Computing Alumni Data with Python libraries to visualize and embed to GitHub pages site

Saiteja Goud Voruganti
sv1081@wildcats.unh.edu
Master Project in Information Technology
Department of Applied Engineering and Sciences
University of New Hampshire
Manchester, New Hampshire, USA

ABSTRACT

The project is the study of computing alumni data of the University of New Hampshire, Manchester since the university maintains the computing alumni data extracted manually into spreadsheets from the program's Linked In group. The project's high interest is to create the visualizations and make them publicly available. The motivation of my study is to work with alumni data to produce graphs and charts to provide evidence of the quality of education at the University of New Hampshire. Since the alumni, data updates every semester Automation of the process is necessary. The goal of the project is to study the alumni data and develop visualizations using python libraries to automate the process of retrieving the data from the GitHub repository updating the visualizations and embed into the GitHub pages site. Data generated may not be accurate as there will be a chance of duplicate and missing values so the validation of data is necessary to analyze the data. Producing interactive visualizations using python libraries and publishing these visualizations into GitHub pages site by converting them into an HTML file. Automation of graphs if data is updated and published.

Achieving data cleansing by clearing out the missing and unnecessary values using appropriate methods and tools such as Tableau-Prep Builder. Tableau-prep builder provides various options to perform data cleansing such as space trim, organizing, joins, and being able to produce a CSV file as output. Developing the interactive visualizations is a high priority, is achieved using python libraries such as seaborn, plotly, plotly express, and folium these libraries provide a plethora of graphs, bars, and charts to produce interactive visualizations. plotly allows python users to create web-based visualizations saved to standalone HTML files. Embedding visualizations to GitHub pages site with the saved HTML file and automating the process of updating graphs.

The project produces interactive visualizations with python libraries. Publicly available through the GitHub pages site. It offers an automation process so no need to create and embed visualizations each time alumni data updated. Achieved creating interactive visualizations and publish them on GitHub pages site. The published visualizations will provide the meaning full insights, which will be helpful for the computing programs and prospective students about the quality of education.

[**?**]

CCS CONCEPTS

• **Human-centered computing** → **Visualization design and evaluation methods**; • **Computing methodologies** → *Knowledge representation and reasoning*; **Search methodologies**.

KEYWORDS

Tableau-Prep Builder
Python Libraries
Plotly, Plotly Express, Seaborn
GitHub pages

1 INTRODUCTION

The project defines the study of computing alumni data of the University of New Hampshire, Manchester. The study involves meaning full insights about alumni achievement. The project works with alumni data, which will be a useful resource to understand the quality of education the university is providing. The study involves the research and development process providing a good understanding of developing graphs and charts using python libraries and publishing them.

The university maintains computing alumni data extracted manually into spreadsheets from the program's Linked In group. The program's high interest is to create graphs and publish them. The project involves creating interactive graphs and charts using python libraries and publishing the visualizations with automating the process. This will help the university computing programs for prospective and current students to understand the quality of the program that the university is providing.

The project identifies the study of alumni data and the development process of creating graphs and publishing visualizations. Making them publicly available for university students. The project study is to understand the quality of education that the university is providing and analyze the patterns of alumni employment. It involves python libraries and GitHub pages. The goal of the project is to study the alumni data and develop visualizations using python libraries to automate the process of retrieving the data from the GitHub repository updating the visualizations and embed into the GitHub pages site.

2 OBJECTIVES

The process of achieving the goal involves data validation, visualization, publishing, and automation of the process. With obtaining the alumni data, I would like to validate the data. The generated

data might not be perfect there is a chance of having duplicate values, missing values, and spelling mistakes.to generate perfect and interactive visualizations the data validation is necessary. I would like to achieve data validation using proper cleaning methods and tools. I have researched tools that are used for data cleansing and I found out Tableau-prep builder provides a variety of cleansing options for data cleansing. Python is another option for data cleansing but it is a time taking process. Tableau-prep Builder is easier and timesaving. I have preferred the Tableau-Prep builder for data cleansing and producing the CSV file.

Creating visualizations is a high priority, it helps you to gain meaningful insights from data and make key decisions from it. I will analyze all the possible factors that are present inside the data to generate interactive visualizations. I have decided to use python libraries to produce visualizations. The libraries I have preferred were Plotly and Seaborn both the libraries have significant features in creating visualizations. Plotly provides more features compared to seaborn using Plotly we can generate dynamic web-based visualizations, save them to a standalone HTML file. Using Plotly I can able to generate visualizations and save them to an HTML file. Plotly has more advantages compared to Seaborn. I have created predefined queries which were able to answer a few things like who are the alumni working as software engineers similarly to other related job positions. Jupiter notebook supports python libraries is helpful for viewing interactive visualizations.

Embedding visualizations into GitHub pages site where it is publicly available for students and organizations to view the visualizations. To embed visualizations first, I need to create a GitHub repository in the GitHub programs organization then I need to add the index.html file, which contains the visualizations, developed using python libraries. Visualizations will be displayed in the provided link by the GitHub page's site.

The alumni data gets updated each and every semester and it will be a time taking process for each and every time creating visualizations and updating the graphs so by automating the process we can able to overcome this. The automation is achieved by using python such as retrieving the data from the GitHub repository updating the visualization, creating an HTML that contains visualizations, and embed into the GitHub page's site. This will provide the option of updating the graphs and charts.

.

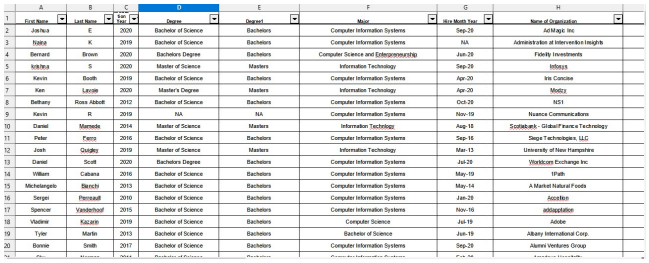
3 RESOURCES AND APPROACH

3.1 Data Cleansing

Having the accurate data is very important it improves the over all productivity of the product and Data cleansing is important factor to consider while working on any data set. Cleansing the data will provide clear information about the work and it is beneficial for analysing the data

The generated data might not be accurate it is possible that there will be still some missing values that need to be validated before performing any kind of visualizations. I have researched various data cleaning tools then decided to use Tableau-Prep Builder [5], which is easy to use and most effective in the data cleansing approach. The Tableau-Prep Builder allows me to sort the N/A values, as there are few N/A values in the data set. Some of the entries in the data set were wrongly entered such as in place of Information Technology it was entered as Info Tech. It allows organizing the similar entries of each column of the data set such that there would not be any confusion while creating the visualizations. The data set contains duplicate values of attributes Degree, Degree1 and State, State1 since both the fields have similar entries so I have deleted the columns Degree1 and State1 from the data set, which is unnecessary. Tableau prep-builder also provides Trim space, used to remove the unnecessary gaps in the data set and align the entries in an organized manner.

Below figure is from the original data set obtained



	A	B	C	D	E	F	G	H
1	First Name	Last Name	Exp Grad	Degree	Degree1	Major	Hire Month Year	Name of Organization
2	Joshua	E	2020	Bachelor of Science	Bachelors	Computer Information Systems	Sep-20	AdMagic Inc
3	Naina	K	2019	Bachelor of Science	Bachelors	Computer Information Systems	NA	Administration at Intervention Insights
4	Bernard	Brown	2020	Bachelors Degree	Bachelors	Computer Science and Entrepreneurship	Jun-20	Fidelity Investments
5	lokitha	S	2020	Master of Science	Masters	Information Technology	Sep-20	Infosys
6	Kevin	Booth	2019	Bachelor of Science	Bachelors	Computer Information Systems	Apr-20	Iris Concise
7	Ken	Lavoie	2020	Master's Degree	Masters	Information Technology	Apr-20	Modzy
8	Bethany	Ross Abbott	2012	Bachelor of Science	Bachelors	Computer Information Systems	Oct-20	NS1
9	Kevin	R	2019	NA	NA	Computer Information Systems	Nov-19	Nuance Communications
10	Daniel	Marmeda	2014	Master of Science	Masters	Information Technology	Aug-18	Scotiabank - Global IT
11	Peter	Ferro	2016	Bachelor of Science	Bachelors	Computer Information Systems	Sep-16	Sega Technologies, LLC
12	Josh	Quigley	2019	Master of Science	Masters	Information Technology	Mar-13	University of New Hampshire
13	Daniel	Scott	2020	Bachelors Degree	Bachelors	Computer Information Systems	Jul-20	Workicom Exchange Inc
14	William	Cabana	2016	Bachelor of Science	Bachelors	Computer Information Systems	May-19	1Path
15	Michelangelo	Bianchi	2013	Bachelor of Science	Bachelors	Computer Information Systems	May-14	A Market Natural Foods
16	Sergei	Perreault	2010	Bachelor of Science	Bachelors	Computer Information Systems	Jan-20	Accellion
17	Spencer	Vanderhoof	2015	Bachelor of Science	Bachelors	Computer Information Systems	Nov-16	addappitation
18	Vladimir	Kazanin	2019	Bachelor of Science	Bachelors	Computer Science	Jul-19	Adobe
19	Tyler	Martin	2013	Bachelor of Science	Bachelors	Bachelor of Science	Jun-19	Albany International Corp.
20	Bonnie	Smith	2017	Bachelor of Science	Bachelors	Computer Information Systems	Sep-20	Alumni Ventures Group

Figure 1: Original data set

I have found out a few alumni are working in Canada and Saudi and they are misplaced in the data fields of the column state. so I have decided to create an extra data field with the column name "Country" so that the data will be perfect. in the organization column some of the data entries are similar but wrongly entered with spelling mistakes and commas with the help of tableau prep-builder I have sorted out the data and organized them so that they look familiar. It has a great option of converting all the workflow into a CSV data file. With all the required changes I have generated a CSV data file.

Below figure are from the cleansed data set

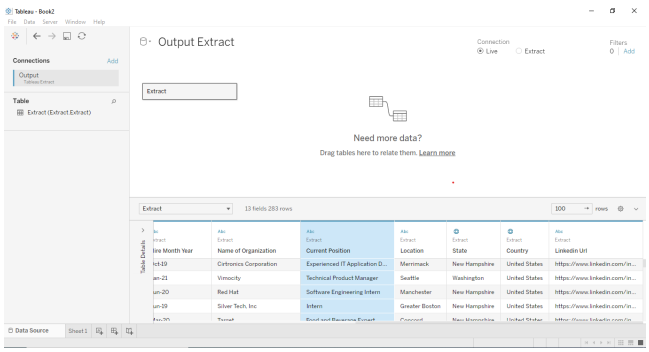


Figure 2: Cleansed data set

3.2 Visualization

Visualization is the major part of the project I have used jupyter notebook to analyze the data and find the subsets to build visualizations. Most of the python libraries support jupyter notebook. Alumni data contains information such as First name, Last name,

Year Graduated, Masters, Hiring month, Name of organization, Location, and Linked in Url. I have started creating subsets of data by grouping some of the attributes from the original data frame to find the best possible inter-activeness and also for implementing queries. I have started using Seaborn[6] Python libraries to build visualizations. Most of the visualizations were static but I was able to discover interesting things from the data set regarding the alumni location.

Then I started searching another library and found Plotly[3], which is an open-source plotting library that provides more than 40 charts for visualization. It is built on plotly JavaScript library that allows the users to create interactive web-based visualizations and which can be displayed in the jupyter notebook. It also provides standalone HTML pages. I have created a few more subsets for my better understanding, which also helped me to develop predefined queries. I was able to create interactive visualizations using plotly graphing library such as bar, scatter3d, sunburst, and pie.

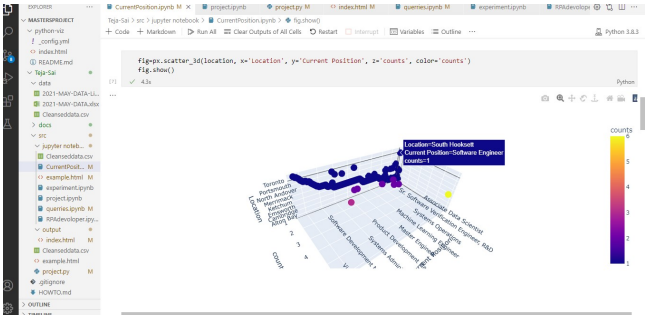


Figure 3: scatter3d

For the queries

Regarding the queries I have created 10 [2] bucket list from the Column "Current position" such as 1)Software Engineer 2)RPA developer 3)IT developer 4)Network Engineer 5)Analyst 6)Manager 7)Security 8)Administrators 9)Data handlers 10)Intern Below code is for generating the software engineers of Alumni data set and Visualizing the role with respect to the location.

```
software_engineer=df.loc[(df['Current
    Position'].str.contains("Software","
    software")) &
(~df['Current Position'].str.contains("
    Developer", "Robotics"))]

software_engineer.head()
SE=software_engineer.groupby(['First Name',
    "Hire Month Year" ,'Last Name','Current
    Position','Location', 'Major']).size().
    reset_index(name='counts')
fig = px.bar(SE, x= 'Current Position', y='
    Location',)

fig.show()
```

Figure 4: bar graph with Software engineers and Location

With the help of the these bucket lists i have generated few queries which will be helpful to understand the Job roles of alumni from my observations most of the software engineers are from Manchester and majority of them did major in Computer information systems. Similarly there are some interesting results in other bucket lists such as all of the RPA developers are from the organisation Colburn Hill Group.

3.3 Creating HTML File

Now to publish these visualizations into the GitHub page site requires an HTML file. I have to create an HTML file that contains the visualizations and it is possible using plotly [4]

```
plotly.io.to_html

f.write(fig.to_html(full_html=False,
    include_plotlyjs='cdn'))
        f.write(fig1.to_html(full_html=False
            , include_plotlyjs='cdn'))
        f.write(fig2.to_html(full_html=False
            , include_plotlyjs='cdn'))
        f.write(fig3.to_html(full_html=False
            , include_plotlyjs='cdn'))
```

which converts the figure into HTML string representation. with the following code i have converted all the fours charts to single HTML file.

3.4 Automation

Alumni data gets updated each and every semester so automating the process will be helpful such that if new data gets added to the program. Then the automation process is able to perform visualizations and convert them into an HTML file and embed them into GitHub pages site. The automation works with retrieving the updated data from the GitHub repository by cloning it to the local system. Reading the updated data file and then performing visualizations with updated data and creating an updated HTML file. Then committing the changes and pushing it back to the repository with updated HTML file will able publish the visualization on GitHub pages site.

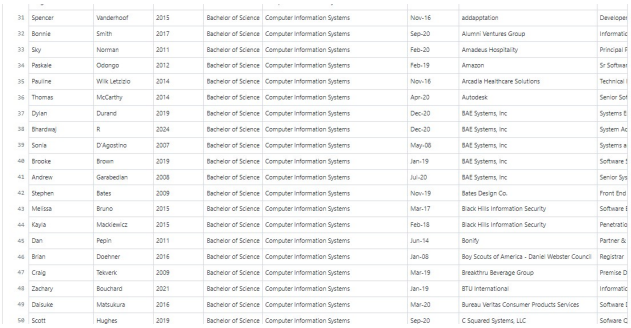
```
#pulling the repository
def run_pull_command():
    return subprocess.check_output(["git","
        pull"])


# reading the datafile
def read_csv():
    df= pd.read_csv("Cleanseddata.csv")
    return df

#adding the changes
def run_add_command():
    return subprocess.check_output(["git","
        add","-A"])

#changes need to be committed
def run_commit_command():
    today= datetime.now().date()
    return subprocess.check_output(["git","
        commit","-m",f"added images to the
        docs"])


# pushing to the repo
def run_push_command():
    return subprocess.check_output(["git","
        push"])


# defining subsets to implement
    visualizations

def do_visualize(df):
    df.dropna(inplace=True)

    df["Graduation Year"] =pd.to_numeric(df
        ['Graduation Year'])

    # creating a new datafram Location by
        joining attributes from original
        data sets
    location = df.groupby(['Location','
        Current Position']).size().
        reset_index(name='counts')

    Current_position= df.groupby(['Current
        Position']).size().reset_index(name
        ='counts')

    degree = df.groupby(['Degree','Location
        ','Current Position','Graduation
        Year', 'Major']).size().reset_index(
        name='counts')
    #plotting the visualizations
    #scatter_3d, bar, sunburst, pie
    fig=px.scatter_3d(location, x='Location
        ', y='Current Position', z='counts',
         color='counts')
    fig1 = px.bar(Current_position, x= '
        Current Position', y='counts')
    fig2=px.sunburst(degree, path=[ 'Major',
         'Location', 'Current Position'],
        values ='counts')
    fig3=px.pie(degree, values='counts',
        names='Degree')

     with open('index.html', 'a') as f:

        #HTML
        f.write(fig.to_html(full_html=False,
            include_plotlyjs='cdn'))

    if __name__ =="__main__":

    run_pull_command()
    print("pulled successfully")
    df=read_csv()
    print("csv file read.")
    do_visualize(df)
    print("Plotting done")
    run_add_command()
    run_commit_command()
    run_push_command()
```

3.5 Embedding Visualisations

After developing interactive visualization, I would like to embed all the visualizations Into GitHub where all the visualizations will change time to time after the updating of the data. Will produce a GitHub page as a product of the project where I would embed all the visualizations that I have obtained from the data.These visualizations are interactive and able to answer the queries that are predefined.To start with embedding the visualisations in GitHub pages site I have created a repository with the name "python-viz" in the organisation called unh-computing-alumni and i have selected GitHub minimal page theme. Now I need add the index.html file that contains the converted visualizations to the repository and the visualizations are published into GitHub pages site at the provided link.

4 RESULTS

I was able to develop the interactive visualizations using python libraries seaborn and plotly. Visualizations are helpful to gain meaningful insights from the data. The university maintains the alumni data, The project involves creating graphs and charts to support the quality of education the university is providing for the students. I have gained lots of insights from the alumni data which supports how successful the university is. I was able to make the visualizations publicly available to the organization using GitHub pages site[1].



31	Spencer	Vanderhoof	2015	Bachelor of Science	Computer information Systems	Nov-16	addaptation	Developer
32	Bonnie	Smith	2017	Bachelor of Science	Computer information Systems	Sep-20	Alumni Ventures Group	Informatic
33	Sky	Norman	2011	Bachelor of Science	Computer information Systems	Feb-20	Amadeus Hospitality	Principal F
34	Peskale	Odongo	2012	Bachelor of Science	Computer information Systems	Feb-19	Amazon	Sr Softwar
35	Pauline	Wilk Letizzio	2014	Bachelor of Science	Computer information Systems	Nov-16	Arcadia Healthcare Solutions	Technical
36	Thomas	McCarthy	2014	Bachelor of Science	Computer information Systems	Apr-20	Autodesk	Senior Sof
37	Dylan	Durand	2019	Bachelor of Science	Computer information Systems	Dec-20	BAE Systems, Inc	Systems E
38	Bhardwaj	R	2024	Bachelor of Science	Computer information Systems	Dec-20	BAE Systems, Inc	System Ac
39	Sonia	D'Agostino	2007	Bachelor of Science	Computer information Systems	May-08	BAE Systems, Inc	Systems a
40	Brooke	Brown	2019	Bachelor of Science	Computer information Systems	Jan-19	BAE Systems, Inc	Software
41	Andrew	Gerabedian	2008	Bachelor of Science	Computer information Systems	Jul-20	BAE Systems, Inc	Senior Sys
42	Stephen	Bates	2009	Bachelor of Science	Computer information Systems	Nov-19	Bates Design Co.	Front End
43	Melissa	Bruno	2015	Bachelor of Science	Computer information Systems	Mar-17	Black Hills Information Security	Software E
44	Kayla	Mackiewicz	2015	Bachelor of Science	Computer information Systems	Feb-18	Black Hills Information Security	Penetratio
45	Dan	Pepin	2011	Bachelor of Science	Computer information Systems	Jun-14	Bonify	Partner &
46	Brian	Doehner	2016	Bachelor of Science	Computer information Systems	Jan-08	Boy Scouts of America - Daniel Webster Council	Registrar
47	Craig	Tekverk	2009	Bachelor of Science	Computer information Systems	Mar-19	Breakthru Beverage Group	Premise D
48	Zachary	Bouchard	2021	Bachelor of Science	Computer information Systems	Jan-19	BTU International	Informatic
49	Daisuke	Matsukura	2016	Bachelor of Science	Computer information Systems	Mar-20	Bureau Veritas Consumer Products Services	Software E
50	Scott	Hughes	2019	Bachelor of Science	Computer information Systems	Sep-20	C Squared Systems, LLC	Software C

Figure 5: Cleansed data

from the figure we can say that most of the alumni completed majors in Computing Information systems. Similarly from other observations most of the software engineers are from Manchester and completed degree in Bachelor of science and took major in computer information systems. The highest percentage of hiring the alumni is from BAE Systems, Collburn Hill Group, liberty mutual

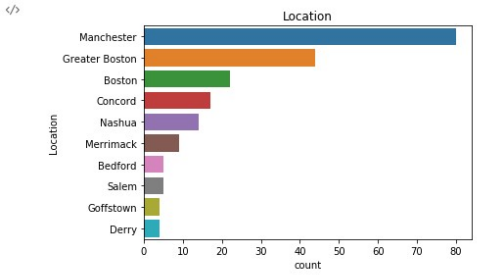


Figure 6: Majors details of alumni

visualizations developed using plotly[3], I have created subsets of tha data frame before producing visualizations. Below figures represents the sunburst graph with the path Major, Location and Current position so first circle represents the alumni major and if we select the any major then it will pop the location of alumni working with that major. if i select a location then it will display the position of alumni working in that location.

The above figure describes the implementation of python code with retrieving the data, reading the data file, producing visualizations in a HTML file and pushing it to the repository for embedding visualizations.

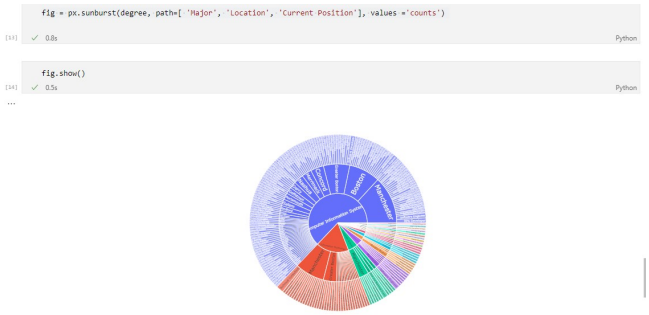


Figure 7: sunburst

```
C:\Users\unhmguest\Desktop\mastersproject\Teja-Sai\src>python project.py
pulled sucessfully
csv file read.
Plotting done
warning: LF will be replaced by CRLF in src/jupyter notebook/CurrentPosition.ipynb.
The file will have its original line endings in your working directory
Enumerating objects: 23, done.
Counting objects: 100% (23/23), done.
Delta compression using up to 4 threads
Compressing objects: 100% (12/12), done.
Writing objects: 100% (14/14), 208.16 KiB | 2.51 MiB/s, done.
Total 14 (delta 6), reused 0 (delta 0)
remote: Resolving deltas: 100% (6/6), completed with 6 local objects.
To https://github.com/master-projects-theses/Teja-Sai.git
   0fddff9..2d8d0fa  main -> main

C:\Users\unhmguest\Desktop\mastersproject\Teja-Sai\src>
```

Figure 8: Implementing the python code for automation

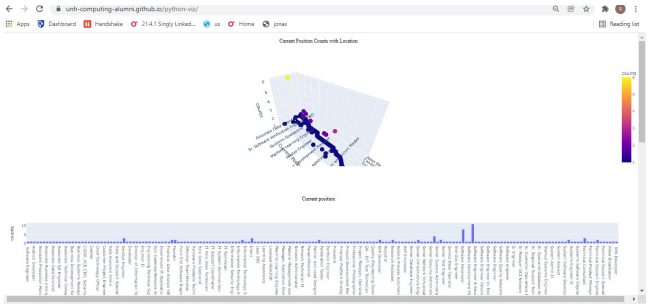


Figure 9: Visualizations in GitHub pages site

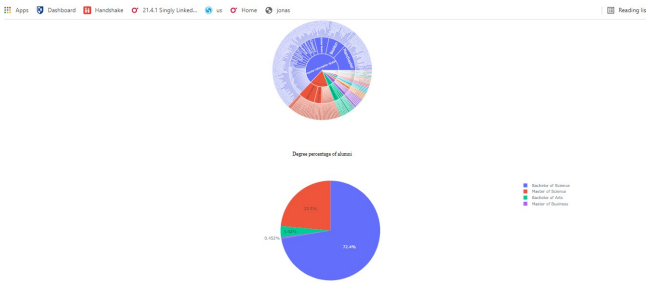


Figure 10: Visualizations in GitHub pages site

Publishing the visualizations in the GitHub pages site[1] and you can view the visualizations in the following link https://unh-computing-alumni.github.io/python-viz/

5 EVALUATION

The visualizations created using python libraries are successfully embedded into the GitHub page's site [1]. The visualizations are publicly available for the organization and students to view at https://unh-computing-alumni.github.io/python-viz/. Published visualizations are interactive and dynamic where you can use click, Hover and selection options for better experience.

I have successfully created interactive graphs and charts using alumni data to support the quality of education that the university is providing.

I was able to automate the process of updating the visualizations by automate the process of retrieving the data from the GitHub repository updating the visualizations and embed into the GitHub pages site.

All the visualizations and queries will help you to gain the meaningful insights of alumni and all the graphs and charts are developed by the python libraries will help you to understand the data in a easy way rather than the looking at a csv file. the predefined queries will provide the more information about the alumni.

REFERENCES

[1] Saiteja goud voruganti. 2021. Embedding | Visualizations Format: acmrt.cls. ://unh-computing-alumni.github.io/python-viz/.
[2] GreekforGreeks. 2021. Python | Pandas Series.str.contains() Typesetting Format: acmrt.cls. https://www.geeksforgeeks.org/python-pandas-series-str-contains/.
[3] plotly. 2021. Python | plotly graphing library Typesetting Format: acmrt.cls. https://plotly.com/python/getting-started/.
[4] plotly. 2021. Python Library | python-api-reference Format: acmrt.cls. https://plotly.com/python-api-reference/generated/plotly.io.to_html.html.
[5] Tableau prep Builder. 2021. Tableau | Data cleansing Typesetting Format: acmrt.cls. https://help.tableau.com/current/prep/en-us/prep_about.htm.
[6] Michael Waskom. 2021. Python Library | VisualizationsFormat: acmrt.cls. https://seaborn.pydata.org/.