CS432/532: Final Project Report

Project Title: Data analysis of Stack Overflow Developers Survey 2022

Team Member(s): Patil Shubham, Shetty Disha

I. Problem

The Stack Overflow Developer Survey 2022 provides a comprehensive dataset that can be analyzed to gain insights into various aspects of software development, including programming languages, frameworks, platforms, tools, education, career satisfaction, diversity, and inclusion. As the demand for software developers continues to grow, understanding the experiences, preferences, and trends of developers worldwide is crucial for educators, employers, policymakers, and technology companies. By analyzing and interpreting the survey data, we can identify patterns, correlations, and discrepancies that can inform decisions related to hiring, training, retention, and innovation in the software industry.

II. Software Design and Implementation

A. Software Design and NoSQL-Database and Tools Used:

MongoDB was majorly utilized to implement the project, with the assistance of tools such as MongoDB Compass and Jupyter Notebook. In addition, we have utilized several libraries including PyMongo for database interaction and matplotlib for data visualization.

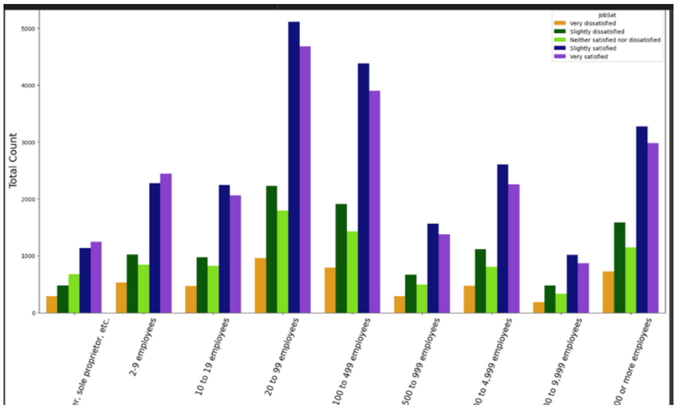
B. Parts that you have implemented:

To perform the analysis, we first created an aggregation pipeline in MongoDB Compass, and then connected it to a Python file using the PyMongo library. After loading the data set into the Python file, we utilized various Python libraries including Matplotlib, Pandas, and NumPy to conduct our analysis.

III. Project Outcome

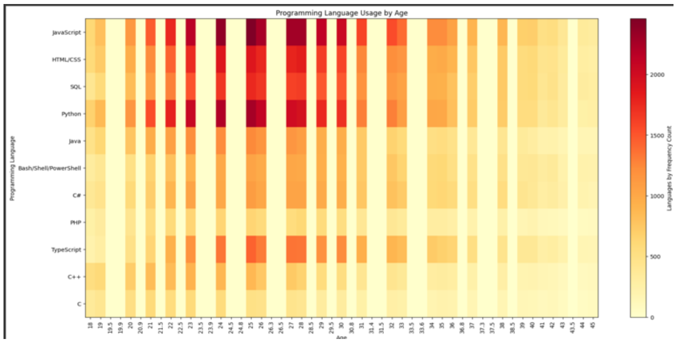
Below are the snapshots of the said analysis:

Analysis I - To analyze what is the impact of company size (i.e., MNC's or Startups or medium scaled IT business) on the job satisfaction of developers/employees.



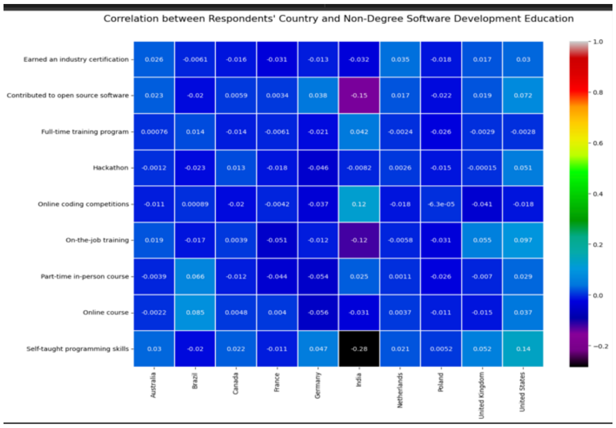
In this analysis it's interesting to note that freelancers/sole proprietors and people in smaller companies of 2-9 employees have the highest proportion of job satisfaction. Also, in larger organizations ranging from 20-99 employees and 100-499 employees still reported a high level of job satisfaction, with more respondents reporting being 'Very satisfied' or 'Slightly satisfied' than neutral or dissatisfied. This suggests that larger organizations can still provide a positive work environment and culture that fosters job satisfaction, despite potential challenges such as distracting work environments and frequent meetings.

Analysis II - To analyze whether is there a correlation between the age of developers and the programming languages they are proficient in or interested in learning.



It is an interesting to note that the trend of younger respondents showing a greater preference for Python, C, C++, and Java. One possible explanation for this trend is that these languages are widely used in computer science and engineering programs at universities and colleges, and younger respondents may have been exposed to them through their coursework or extracurricular activities. Additionally, these languages have a strong presence in the tech industry and are frequently used for building complex software systems and applications, which may make them more attractive to younger developers who are looking to build their skills in these areas.

Analysis III – To analyze Are the methods of learning software development different across various countries?



The correlation analysis reveals interesting trends among the respondents from different countries. For example, Indian respondents are less inclined towards self-teaching and contributing to open-source software but are more likely to participate in online coding competitions and full-time training. On the other hand, respondents from the US are more likely to have self-taught skills, on-the-job training, and open-source contributions, and to participate in hackathons. Brazilian respondents have a higher tendency to learn from online courses, while UK respondents prefer on-the-job training and self-teaching. Interestingly, French respondents are less likely to learn from courses, while German respondents tend to prefer self-teaching and open-source contributions. These findings suggest that different regions may have unique approaches to learning and skill development in the tech industry, which could be influenced by cultural, educational, and economic factors.

Source Code:

For Analysis I.

```
Analysis1_pipeline = [
  {
    "$match": {"$and": [{"OrgSize": {"$ne": None}}, {"JobSat": {"$ne": None}}]}
  },
  {
    "$group": {
      "_id": {
        "OrgSize": "$OrgSize",
        "JobSat": "$JobSat"
      },
      "count": {"$sum": 1}
    }
  },
  {
    "$sort": {"_id.OrgSize": 1}
  },
  {
    "$project": {
      "_id": 0,
      "OrgSize": "$_id.OrgSize",
      "JobSat": "$_id.JobSat",
      "count": 1
    }
  }
]

results = list(stack_overflow_data.aggregate(Analysis1_pipeline))

df = pd.DataFrame(results)

# Define the order of categories for the x-axis of the chart
order= ["Just me - I am a freelancer, sole proprietor, etc.",
    "2-9 employees", "10 to 19 employees", "20 to 99 employees",
    "100 to 499 employees", "500 to 999 employees",
    "1,000 to 4,999 employees", "5,000 to 9,999 employees",
    "10,000 or more employees"]

# Define the order of categories for the hue of the chart
hue_order= ['Very dissatisfied', 'Slightly dissatisfied',
      'Neither satisfied nor dissatisfied',
      'Slightly satisfied', 'Very satisfied']

# Create a clustered bar chart using Seaborn
```

```
palette= {'Very dissatisfied':'orange', 'Slightly dissatisfied':'darkgreen',
        'Neither satisfied nor dissatisfied':'chartreuse', 'Slightly satisfied':'darkblue',
        'Very satisfied':'blueviolet'}



fig, ax = plt.subplots(figsize=(20, 10))
sns.barplot(ax=ax, x='OrgSize', y='count', hue='JobSat', data=df,
          order=order, hue_order=hue_order, palette=palette)
ax.set_xticklabels(ax.get_xticklabels(), fontsize=15, rotation=70)
ax.set_title('Grouping developer satisfaction levels by organization size.', fontsize=20)
ax.set_xlabel('Organization Size', fontsize=18)
ax.set_ylabel('Total Count', fontsize=18);

# Show the chart
plt.show()

df = pd.DataFrame(results)

df.head()
```

For Analysis II.

```
Analysis2_pipeline = [
   {
      "$match": {
         "Age": {"$gte": 18, "$lte": 45}
      }
   },
   {
      "$group": {
         "_id": "$Age",
         "LanguageWorkedWith": {"$push": "$LanguageWorkedWith"}
      }
   },
   {
      "$project": {
         "_id": 0,
         "Age": "$_id",
         "LanguageWorkedWith": {
            "$reduce": {
               "input": "$LanguageWorkedWith",
               "initialValue": "",
               "in": {"$concat": ["$$value", "$$this", ";"]}
            }
         }
      }
   },
   {
      "$project": {
         "Age": 1,
         "LanguageWorkedWith": 1,
         "counts_per_num": {"$split": ["$LanguageWorkedWith", ";"]}
      }
   },
   {
      "$unwind": "$counts_per_num"
   },
   {
      "$group": {
         "_id": {"Age": "$Age", "counts_per_num": "$counts_per_num"},
         "count": {"$sum": 1}
      }
   },
   {
      "$group": {
         "_id": "$_id.Age",
         "age_counts": {"$push": {"Age": "$_id.Age", "count": "$count", "LanguageWorkedWith": "$_id.counts_per_num"}}
      }
   },
   {
      "$project": {
         "_id": 0,
         "Age": "$_id",
         "age_counts": 1
      }
   },
    {
        "$sort": {"Age": 1}
    }
]

# List of all programming languages
languages = ['JavaScript', 'HTML/CSS', 'SQL', 'Python', 'Java', 'Bash/Shell/PowerShell',
            'C#', 'PHP', 'TypeScript', 'C++', 'C']

# Execute the pipeline and extract the data
result = stack_overflow_data.aggregate(Analysis2_pipeline)

data = list(result)

# Create an array to hold the counts for each language and age
counts = np.zeros((len(languages), len(data)))

# Loop through all programming languages and fill in the counts array
for i, language in enumerate(languages):
    for j, d in enumerate(data):
        count = next((ac['count'] for ac in d['age_counts'] if ac['LanguageWorkedWith'] == language), 0)
        counts[i, j] = count

# Set the figure size
plt.figure(figsize=(20, 10))

# Create the heatmap
im = plt.imshow(counts, cmap='YlOrRd', aspect='auto')

# Set the x-tick labels
plt.xticks(np.arange(len(data)), [d['Age'] for d in data], rotation=90)

# Set the y-tick labels
plt.yticks(np.arange(len(languages)), languages)

# Add a colorbar
cbar = plt.colorbar(im)

# Set the colorbar label
cbar.set_label('Languages by Frequency Count')

# Add chart title and axis labels
plt.title("Programming Language Usage by Age")
plt.xlabel("Age")
plt.ylabel("Programming Language")

# Show the chart
plt.show()
```

For Analysis III.

```
Analysis3_pipeline = [
    {"$match":
     {"$or":
      [
        {"Country":
                "United States"},
        {"Country":
                "India"},
        {"Country":
                "Germany"},
        {"Country":
                "United Kingdom"},
        {"Country":
                "Canada"},
        {"Country":
                "France"},
        {"Country":
                "Brazil"},
        {"Country":
                "Poland"},
        {"Country":
                "Australia"},
        {"Country":
                "Netherlands"}
    ]
    }
     },
    {"$project":
     {"_id": 0,
        "Country": 1,
        "EduOther": 1
        }
       }
]

data = list(stack_overflow_data.aggregate(Analysis3_pipeline))
df = pd.DataFrame(data)

# Split the EduOther values into columns of dummy variables
df_edu = df['EduOther'].str.get_dummies(sep=';')
df = df.drop('EduOther', axis=1)

# Split the Country values into columns of dummy variables
df_country = df['Country'].str.get_dummies()
df = df.drop('Country', axis=1)

# Concatenate the dummy variable columns
df = pd.concat([df, df_country, df_edu], axis=1)

# Make a correlation matrix, and then use the correlation matrix as a new dataframe
df3_corr = df.corr()


# Drop the Country rows from one axis, and the EduOther columns from the other
df3_corr.drop(['France', 'India', 'Canada', 'Australia', 'Germany', 'Brazil',
        'United Kingdom', 'Poland', 'Netherlands', 'United States', 'NA'], axis=0, inplace=True)
df3_corr.drop(['Completed an industry certification program (e.g. MCPD)',
        'Contributed to open source software',
        'Participated in a full-time developer training program or bootcamp',
        'Participated in a hackathon',
        'Participated in online coding competitions (e.g. HackerRank, CodeChef, TopCoder)',
        'Received on-the-job training in software development',
        'Taken a part-time in-person course in programming or software development',
        'Taken an online course in programming or software development (e.g., a MOOC)',
        'Taught yourself a new language, framework, or tool without taking a formal course',
        'NA'], axis=1, inplace=True)

# Rename the EduOther rows to conserve space on the graph
df3_corr.rename(index={'Completed an industry certification program (e.g. MCPD)': 'Earned an industry certification',
            'Participated in a full-time developer training program or bootcamp': 'Full-time training program',
            'Participated in a hackathon': 'Hackathon',
            'Participated in online coding competitions (e.g., HackerRank, CodeChef, TopCoder)': 'Online coding competitions',
            'Received on-the-job training in software development': 'On-the-job training',
            'Taken a part-time in-person course in programming or software development': 'Part-time in-person course',
            'Taken an online course in programming or software development (e.g., a MOOC)': 'Online course',
            'Taught yourself a new language, framework, or tool without taking a formal course': 'Self-taught programming skills'},
        inplace=True)

# Plot the results on a heatmap
colormap = plt.cm.nipy_spectral
plt.figure(figsize=(14, 10))
plt.title("Correlation between Respondents' Country and Non-Degree Software Development Education", y=1.05, size=15)
sns.heatmap(df3_corr, linewidths=0.1, vmax=1, square=True, cmap=colormap, linecolor='white', annot=True);
```

IV. REFERENCES

1. https://insights.stackoverflow.com/survey
2. https://pandas.pydata.org/docs/
3. https://www.geeksforgeeks.org/plot-a-pie-chart-in-python-using-matplotlib/
4. https://matplotlib.org/stable/tutorials/colors/colorbar_only.html#sphx-glr-tutorials-colors-colorbar-

only-py

5. https://www.geeksforgeeks.org/seaborn-heatmap-a-comprehensive-guide/
6. https://www.geeksforgeeks.org/matplotlib-pyplot-imshow-in-python/

CS432/575 (based on the IEEE conference template)