor nisl luctus sem, quis blandit tortor dui eu dui.

Praesent sagittis, leo vitae sodales cursus, purus orci rhoncus ligula, quis facilisis ante est nec lectus. Curabitur augue quam, ultricies at arcu eget, molestie eleifend neque. Sed interdum tempor purus, et luctus tellus feugiat imperdiet. Maecenas scelerisque viverra orci tincidunt luctus. Vestibulum sodales eros ut ex luctus tempor. Curabitur eget leo vehicula, malesuada urna ut, eleifend nisi. Quisque sapien tellus, ornare ac magna quis, ultricies consectetur mauris. Nunc erat ligula, mattis id tempor ut, venenatis ac mi. Sed sit amet odio ac ligula tincidunt iaculis sit amet vel ipsum.

Etiam lobortis aliquam metus, eu aliquam ante aliquam ut. Nam tristique sagittis mauris vel tempor. Quisque rhoncus, justo sed lobortis porta, nulla libero pulvinar tortor, ut ullamcorper justo enim ut erat. Praesent lobortis ut leo in aliquet. Suspendisse aliquet velit nulla, a rhoncus nibh vestibulum iaculis. Praesent mollis nibh nec ultrices blandit. Pellentesque felis elit, pretium at risus in, commodo consectetur tortor. Etiam fringilla mi quis erat mollis ultricies. Phasellus vestibulum elementum commodo. Sed congue vulputate orci in porta. Pellentesque scelerisque facilisis justo, a bibendum ligula tempus quis. Aenean efficitur eleifend lorem, et tempus risus consequat quis. Cras varius metus sapien, ac malesuada sem volutpat sed.

Sincerely,

[Sign here: e.g., Jane Smith]

[Your name, title]

# 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

* Describe the problem.
* Summarize the client and their needs as related to the problem.
* Provide descriptions of all deliverables. For example, the finished application and a user guide.
* Provide a summary justifying how the application will benefit the client.

## Data Summary

* Provide the source of the raw data, how the data will be collected, or how it will be simulated.
* 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

* 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 |
|  |  |  |  |
|  |  |  |  |

## 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

* Itemize hardware and software costs.
* Itemize estimated labor time and costs.
* Itemize estimated environment costs of the application, e.g., deployment, hosting, maintenance, etc.

# 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.

**Submitted Files**

* **C964\_Capstone\_Project.ipynb**

*I am including the local version of the JupyterLab notebook file as well in case the JupyterLite version doesn’t work. This one requires additional libraries but works fine with the Anaconda Distribution of Python found* [*here*](https://www.anaconda.com/download)*. Once this is installed, search for Anaconda Navigator and launch it. Open JupyterLab, then load this file as well as the csv file. Select Run -> Restart Kernel and Run All Cells.*

* **C964\_Capstone\_Project\_JupyterLite.ipynb**

*This should be uploaded to the JupyterLite online platform along with the csv file per the user manual below.*

* **ds\_salaries.csv**

*This data file from Kaggle.com contains the data used for the training and test sets. This should be uploaded to JupyterLab or JupyterLite so that it can be read by pandas.*

# Part D: Post-implementation Report

## Solution Summary

* **Summarize the problem and solution.**

*The problem stems from a challenge faced by the company, QuantumCipher Analytics Co. The company needed to determine competitive salaries for their rapidly expanding team of data scientists. Because this is such a large company with many competitors, the company recognized the necessity of offering competitive compensation to attract and retain top talent. To tackle this issue, the company turned to a data-driven approach, utilizing a comprehensive dataset of data science job salaries. This dataset provided the data necessary for developing a robust solution.*

*The solution to this problem was the development and application of a multiple regression machine learning model. This was completed and is operational and provides prediction of data scientist salaries based on a variety of independent variables, such as experience level, location, and company size.*

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

*The application allows the company to provide competitive and fair compensation to their employees by aiding in the strategic decision-making regarding employee salaries and enhancing the company’s ability to maintain its position as a leader in the industry. The model’s predictive capabilities, grounded in rigorous data analysis, position the company to effectively navigate the intricacies of salary determination in the field of data science.*

## Data Summary

* **Provide the source of the raw data, how the data was collected, or how it was simulated.**

*The raw data for this project was sourced from Kaggle.com, specifically the "Data Science Job Salaries" dataset. This dataset is a compilation of over 600 entries, each entry detailing various aspects of data science job positions. The data encompasses a range of columns: Work Year, Experience Level, Employment Type, Job Title, Salary, Salary Currency, Salary in USD, Employee Residence, Remote Ratio, Company Location, and Company Size. The original data was found on https://ai-jobs.net/salaries/, gathered through a combination of surveys and job listings with disclosed salary information. This comprehensive dataset provides a broad view of the data science job market, particularly focusing on salary variations across different factors such as experience, location, and company size.*

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

*In the preprocessing stage, the dataset underwent several modifications to align with the project’s specific needs and to ensure data quality. Firstly, the 'Job Title' column was removed. This decision was based on the assessment that job titles, in this context, do not significantly correlate with salary variations and could introduce unnecessary complexity into the analysis. There are some titles that do convey an increased salary, such as those containing “Director”, but most do not.*

*To maintain relevance to the project’s business case, which is centered on a U.S. company, the data was filtered to include only those entries where the company location is in the United States. This filtering helps in focusing the analysis on the geographical context of interest. Additionally, data points corresponding to part-time employment were excluded due to their insufficient representation in the dataset. Finally, a critical step in the preprocessing was filtering out outliers. These outliers were filtered based on groups in each column. The extreme salaries were removed and the accuracy increased.*

*To streamline the data and avoid redundancy, the 'Salary' and 'Salary Currency' columns were removed, as these were deemed superfluous given the presence of a 'Salary in USD' column which provided a standardized measure of salaries across different geographies. This standardization is crucial for consistent and comparable analysis, especially when dealing with a global dataset.*

*Additionally, an unnamed index column present in the dataset was also removed. This column, typically used for identification purposes, was not necessary for our analysis and its removal helped in simplifying the dataset. These steps were instrumental in ensuring that the dataset was not only relevant to the specific context of the study but also free from unnecessary complexities that could potentially skew the analysis or interpretation of the results.*

*Over the lifecycle of the application, more data will be used to increase effectiveness and this may lead to more categories and columns of data. This can also lead to previously unused portions of data being used (with more part-time data, that option can be re-admitted). Each new potential data category will need to be reviewed thoroughly to be accepted. This will happen during development and maintenance.*

## Machine Learning

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

* **Identify the method and what it does (the “what”).**

*The method used for machine learning on this project was a multiple regression model. This was implemented through sklearn.linear\_model import LinearRegression. This method is used to understand the relationship between one dependent variable (salary in USD) and two or more independent variables. The value of the dependent variable is predicted based on the combined influence of the independent variables.*

* **Describe how the method was developed (the “how”).**

*The model calculates the coefficients for each independent variable, providing a number of their individual contribution to the dependent variable. These coefficients are derived from using training data and fitting the model to the data using the least squares method, which “minimizes the sum of squared residuals” (Introduction to Multiple Linear Regression, 2020).*

* **Justify the selection and development of the method (the “why”).**

*The method was chosen based on its ability to analyze and quantify relationships between multiple independent variables and a single dependent variable. This did exactly what was proposed and provided a prediction of a salary based on several variables. There is no other method that would fit better in this instance. This doesn’t mean that the method cannot be refined and more variables could be added to ensure greater accuracy.*

*Another option that could have been used is a simple linear regression with just one independent variable and one dependent variable, but this would make for a fairly useless model since the accuracy would be very low.*

## Validation

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

* **An appropriate validation method.**

*For validation, I used a test set with 20% of the cleaned data to evaluate the model’s performance. The main validation metrics used were Mean Squared Error and R-squared. MSE is a metric used to quantify the average squared difference between the estimated values and actual values, giving a picture of the model’s prediction accuracy. R-squared “is the proportion of the variance in the response variable that can be explained by the explanatory variables” (Introduction to Multiple Linear Regression, 2020) and provides a number on how good a fit the model is with the independent variables.*

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

*The results indicated moderate accuracy levels suggesting that the model has potential to scale with more data. Although the accuracy wasn’t high, the results look promising for future applications with increased data volume This validation process serves as a foundation for further refining and enhancing the model’s predictive capabilities.*

## Visualizations

*The three visualization are included in the notebook along with the tool:*

1. *A histogram of the salary ranges used in the algorithm.*
2. *A scatterplot of experience vs. salary.*
3. *A bar chart of average salaries of different company sizes.*

* Chart, histogram

  Description automatically generated
* Chart, line chart

  Description automatically generated
* Chart, bar chart

  Description automatically generated

## User Guide

1. Download the project zip folder with two notebooks and a csv file. Extract the files.
2. Navigate to [Project Jupyter | Home](https://jupyter.org/)
3. Click on “Try” on the top bar.

A picture containing diagram

Description automatically generated

1. Click on “JupyterLab”.

Graphical user interface, text, application, chat or text message

Description automatically generated

1. Click on the upload button.



1. Upload the file “C964\_Capstone\_Project\_JupyterLite”.
2. Repeat step 5 to additionally upload the file “ds\_salaries”.
3. Double-click on the “C964\_Capstone\_Project\_JupyterLite” file in JupyterLite.
4. Select Run -> Restart Kernel and Run All Cells and wait.

.

1. Click the “No Kernel” button in the top-right corner when it appears.
2. Select Python.
3. Select Run -> Run All Cells
4. The visualizations and markdowns are provided above the tool. To use the tool, scroll to the bottom of the screen.
5. A work year can be selected with the slider.
6. The rest of the items are dropdown menus.
7. Select “Predict Salary”
8. The predicted salary will appear in the box below.

# Reference Page

Following APA guidelines, include references for any cited works, e.g., (Author, year). References are not required, and this page can be removed if no references are used. To cite sources used for code, you should include the references as code comments within the source code.

*Introduction to Multiple Linear Regression*. (2020, October 27). Retrieved from Statology: https://www.statology.org/multiple-linear-regression/