

“STATE OF DATA SCIENCE & MACHINE LEARNING 2020”

DIVYANSHI GARG

MTECH DA

2022 batch

INDEX

• Introduction	3
• Problem Statement	4
• Literature Review	5
• Methodology	7
• Data Collection & Preprocessing	9
• Exploratory Data Analysis & Visualization	14
• Feature Engineering	19
• Association Rule Mining	27
• Prediction/Results	28
• References	32

INTRODUCTION

Some years back, when there was not much knowledge about what technology is or how to use it, most of work process in any organizations or business was being done manually i.e., data collection, evaluation, visualization and processing which consumed so much time leading to loss of time as well as cost. Secondly, it also used software which couldn't handle such vast data due to which it caused lot of data redundancy and data inconsistency. Huge losses in business would occur and people started to lose their trust and worthiness on any form of business and organization, therefore, causing the downpours of various business and organizations.

But after few years, as people started to be more passionate towards their work and knowledge basically in technical field, technology also started to rise and thus paved the way of digital era in which now the special technologies like Python, Machine Learning and Data Science are being used by the business area to increase their profit and be the number one in the market as it helps them to understand the capabilities and drawbacks not only in technical but statistical manner also.

“Why Python, Machine learning and data science only can be used in business growth?” is the question which always comes around the mind of people and people get confused. So, the best way to understand this is by making them understand about how helpful it really is. So, Machine learning is basically a technique used, through which meaningful insights of any raw data is extracted with the help of a specific data driven shaping tool such as python for solving the complex data rich behavior and business problems. Because of its various advantages such as fast computational processing technique, affordable data storage, low cost and ease to use, organizations have now started to implement it.

Data science helps businesses in better decision-making strategies on the basis of numbers, facts and statistics. It also helps organizations to understand deeply how market analysis is best strategy for their company's rise. Because of all these factors, the organizations have started to use this technology and to make it more successful they want some experienced and skilled people such as Data scientist, Data Analyst and other job titles which support data science and machine learning and can handle it well to evolve their business. But due to certain parameters such as gender imbalance, age, qualification and technology the evolvement of machine learning and data science has shown very slow progress.

So, to understand the factors behind it we have downloaded the dataset from the Kaggle website of “State of Machine Learning and data science” which includes the survey from various applicants, employees of Data scientist profile and asked varied questions regarding various parameters like Gender, Age, Country, Education, Experience, Employment, Technology, Tools.

So that on basis of the gathered insights of the demographic, technology, employment and education factors we can understand why there is a difference in no. of men/woman in the industry, if higher education is required for such kind of industries or not, how much age difference is there and which particular tools/programming language would be more beneficial and also it would check where we are lacking and how can we overcome it to make more people ready for this job role. The main aim of our project is to understand where business and people are lagging behind in terms of machine learning and data science so that we could overcome that factors and rise above it.

PROBLEM STATEMENT

The data is becoming the new emerging field. As companies are entering the digital world, data science and machine learning is an important aspect for the development of the companies as datasets can predict and shape the future. It is the data scientist's role to transform organizations from reactive environments with static and aged data, to automated ones that continuously learn in real time. Kaggle 2020 Data Science & Machine Learning Survey is used to extrapolate survey responses to general population, including the multiple geographical locations, all genders and adult age groups and diverse academic & professional background, to become familiar with the emerging technical advance in tools, diversity in data science and machine learning. It will capture the comprehensive and association view of evolving state of data science and machine learning over years through a combination of narrative text, data exploration and statistical measures.

LITERATURE REVIEW

The research papers focused on the [3] pedagogic experience, skills and education holds by Data Analyst and Machine Learning Engineers belonging to the different positions and industry level. A pedagogy for the use [4] of a Kaggle competition to teach machine learning is suggested, based on the theories of game-based learning and social constructivism and experience of teaching a master level machine learning subject at our university. Pedagogy is conceived as a 7-step teaching model focused on online competition. It is intended to provide game-like features to make learning more enjoyable, as well as to have a social learning atmosphere to encourage students' peer participation and learning. As many organizations are seeking both of the professionals in many distinct disciplines including statistics, programming, databases and more. Such demand led to the rise of such professions of major relevance. The [8] analysis revealed the impressive growth in the professionals over time including the advances made in machine learning in terms of Information Technology sector on integrating AI capabilities along with software and services to evolve their development process. In reference to it, the study had described with the various Microsoft Teams in which they learned how software applications were built with customer focused Artificial Intelligence features.

There is growing mainstream and academic attention to the potential for machine learning (ML) systems to intensify social inequality and unfairness. A boom in recent work has centered on the development of algorithmic tools for evaluating [2] and mitigate unfairness. However, if these instruments are to have a positive effect on industry practice, it is critical that their design is guided by an understanding of practical needs of the world. [7] Via 35 semi-structured interviews and an online survey of 267 ML practitioners, we undertook the comprehensive investigation of the challenges and needs of commercial product teams to establish fairer ML systems. In the ML [1]research literature, the data collected from two main phases i.e., Interviews and Survey discussed about various Applications of Artificial Intelligence in traditional areas such as search, advertising, machine translation, predicting customer purchases, voice recognition, and image recognition and being used in novel areas, such as identifying customer leads, providing design advice for presentations and word processing texts, which offer different drawing features, health care, and improved gameplay. And in response, the respondents used a broad spectrum of ML approaches to build their applications, from classification, clustering, dynamic programming, and statistics, to user behavior modeling, social networking analysis, and collaborative filtering and many more.

To do so, the [11] first concept is “A Business-first research approach” which basically focuses on the basis of deep learning and tells us about factors of it. For instance, various architectures such as Google Net, etc. along with its framework like TensorFlow, some GPU's such as Tesla V100 and many more such as cost, etc. Second concept is about datasets with a special motto that ‘Every data set is unique’ which we need to keep in mind during solving the problems related to business. [9] Third concept contains “Planning for an End-to-End solution” tells us of various methodologies such as Data Complexity, Productization, System Performance Metrics and Performance Monitoring and User Feedback so that by using his we can carry our business in an efficient and healthy manner.

So, to[13] further increase the business and gain profit over it, organizations mostly prefer to have those candidates who are Master or Ph.D. degree holder in the field of cognitive computing, ML, AI, computer vision or NLP. Secondly, the scientist needs to be equipped with sets of skills such as problem solving, literature review, analysis, writing and presentations so that they could easily handle any overcoming situations. [12] Thirdly, the various concepts and methodologies are defined on the basis of which one can be able to become an effective applied scientist in corporate environment on basis of ML based system design which tells about requirements for ML systems such as Quantitative Targets/ Functional

requirements, Explain ability, Freedom from discrimination, Legal and Regulatory requirements and Data requirements such as data quantity, data quality.

Finally, after further discovering and researching,[10] the three aspects of the AI domain were identified that make it fundamentally different from prior software application domains: Discovering, managing, and versioning the data required for machine learning applications is far a lot of advanced and troublesome than different sorts of software engineering.[5] The research findings clearly shows that it can be a highly successful approach for teaching machine learning to use the online competition teaching model to run a Kaggle competition. All Machine Learning teachers are supposed to benefit from this teaching model. Students became more inspired to explore new approaches and apply advanced methods to develop their ideas further, found the experience exciting and enjoyable, and had a better understanding of the application of the principles and algorithms they studied to solve problems in the real world.[6] The emphasis on symbolic representations of learned knowledge, such as production laws, decision trees and logical formulas, was also characterized by the early research effort on machine learning.

In contrast [14] with more complex tasks, such as reasoning, problem solving, and language comprehension that had traditionally played important roles, one of the major changes implemented in the field included an increased focus on classification and regression tasks. The teaching model can be applied to the teaching of other subjects related to computer science and engineering, such as bioinformatics. Apart15] from this, graduates across fields indulge into different jobs upon graduation, and people with CS, stats or electrical engineering background tend more in the field of data science and ML due to the wide range of opportunities. There exists a strong connection between the mindset of physics and data science. As people with other disciplines work with empirical data which is messy and practical that encourages the mindset much needed for the data scientist. On the basis of these results, we highlight guidelines for future ML and HCI research which will better address the needs of practitioners.

METHODOLOGY

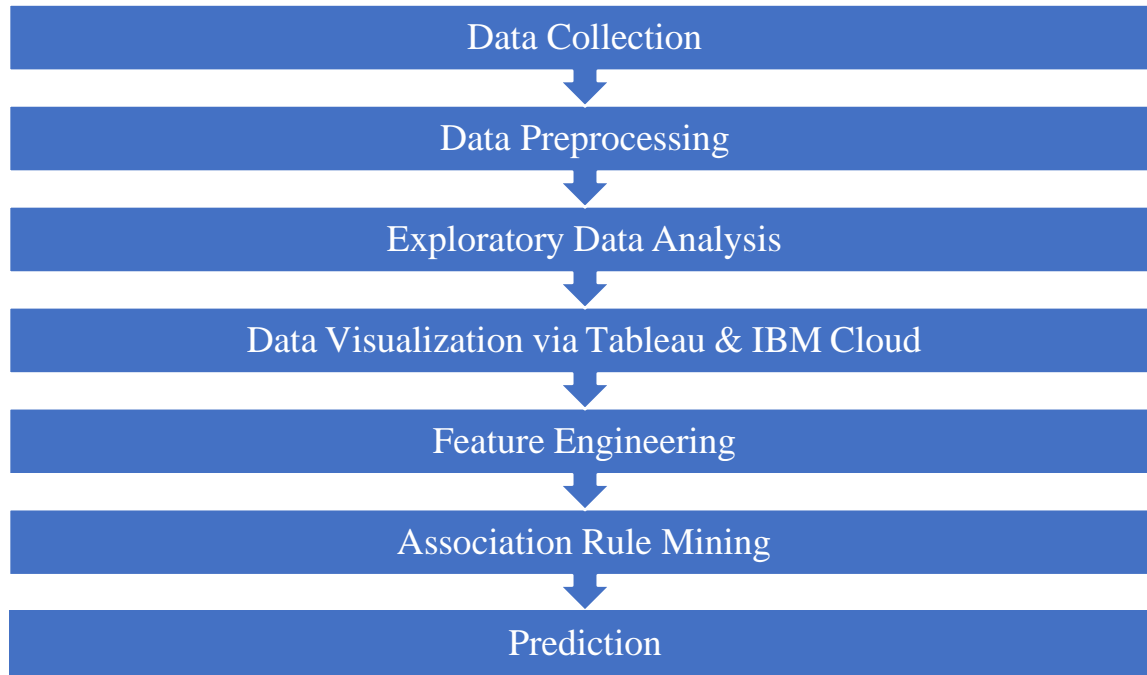


Fig 1. Flow Chart of Proposed framework

Data Collection:

The first step in our project is data collection and then processing is performed on it. We have downloaded the dataset from Kaggle website titled “2020 Kaggle Machine Learning & Data Science Survey” which includes the survey from various applicants who are of Data scientist profile and asked varied questions regarding multiple parameters like Gender, Age, Country, Education, Experience, Employment, Technology, Tools.

The link of the dataset is provided, <https://www.kaggle.com/c/kaggle-survey-2020/data>

Data Preprocessing:

To get deeper knowledge, preprocessing is done on dataset. The dataset is checked on various parameters: the shape of the dataset, the number of null values, irrelevant rows/columns are dropped and new features are added. It makes the data ready for the next step, which is Exploratory Data Analysis (EDA) and Visualization using graphs.

Exploratory Data Analysis:

Exploratory data analysis is a statistical approach of data sets analyzations to summarize their main characteristics, by using statistical representations and other visualization methods. So, to analyze the data sets on basis of their main features, we have created various graphs on basis of features like line, bar, heat, pie chart and many more.

Data Visualization:

Data visualization is the graphical illustration of records and statistics which is performed by the BI tool Tableau and IBM Cloud Model. By the use of visible factors like charts, graphs, and maps, statistics visualization gear offers an available manner to peer and apprehend trends, outliers, and patterns in statistics of different attributes of data such as machine learning experience, age, gender, the highest level of education attained etc.

Feature Engineering:

Feature engineering is the system of the use of area information to extract features from uncooked data. These features may be used to enhance the overall performance of device studying algorithms. Feature engineering may be taken into consideration as implemented device studying itself. Three separate new data frames were created along with varied features to find out the correlation between them and drop the irrelevant variables for better prediction.

Association Rule Mining:

Association Rule Mining is a type of unsupervised learning that helps to find relationships between seemingly independent data or other data repositories. It is based on different rules to find an interesting relationship between variables in the database. It aims to view recurring patterns, interactions, or associations from data sets found in a variety of information topics such as relationship-related information, transaction information, and other types of storage. A typical example is Market Based Analysis. Market Based Analysis is used to show associations between items.

Prediction:

In this through Association Rule Mining, we will first find out the three main metrics such as support, confidence and lift on the basis of which we would be predicting the final results on basis of which we can get to know that which features have more associations with each other.

DATA COLLECTION & PREPROCESSING

To get a high-level understanding of data (columns, nulls, shape etc.) and to identify data scientists and machine learning engineer based on the various diversified factors, the first step is Data Collection and then perform preprocessing on it.

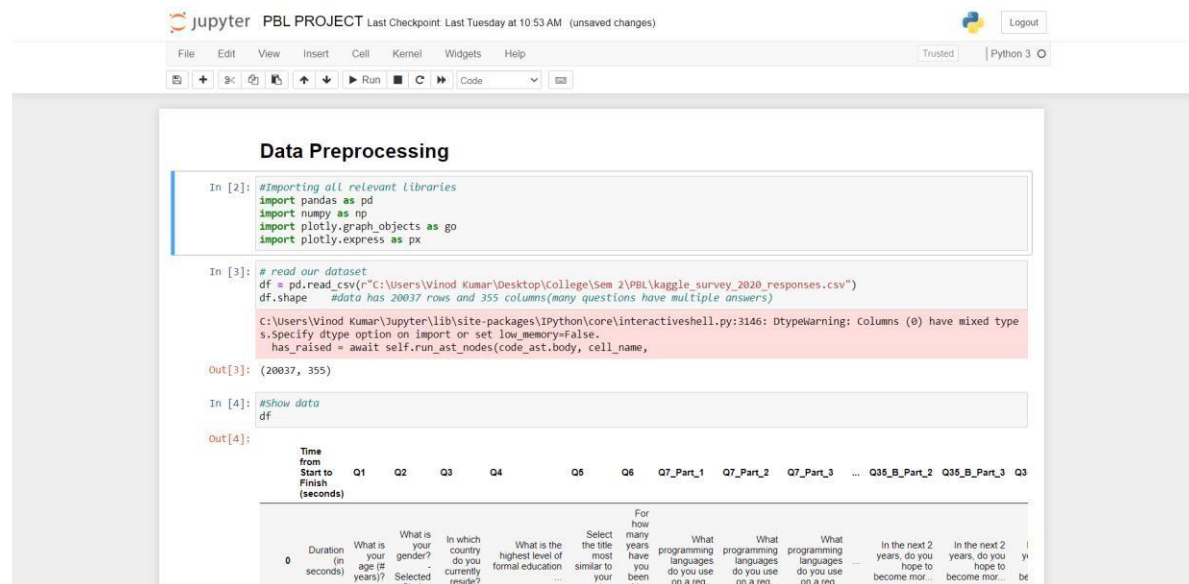


Fig 2. Load dataset

Fig.2 represents the import of relevant libraries, load the dataset and display it. The dataset is downloaded from Kaggle.

Link: <https://www.kaggle.com/c/kaggle-survey-2020>

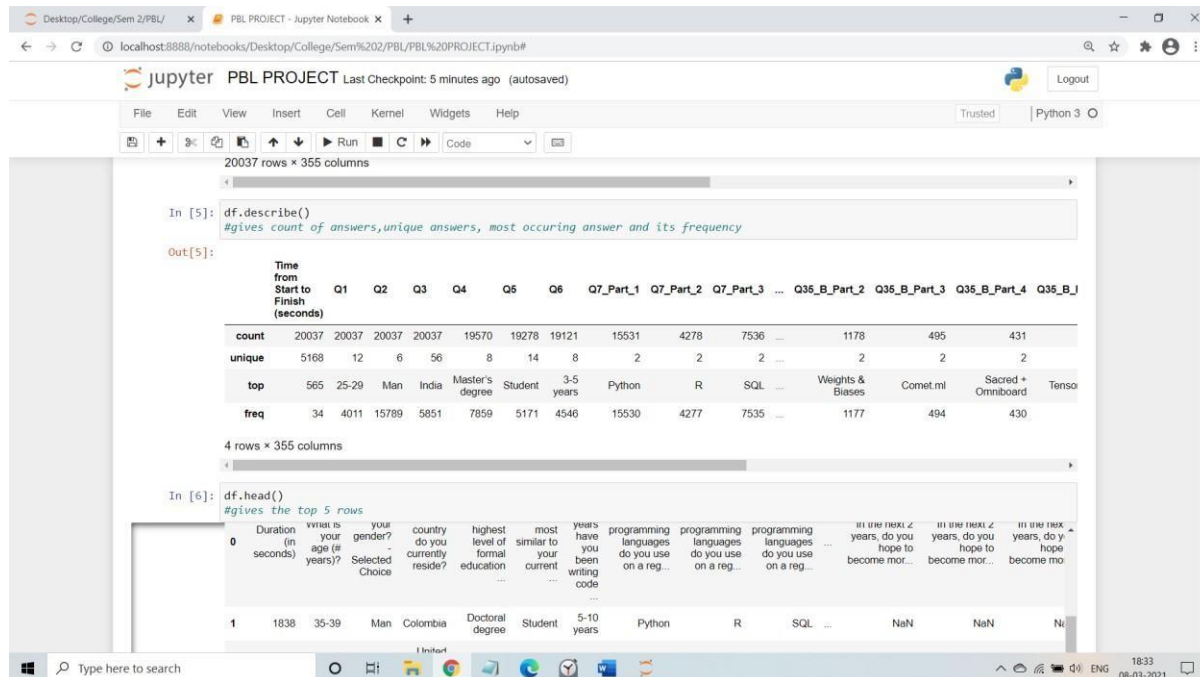


Fig 3. Summary of dataset

Fig. 3 describes the summary of all the features of data and display the top 5 rows of the dataset via head() function.

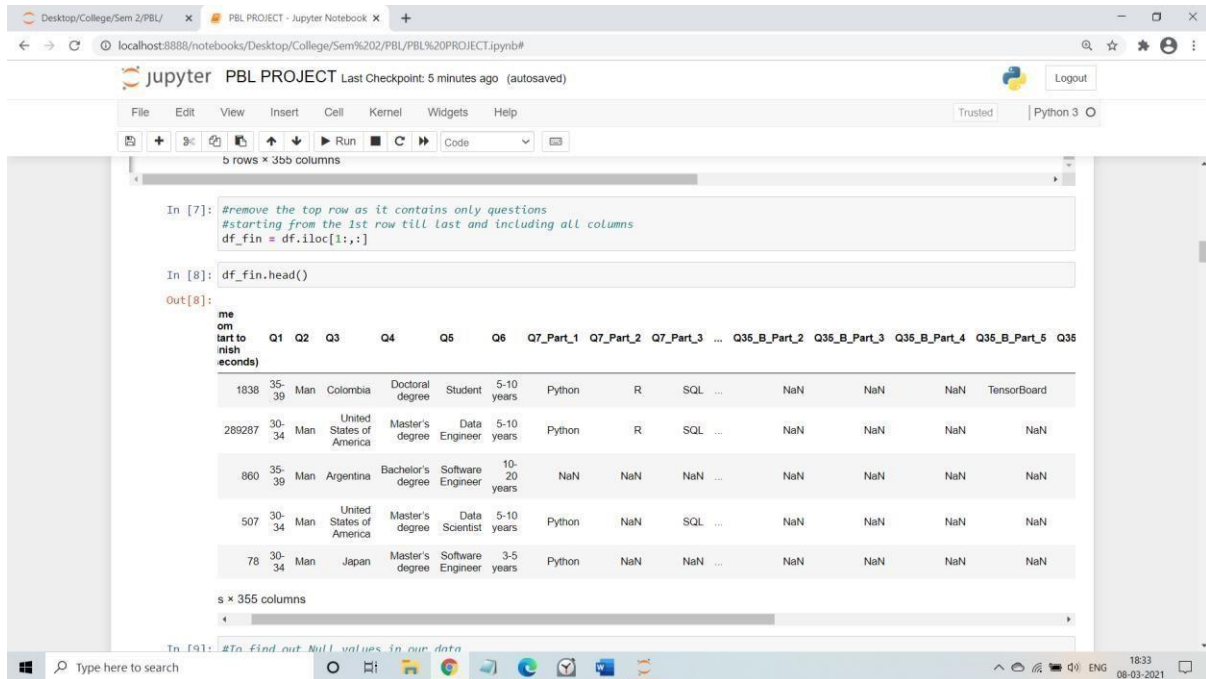


Fig 4. Exclude Top row

Then, remove the top row (Fig.4) as it doesn't contain useful information.

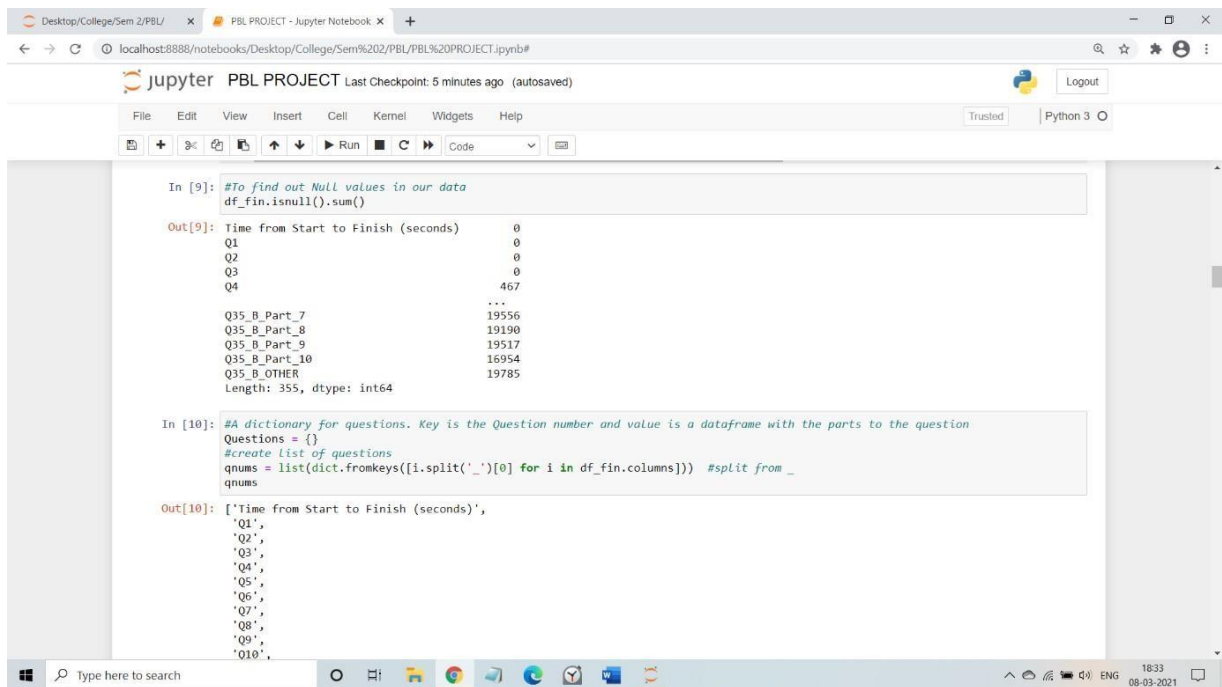
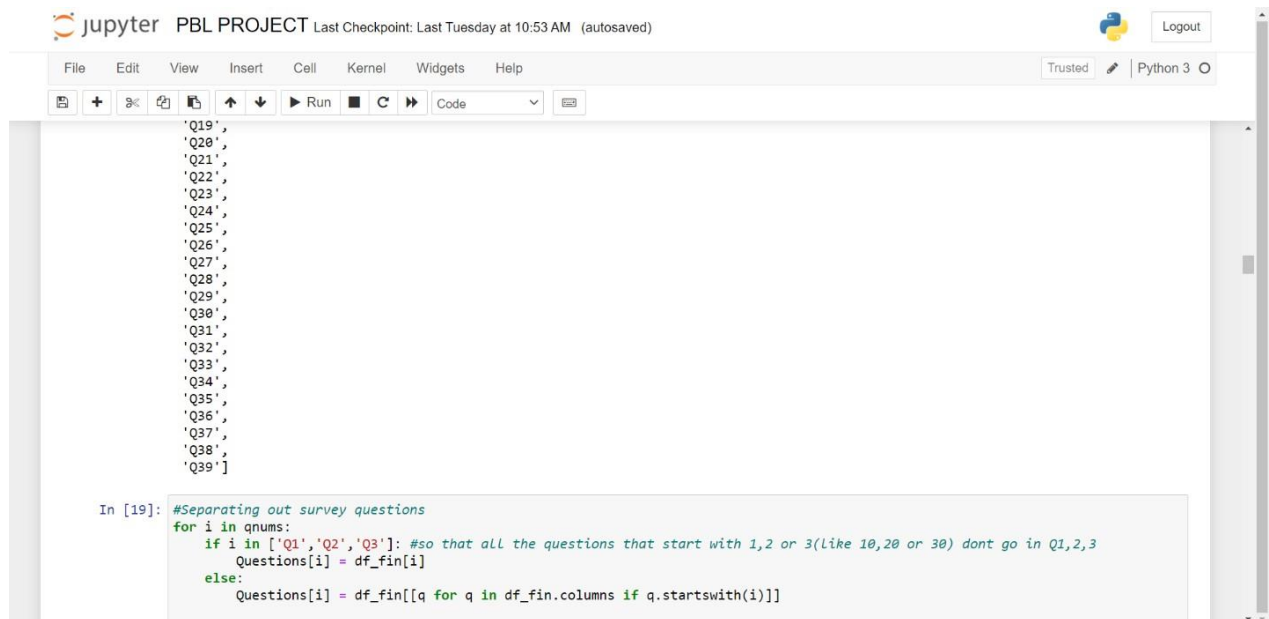


Fig 5. Look for Null Values

Fig.5 shown above finds out the number of null values in our dataset. As it can be seen that Q1, Q2 and Q3 have no null values.

Then we create an empty dictionary. We store the questions in a list and add them to this dictionary in the next step. We do this so that it becomes easy to access all the questions and can analyze each question in isolation.



The screenshot shows a Jupyter Notebook titled 'PBL PROJECT'. The top bar indicates the last checkpoint was at 10:53 AM. The interface includes a menu bar (File, Edit, View, Insert, Cell, Kernel, Widgets, Help) and a toolbar with icons for file operations, running, and code execution. The code cell shows a list of questions from 'Q19' to 'Q39' and a loop that adds them to a dictionary named 'Questions'.

```

'Q19',
'Q20',
'Q21',
'Q22',
'Q23',
'Q24',
'Q25',
'Q26',
'Q27',
'Q28',
'Q29',
'Q30',
'Q31',
'Q32',
'Q33',
'Q34',
'Q35',
'Q36',
'Q37',
'Q38',
'Q39']

In [19]: #Separating out survey questions
for i in qnums:
    if i in ['Q1','Q2','Q3']: #so that all the questions that start with 1,2 or 3(Like 10,20 or 30) dont go in Q1,2,3
        Questions[i] = df_fin[i]
    else:
        Questions[i] = df_fin[[q for q in df_fin.columns if q.startswith(i)]]

```

Fig 6. Adding list to dictionary

Now we add the list of questions to the dictionary created in the previous step.

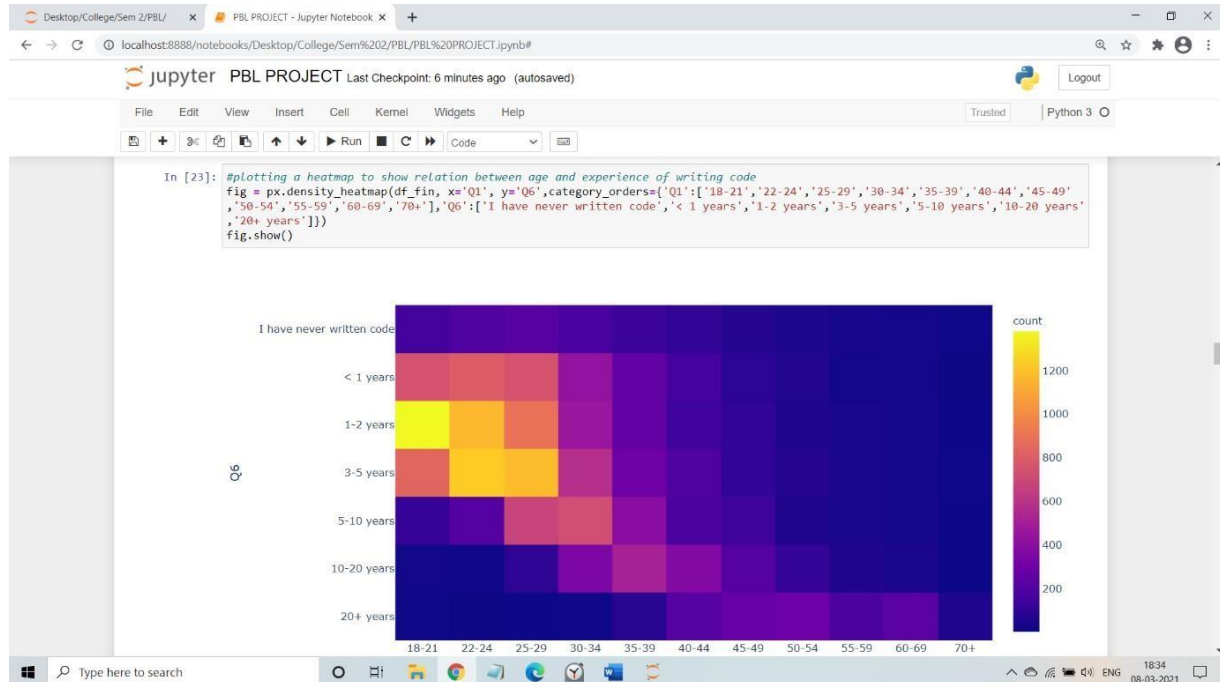


Fig 7. Heatmap between Age groups & code experience

The heatmap between Q1 and Q6 shows the correlation between the ages and experience of writing code. It shows that people of age 18-21 have 1-2 years of experience, people of age 22-24 have around 3-5 years of experience, and so on.

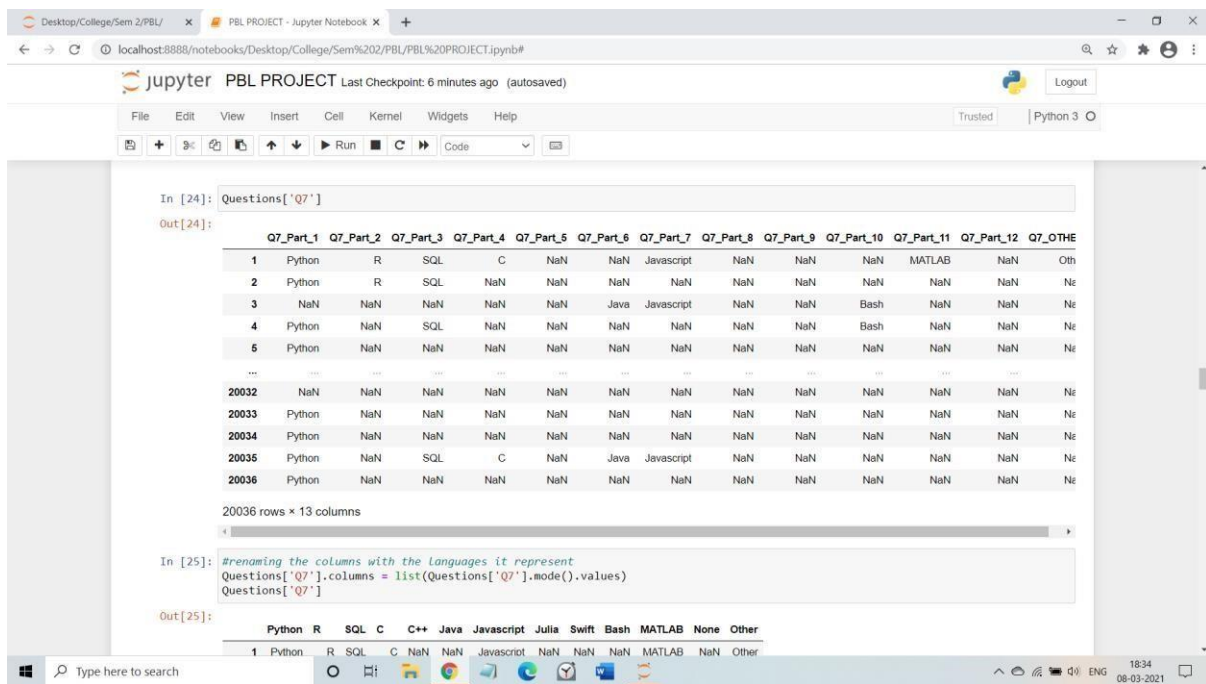
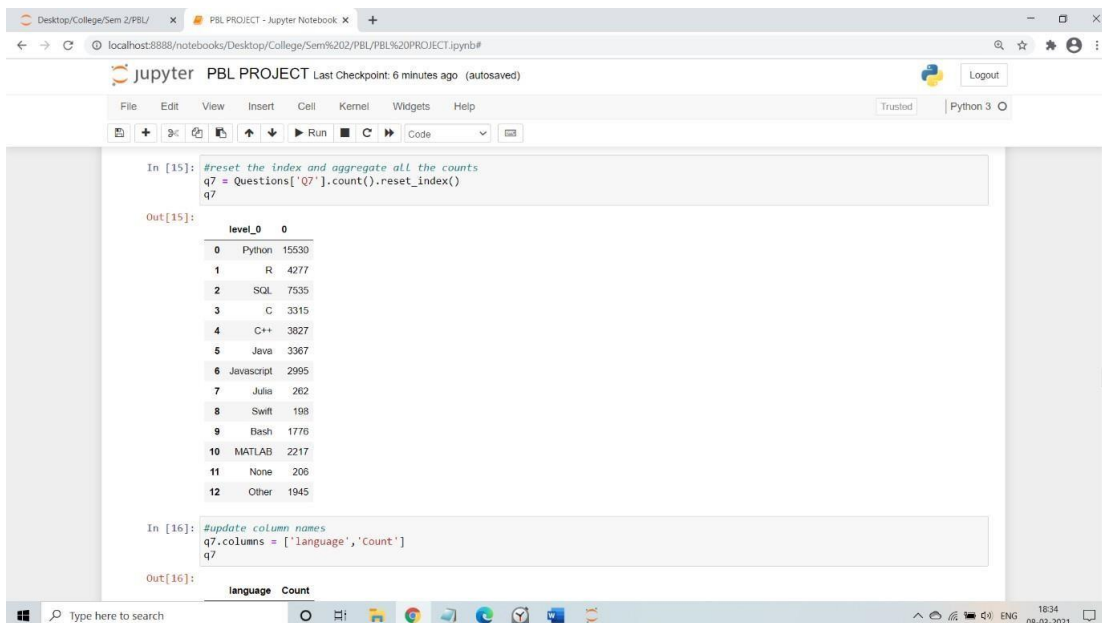


Fig 8. Rename columns

Now we try to explore Q7, the answer of which is stored in a different format. Firstly, rename the columns of Q7 by all the programming languages.



To format the data frame differently, take the count of each column and display it along with the languages used such that we are grouping by the column names.

Fig 9. Count languages

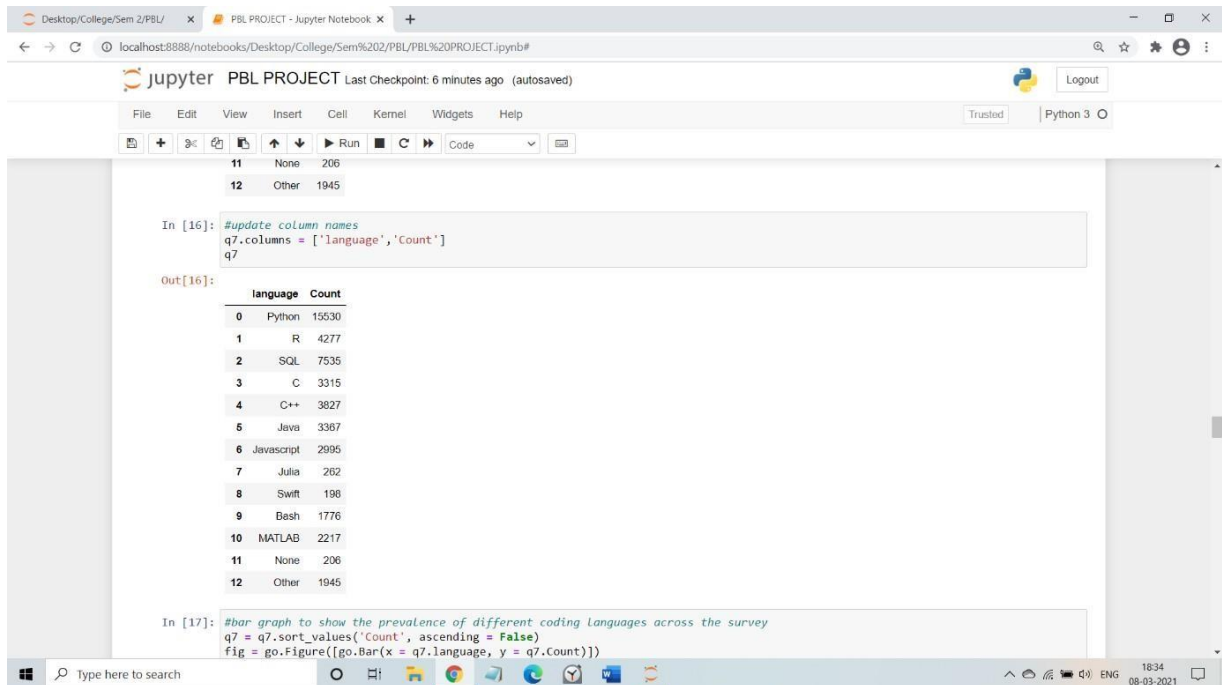


Fig 10. Rename Columns Names

Here, we update the column names and now data is stored meaningfully and it is easy to plot various graphs.

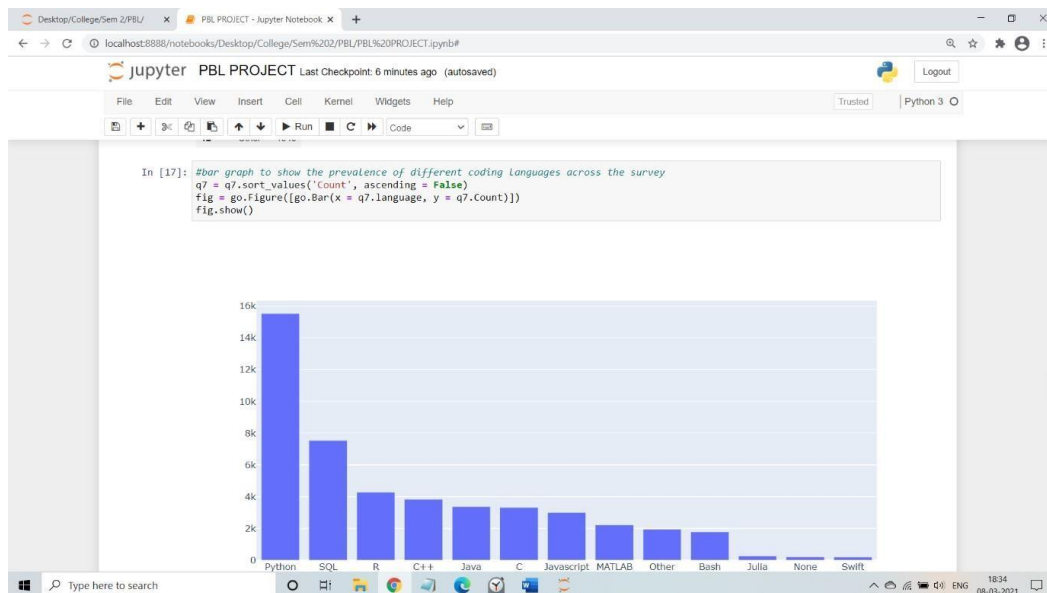
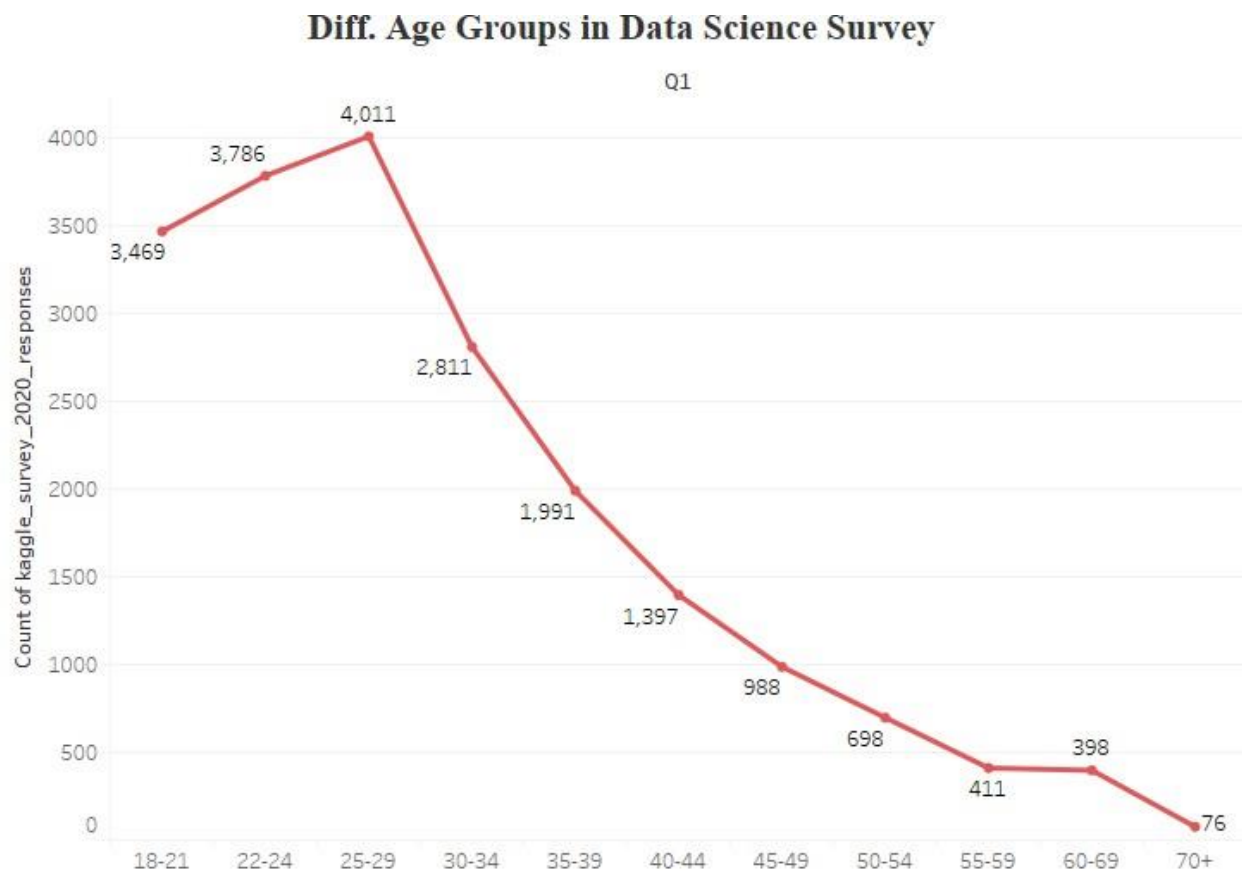


Fig 11. Visualization of programming languages as per count

From the bar graph of Q7, we can infer that Python is the most used language by the respondents.

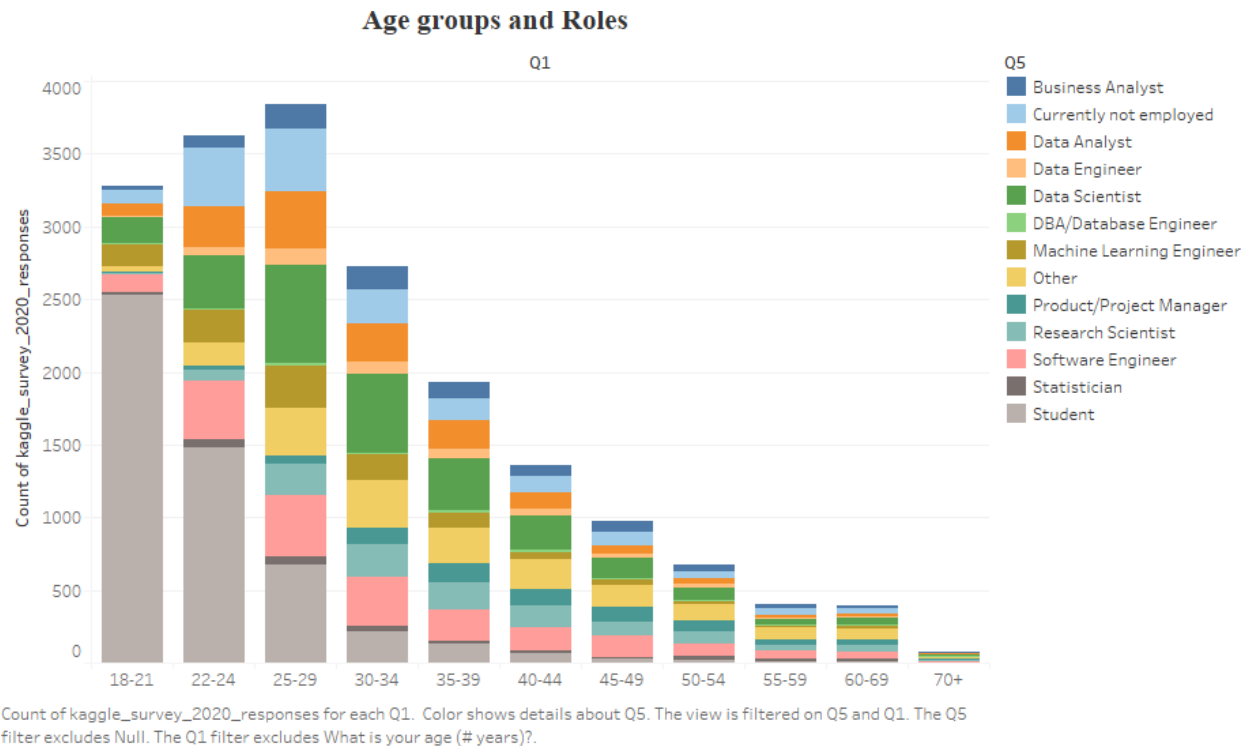
EXPLORATORY DATA ANALYSIS & DATA VISUALIZATION

Age groups of practitioners in Data Science Survey:

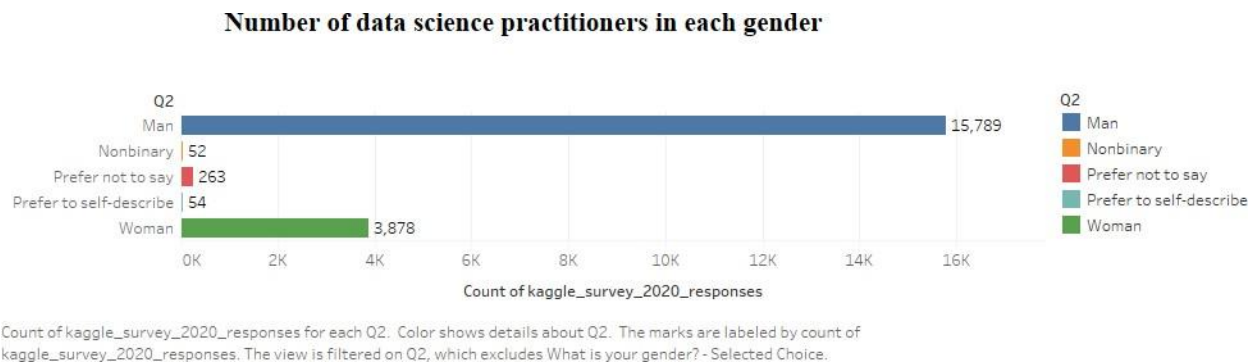


The trend of count of kaggle_survey_2020_responses for Q1. The marks are labeled by count of kaggle_survey_2020_responses. The view is filtered on Q1, which excludes What is your age (# years)?.

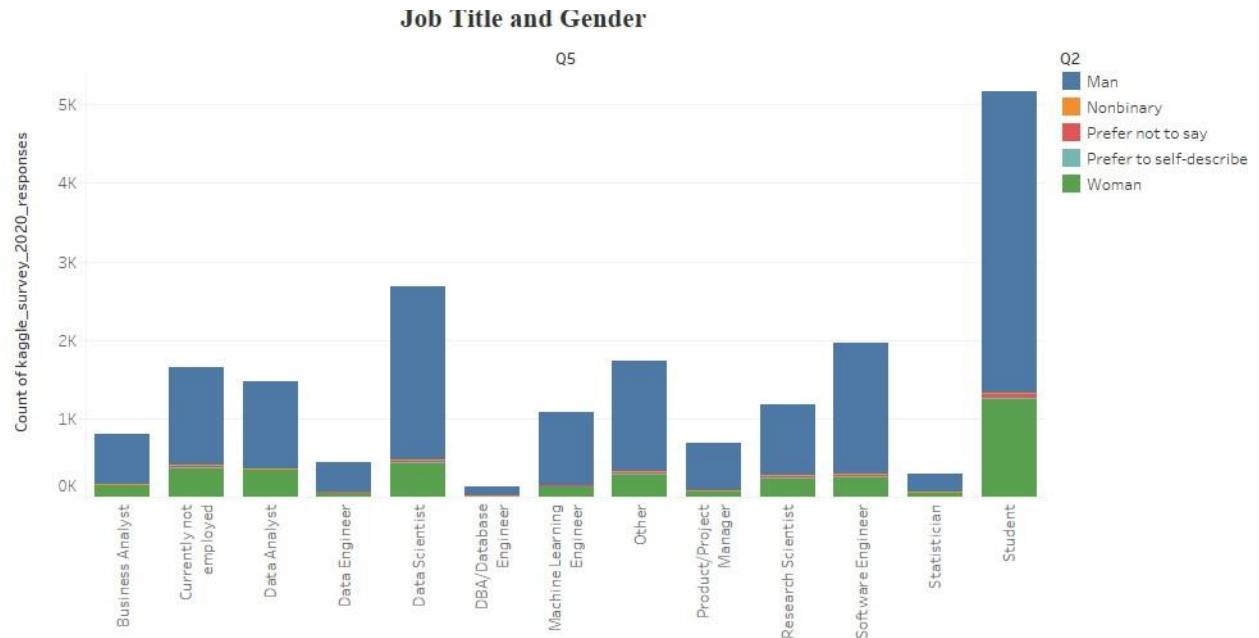
As we can see through the line plot, it seems that majority of the respondents are in the age group of 25-29 years. It's obvious that there some Kagglers under 18 but probably the minority. Also, we see that there are a large number of people from 40 years and older, which is a very good information.



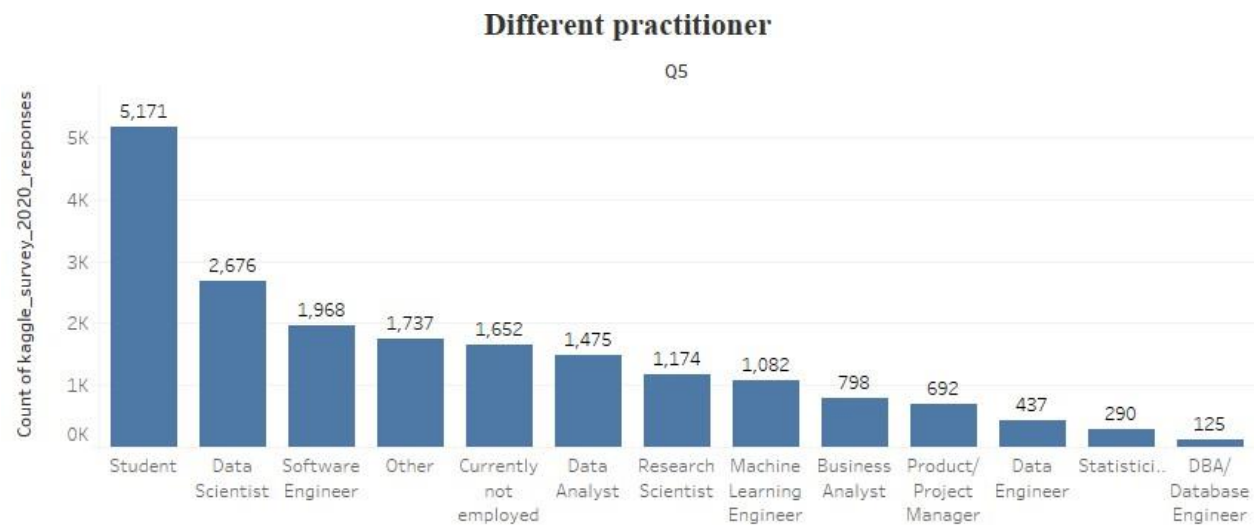
The graph shows the distributions of job titles to different age group. It's clearly visible that the most job titles resemble to Data Science and ML are from 25-29 years of age.



As it happens in the engineering degrees, there are more men than women.



Count of kaggle_survey_2020_responses for each Q5. Color shows details about Q2. The view is filtered on Q2 and Q5. The Q2 filter excludes What is your gender? - Selected Choice. The Q5 filter excludes Null.



Count of kaggle_survey_2020_responses for each Q5. The marks are labeled by count of kaggle_survey_2020_responses. The data is filtered on Q1 and Exclusions (Q1,Q5). The Q1 filter excludes What is your age (# years)?. The Exclusions (Q1,Q5) filter keeps 152 members. The view is filtered on Q5, which excludes Null.

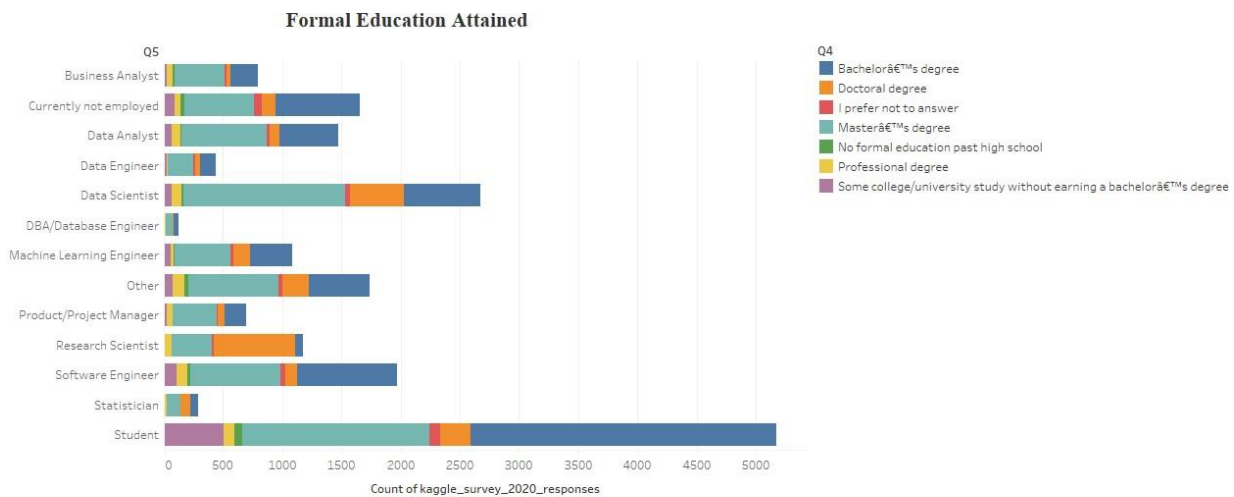
The majority of respondents are Students followed by Data Scientists and Software Engineer in the survey.

Highest level of formal education attained by data science practitioners



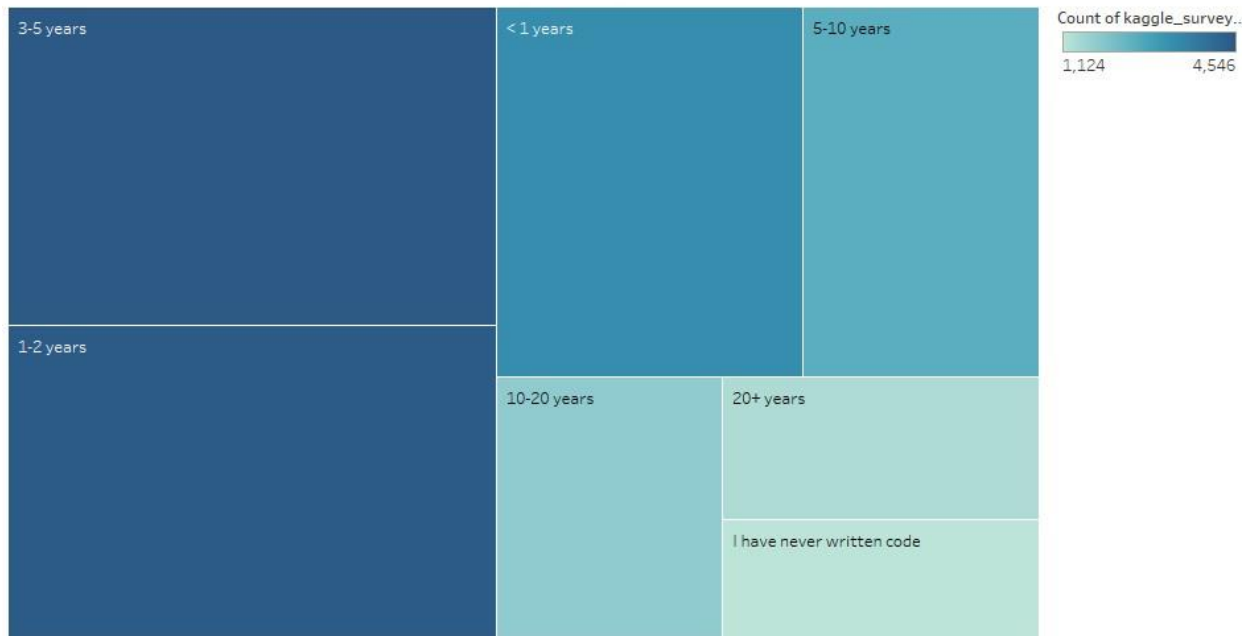
Count of kaggle_survey_2020_responses broken down by Q5 vs. Q4. Color shows count of kaggle_survey_2020_responses. The marks are labeled by count of kaggle_survey_2020_responses. The view is filtered on Q4 and Q5. The Q4 filter excludes Null and What is the highest level of formal education that you have attained or plan to attain within the next 2 years?. The Q5 filter excludes Null.

As this is a scientific community, most of them have bachelors or master's degree. Still ratio of master's degree is more than bachelor's degree. There are also practitioners with No formal education past high school.



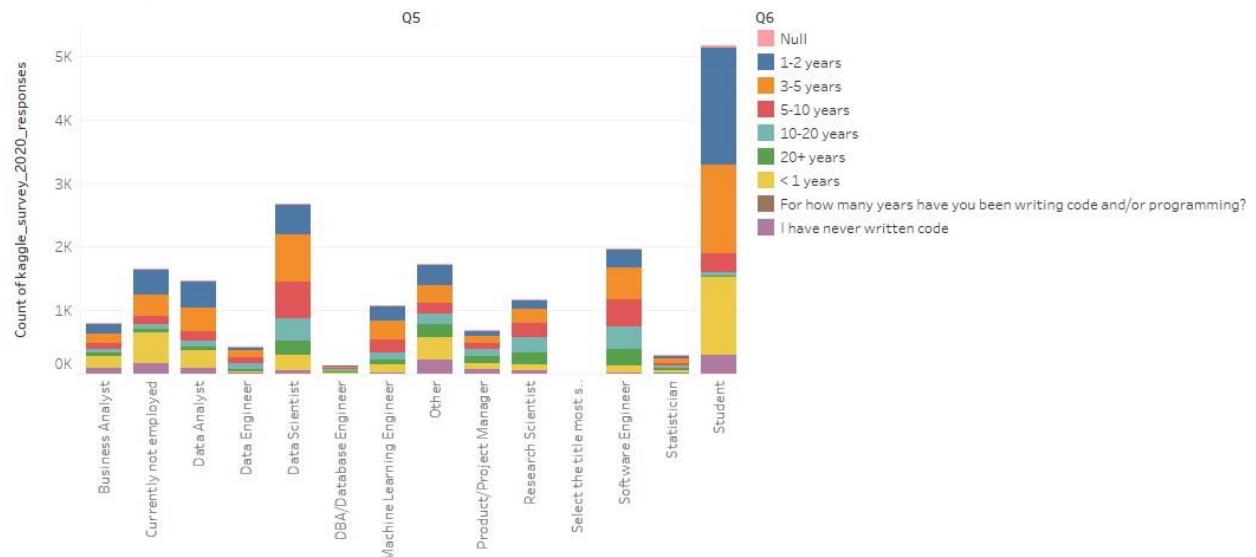
Count of kaggle_survey_2020_responses for each Q5. Color shows details about Q4. The view is filtered on Q5, which excludes Null and Select the title most similar to your current role (or most recent title if retired): - Selected Choice.

Kagglers programming experience



Q6. Color shows count of kaggle_survey_2020_responses. Size shows count of kaggle_survey_2020_responses. The marks are labeled by Q6. The view is filtered on Q6, which excludes Null and For how many years have you been writing code and/or programming?.

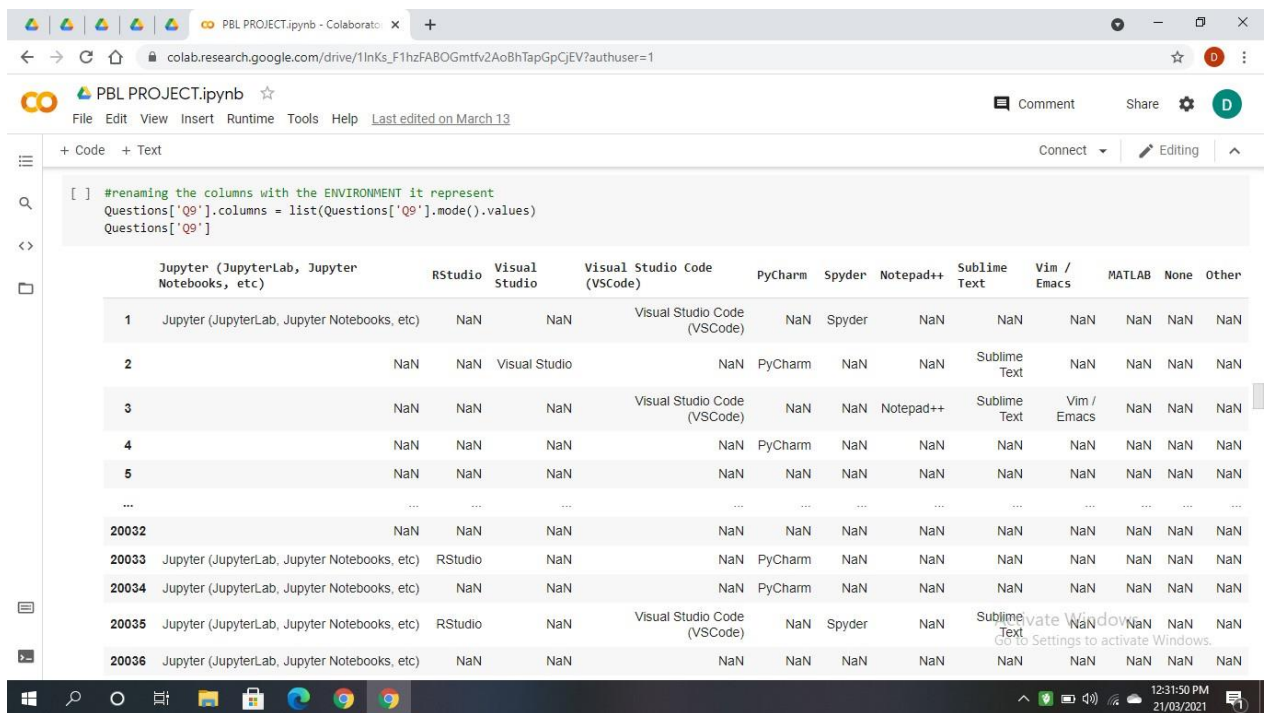
Kagglers programming experience as per their roles



Count of kaggle_survey_2020_responses for each Q5. Color shows details about Q6. The view is filtered on Q5, which excludes Null.

The majority of data science practitioners have less than 5 years of coding expertise. Many practitioners have coding experience of less than 1 year. This shows that more and more students are breaking into the field. Only few practitioners have coding experience of more than 20 years in the survey.

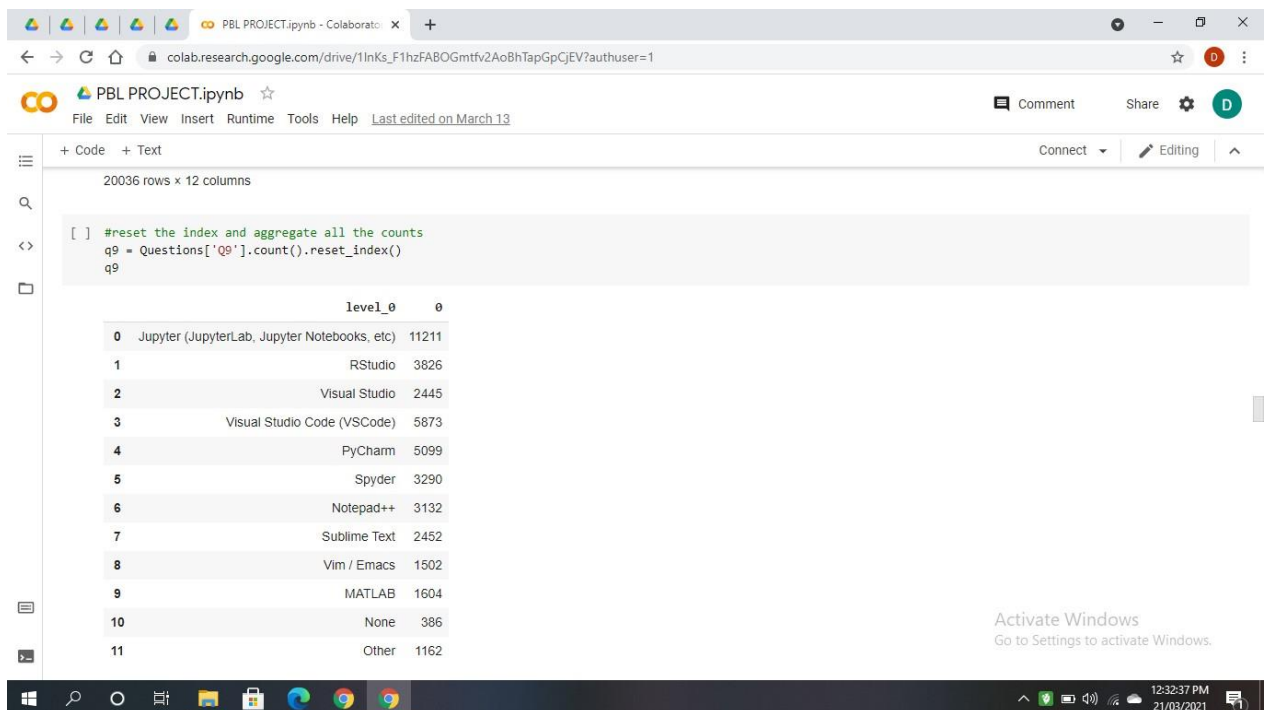
FEATURE ENGINEERING



The screenshot shows a Google Colab notebook titled "PBL PROJECT.ipynb". The code cell contains a comment: "#renaming the columns with the ENVIRONMENT it represent" and a line of code: `Questions['Q9'].columns = list(Questions['Q9'].mode().values)`. Below the code, a table is displayed with 13 columns representing different IDEs: Jupyter (JupyterLab, Jupyter Notebooks, etc), RStudio, Visual Studio, Visual Studio Code (VSCode), PyCharm, Spyder, Notepad++, Sublime Text, Vim / Emacs, MATLAB, None, and Other. The table has 20036 rows, with the first 5 rows showing the distribution of IDEs used by the dataset.

	Jupyter (JupyterLab, Jupyter Notebooks, etc)	RStudio	Visual Studio	Visual Studio Code (VSCode)	PyCharm	Spyder	Notepad++	Sublime Text	Vim / Emacs	MATLAB	None	Other
1	Jupyter (JupyterLab, Jupyter Notebooks, etc)	NaN	NaN	Visual Studio Code (VSCode)	NaN	Spyder	NaN	NaN	NaN	NaN	NaN	NaN
2		NaN	NaN	Visual Studio	NaN	PyCharm	NaN	NaN	Sublime Text	NaN	NaN	NaN
3		NaN	NaN	NaN	Visual Studio Code (VSCode)	NaN	NaN	Notepad++	Sublime Text	Vim / Emacs	NaN	NaN
4		NaN	NaN	NaN	NaN	PyCharm	NaN	NaN	NaN	NaN	NaN	NaN
5		NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Fig 12. Rename columns



The screenshot shows a Google Colab notebook titled "PBL PROJECT.ipynb". The code cell contains a comment: "#reset the index and aggregate all the counts" and a line of code: `q9 = Questions['Q9'].count().reset_index()`. Below the code, a table is displayed with 2 columns: "level_0" and "count". The table has 20036 rows, with the first 11 rows showing the distribution of IDEs used by the dataset.

level_0	count
0 Jupyter (JupyterLab, Jupyter Notebooks, etc)	11211
1 RStudio	3826
2 Visual Studio	2445
3 Visual Studio Code (VSCode)	5873
4 PyCharm	5099
5 Spyder	3290
6 Notepad++	3132
7 Sublime Text	2452
8 Vim / Emacs	1502
9 MATLAB	1604
10 None	386
11 Other	1162

Fig 13. Reset index and aggregate all counts

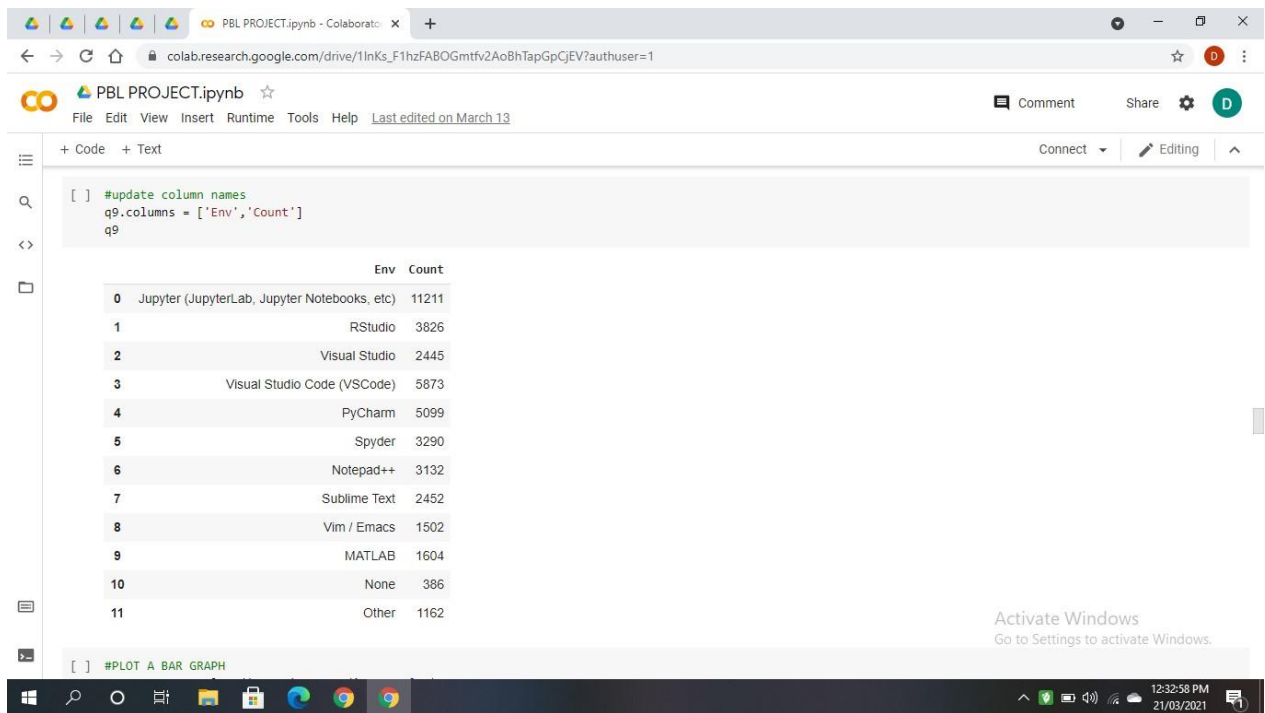


Fig 14. Update column names

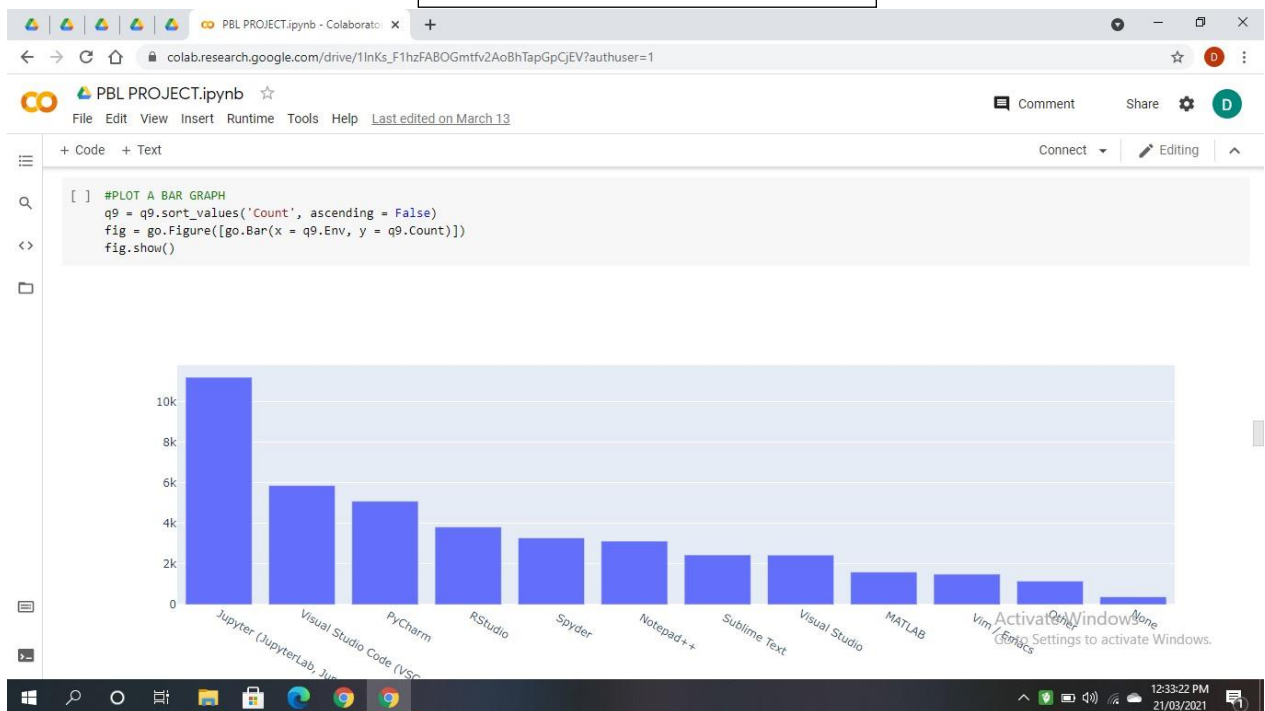


Fig 15. Code editor visualization

Jupyter Lab/Notebook is widely used among data scientists followed by Visual Studio Code.



Fig 16. Notebooks

Colab notebooks and Kaggle notebooks are similar in usage by data scientists.

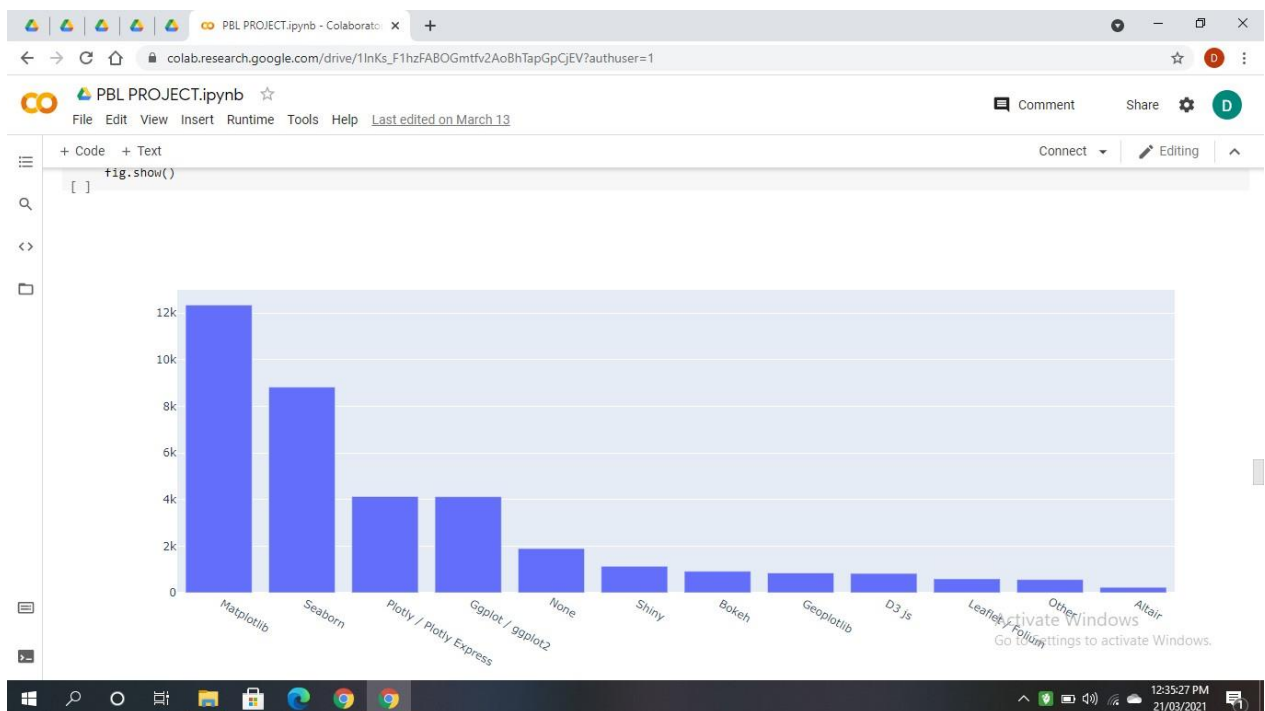


Fig 17. Data Visualization Libraries

Matplotlib is widely used as a data visualization library by data scientists and Machine learning engineers.



Fig 18. Cloud Computing

Cloud computing is a tool that delivers a variety of services through the internet. AWS Cloud and Google Cloud Platform have an almost similar count.

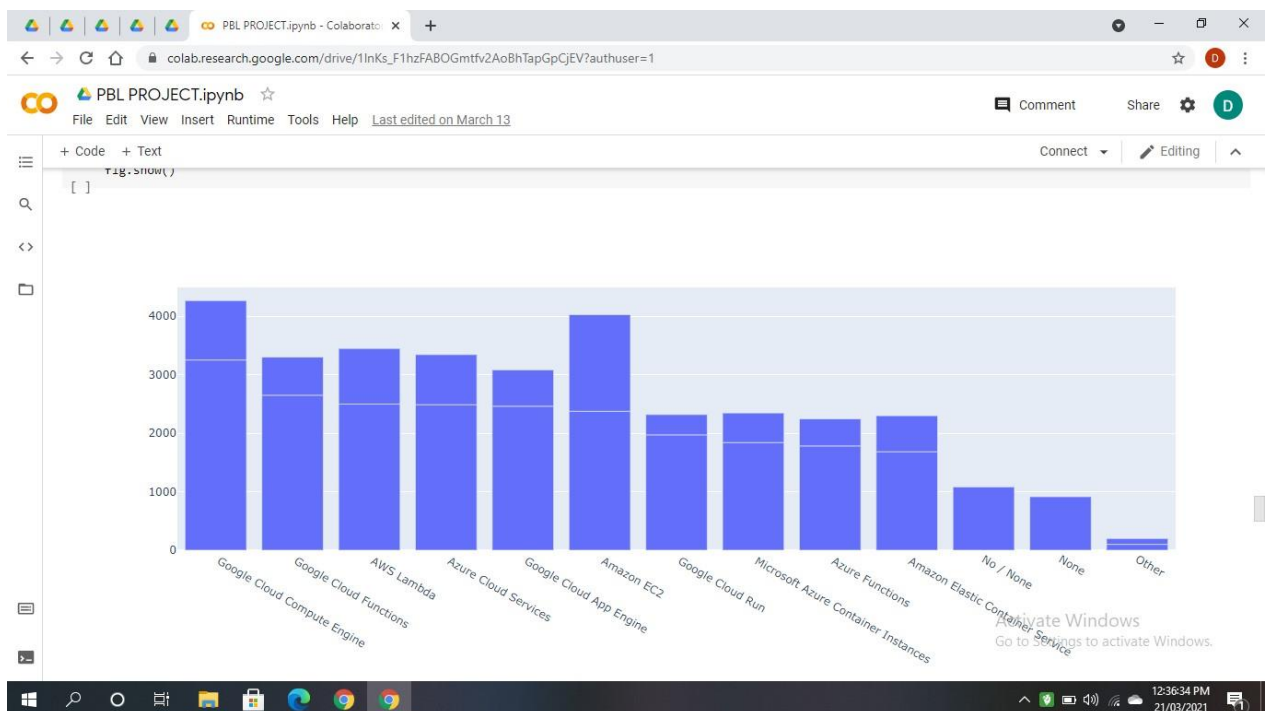


Fig 19. Cloud Services

colab.research.google.com/drive/1lnKs_F1hzFABOGmtfv2AoBhTapGpCjEV?authuser=1

PBL PROJECT.ipynb

File Edit View Insert Runtime Tools Help Last edited on March 13

+ Code + Text

Connect Editing

Source	Count (approx.)
Coursera	5500
Kaggle Learn Courses	4500
Udemy	4200
University Courses (resulting in a university degree)	3500
DataCamp	2800
edX	2200
Udacity	1800
Other	1500
LinkedIn Learning	1200
None	1000
Cloud-certification programs (direct from AWS, Azure, GCP, or similar)	800
Fast.ai	700

Fig 20. Learning platforms

The screenshot shows a Google Colab notebook titled "PBL PROJECT.ipynb". The notebook is open in a web browser, displaying the code editor. The code defines a list of categorical variables and their corresponding codes. The variables are grouped into two sections: "df1-df_fin.copy()" and "df1-df_fin.copy()". The codes are listed for each variable, such as "Q1_C", "Q2_C", "Q3_C", etc. The notebook interface includes a menu bar with options like "File", "Edit", "View", "Insert", "Runtime", "Tools", and "Help". The status bar at the bottom shows the time as 12:37:25 PM on 21/03/2021.

```
[ ] df1-df_fin.copy()
df1['Q1_C']=df1['Q1'].astype('category').cat.codes
df1['Q2_C']=df1['Q2'].astype('category').cat.codes
df1['Q3_C']=df1['Q3'].astype('category').cat.codes
df1['Q4_C']=df1['Q4'].astype('category').cat.codes
df1['Q5_C']=df1['Q5'].astype('category').cat.codes
df1['Q6_C']=df1['Q6'].astype('category').cat.codes
df1['Q71_C']=df1['Q7_Part_1'].astype('category').cat.codes
df1['Q72_C']=df1['Q7_Part_2'].astype('category').cat.codes
df1['Q73_C']=df1['Q7_Part_3'].astype('category').cat.codes
df1['Q74_C']=df1['Q7_Part_4'].astype('category').cat.codes
df1['Q75_C']=df1['Q7_Part_5'].astype('category').cat.codes
df1['Q76_C']=df1['Q7_Part_6'].astype('category').cat.codes
df1['Q77_C']=df1['Q7_Part_7'].astype('category').cat.codes
df1['Q78_C']=df1['Q7_Part_8'].astype('category').cat.codes
df1['Q79_C']=df1['Q7_Part_9'].astype('category').cat.codes
df1['Q710_C']=df1['Q7_Part_10'].astype('category').cat.codes
df1['Q711_C']=df1['Q7_Part_11'].astype('category').cat.codes
df1['Q712_C']=df1['Q7_Part_12'].astype('category').cat.codes
df1['Q70_C']=df1['Q7_OTHER'].astype('category').cat.codes
df1['Q8_C']=df1['Q8'].astype('category').cat.codes
df1['Q91_C']=df1['Q9_Part_1'].astype('category').cat.codes
df1['Q92_C']=df1['Q9_Part_2'].astype('category').cat.codes
df1['Q93_C']=df1['Q9_Part_3'].astype('category').cat.codes
df1['Q94_C']=df1['Q9_Part_4'].astype('category').cat.codes
df1['Q95_C']=df1['Q9_Part_5'].astype('category').cat.codes
df1['Q96_C']=df1['Q9_Part_6'].astype('category').cat.codes
df1['Q97_C']=df1['Q9_Part_7'].astype('category').cat.codes
df1['Q98_C']=df1['Q9_Part_8'].astype('category').cat.codes
df1['Q99_C']=df1['Q9_Part_9'].astype('category').cat.codes
```

Fig 21. New Data frame

To visualize the correlation between features of the original dataset via heatmap. A new data frame along with the new features are created.

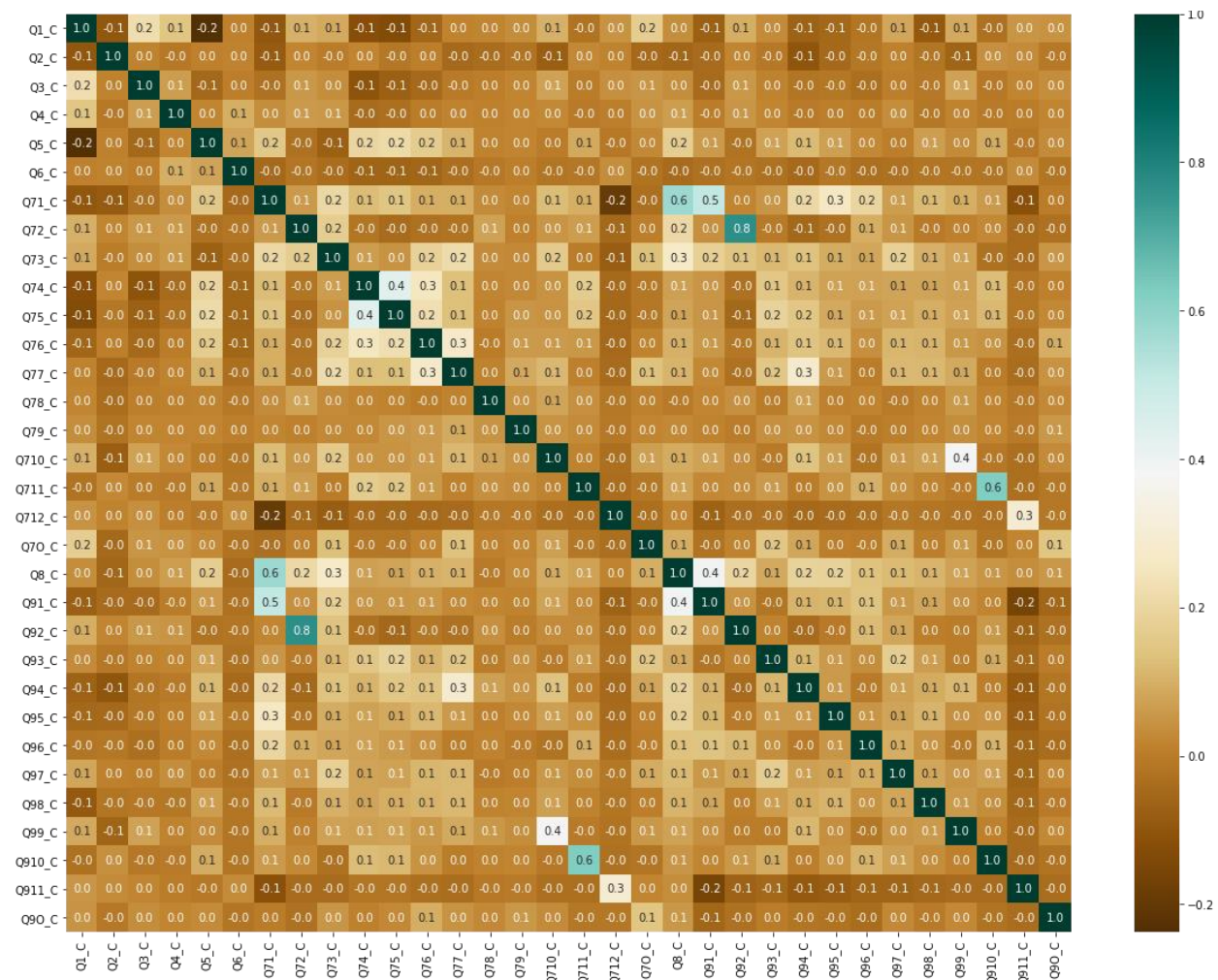


Fig 22. Heatmap of Df1

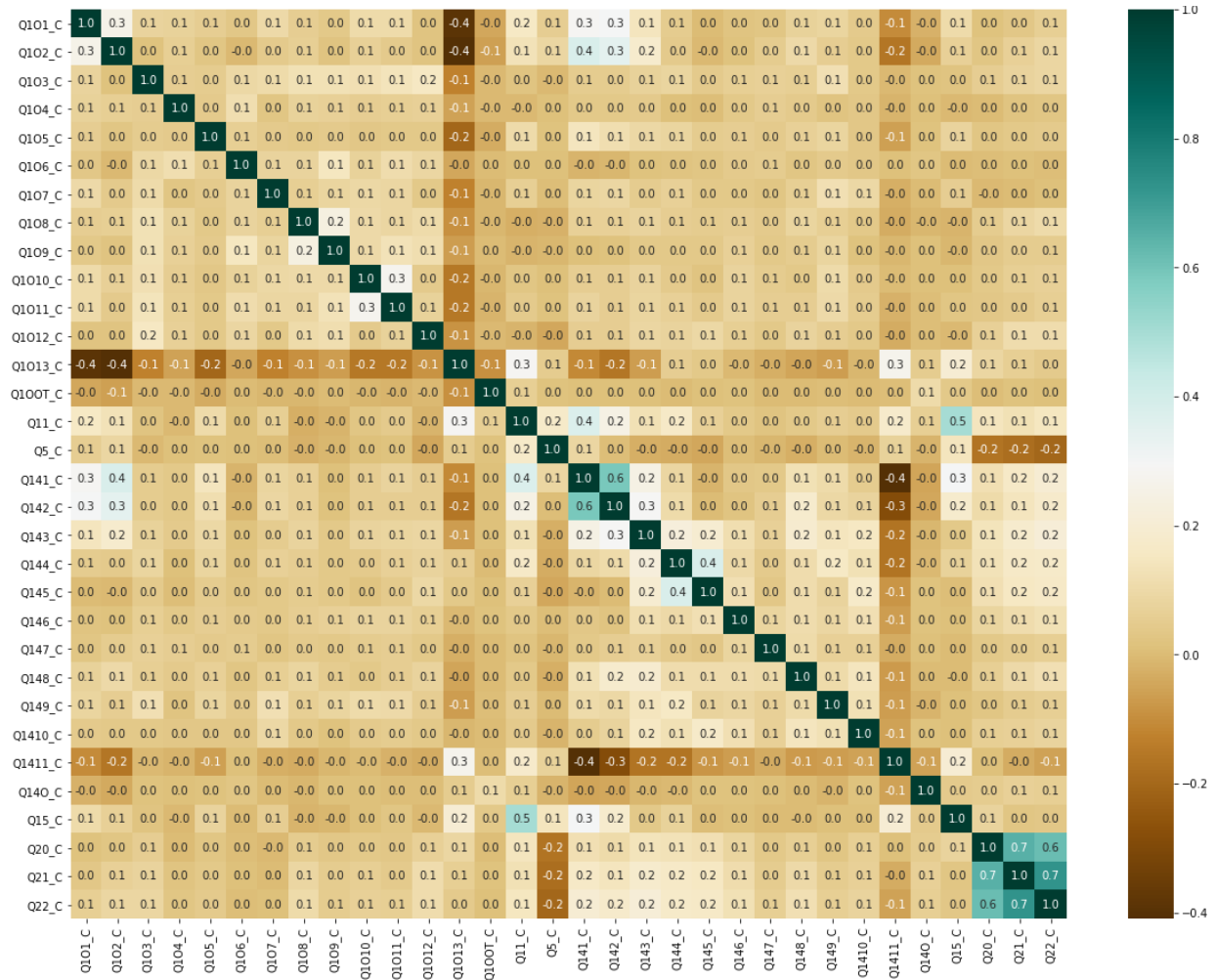


Fig 23. Heatmap of Df2

ASSOCIATION RULE MINING

We have generated rules using three main metrics such as Min. Support = 0.02, Min. Confidence = 0.6 and Min. Lift = 3 to take out association rule mining results.

```
In [46]: MIN_SUPPORT = 0.02
MIN_CONFIDENCE = 0.6
MIN_LIFT = 3
MAX_LENGTH = 2

rules = apriori(product_list,min_support=MIN_SUPPORT,min_confidence=MIN_CONFIDENCE,min_lift=MIN_LIFT,max_length=MAX_LENGTH)
rules = list(rules)

rules_df = pd.DataFrame()
for i in range(len(rules)):
    rules_df = rules_df.append(pd.DataFrame({"Antecedent": (list(rules[i][2][0][0])[0]).strip(),
                                             "Consequent": (list(rules[i][2][0][1])[0]).strip(),
                                             "Support": np.round(float(list(rules[i][1])[0]),3),
                                             "Confidence": np.round(float(list(rules[i][2][0][2]),3),
                                             "Lift": np.round(float(list(rules[i][2][0][3]),3),index=[i]))

rules_df = rules_df.loc[(rules_df["Consequent"]!='') & (rules_df["Antecedent"]!='')].sort_values(by=["Lift","Confidence"],ascending=False)
rules_df.head(10)
```

```
Out[46]:
```

	Antecedent	Consequent	Support	Confidence	Lift
0	Auto-Sklearn	Automated model selection (e.g. auto-sklearn, ...	0.021	0.710	17.445
1	Generative Adversarial Networks	Generative Networks (GAN, VAE, etc)	0.035	0.692	12.704
2	Google Cloud AI Platform / Google Cloud ML Engine	Google Cloud Compute Engine	0.022	0.612	12.085
3	Transformer Networks (BERT, gpt-3, etc)	Transformer language models (GPT-3, BERT, XLne...	0.053	0.817	11.458
4	Azure Functions	Microsoft Azure	0.021	0.926	10.887
5	Microsoft Azure Container Instances	Microsoft Azure	0.023	0.911	10.704
6	Azure Cloud Services	Microsoft Azure	0.038	0.900	10.571
7	Azure Machine Learning Studio	Microsoft Azure	0.024	0.876	10.299
8	Amazon Elastic Container Service	Amazon EC2	0.025	0.802	9.728
9	Amazon SageMaker	Amazon EC2	0.023	0.740	8.983

```
In [47]: rules_df.sort_values(by='Confidence',ascending=False,inplace = True)
n = 1
for i in range(len(rules_df)):
    if (rules_df.iloc[i,3]*100) < 70:
        break
    print(f'{n}. {colored(np.round(rules_df.iloc[i,3]*100,1),"grey","on_cyan")}{colored("%", "grey", "on_cyan")} of the respondents
    n += 1
```

```
1. 98.8% of the respondents who use Image classification and other general purpose networks (VGG, Inception, ResNet, ResNeXt, NASNet, EfficientNet, etc) also use Convolutional Neural Networks
2. 98.7% of the respondents who use Image segmentation methods (U-Net, Mask R-CNN, etc) also use Convolutional Neural Networks
3. 98.6% of the respondents who use General purpose image/video tools (PIL, cv2, skimage, etc) also use Convolutional Neural Networks
4. 98.5% of the respondents who use Object detection methods (YOLOv3, RetinaNet, etc) also use Convolutional Neural Networks
5. 96.9% of the respondents who use Amazon Elastic Container Service also use Amazon Web Services (AWS)
6. 96.1% of the respondents who use Amazon EC2 also use Amazon Web Services (AWS)
7. 95.9% of the respondents who use Amazon SageMaker also use Amazon Web Services (AWS)
8. 95.6% of the respondents who use AWS Lambda also use Amazon Web Services (AWS)
9. 94.8% of the respondents who use Generative Networks (GAN, VAE, etc) also use Convolutional Neural Networks
```

Fig 25. Apriori Algorithm

PREDICTION/RESULTS

Features	Inference
Age group	22-29 years
Gender	Male
Diff. practitioners	Student, Data Scientists
Highest level of formal education	Master's degree
Programming language Used	Python, SQL
Programming language first recommended	Python
Programming experience	Less than 5 years
Machine Learning experience	0-2 years
Employment in company	10,000 Or More Employees & 0-49 Employees
Current use of ML in business	Well established ML methods
Pay distribution of respondents	\$100,000-\$124,99
Money spent on Cloud Computing & ML	Highest by Data Scientists & least by statisticians
Cloud database service	SQL &Postgress SQL
Business Intelligence Tools	Tableau, Microsoft Power BI
Country	India, United States
Code Editor	Jupyter, Visual Studio
Notebooks	Colab, Kaggle
Data Visualization libraries	Matplotlib, Seaborn
Cloud Computing	AWS, Google Cloud Platform
Cloud Services	AWS EC2, Google Cloud Compute Engine
Learning Platforms	Coursera

Table 1. Inference of features

Feature 1	Feature 2	Correlation Coeff.
Programming language (Python)	Recommended language	0.6
Matplotlib	Seaborn	0.6
Individuals working in DS	ML methods in business	0.7
Money spent on Cloud & ML	Compensation	0.5
AWS	AWS EC2	0.7
Google cloud platform (GCP)	GCP Computing Engine	0.6
Microsoft Azure	Azure Machine Learning	0.6

Table 2. Correlation between features

For rules generation, Min. Support = 0.02, Min. Confidence = 0.6, Min. Lift = 3

Rules (A → B)	Support	Confidence	Lift
Generative Adversarial Networks → Generative Networks (GAN, VAE, etc.)	0.035	0.692	12.704
Google Cloud AI Platform/ML Engine → Google Cloud Compute Engine	0.22	0.612	12.085
Transformer Networks (BERT, gpt-3, etc.) → Transformer language models (GPT-3, BERT, XLnet, etc.)	0.53	0.817	11.458
Microsoft Azure Container Instances → Microsoft Azure	0.023	0.911	10.704
Azure Machine Learning Studio → Microsoft Azure	0.024	0.876	10.299
Amazon Elastic Container Service → Amazon EC2	0.025	0.802	9.728
Amazon Sagemaker → Amazon EC2	0.023	0.740	8.983

Table 3. Association Rule Mining Results

Tools/Technologies	Platforms	Association %
Image segmentation methods (U-Net, Mask R-CNN, etc)	Convolutional Neural Networks	98.7
General purpose image/video tools (PIL, cv2, skimage, etc)	Convolutional Neural Networks	98.6
Object detection methods (YOLOv3, RetinaNet, etc)	Convolutional Neural Networks	98.5
Generative Networks (GAN,VAE,etc)	Convolution Neural Networks	94.8
Amazon Elastic Container Service	Amazon Web Services (AWS)	96.9
Amazon EC2	Amazon Web Services (AWS)	96.1
Amazon SageMaker	Amazon Web Services (AWS)	95.9
AWS Lambda	Amazon Web Services (AWS)	95.6
Google cloud compute Engine	Google cloud platform (GCP)	92.6
Tidymodels	Ggplot/ ggplot2	91.9
Amazon Redshift	Amazon Web Services(AWS)	91.9
LightGBM	Gradient Boosting Machines (xgboost, lightgbm, etc)	91.8
Microsoft Azure Container Instances	Microsoft Azure	91.1
Encoder-decoder models (seq2seq, vanilla transformers)	Recurrent Neural Networks	90.1
CatBoost	Gradient Boosting Machines (xgboost, lightgbm, etc)	89.9
Google Cloud AI Platform / Google Cloud ML Engine	Google Cloud Platform (GCP)	89.9
Tidymodels	RStudio	89.5
Caret	R	87.9
Azure Machine Learning Studio	Microsoft Azure	87.6
Contextualized embeddings (ELMo, CoVe)	Recurrent Neural Networks	87.4
Shiny	R	86.5
Contextualized embeddings (ELMo, CoVe)	Word embeddings/vectors (GLoVe, fastText, word2vec)	85.8
CatBoost	Xgboost	85.6
Caret	Ggplot/ggplot2	85.1
Shiny	Ggplot/ggplot2	83.6
Caret	RStudio	83.5

LightGBM	Xgboost	83.3
Transformer Networks (BERT, gpt-3, etc)	Transformer language models (GPT-3, BERT, XLnet, etc)	81.7
Colab	GitHub	81.1
Amazon Elastic Container Service	Amazon EC2	80.2
Kaggle	GitHub	78.2
Object detection methods (YOLOv3, RetinaNet, etc)	Image classification and other general purpose networks (VGG, Inception, ResNet, ResNeXt, NASNet, EfficientNet, etc)	77.1
General purpose image/video tools (PIL, cv2, skimage, etc)	Image classification and other general purpose networks (VGG, Inception, ResNet, ResNeXt, NASNet, EfficientNet, etc)	74.8
Fast.ai	PyTorch	74.4
Amazon SageMaker	Amazon EC2	74.0
Auto-Sklearn	GitHub	73.9
Encoder-decoder models (seq2seq, vanilla transformers)	Word embeddings/vectors (GLoVe, fastText, word2vec)	73.8
Transformer language models (GPT-3, BERT, XLnet, etc)	Recurrent Neural Networks	73.4
CatBoost	LightGBM	72.8
AWS Lambda	Amazon EC2	72.3
Google Data Studio	Google Cloud Platform (GCP)	71.0

Table 4: Confidence Table

REFERENCES

1. Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar Nachiappan Nagappan, Besmira Nushi, Thomas Zimmermann, "Software engineering for Machine Learning: A case study", IEEE, August, 2019
2. Fazel Ansari, Selim Erol, Wilfried Sihm, "Rethinking Human-Machine Learning in Industry 4.0: How Does the Paradigm Shift Treat the Role of Human Learning?", Elsevier, Pages 117-122, Volume 23, 2018.
3. Daniel Adomako Asamoah, Ramesh Sharda, Amir Hassan Zadeh, Pankush Kalgotra, "Preparing a Data Scientist: A Pedagogic Experience in Designing a Big Data Analytics Course", April, 2017.
4. Winn Chow, "A Pedagogy that Uses a Kaggle Competition for Teaching Machine Learning: an Experience Sharing", IEEE, Dec, 2019.
5. Monica Ciolacu, Ali Fallah Tehrani, Leon Binder, Paul Mugur Svasta, "Education 4.0 - Artificial Intelligence Assisted Higher Education: Early recognition System with Machine Learning to support Students' Success", IEEE, Jan, 2019.
6. Carlos Costa Maribel, Yasmina Santos, "The data scientist profile and its representativeness in the European e-Competence framework and the skills framework for the information age", Elsevier, Dec 2017.
7. Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudík, Hanna Wallach, "Improving fairness in machine learning systems: What do industry practitioners need?", CHI, Jan, 2019.
8. Vijaya B. Kolachalama, Priya S. Garg, "Machine learning and medical education", Dec, 2018.
9. Melaine A Meyer, "Healthcare data scientist qualifications, skills, and job focus: a content analysis of job postings", *Journal of the American Medical Informatics Association*, Volume 26, Issue 5, May 2019.
10. Patrick Mikalef, Michail N. Giannakos, Ilias O. Pappas, John Krogstie, "The human side of big data: Understanding the skills of the data scientist in education and industry", IEEE, 2018.
11. Nikhil Rasiwasia, "Perspectives on Becoming an Applied Machine Learning Scientist", IEEE, May, 2019.
12. The state of data Science: STITCH benchmark report. (n.d.). Retrieved March 12, 2021.
13. Andreas Vogelsang, Markus Borg, "Requirements Engineering for Machine Learning: Perspectives from Data Scientists", Aug, 2019.
14. "Data Scientist: The Sexiest Job of the 21st Century", Spotlight on Big Data.
15. Saša Baškarada, Andy Koronios, "Unicorn data scientist: the rarest of breeds", Emerald insight, April, 2017.