Project Title
CIC Diagnostics Grant Workbook

Tajrin Kashem
Section:VC1C
Email:tajrinkashem@gmail.com

CISC 4900
Spring 2023

Project Supervisor
Fendja Larivaux
CUNY 2X Tech CIS
Program Manager
Email:Fendja.Larivaux@brooklyn.cuny.edu

Table of Contents

1. Executive Summary
2. Project Detail
   2.1 Project Objective
   2.2 What is the CIC Diagnostic Grant?
   2.3 Why is it Necessary?
   2.4 What Data is Collected?
   2.5 Who is Going to Use the Solution/Program?
3. Technical Overview
   3.1 How to Analyze/Query Data for the CIC Workbook?
   3.2 Software Requirements
   3.3 Flowchart Explanation
   3.4 Software Construction
   3.5 Design Pattern
4. Milestone and Reporting
   4.1 Project Management
   4.2 Tasks
   4.3 Revision and Review
   4.4 Estimated Time
   4.5 Actual Time
5. Delivery
5.1 Testing
5.2 Documentation and Source Code
5.3 Additional Project Documents
6. Obstacles and Challenges
7. References
8. Project Log

1.Executive Summary

The CIC Diagnostics Grant Workbook project aims to analyze and query data to complete the CIC grant workbook at Brooklyn College. The project utilizes Python and pandas for data analysis, specifically focusing on a provided CSV file. The objective is to identify areas for interventions and improvements in student retention by analyzing demographic data. The project follows a flowchart-based approach and utilizes agile project management principles. The team successfully delivered the expected outcome despite challenges such as unavailable data. The project contributed to the growth and learning of the team members in data analysis and project management. Overall, the project assists Brooklyn College in evaluating student success interventions and improving student retention.

2.Project Detail:

2.1 Project objective:

To query/analyze the data for completion of the CIC grant workbook .

2.2 What is the cic diagnostic grant?

The Diagnostic Grant from the Center for Inclusive Computing at Northeastern university allows eligible undergraduate institutions to provide deeper insight into the quantitative data of student persistence and retention by collecting and analyzing demographic data .

The report aims to make cultural and institutional change by using the data to identify areas where interventions are necessary and improvements can be made.

2.3 Why is it necessary?

Brooklyn College has received this grant to evaluate student success interventions and maximize student achievement and retention.The Department seeks to understand not just what is happening with regards to student retention, but why it is happening. The Department plans to establish an ongoing data-collection effort to effectively and continuously track progress on these statistics, and identify specific courses of action.

2.4 What data is collected?

Schools collect and submit historic and current persistence, retention, and graduation data to the CIC Data Collection Portal. Specifically:

- Student enrollment, completion, and faculty/TA support in introductory CS courses
- Term-to-term retention and graduation of computing majors
- Gender and race/ethnicity of students and introductory course faculty and teaching assistants

Data collection overview:

https://cic.northeastern.edu/wp-content/uploads/sites/5/2021/03/CIC_Data-Collection-Overview.pdf

2.5 Who is going to use the solution/program?

The department collects the data and submits to the CIC data collection portal .Through the grant provided by CIC it allows the department to conduct further research, investigate issues to opt for better solutions.

3.Technical overview:

3.1 How to analyze/query data for the CIC workbook?

We were provided with a CSV file named ‘roster_anon_cohorts.csv’ by the department which contained necessary data for us to report the CIC Grant workbook with appropriate tabulation on various questions .

COHORTS

Table 1. Cohort Table

		Terms / Semesters			
Class	**Courses**	**Spring 2021**	**Fall 2021**	**Spring 2022**	**Fall 2022**
CS1	1115, 1170	*A CS1xS21*	*B CS1xF21*	*D CS1xS22*	*G CS1xF22*
CS2	3115		*C CS2xF21*	*E CS2xS22*	*H CS2xF22*
CS3	3130			*F CS3xS22*	*I CS3xF22*
R&G					

Anonymized Data

The fields in the `roster_anon_cohorts.csv` file

Table 2. Fields

cohort	*Corresponds to the 10 groups in cohort table (see Table 1 above)*
semester	*Each semester*
class	*The course number grouped into CS1 / CS2 / CS3*
course	*The course number 1115, 1170, 3115, 3130*
registration_date	*Date student registered for that class/section*
division	*Graduate (G) vs Undergraduate (U)*
level	*Freshman, Sophomore, Junior, Senior, etc*
admission	*Used to group transfer vs continuing students. 0 means unknown*
Major	*CS, IS, MMC, Undeclared, or Other*
passfaildrop	*Grades grouped in pass / fail / drop*
grade	*The letter grade that the student received*
gender	*Student gender: Male/Female/Unknown*
eid	*Student’s emplid*
section	*Section of the course*
Instructor	*Instructor of the section*
Instructor type	*Tenure Track (TT) vs part time (PT)*

3.2 Software Requirement:

We used python as our main programming language, more specifically pandas’ features to implement data processing,organizing and cleaning to build query functionality for the data and conduct analysis of the data .We choose to work with python because of its easy approach and versatility .

We also used Google collab, a free Jupyter notebook environment that runs entirely in the cloud. Most importantly, it does not require a setup and the notebooks can be simultaneously edited by other team members,which is a great feature esp when working in a group .Also provides automatic version control.

Although Sql or Excel could have resulted in the same outcome, we prefer to work with python .

3.3 Flow chart:

```
start
|
v
searchFlags = (condition1) & (condition2) & (condition3) & ...
|
v
df.loc[searchFlags]
|
v
df.loc[searchFlags].groupby('grouping_column')['count_column'].nunique()
|
v
count_grouped
|
v
df.loc[searchFlags]['count_column'].nunique()
|
v
count_total
|
v
end
```

3.4 How does the code work?

From the above flowchart we see that it starts with the `searchFlags` variable, which is created by combining multiple boolean conditions. The `searchFlags` variable is then used to filter the DataFrame `df` using the `loc` method. The filtered DataFrame is then grouped by a `grouping_column` using the `groupby` method, and the number of unique values in a `count_column` for each group is counted using the `nunique` method. The count of unique values in the grouped DataFrame is stored in the `count_grouped` variable. The total number of unique values in the filtered DataFrame is also counted using the `nunique` method, and the count is stored in the `count_total` variable. Thus,providing required/desired results from the query.Depending on the questions our search flags vary/differ.

3.5 Software construction:

The project's code and construction were organized into three main sections representing CS1, CS2, and CS3 classes, each with sub-sections for different semesters. Approximately 50 questions were addressed within each section. Standard python conventions code style was used in this project.

Design pattern

The design approach for query function in Pandas falls under the behavioral design pattern.The query function allows for communication between the data frame and the user by filtering the data frame based on the specified condition. The condition can be a boolean expression that returns True or False.

4.Milestone and reporting

4.1 Project management:

On this project a team/group of 2 people worked together.We both worked as group mates combining our ideas, discussing our expectations and shared responsibilities to get our project up and running .We checked in with our supervisor on a weekly basis(mostly every Wednesdays from 5pm-6pm) with an update on the progress of the project and to further clarify on issues with uncertainty or unavailable data .We mainly used slack and email to communicate with each other.

Our approach to manage this project more likely agile methodology emphasizes flexibility and collaboration, with a focus on delivering a working product in short iterations.

4.2 Tasks:

We divided the workload among ourselves .We individually took responsibility of completing our own parts after discussing our approach and requirements of the task.

4.3 Revision and Review

At every iteration after meeting with our supervisor on a weekly basis we would review the code to ensure quality and readability.Google collab is embedded with tracking revision so at every change it would have a record.

4.4 Estimated time:

By the deadline to submit the CIC workbook.

*4.5 Actual time***:**

It took more time than anticipated since there was unavailable data and changes in the CIC workbook.

5.Delivery:

5.1 Expected outcome:

To find exact numbers/tabulation for the questions asked in the CIC grant workbook with the provided data in the csv file .

5.2 Actual outcome:

We were able to successfully query data to solve questions .Although there were few questions which couldnt be answered since there was not enough information.
Some of the questions involved information about students and faculties’s race/ethnicity and gender,which was unavailable. Also there were some questions about the previous or next semester’s data.Since we have information only for four semesters,namely, Fall 21,Spring 21,Fall 22,Spring 22, we can not answer for the previous semester of spring 21 and next semester of Fall 22 .

5.3 Testing:

We checked for null values and duplicate values.we would only count for unique values to ensure no duplicate values are accounted for in the results.

5.4 Documentation and source code:

< link to google colab >
https://colab.research.google.com/drive/1jcuGQAV_ZeBZjqbsAVX6ou-WGQuLz66H?usp=sharing

5.5 Additional Project Documents

<a demo of the workbook:google sheet>
https://docs.google.com/spreadsheets/d/1H69rbMsp1LkXSUzQE4xW29FnqsH4YXh-/edit?usp=sharing&ouid=116251180674884202592&rtpof=true&sd=true

6.Challenges and Growth:

Through this project we got to learn about the CIC diagnostics grant and its importance. From Technical perspective, we have learned to query and analyze data through various pandas' features. From project management perspective, there are few things we have learned along the way .It is always a good idea keep a backup file just in case the master file is deleted accidentally or get corrupted for some reason, to be adaptive and flexible to any uncertainty /changes or to accommodate the needs of group member and establish ground rule to communicate for efficient operation and to be precise on shared responsibilities among group members.

7.References **:**

CIC website:
https://cic.northeastern.edu/grants/diagnostic-grants/#:~:text=Overview,are%20losing%20or%20gaining%20students

Pandas:
https://pandas.pydata.org/docs/

8.Project Log:

<Google sheet:Tajrin>
https://docs.google.com/spreadsheets/d/1qqUK8tRhIlDgdrnw1EH0hntFF887ylpR_9z-flYRUm0/edit?usp=sharing