BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

Work Integrated Learning Programmes Division

Post Graduate Programme

In

Artificial Intelligence and Machine Learning

NLP System for Creating Job Description

CAPSTONE PROJECT

Submitted in partial fulfillment of the requirements of the

Post Graduate Certification Programme

in

Artificial Intelligence and Machine Learning

By

S.No	Student Name	BITS ID
1	Akanksha Nagar	2019AIML592
2	Chandrika	2019AIML673
3	Manoj Avirineni	2019AIML684
4	Sriram P R	2019AIML511
5	Jayashree Namasivayam	2019AIML585

Under the supervision of

DESHMUKH SUDARSHAN S

Project work carried out at

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE

Pilani (Rajasthan) INDIA

(April, 2021)

Acknowledgements

We all came from different professional backgrounds however we eventually became a team with a shared goal since we began this project. While we learnt a lot as a team during this project, it wouldn't be possible for us to succeed in this project accomplishment without our key stakeholders.

Mentoring is a brain to pick, an ear to listen, and a push in the right direction – John C. Crosby

Yes, our mentor Mr. Sudarshan Deshmukh helped us understand the goal better, guided us in taking the right approach, pushed us in the right direction. He pulled all our ideas by asking pointed questions, encouraged the individuals as well as motivated the team spirit.

He played an important role in inculcating clarity in terms of the problem statement, ideating to come up with approaches, evaluating the options and keep moving ahead to achieve each milestone. We would like to pay our gratitude to Mr. Sudarshan who is the backbone of our team.

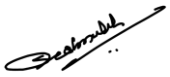
Needless to mention the entire BITS WILP team - all the professors who introduced us to the concepts through contact sessions, professors who delivered digital content, TAs who helped us throughout the program. All that laid a strong base for our CAPSTONE.

Many people in our batch have shared valuable inputs which helped us move ahead when we had any blocker. We thank all the people for their help directly and indirectly to complete our assignment.

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

CERTIFICATE

I hereby certify that the work which is being presented in the capstone project entitled **" NLP System for Creating Job Description"** is satisfactorily completed and submitted by Akanksha Nagar (2019AIML592), Chandrika (2019AIML673), Manoj Avirineni (2019AIML684), Sriram P R (2019AIML511), Jayashree Namasivayam (2019AIML585) in partial fulfillment of the requirements of **PCAM ZC321 Capstone Project**, which embodies the work carried out by him/her under my guidance.



Date : 12th April 2021 Signature of Mentor

Place : Bangalore Name: SUDARSHAN S DESHMUKH

Abstract

Natural language processing, or NLP, is a type of artificial intelligence that deals with analyzing, understanding, and generating natural human languages so that computers can process written and spoken human language without using computer-driven language. Natural language processing sometimes also called “computational linguistics,” uses both semantics and syntax to help computers understand how humans talk or write and how to derive meaning from what they say. This field combines the power of artificial intelligence and computer programming into an understanding so powerful that programs can even translate one language into another reasonably accurately. This field also includes voice recognition, the ability of a computer to understand what you say well enough to respond appropriately.

Programmers use a variety of techniques to help machines understand natural language. For example, automatic summarization consists of two techniques, extraction or abstraction. Extraction is a technique that attempts to extract the most important segments of the text and make a summary list of it. Abstraction, which is much more complex, involves writing a summary of the information.

Text generation is a subfield of natural language processing. It leverages knowledge in computational linguistics and artificial intelligence to automatically generate natural language texts, which can satisfy certain communicative requirements. There are a lot of subtasks - Dialogue generation, Data to Text conversion, Multi-document Summarization, Text style transfer, Paraphrase generation, conditional Text Generation, Spelling correction etc., Examples of text generation include machines writing entire chapters of popular novels like Game of Thrones and Harry Potter, with varying degrees of success, Automatic Text Summaries (Text Generation) of Damage Descriptions for a Property Insurance.

Our project aims at implementing a Natural Language based solution which would deliver job descriptions based on a few criteria like preparing the transformed dataset, performing NLP processing to rank the best features in the Job Description, training transformer model and implementing an API endpoint which can be consumed over the web.

The input data has been refined and data dictionaries have been created to generate Job description based on job title and skills. We have trained the GPT2 decoder transformer model where the last token embedding of the input sequence is used to make predictions about the next token that should follow the input. This project will walk you through the detailed insights on the use case and implementation thereby keeping into consideration the performance aspects.

Symbols and Abbreviations used

NLP	-	Natural Language Processing
GPT2	-	Generative Pre-trained Transformer 2
POS	-	Part of Speech
LSTM	-	Long Short-Term Memory

LIST OF TABLES

LIST OF FIGURES

TABLE OF CONTENTS

Synopsys

Business Understanding/Problem Statement

A well-written job description will establish a solid set of expectations for employers to communicate to their employees which helps to work more efficiently in their roles. Inconsistent job descriptions lead to improper selection which affects the ROI. To overcome the above-mentioned problem, we need an AI based solution to write the job description.

Objective of the project

To Implement a Natural Language based solution which will write the job description based on some given criteria. For example, given a Roles, Responsibilities and Technology, Tools data, the model should be able to generate the relevant job description.

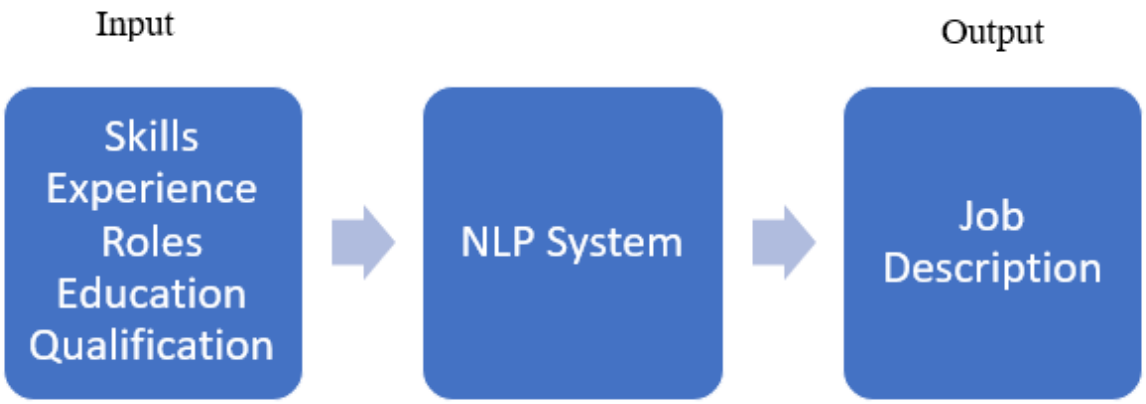


Fig 1: Objective of the Project

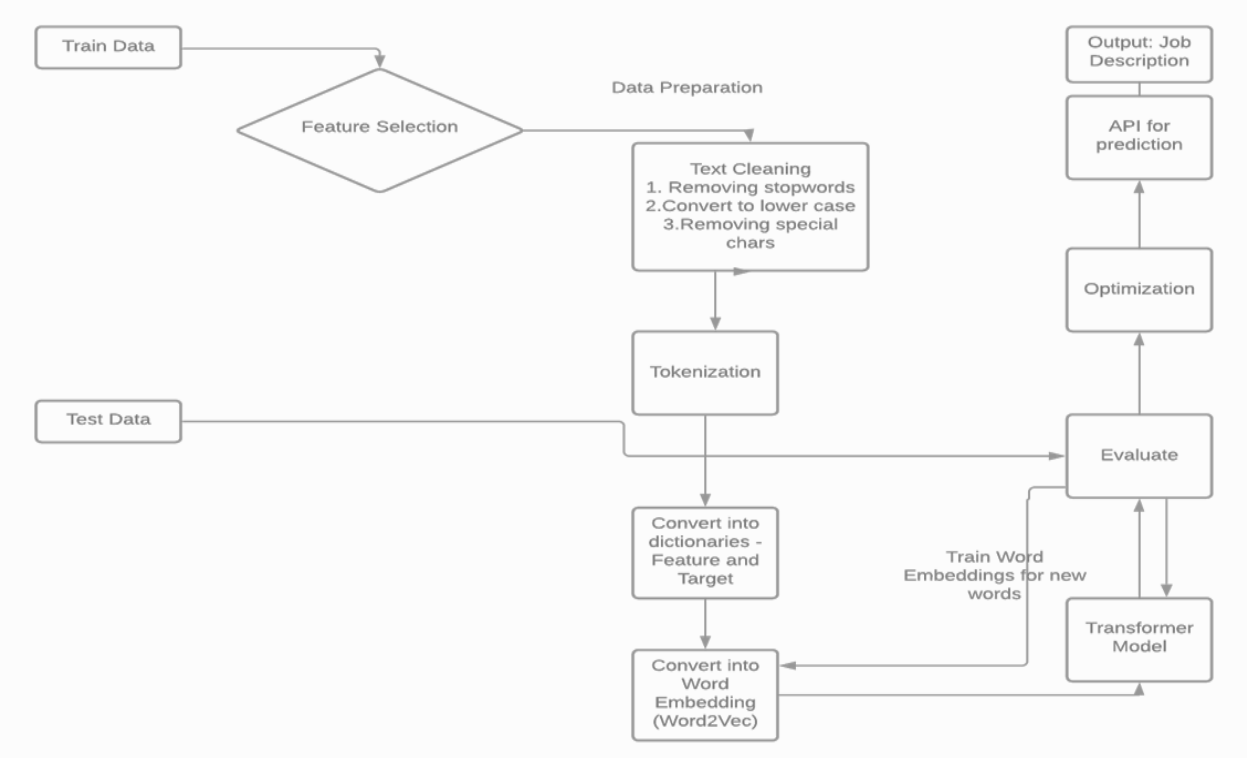
Background of previous work done in the chosen area:

In the early stages of the project, we explored the internet for resources on getting a better understanding on how machine learning concepts are applied for Text Generation. Listed below are some of the resources that we have found relevant to the context of our project:

- Data Mining Medical Records With Machine Learning - 5 Current Applications

- https://towardsdatascience.com/how-to-use-nlp-in-python-a-practical-step-by-step-example-bd82ca2d2e1e
- https://machinelearningmastery.com/text-generation-lstm-recurrent-neural-networks-python-keras/
- https://towardsdatascience.com/how-our-device-thinks-e1f5ab15071e
- https://towardsdatascience.com/how-our-device-thinks-e1f5ab15071e

Process flow (Consolidated Approach / Solution Architecture)



PROCESS FLOW

Train Data :

Training data is the data you use to train an algorithm or machine learning model to predict the outcome you design your model to predict. Test data is used to measure the performance, such as accuracy or efficiency, of the algorithm you are using to train the machine.

In Our Process the Training plays a very important role as without it the machine learning model would be very inefficient and incoherent in delivering accurate output.

Feature Selection : In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. Feature selection techniques are used for several reasons:

- simplification of models to make them easier to interpret by researchers/users
- shorter training times
- to avoid the curse of dimensionality,
- enhanced generalization by reducing overfitting (formally, reduction of variance)

Text Cleaning: The main aim of Text Cleaning is to identify and remove errors & duplicate data, in order to create a reliable dataset. This improves the quality of the training data for analytics and enables accurate decision-making.

Removing Stop words:

Words such as articles and some verbs are usually considered stop words because they don't help us to find the context or the true meaning of a sentence. These are words that can be removed without any negative consequences to the final model that you are training.

Tokenization : Tokenization is the process of dividing text into a set of meaningful pieces. These pieces are called tokens. For example, we can divide a chunk of text into words, or we can divide it into sentences.

Convert Data into Dictionaries Features and Targets :

This is the most vital step of the process and is the soul of Natural Language Processing. This is used as a reference for the Machine Learning model to deliver accurate output.

TOKENIZATION : Tokenization is the process of dividing text into a set of meaningful pieces. These pieces are called tokens. For example, we can divide a chunk of text into words, or we can divide it into sentences.

GPT2 : Generative Pre-trained Transformer 2 (GPT-2) is an open-source artificial intelligence created by OpenAI.

GPT-2 is a very large language model with 1.5 billion parameters, trained on a dataset of 8 million web pages. Due to the diversity of the training dataset, it is capable of generating conditional synthetic text samples of unprecedented quality.

Evaluate : Model evaluation aims to estimate the generalization accuracy of a model on future (unseen/out-of-sample) data. Methods for evaluating a model's performance are divided into 2

categories: namely, holdout and Cross-validation. Both methods use a test set (i.e data not seen by the model) to evaluate model performance

API For Prediction :

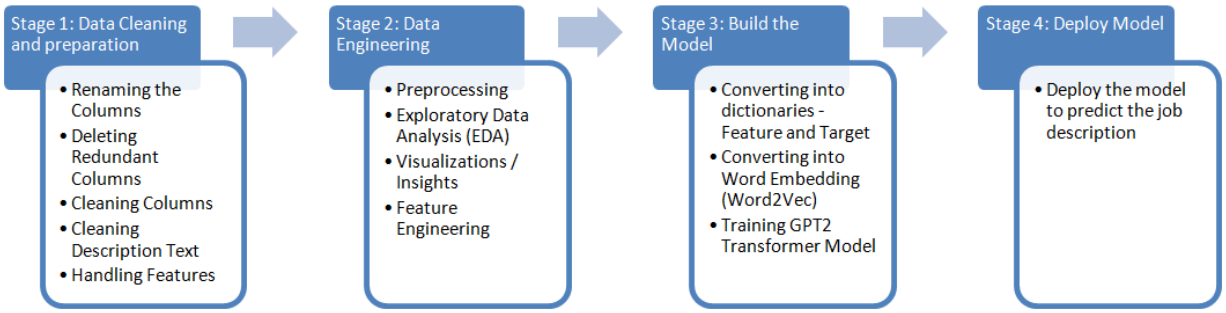
An API, or Application Programming Interface, is a server that you can use to retrieve and send data to using code. APIs are most used to retrieve data, and that will be the focus of this beginner tutorial. When we want to receive data from an API, we need to make a request. Requests are used all over the web.

FastAPI is a modern, fast (high-performance), web framework for building APIs with Python based on standard python type hints.

It enables the user to access the model whenever there is an inference call.

The Latency of Fast API is very good and user friendly too with an interactive UI.

Approach:



Resources needed for the project

S. No	Resource Type	Details
1	People	● Akanksha Nagar (2019AIML592) ● Chandrika (2019AIML673) ● Manoj Avirineni (2019AIML684) ● Sriram P R (2019AIML511) ● JayashreeNamasivayam (2019AIML511)
2	Hardware	● Nvidia Tesla K80 GPU ● Model name: Intel(R) Xeon(R) CPU @ 2.20GHz ● Processors: 2 ● CPU cores: 1 ● CPU MHz : 2200.000 ● Address sizes: 46 bits physical, 48 bits virtual ● Cache size: 46080 KB
3	Software	● Python 3.6 ● Microsoft Excel 2016 ● Microsoft Word 2016
4	Communication Channel	● Canvas ● Email ● WhatsApp ● Google Meetings

1.1 Data challenges & risks in doing the project

Most of the real-world data is messy, some of these types of data are:

- ***Missing data*:** Missing data can be found when it is not continuously created or due to technical issues in the application. Many columns had more than 80% missing data. These columns were deleted. Few columns such as Company, Industry also have a significant number of missing values, but since we needed this column as an input, we created another category - 'Other' and filled missing rows with this value.

- ***Inconsistent data*:** This type of data might be collected due to human errors (mistakes with the name or values) or duplication of data. For example, for Job_Title, few values were similar, such as Sr. and Senior. So to avoid biases in our model, we cleaned this column to standardize our column values.

- ***Redundant data:*** There were several highly correlated columns in our dataset, such as No of Skills, One hot encoded skills value. So, to reduce the dimensionality of our dataset and decrease redundancy, we deleted these columns.

1.2 Detailed Plan of Work

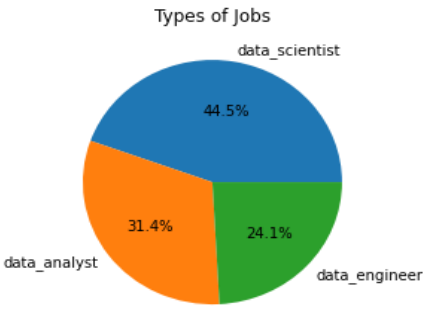
Week #	List of Activities	Status
Week# 1 & 2	1. Review the requirement documents. 2. Come up with a list of questions to clarify the requirement with SME. 3. Understand the project data. 4. Perform data analysis using Python libraries. 5. Identify the project environment. 6. Develop preliminary python code to perform feature engineering activities. 7. Perform text preprocessing steps like Data Cleaning, Data Normalization and Standardization.	Completed
Week# 3 & 4	1. Finalize the feature engineering python code with all the pre-processing, data cleansing and the transformation steps. 2. Segregate Job Description into relevant sections - Skills, Role and responsibilities and About company 3. Find the best features needed to generate the 'skills' section of the Job Description. 4. Find the best features needed to generate the 'Roles and responsibilities' section of the Job Description.	Completed

	5. Find the best features needed to generate the 'About company' section of the Job Description.	
Week# 5, 6 & 7	1. Train 3 transformer models for 3 sections of Job Description. 2. Optimize the result of the model with different hyperparameters. 3. Develop insights about the average package, location wise open positions for the given job role etc. 4. Review the project progress with SME during the sync-ups and incorporate the suggestions. 5. Implement an API endpoint which can be consumed over the web. 6. Create a PPT to provide updates to the Reviewer during the bi-weekly review.	Completed

Exploratory Data Analysis / Insights:

1. **Job Role in Demand**

	Job_Type	open_positions
0	data_scientist	2543
1	data_analyst	1793
2	data_engineer	1379



Conclusion: From the above visualization, we can draw that most of the jobs fall into the Data Scientist category followed by Data Analyst and Data Engineer. We can also note that there were 750-1100 more data scientist jobs than other job roles.

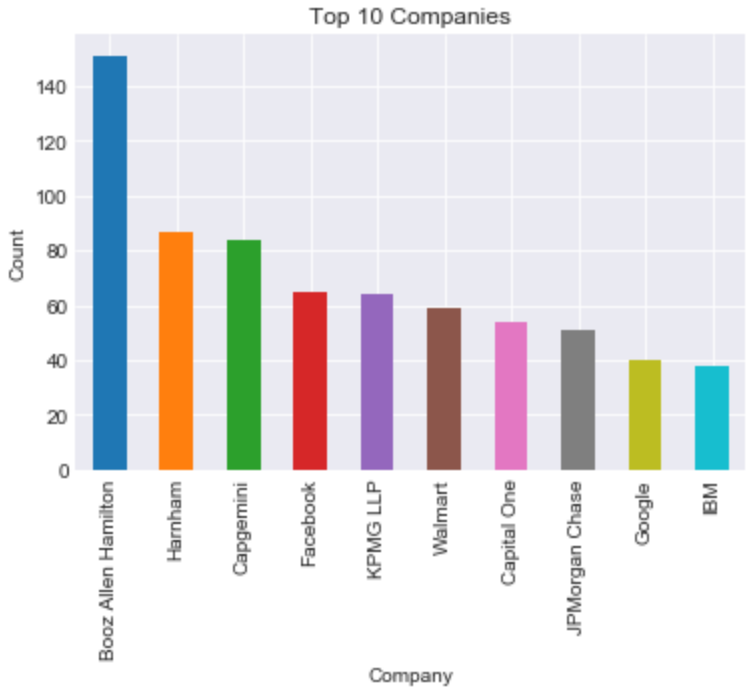
2. **Top 5 Locations having highest no of Job openings**

	Location	No_Of_Jobs	PCT
0	CA	1395	25%
1	NY	601	11%
2	VA	334	6%
3	TX	329	6%
4	MA	271	4%

Conclusion: California (CA) by far has the largest job market overall with 25% occupancy of the overall job listings followed by NewYork (NY), Virginia (VA), and Texas (TX).

3. **Top Hiring Companies**

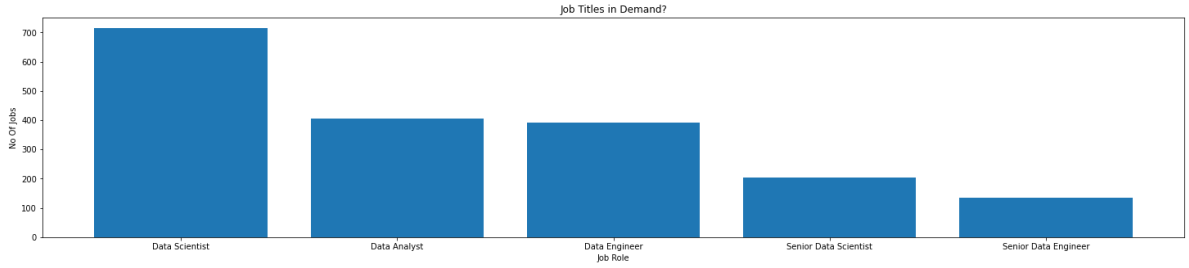
	Company
Booz Allen Hamilton	151
Harnham	87
Capgemini	84
Facebook	65
KPMG LLP	64
Walmart	59
Capital One	54
JPMorgan Chase	51
Google	40
IBM	38



Conclusion: From the chart above, Booz Allen Hamilton Company seems to have the most job openings with 151 job listings whereas Harnham and Capgemini companies have 87 and 84 job listings respectively.

4. Job Position in Demand?

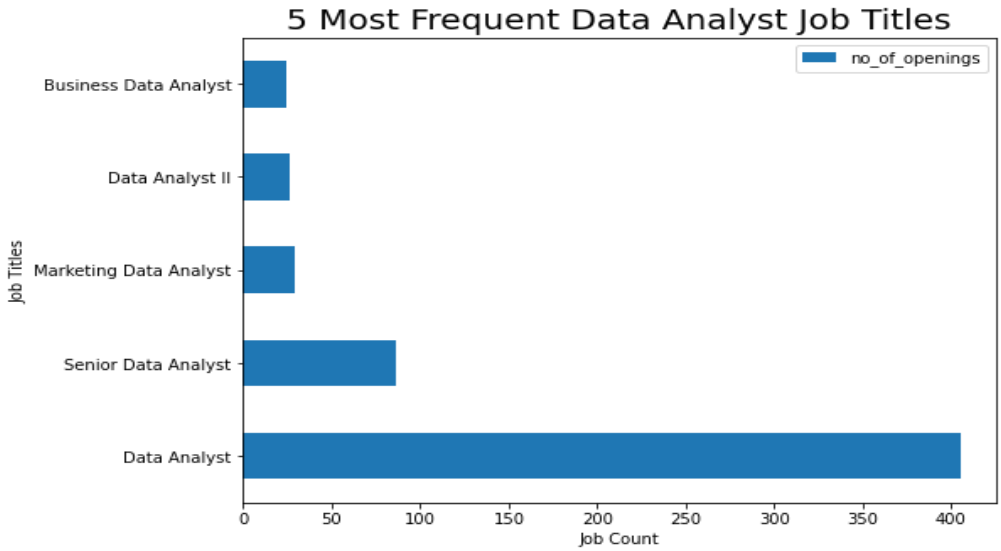
	Job_Title	No_Of_Jobs	PCT
0	Data Scientist	715	12%
1	Data Analyst	405	7%
2	Data Engineer	391	6%
3	Senior Data Scientist	205	3%
4	Senior Data Engineer	136	2%



Conclusion: Data Scientist has more demand across all job positions with 12%, followed by Data Analyst (7%) and Data Engineer (6%).

5. Top Data Analyst Positions

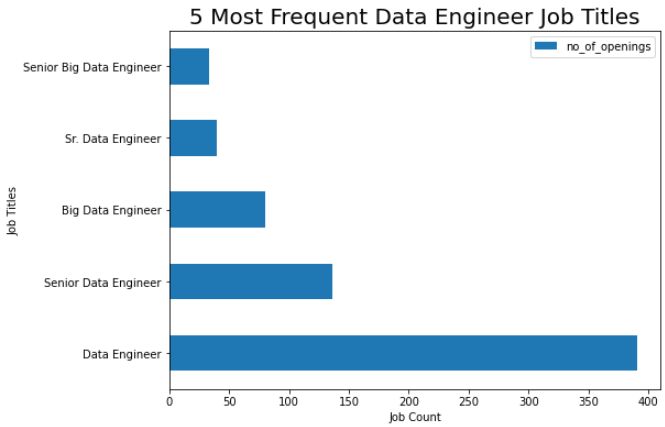
	Job_Title	no_of_openings
0	Data Analyst	405
1	Senior Data Analyst	86
2	Marketing Data Analyst	29
3	Data Analyst II	26
4	Business Data Analyst	24



Conclusion: Data Analyst position by far has huge openings when compared to Senior Data Analyst and Marketing Data Analyst.

6. Top Data Engineer Positions

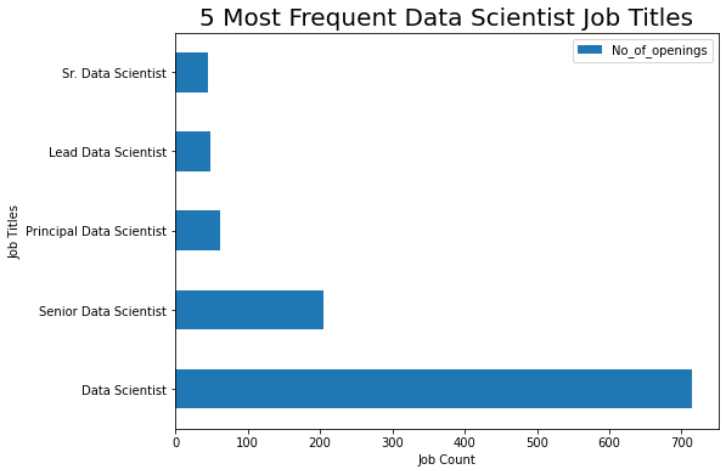
	Job_Title	no_of_openings
0	Data Engineer	391
1	Senior Data Engineer	136
2	Big Data Engineer	80
3	Sr. Data Engineer	40
4	Senior Big Data Engineer	33



Conclusion: From the above visualization we can conclude that entry level data engineer has more number of job openings than senior data engineer and big data engineer.

7. Top Data Scientist Job Positions

	Job_Title	No_of_openings
0	Data Scientist	715
1	Senior Data Scientist	205
2	Principal Data Scientist	62
3	Lead Data Scientist	49
4	Sr. Data Scientist	45



Conclusion: Data scientist by far have a large number of job opportunities when compared to Senior data scientist and principal data scientist.

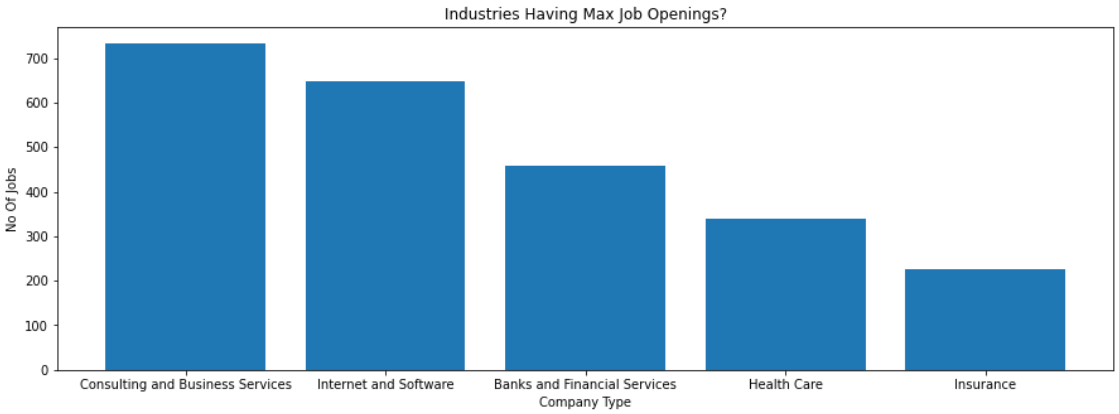
8. Is Job Related to Location?

	Job_Title	Location	No_Of_Jobs
0	Data Scientist	DC	713
1	Data Analyst	NY	405
2	Data Engineer	CA	294
3	Senior Data Scientist	MA	205
4	Senior Data Engineer	NY	111
5	Senior Data Analyst	NY	86

Conclusion: From the observation, Washington DC has the most job positions for the role of Data Scientist followed by Newyork for the role of data analyst.

9. Any Industry on a Hiring Surge?

	Company_Industry	No_Of_Jobs	PCT
0	Consulting and Business Services	733	12%
1	Internet and Software	647	11%
2	Banks and Financial Services	459	8%
3	Health Care	339	5%
4	Insurance	227	3%



Conclusion: Consulting and Business Services, Internet and Software & Banks and Financial Services are top industries who are recruiting for data science jobs.

10. Total Job Openings for Data Scientist Designation

	No_Of_DataScience_Jobs	PCT
0	2385	41%

Conclusion: Out of the total job postings in the Data Science Field, Data scientist positions have a higher amount of job openings with 41%.

11. Most Recent Job Openings

	Company	Job_Type	Date_Since_Posted	count(*)
0	Cognizant	data_engineer	1.0	8
1	Capgemini	data_analyst	1.0	7
2	Centene	data_analyst	1.0	6
3	GlassDoor	data_engineer	1.0	6
4	All-In Analytics	data_engineer	1.0	5
...	...	...	...	...
523	Virginia Tech	data_engineer	7.0	1
524	WebMD	data_analyst	7.0	1
525	Wish	data_analyst	7.0	1
526	Xactly Corporation	data_scientist	7.0	1
527	eHire, LLC	data_engineer	7.0	1

Conclusion: These are the insights where the companies have posted their requirement less than a week ago, where 528 companies have their requirements across different job designations.

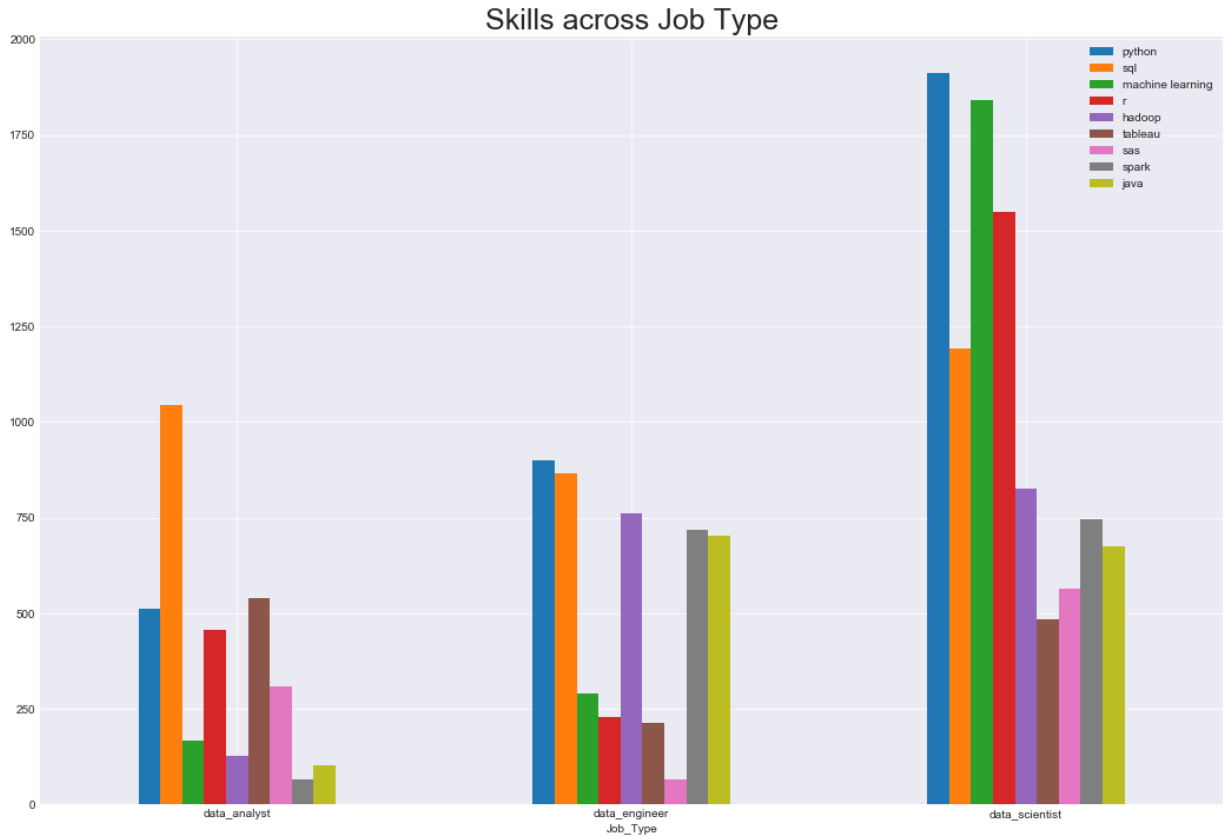
12. Data Analyst Job Openings across Location with average salary

	Location	No_of_openings	Avg_Salary
0	CA	376	65372.340426
1	NY	230	56695.652174
2	TX	117	41880.341880
3	MA	86	41627.906977
4	VA	85	57882.352941
5	IL	66	51818.181818
6	GA	59	38305.084746
7	WA	57	49824.561404
8	PA	56	35000.000000
9	MD	55	45454.545455

Conclusion: California has huge opportunities for the role of data analyst and the average salaries are a bit higher when compared to other locations. If you are someone who is looking for a data analyst job position where salary is a preference you can opt for California, Virginia and Newyork where the average salary is more when compared to other states.

13. Most Popular skills as per job type?

	Job_Type	Python_Jobs	SQL	R	Hadoop	Tableau	SAS	Spark	JAVA	Others
0	data_analyst	512	1044	456	126	538	310	67	101	1462
1	data_engineer	901	867	228	761	214	67	719	704	1338
2	data_scientist	1912	1193	1550	827	484	564	745	675	2352



Conclusion: As per the observation from the above insight, we can conclude that Python is the most popular skill followed by SQL and R for different job designations.

14. Top Skills across Industry:

	Company_Industry	Python_Jobs	SQL	R	Hadoop	Tableau	SAS	Spark	JAVA	Others	Total
0	Consulting and Business Services	426	338	275	270	200	121	235	172	658	2695
1	Internet and Software	432	376	266	242	105	82	212	192	583	2490
2	Banks and Financial Services	264	267	171	165	105	96	160	155	436	1819
3	Health Care	126	180	112	56	82	92	51	53	290	1042
4	Insurance	124	135	103	93	49	59	35	69	194	861

Conclusion: Consulting and Business Services have more job opportunities in the data science field and also python is most preferred skills when compared to other skills.

15. Most preferred industry to work for - as per ratings

	Company_Industry	Queried_Salary	No_of_Reviews	Avg_Rating	Weighted_Rating
0	Retail	140000-159999	157475.0	3.636046	0.027665
1	RetailConsumer Goods and Services	120000-139999	41290.0	3.800000	0.007657
2	Food and Beverages	80000-99999	33082.0	3.700000	0.005973

Conclusion: Retail is the most preferred industry to work for where the weighted average rating is 0.027 which is higher than the Retail Consumer Goods and Services industry.

16. Best Companies to work for - based on ratings

	Company	No_of_Stars
0	The Church of Jesus Christ of Latter-day Saints	4.8
1	AmeriCorps	4.4
2	University of Michigan	4.4
3	Google	4.3
4	SAP	4.3
5	Johnson & Johnson Family of Companies	4.3
6	University of California UCOP	4.3
7	Vivo	4.3
8	usajobs.gov	4.3
9	Bosch Group	4.2

Conclusion: The Church of Jesus Christ of Latter-day Saints (4.8 Ratings) is the most preferred company to work for as per the ratings provided by the employees followed by AmeriCorps (4.4 Ratings) and University of Michigan (4.4 Ratings). This insight is based on the criteria where companies having a number of reviews greater than 1000.

17. Average No. of Reviews related to ratings:

	Company	No_of_Reviews	Avg_Rating
0	Walmart	157475.0	3.6
1	The Home Depot	41290.0	3.8
2	SUBWAY	33082.0	3.7
3	AT&T	31970.0	3.8
4	Wells Fargo	29966.0	3.8
5	Lowe's	29206.0	3.7

Conclusion: Taking more number of reviews as criteria, Walmart have higher number of reviews and has an average rating of 3.6 followed by The Home Depot with number of reviews as 41290 and average rating of 3.8.

18. Are Senior Positions Hard to Fill?

	Job_Title	Company_Industry	Queried_Salary	Date_Since_Posted
0	2019 PhD Data Scientist Internship - UberEverything	Internet and Software	120000-139999	30.0
1	2019 PhD Data Scientist Internship - UberEverything - New Yo...	Internet and Software	120000-139999	30.0
2	23103 Principal Data Scientist - HealthTech / Diagnostics -...	None	140000-159999	30.0
3	23222 Senior Data Scientist (Machine Learning / Computer Vis...	None	140000-159999	30.0
4	36th EWS Mission Data Analyst	Aerospace and Defense	80000-99999	30.0
5	5G+ Wireless Machine Learning Research Scientist	Computers and Electronics	140000-159999	30.0
6	AI Chief Data Scientist - PS12747	Health Care	100000-119999	30.0
7	AI Principal Data Scientist - PS12776	Health Care	80000-99999	30.0
8	AI Scientist - Machine Learning Focus	Internet and Software	80000-99999	30.0
9	AI Scientist - NLP Focus	Internet and Software	120000-139999	30.0

Conclusion: Senior level positions need more time to be filled. As we can see that most of the positions that have not been filled in 30 days, are for niche roles.

19. Jobs Not filled in a month

	count(Job_Title)	Date_Since_Posted
0	2954	30.0

Conclusion: There are 2954 job positions which were not filled within a month from the day the job requirement is posted.

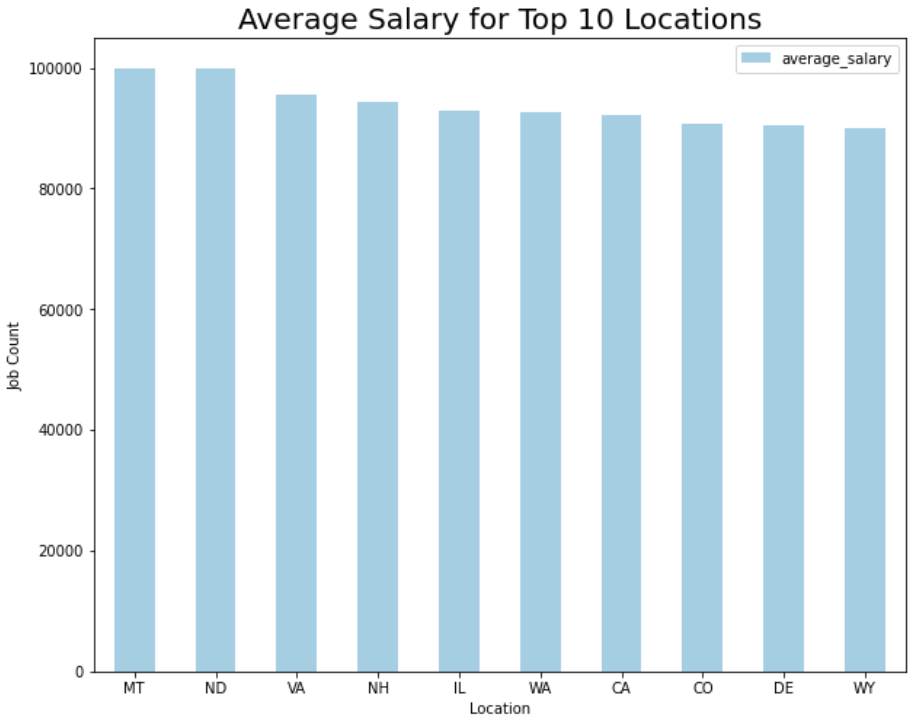
20. Bigger corporations have a harder time filling the job?

	Company	Company_Employees	Date_Since_Posted	No_of_unfilled_Jobs
0	Booz Allen Hamilton	10,000+	30.0	78
1	Harnham	Less than 10,000	30.0	69
2	KPMG LLP	None	30.0	47
3	Walmart	10,000+	30.0	38
4	Capgemini	10,000+	30.0	33
5	Capital One	10,000+	30.0	29
6	Allstate	10,000+	30.0	28
7	Microsoft	10,000+	30.0	25
8	Uber	10,000+	30.0	25
9	JPMorgan Chase	10,000+	30.0	23

Conclusion: Established corporations, having more than 10000+ employees have a hard time filling the job. It can be due to the hiring process or unrealistic expectations in terms of skills required.

21. Average Salary Across different

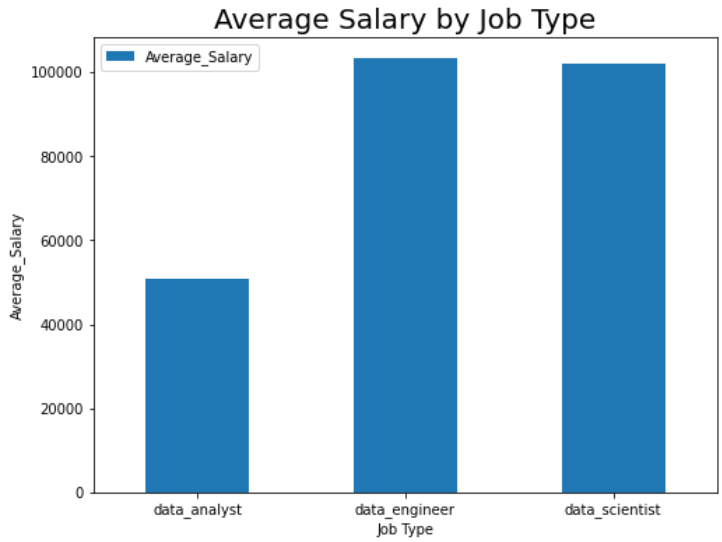
	Location	average_salary
0	MT	100000.000000
1	ND	100000.000000
2	VA	95628.742515
3	NH	94285.714286
4	IL	92916.666667
5	WA	92612.612613
6	CA	92157.706093
7	CO	90825.688073
8	DE	90588.235294
9	WY	90000.000000



Conclusion: Montana and North Dakota locations have the highest average salary with $100000 per year followed by Virginia with $95628 per year.

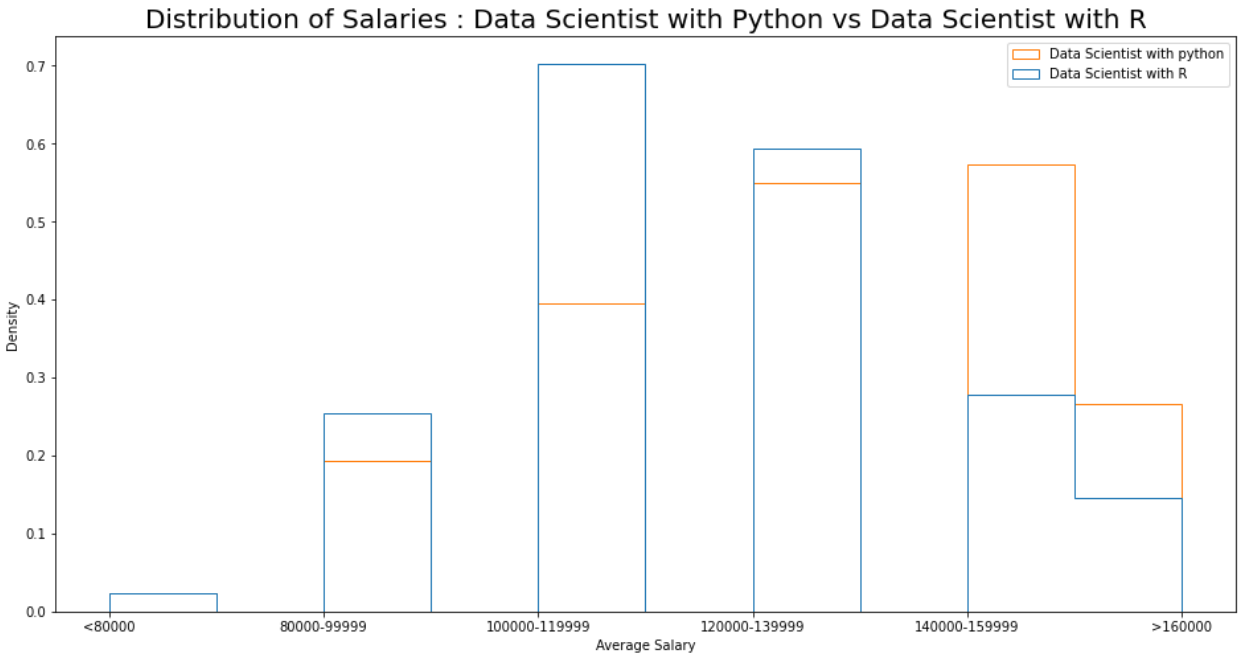
22. Average salary by job designation

	Job_Type	Average_Salary
0	data_analyst	50964.863358
1	data_engineer	103219.724438
2	data_scientist	101918.993315



Conclusion: Data Engineer Designation provides an average salary of $103219 per year which is higher when compared with Data Scientist ($101918) and Data Analyst ($50964) designations.

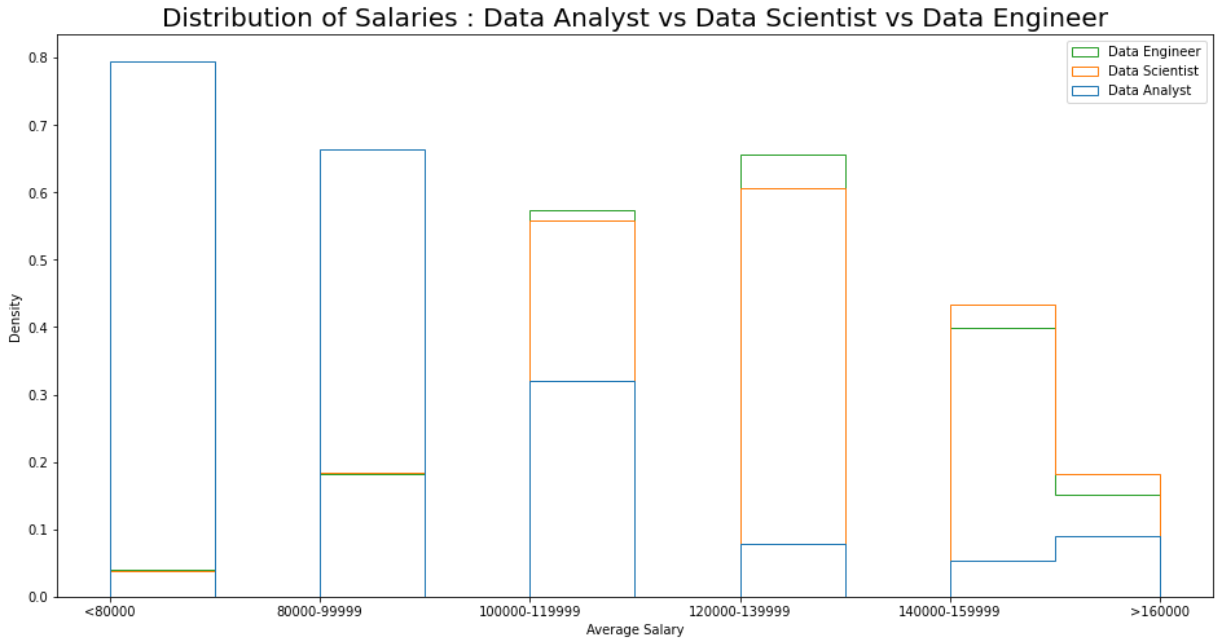
23. High salary- learn Python or R?



Conclusion: For job postings with average salary information provided, comparing the salary distribution for Data Scientist with python vs. data scientist with R and also categorizing the career level based on the salary ranges where less than 8000 to 99999 as entry level, 100000 to 139999 as Mid-level and 140000 to greater than 160000 as senior level.

From the above insight we can say that Data Scientist with R is mostly paid than Data Scientist with Python at the entry level and mid-level career but when coming to senior level career Data scientist with Python is mostly paid when compared with Data Scientist with R.

24. Data Analyst vs Data Scientist vs Data Engineer

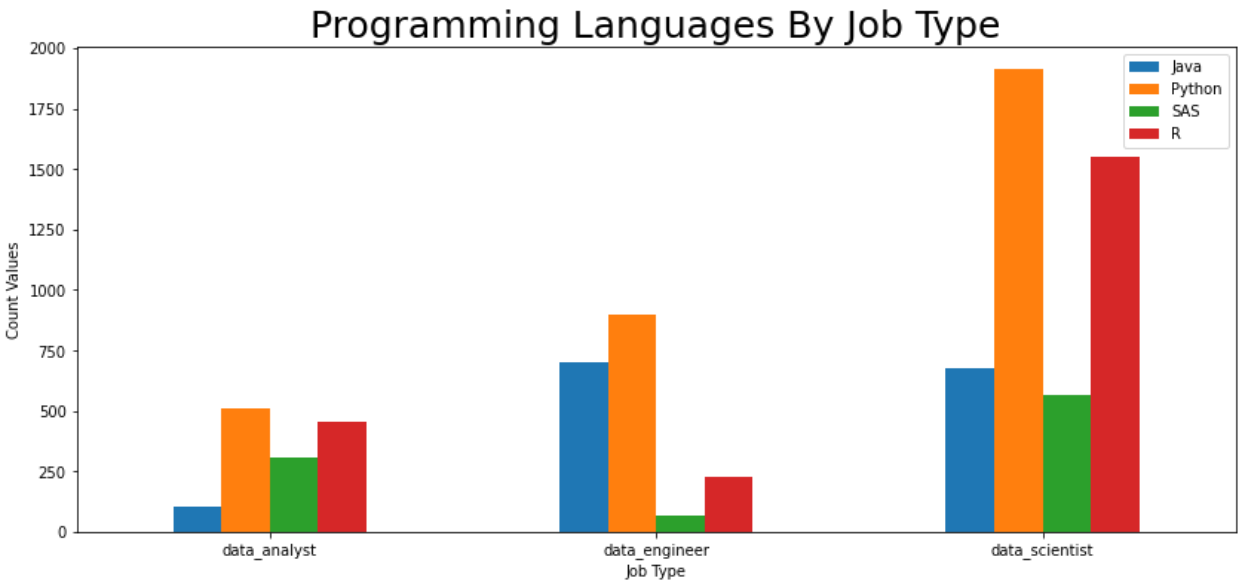


Conclusion: For job postings with average salary information provided, comparing the salary distribution for Data Analyst vs Data Scientist vs Data Engineer and also categorizing the career level based on the salary ranges where less than 8000 to 99999 as entry level, 100000 to 139999 as Mid-level and 140000 to greater than 160000 as senior level.

From the above insight we can come to an assumption that Data Analyst job position is mostly paid at the entry level because of the more number of openings than data scientist and data engineer whereas Data Engineer job position is mostly paid at mid-level career but when coming to senior level career Data scientist is mostly paid job designation when compared with Data Engineer and Data Analyst.

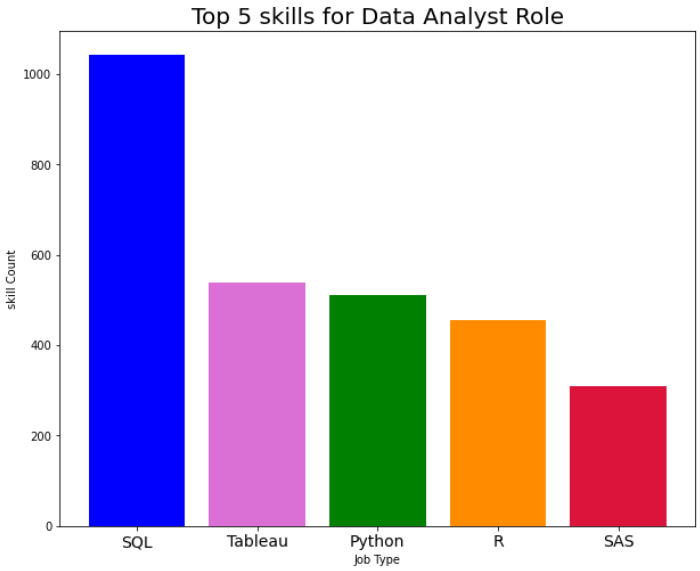
25. Most Preferred Programming language across job roles?

	Job_Type	Java	Python	SAS	R
0	data_analyst	101	512	310	456
1	data_engineer	704	901	67	228
2	data_scientist	675	1912	564	1550



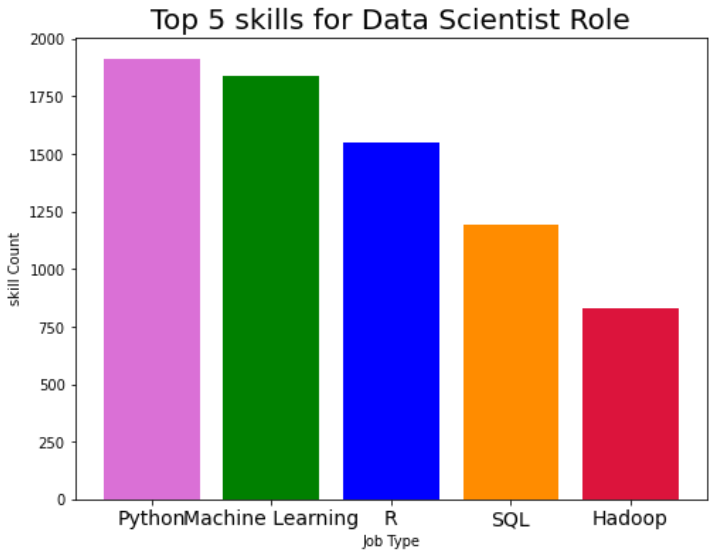
Conclusion: From the observation, we can draw an insight that python is the most common and preferred programming language across different job designations followed by R language and Java programming language.

26. Demand Skills for Data Analyst Role



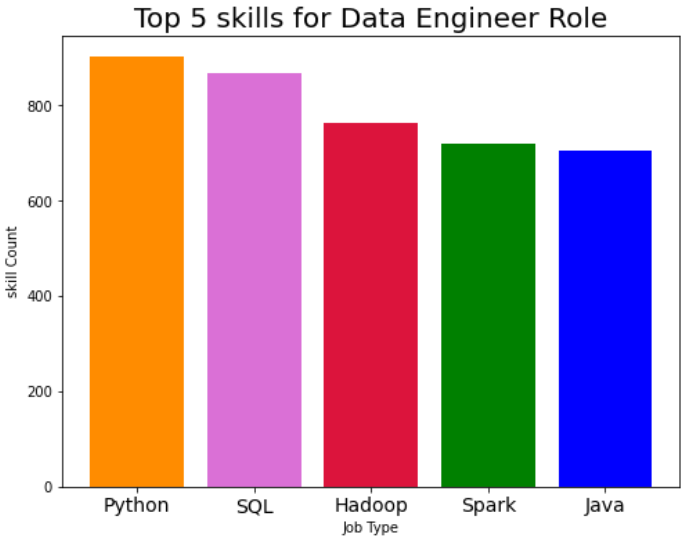
Conclusion: From the above bar chart, we can conclude that SQL is the most preferred Skill for the role of Data Analyst followed by Tableau and Python.

27. Demand Skills for Data Scientist Role



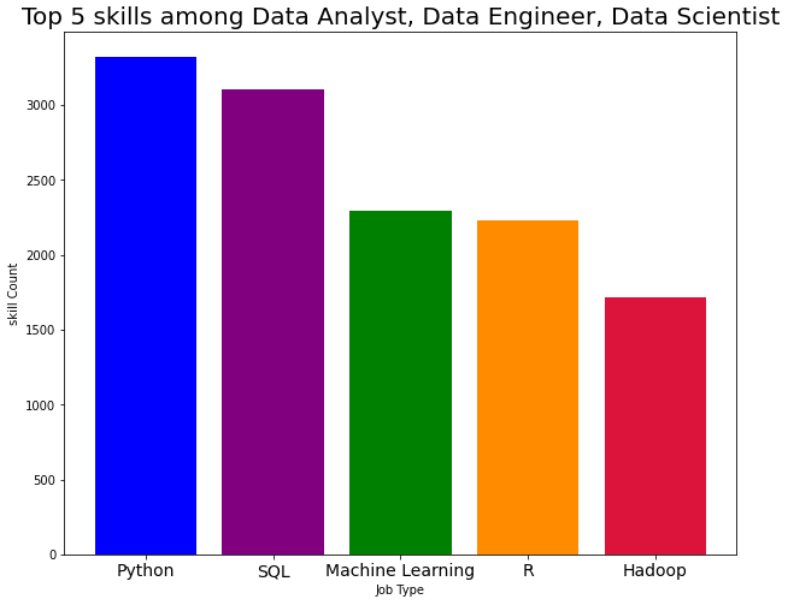
Conclusion: From the above chart, we can conclude that Python and Machine Learning is the most preferred Skill for the role of Data Scientist followed by R, SQL and Hadoop.

28. Demand skills for Data Engineer Role



Conclusion: From the above chart, we can conclude that Python and SQL is the most preferred Skill for the role of Data Engineer followed by Hadoop, Spark and Java.

29. Demand Skills amongst Data Analyst, Data Scientist and Data Engineer



Conclusion: From the above visualization, we can draw that Python and SQL are the most common and preferred Skills for the role of Data Analyst, Data Scientist and Data Engineer.

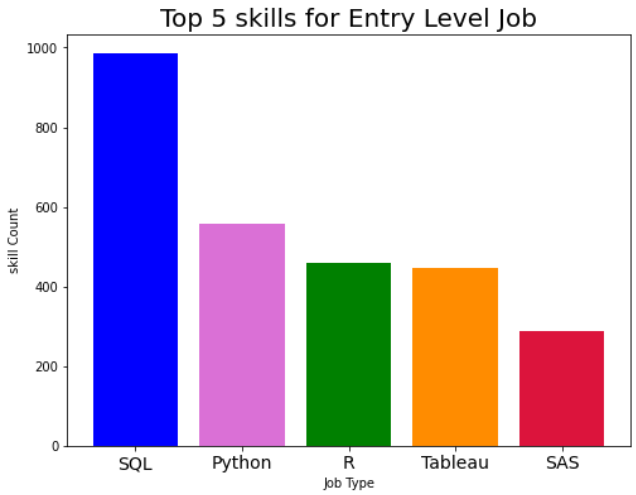
30. Relation Between Skills, Career level and Salary

Queried_Salary	python	sql	machine learning	r	hadoop	tableau	sas	spark	java
100000-119999	861	840	579	648	410	357	288	347	375
120000-139999	957	720	687	601	593	253	189	536	474
140000-159999	674	388	533	376	424	138	124	397	359
80000-99999	412	618	201	311	97	283	156	71	102
<80000	144	367	67	148	12	162	131	9	34
>160000	277	171	230	150	178	43	53	171	136

Job Level	python	sql	machine learning	r	hadoop	tableau	sas	spark	java
Entry Level	556	985	268	459	109	445	287	80	136
Mid Level	1818	1560	1266	1249	1003	610	477	883	849
Senior Level	951	559	763	526	602	181	177	568	495

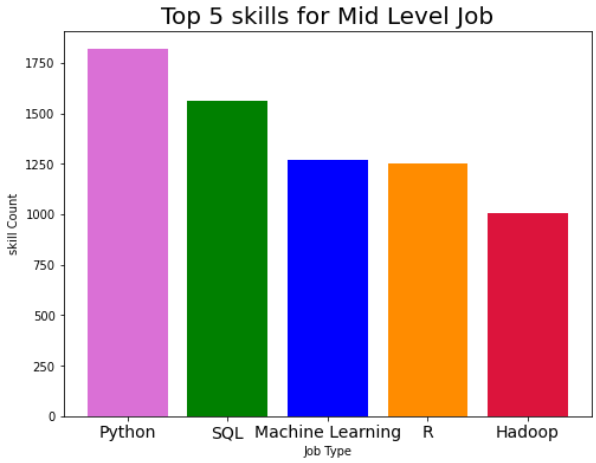
Conclusion: Since we don't have any column based on experience we are categorizing the career level based on the salary ranges where less than 8000 to 99999 as entry level, 100000 to 139999 as Mid-level and 140000 to greater than 160000 as senior level.

31. Demand Skills for Entry Level Job



Conclusion: From the above visualization, we can draw a conclusion that companies are preferring the candidate based on the skills like SQL, Python and R for the entry level job openings.

32. Demand Skills for Mid-Level Job



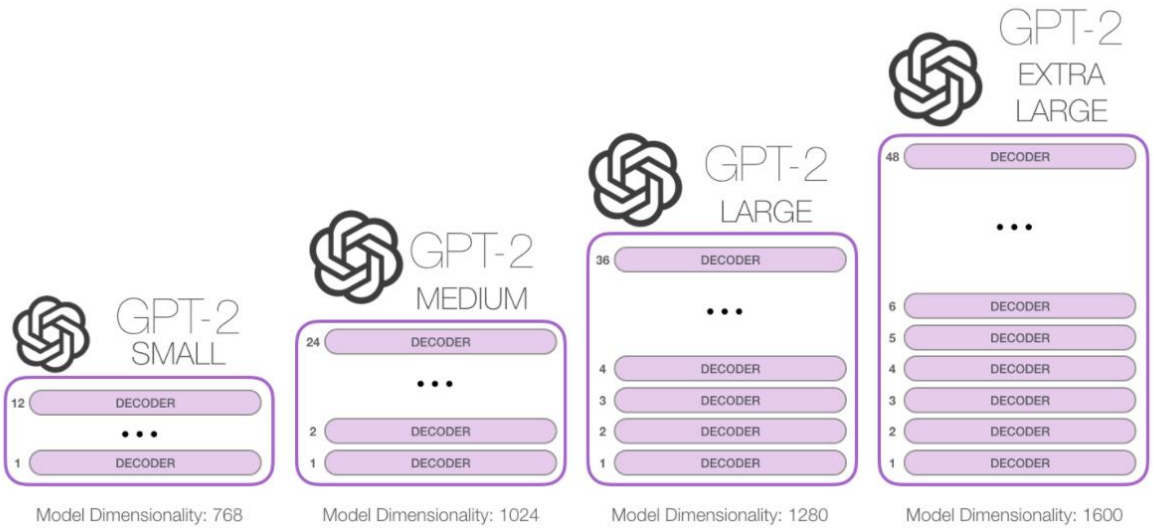
Conclusion: From the above visualization, we can draw a conclusion that companies are hiring the candidates based on the skills like SQL, Python and R for the entry level job openings.

Machine Learning Modelling & Techniques Applied

1. **Text Generation :** The idea to make code write on your behalf is remarkably inspiring. This practice is referred to as Text Generation or Natural Language Generation, which is a subfield of Natural Language Processing (NLP). The fundamentals of text generation can be easily broken down into a simple supervised machine learning problem, wherein, there exists certain features (called x) with their corresponding labels (called y), and using these we can create our own prediction function which will then generate our predicted labels (called ŷ or yhat). We then map these predicted labels to the actual labels to determine the cost and optimize it using an optimization algorithm such as Gradient Descent, RMSprop or even the Adam optimizer.

2. **Model Used:** There are several Auto regressive algorithms that can be used to generate text. For our project, we decided to go with GPT2 due to its simplicity and huge number of parameters it has been trained on and better results.

   GPT2 is a variant of the transformer family. However instead of encoder-decoder architecture, it is based entirely on decoder architecture. GPT2 models are available in different sizes. We decided to go with the smaller version (12 layered), due to limited hardware and time availability.



3. **Approach :** We have separated output into 3 sections.

   JobDescription -

   1. Skills required.
   2. Roles and responsibilities
   3. About Company

We have one GPT2 model for each section. The first model takes Input as Skills and Job Title and Is going to predict the first section. Similarly, the second one takes Input as Skills and Job Title and predicts the second section. And the third model will take Location and company name as in Input and predict the third section.

We combine the outputs from all models to have our desired Job Description ready

Code Screenshots (Full screenshot without cropping the system date-time)

```
JobData = pd.read_csv("/content/drive/My Drive/indeed_job_dataset.csv", index_col = None)
JobData.head(n=2)
```

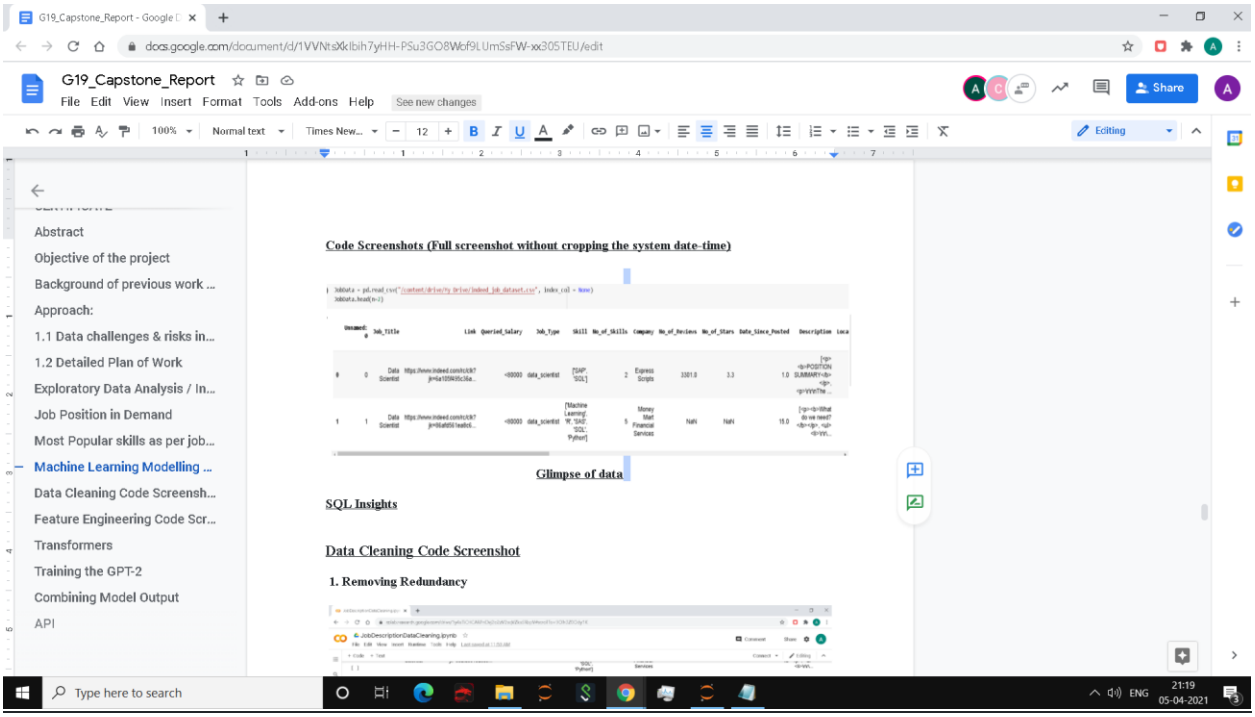
	Unnamed: 0	Job_Title	Link	Queried_Salary	Job_Type	Skill	No_of_Skills	Company	No_of_Reviews	No_of_Stars	Date_Since_Posted	Description	Loca
0	0	Data Scientist	https://www.indeed.com/rc/clk?jk=6a105f495c36a...	<80000	data_scientist	['SAP', 'SQL']	2	Express Scripts	3301.0	3.3	1.0	[POSITION SUMMARY , \r\r\nThe ...	
1	1	Data Scientist	https://www.indeed.com/rc/clk?jk=86afd561ea8c6...	<80000	data_scientist	['Machine Learning', 'R', 'SAS', 'SQL', 'Python']	5	Money Mart Financial Services	NaN	NaN	15.0	[What do we need? , \r\r\...	

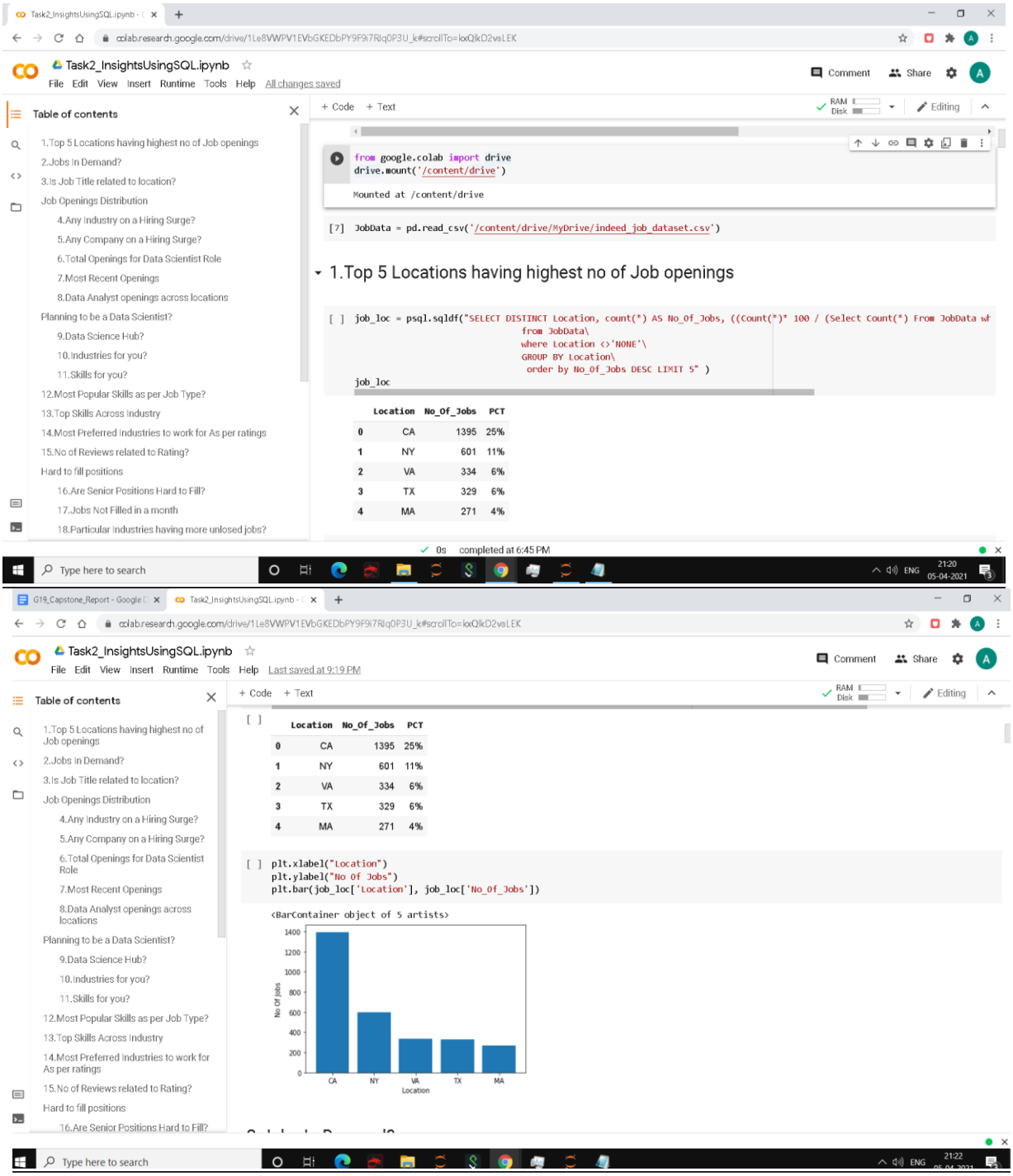
Glimpse of data

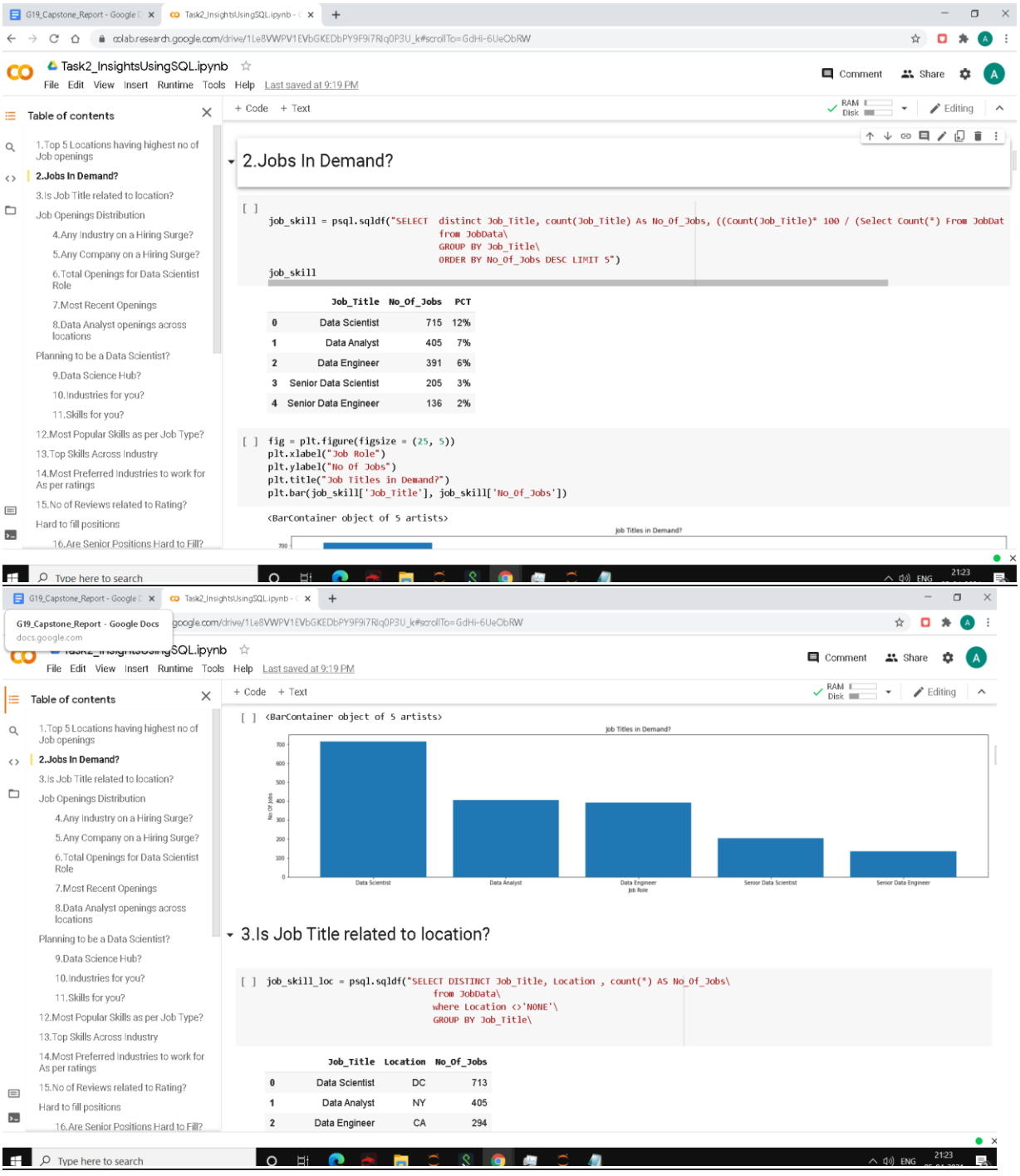
Code Screenshots

Exploratory Data Analysis (EDA)

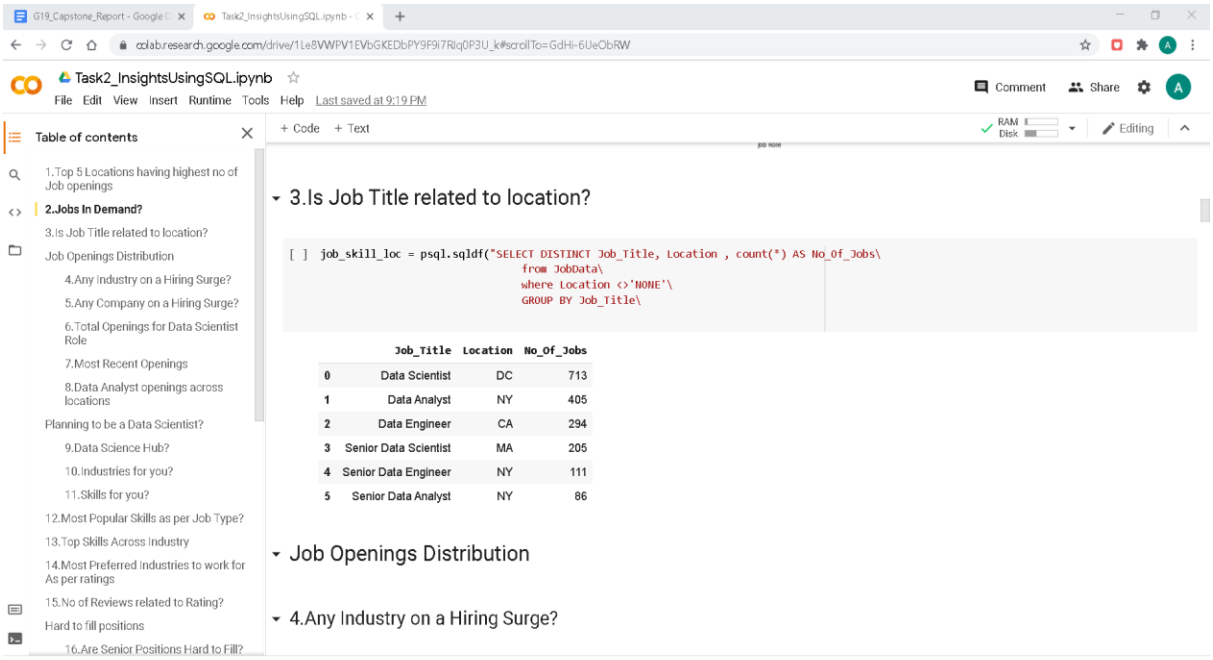
SQL Insights

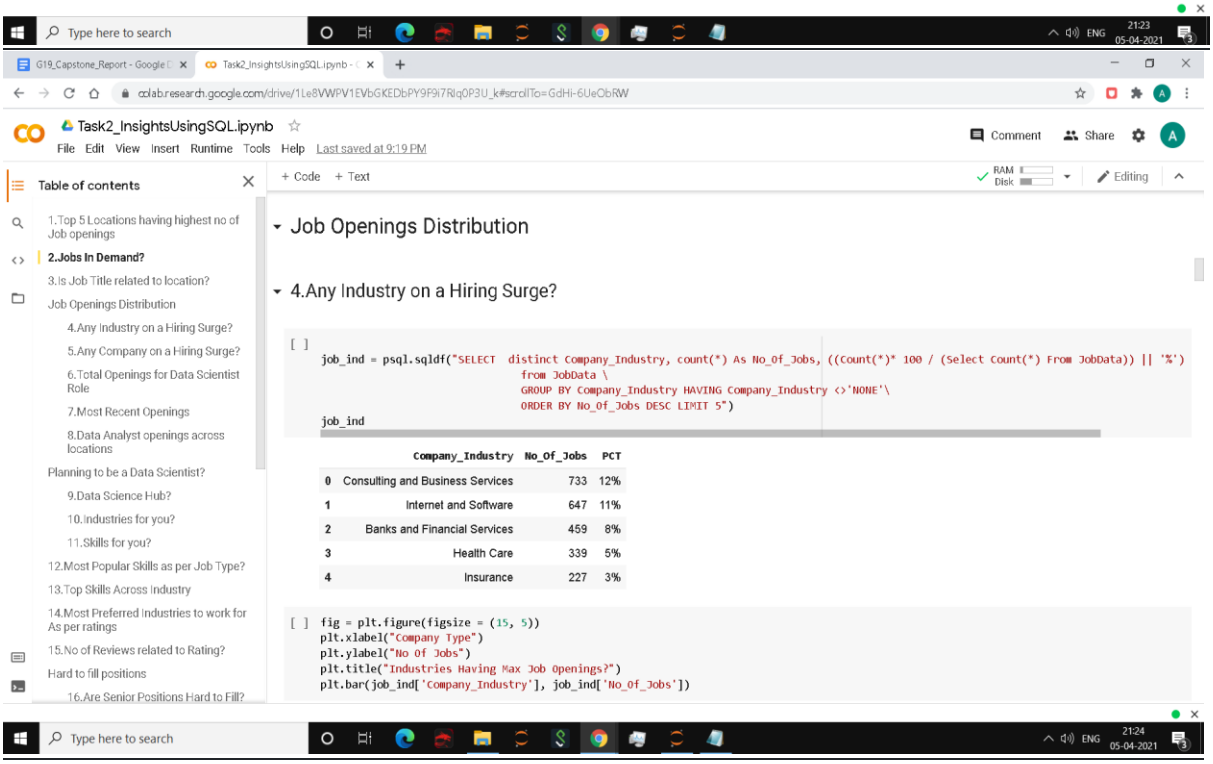


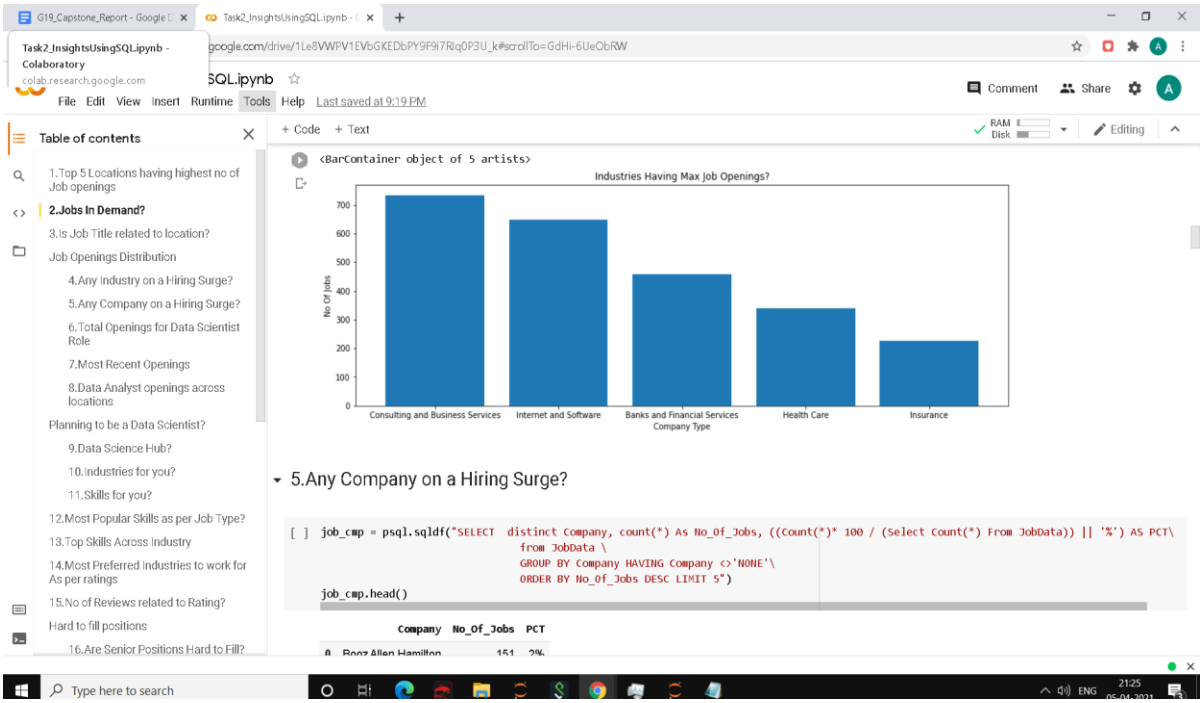

Task2_InsightsUsingSQL.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment Share
Table of contents
1.Top 5 Locations having highest no of Job openings
2.Jobs In Demand?
3.Is Job Title related to location?
Job Openings Distribution
4.Any Industry on a Hiring Surge?
5.Any Company on a Hiring Surge?
6.Total Openings for Data Scientist Role
7.Most Recent Openings
8.Data Analyst openings across locations
Planning to be a Data Scientist?
9.Data Science Hub?
10.Industries for you?
11.Skills for you?
12.Most Popular Skills as per Job Type?
13.Top Skills Across Industry
14.Most Preferred Industries to work for As per ratings
15.No of Reviews related to Rating?
Hard to fill positions
16.Are Senior Positions Hard to Fill?
17.Jobs Not Filled in a month
18.Particular Industries having more unlosed jobs?
+ Code + Text
Editing
from google.colab import drive
drive.mount('/content/drive')
Mounted at /content/drive
[7] JobData = pd.read_csv('/content/drive/MyDrive/indeed_job_dataset.csv')
1.Top 5 Locations having highest no of Job openings
[ ] job_loc = psql.sqldf("SELECT DISTINCT Location, count(*) AS No_Of_Jobs, ((Count(*)* 100 / (Select Count(*) From JobData wh
from JobData\
where Location <>'NONE'\
GROUP BY Location\
order by No_Of_Jobs DESC LIMIT 5")
job_loc
Location No_Of_Jobs PCT
0 CA 1395 25%
1 NY 601 11%
2 VA 334 6%
3 TX 329 6%
4 MA 271 4%
0s completed at 6:45 PM
Type here to search
ENG 21:20 05-04-2021
Task2_InsightsUsingSQL.ipynb
File Edit View Insert Runtime Tools Help Last saved at 9:19 PM
Comment Share
Table of contents
+ Code + Text
Editing
[ ]
Location No_Of_Jobs PCT
0 CA 1395 25%
1 NY 601 11%
2 VA 334 6%
3 TX 329 6%
4 MA 271 4%
[ ] plt.xlabel("Location")
plt.ylabel("No Of Jobs")
plt.bar(job_loc['Location'], job_loc['No_Of_Jobs'])
<BarContainer object of 5 artists>
No Of Jobs
Location
CA
NY
VA
TX
MA
Type here to search
ENG 21:22

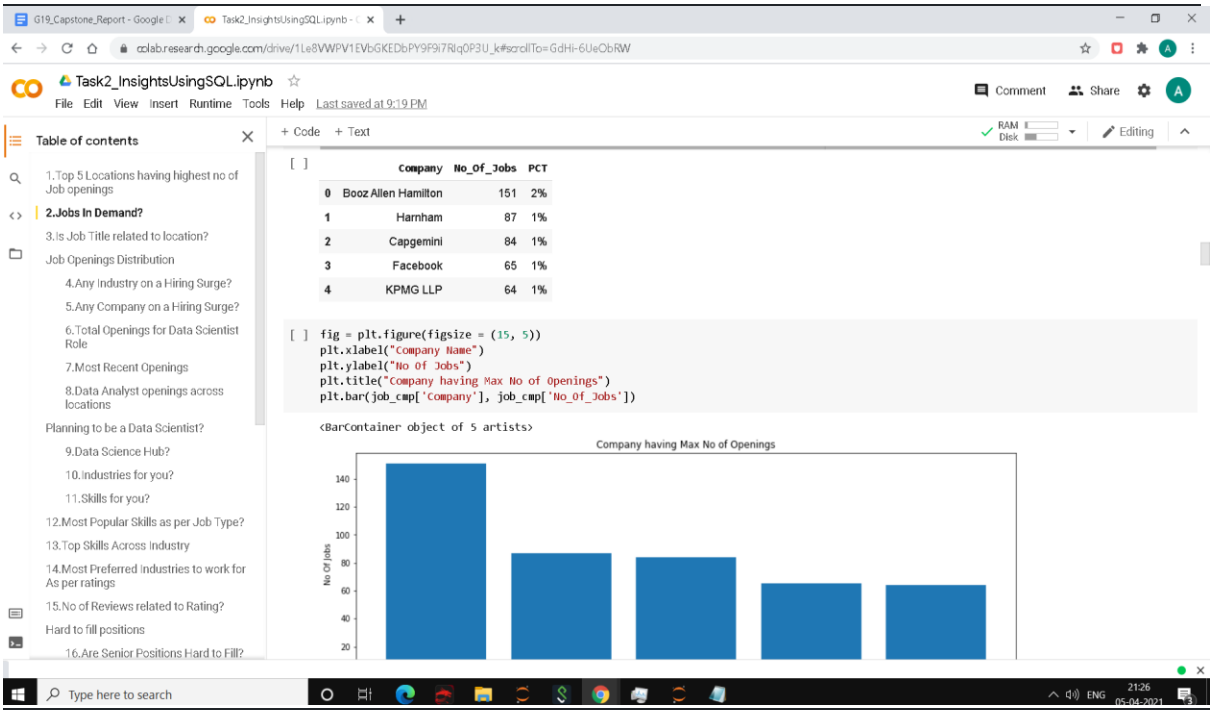

Task2_InsightsUsingSQL.ipynb
File Edit View Insert Runtime Tools Help Last saved at 9:19 PM
Comment Share
Table of contents
+ Code + Text
RAM Disk
Editing
1.Top 5 Locations having highest no of Job openings
2.Jobs In Demand?
3.Is Job Title related to location?
Job Openings Distribution
4.Any Industry on a Hiring Surge?
5.Any Company on a Hiring Surge?
6.Total Openings for Data Scientist Role
7.Most Recent Openings
8.Data Analyst openings across locations
Planning to be a Data Scientist?
9.Data Science Hub?
10.Industries for you?
11.Skills for you?
12.Most Popular Skills as per Job Type?
13.Top Skills Across Industry
14.Most Preferred Industries to work for As per ratings
15.No of Reviews related to Rating?
Hard to fill positions
16.Are Senior Positions Hard to Fill?
2.Jobs In Demand?
job_skill = psql.sqldf("SELECT distinct Job_Title, count(Job_Title) As No_Of_Jobs, ((Count(Job_Title)* 100 / (Select Count(*) From JobDat
from JobData\
GROUP BY Job_Title\
ORDER BY No_Of_Jobs DESC LIMIT 5")
job_skill
Job_Title No_Of_Jobs PCT
0 Data Scientist 715 12%
1 Data Analyst 405 7%
2 Data Engineer 391 6%
3 Senior Data Scientist 205 3%
4 Senior Data Engineer 136 2%
fig = plt.figure(figsize = (25, 5))
plt.xlabel("Job Role")
plt.ylabel("No Of Jobs")
plt.title("Job Titles in Demand?")
plt.bar(job_skill['Job_Title'], job_skill['No_Of_Jobs'])
<BarContainer object of 5 artists>
Job Titles in Demand?
Type here to search
ENG
21:23

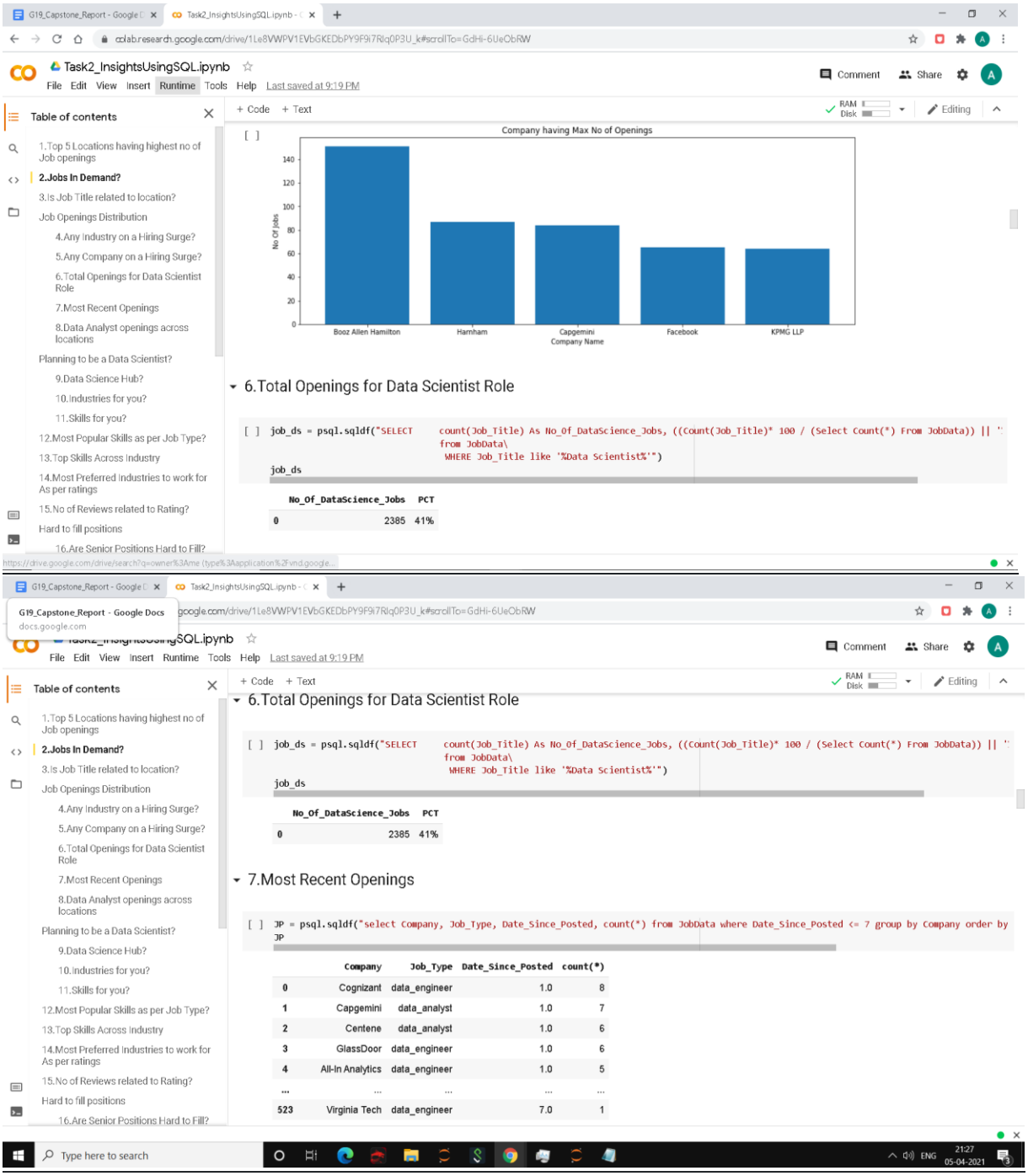
G19_Capstone_Report - Google Docs
docs.google.com
Task2_InsightsUsingSQL.ipynb
File Edit View Insert Runtime Tools Help Last saved at 9:19 PM
Comment Share
Table of contents
+ Code + Text
RAM Disk
Editing
<BarContainer object of 5 artists>
Job Titles in Demand?
No Of Jobs
Job Role
Data Scientist
Data Analyst
Data Engineer
Senior Data Scientist
Senior Data Engineer
3.Is Job Title related to location?
job_skill_loc = psql.sqldf("SELECT DISTINCT Job_Title, Location , count(*) AS No_Of_Jobs\
from JobData\
where Location <>'NONE'\
GROUP BY Job_Title\
Job_Title Location No_Of_Jobs
0 Data Scientist DC 713
1 Data Analyst NY 405
2 Data Engineer CA 294
Type here to search
ENG
21:23


Task2_InsightsUsingSQL.ipynb
File Edit View Insert Runtime Tools Help Last saved at 9:19 PM
Comment Share
+ Code + Text
RAM Disk
Editing
Table of contents
1.Top 5 Locations having highest no of Job openings
2.Jobs In Demand?
3.Is Job Title related to location?
Job Openings Distribution
4.Any Industry on a Hiring Surge?
5.Any Company on a Hiring Surge?
6.Total Openings for Data Scientist Role
7.Most Recent Openings
8.Data Analyst openings across locations
Planning to be a Data Scientist?
9.Data Science Hub?
10.Industries for you?
11.Skills for you?
12.Most Popular Skills as per Job Type?
13.Top Skills Across Industry
14.Most Preferred Industries to work for As per ratings
15.No of Reviews related to Rating?
Hard to fill positions
16.Are Senior Positions Hard to Fill?
3.Is Job Title related to location?
[ ] job_skill_loc = psql.sqldf("SELECT DISTINCT Job_Title, Location , count(*) AS No_Of_Jobs\
from JobData\
where Location <>'NONE'\
GROUP BY Job_Title\
Job_Title Location No_Of_Jobs
0 Data Scientist DC 713
1 Data Analyst NY 405
2 Data Engineer CA 294
3 Senior Data Scientist MA 205
4 Senior Data Engineer NY 111
5 Senior Data Analyst NY 86
Job Openings Distribution
4.Any Industry on a Hiring Surge?
Type here to search
ENG 21:23 05-04-2021


Task2_InsightsUsingSQL.ipynb
File Edit View Insert Runtime Tools Help Last saved at 9:19 PM
Comment Share
+ Code + Text
RAM Disk
Editing
Table of contents
1.Top 5 Locations having highest no of Job openings
2.Jobs In Demand?
3.Is Job Title related to location?
Job Openings Distribution
4.Any Industry on a Hiring Surge?
5.Any Company on a Hiring Surge?
6.Total Openings for Data Scientist Role
7.Most Recent Openings
8.Data Analyst openings across locations
Planning to be a Data Scientist?
9.Data Science Hub?
10.Industries for you?
11.Skills for you?
12.Most Popular Skills as per Job Type?
13.Top Skills Across Industry
14.Most Preferred Industries to work for As per ratings
15.No of Reviews related to Rating?
Hard to fill positions
16.Are Senior Positions Hard to Fill?
Job Openings Distribution
4.Any Industry on a Hiring Surge?
[ ] job_ind = psql.sqldf("SELECT distinct Company_Industry, count(*) As No_Of_Jobs, ((Count(*)* 100 / (Select Count(*) From JobData)) || '%')
from JobData \
GROUP BY Company_Industry HAVING Company_Industry <>'NONE'\
ORDER BY No_Of_Jobs DESC LIMIT 5")
job_ind
Company_Industry No_Of_Jobs PCT
0 Consulting and Business Services 733 12%
1 Internet and Software 647 11%
2 Banks and Financial Services 459 8%
3 Health Care 339 5%
4 Insurance 227 3%
[ ] fig = plt.figure(figsize = (15, 5))
plt.xlabel("Company Type")
plt.ylabel("No Of Jobs")
plt.title("Industries Having Max Job Openings?")
plt.bar(job_ind['Company_Industry'], job_ind['No_Of_Jobs'])
Type here to search
ENG 21:24 05-04-2021


Task2_InsightsUsingSQL.ipynb - Colaboratory
colab.research.google.com
Task2_InsightsUsingSQL.ipynb
File Edit View Insert Runtime Tools Help Last saved at 9:19 PM
Comment
Share
+ Code + Text
RAM
Disk
Editing
Table of contents
1.Top 5 Locations having highest no of Job openings
2.Jobs In Demand?
3.Is Job Title related to location?
Job Openings Distribution
4.Any Industry on a Hiring Surge?
5.Any Company on a Hiring Surge?
6.Total Openings for Data Scientist Role
7.Most Recent Openings
8.Data Analyst openings across locations
Planning to be a Data Scientist?
9.Data Science Hub?
10.Industries for you?
11.Skills for you?
12.Most Popular Skills as per Job Type?
13.Top Skills Across Industry
14.Most Preferred Industries to work for As per ratings
15.No of Reviews related to Rating?
Hard to fill positions
16.Are Senior Positions Hard to Fill?
<BarContainer object of 5 artists>
Industries Having Max Job Openings?
No Of Jobs
0
100
200
300
400
500
600
700
Consulting and Business Services
Internet and Software
Banks and Financial Services
Health Care
Insurance
Company Type
5.Any Company on a Hiring Surge?
job_cmp = psql.sqldf("SELECT distinct Company, count(*) As No_Of_Jobs, ((Count(*)* 100 / (Select Count(*) From JobData)) || '%') AS PCT\
from JobData \
GROUP BY Company HAVING Company <>'NONE'\
ORDER BY No_Of_Jobs DESC LIMIT 5")
job_cmp.head()
Company No_Of_Jobs PCT
0 Booz Allen Hamilton 151 2%
Type here to search
ENG
21:25
05-04-2021


G19_Capstone_Report - Google D
Task2_InsightsUsingSQL.ipynb - C
colab.research.google.com/drive/1Le8VWPV1EVbGKEDbPY9F9i7RIq0P3U_k#scrollTo=GdHi-6UeObRW
Task2_InsightsUsingSQL.ipynb
File Edit View Insert Runtime Tools Help Last saved at 9:19 PM
Comment
Share
+ Code + Text
RAM
Disk
Editing
Table of contents
1.Top 5 Locations having highest no of Job openings
2.Jobs In Demand?
3.Is Job Title related to location?
Job Openings Distribution
4.Any Industry on a Hiring Surge?
5.Any Company on a Hiring Surge?
6.Total Openings for Data Scientist Role
7.Most Recent Openings
8.Data Analyst openings across locations
Planning to be a Data Scientist?
9.Data Science Hub?
10.Industries for you?
11.Skills for you?
12.Most Popular Skills as per Job Type?
13.Top Skills Across Industry
14.Most Preferred Industries to work for As per ratings
15.No of Reviews related to Rating?
Hard to fill positions
16.Are Senior Positions Hard to Fill?
Company No_Of_Jobs PCT
0 Booz Allen Hamilton 151 2%
1 Harnham 87 1%
2 Capgemini 84 1%
3 Facebook 65 1%
4 KPMG LLP 64 1%
fig = plt.figure(figsize = (15, 5))
plt.xlabel("Company Name")
plt.ylabel("No Of Jobs")
plt.title("Company having Max No of Openings")
plt.bar(job_cmp['Company'], job_cmp['No_Of_Jobs'])
<BarContainer object of 5 artists>
Company having Max No of Openings
No Of Jobs
20
40
60
80
100
120
140
Type here to search
ENG
21:26
05-04-2021


Task2_InsightsUsingSQL.ipynb
File Edit View Insert Runtime Tools Help Last saved at 9:19 PM
Comment Share
Table of contents
1.Top 5 Locations having highest no of Job openings
2.Jobs In Demand?
3.Is Job Title related to location?
Job Openings Distribution
4.Any Industry on a Hiring Surge?
5.Any Company on a Hiring Surge?
6.Total Openings for Data Scientist Role
7.Most Recent Openings
8.Data Analyst openings across locations
Planning to be a Data Scientist?
9.Data Science Hub?
10.Industries for you?
11.Skills for you?
12.Most Popular Skills as per Job Type?
13.Top Skills Across Industry
14.Most Preferred Industries to work for As per ratings
15.No of Reviews related to Rating?
Hard to fill positions
16.Are Senior Positions Hard to Fill?
+ Code + Text
Editing
Company having Max No of Openings
No Of Jobs
Booz Allen Hamilton
Harnham
Capgemini
Facebook
KPMG LLP
Company Name
6.Total Openings for Data Scientist Role
job_ds = psql.sqldf("SELECT count(Job_Title) As No_Of_DataScience_Jobs, ((Count(Job_Title)* 100 / (Select Count(*) From JobData)) || '
from JobData\
WHERE Job_Title like '%Data Scientist%'")
job_ds
No_Of_DataScience_Jobs PCT
0 2385 41%
7.Most Recent Openings
JP = psql.sqldf("select Company, Job_Type, Date_Since_Posted, count(*) from JobData where Date_Since_Posted <= 7 group by Company order by JP
Company Job_Type Date_Since_Posted count(*)
0 Cognizant data_engineer 1.0 8
1 Capgemini data_analyst 1.0 7
2 Centene data_analyst 1.0 6
3 GlassDoor data_engineer 1.0 6
4 All-In Analytics data_engineer 1.0 5
523 Virginia Tech data_engineer 7.0 1
Type here to search
ENG 21:27 05-04-2021

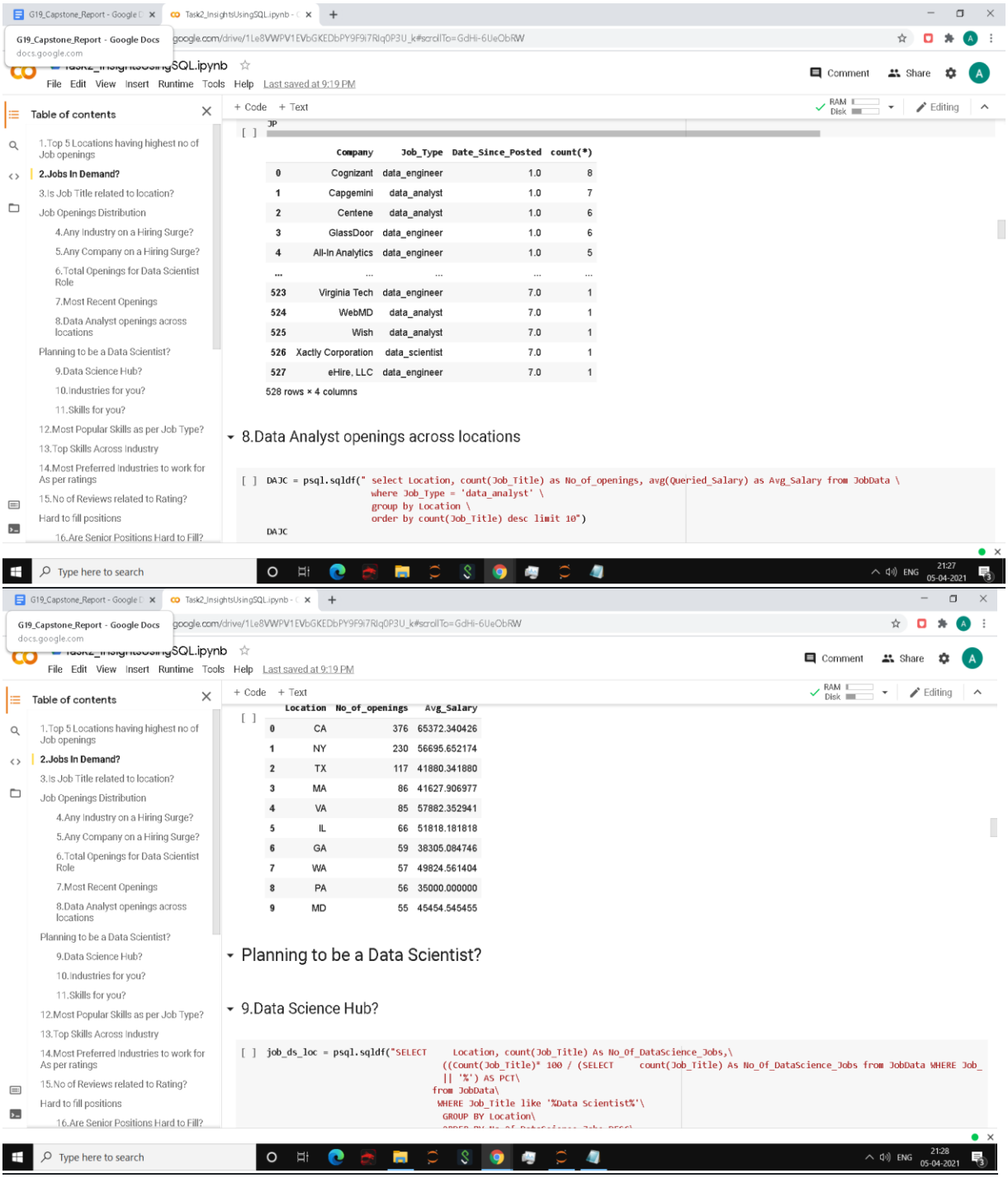


Table of contents

1.Top 5 Locations having highest no of Job openings
2.Jobs In Demand?
3.Is Job Title related to location?
Job Openings Distribution
4.Any Industry on a Hiring Surge?
5.Any Company on a Hiring Surge?
6.Total Openings for Data Scientist Role
7.Most Recent Openings
8.Data Analyst openings across locations
Planning to be a Data Scientist?
9.Data Science Hub?
10.Industries for you?
11.Skills for you?
12.Most Popular Skills as per Job Type?
13.Top Skills Across Industry
14.Most Preferred Industries to work for As per ratings
15.No of Reviews related to Rating?
Hard to fill positions
16.Are Senior Positions Hard to Fill?

	Company	Job_Type	Date_Since_Posted	count(*)
0	Cognizant	data_engineer	1.0	8
1	Capgemini	data_analyst	1.0	7
2	Centene	data_analyst	1.0	6
3	GlassDoor	data_engineer	1.0	6
4	All-In Analytics	data_engineer	1.0	5
...	...	...	...	...
523	Virginia Tech	data_engineer	7.0	1
524	WebMD	data_analyst	7.0	1
525	Wish	data_analyst	7.0	1
526	Xactly Corporation	data_scientist	7.0	1
527	eHire, LLC	data_engineer	7.0	1

528 rows × 4 columns

8.Data Analyst openings across locations

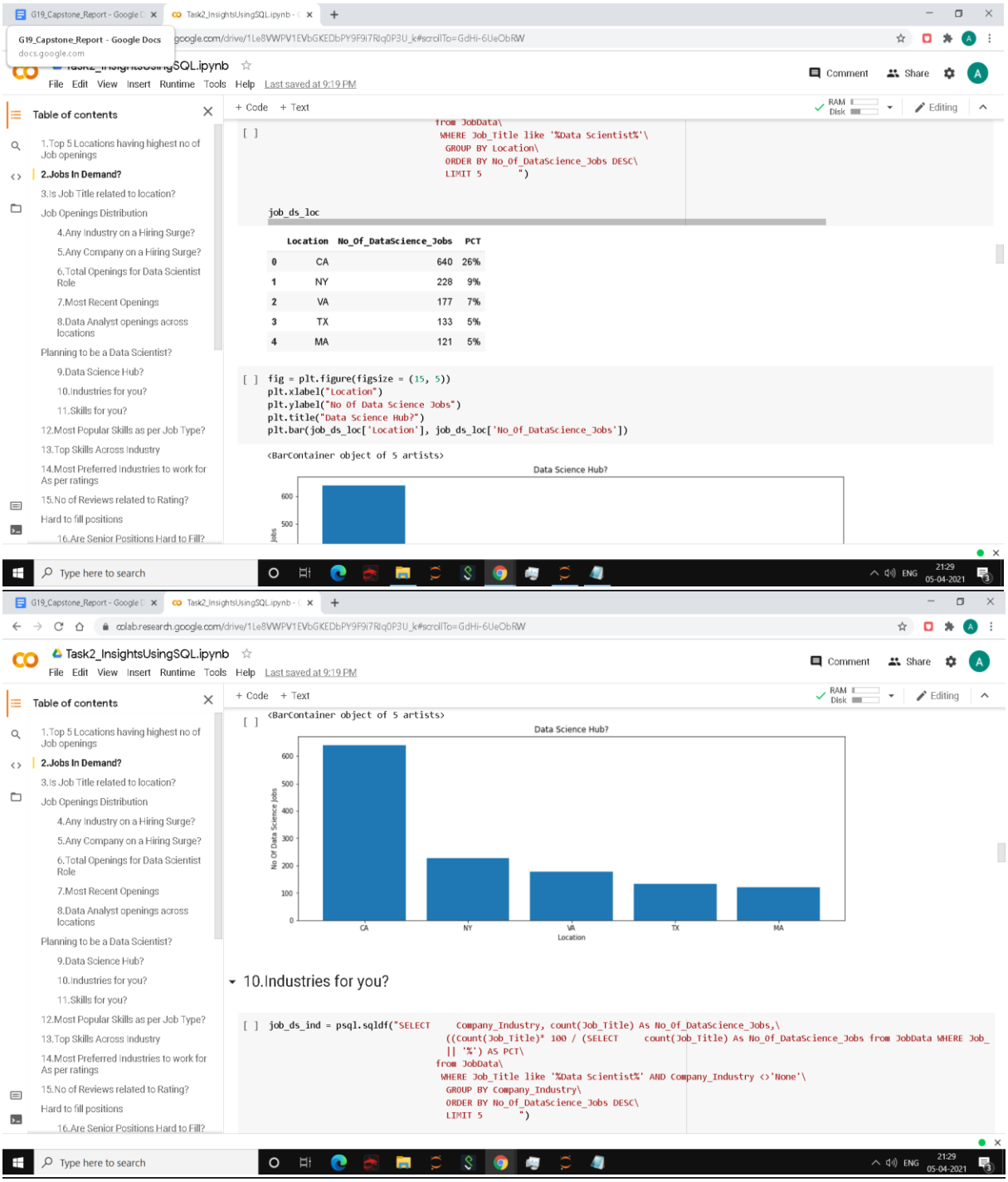
```
DAJC = psql.sqldf(" select Location, count(Job_Title) as No_of_openings, avg(Queried_Salary) as Avg_Salary from JobData \
                    where Job_Type = 'data_analyst' \
                    group by Location \
                    order by count(Job_Title) desc limit 10")
DAJC
```

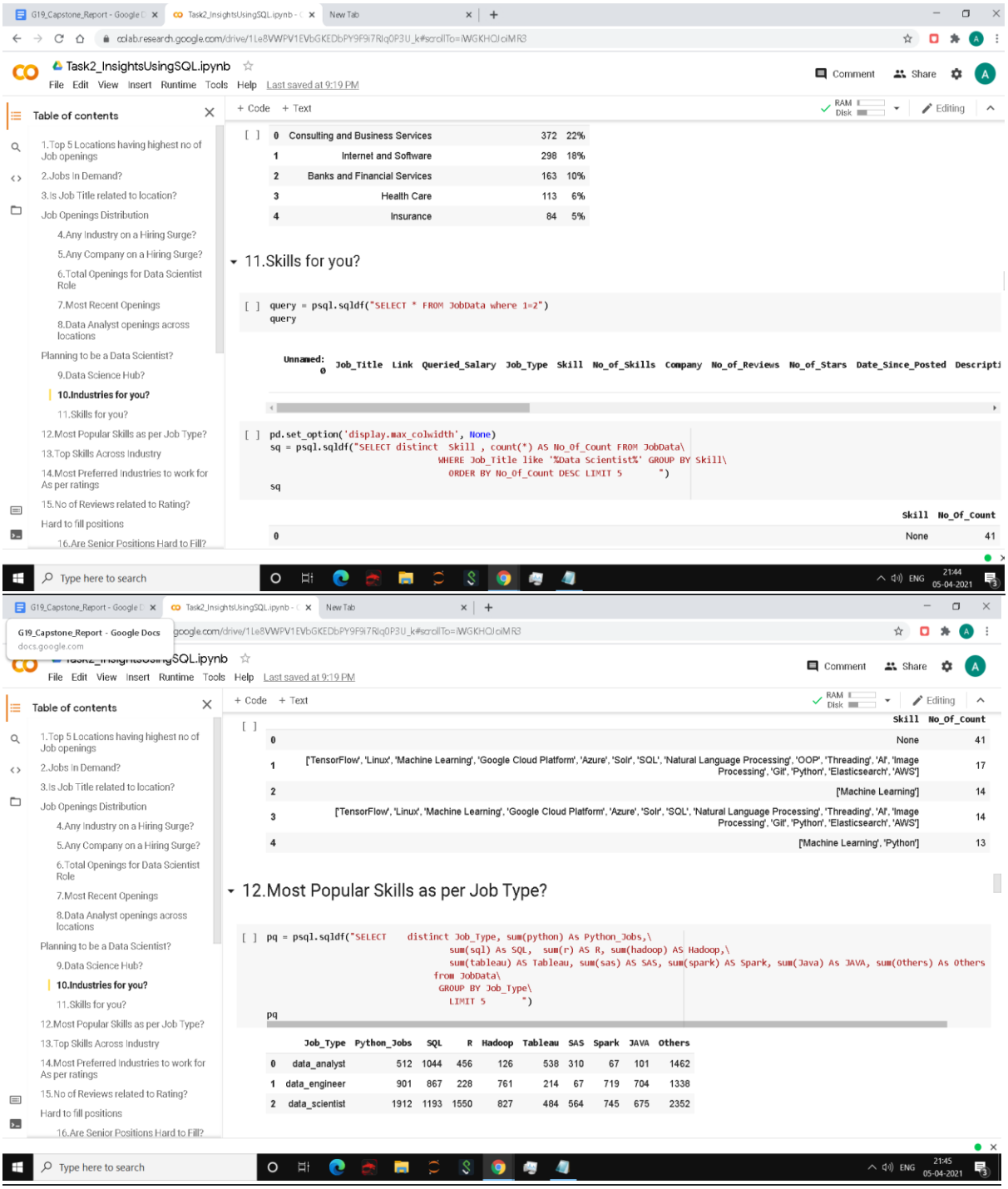

	Location	No_of_openings	Avg_Salary
0	CA	376	65372.340426
1	NY	230	56695.652174
2	TX	117	41880.341880
3	MA	86	41627.906977
4	VA	85	57882.352941
5	IL	66	51818.181818
6	GA	59	38305.084746
7	WA	57	49824.561404
8	PA	56	35000.000000
9	MD	55	45454.545455

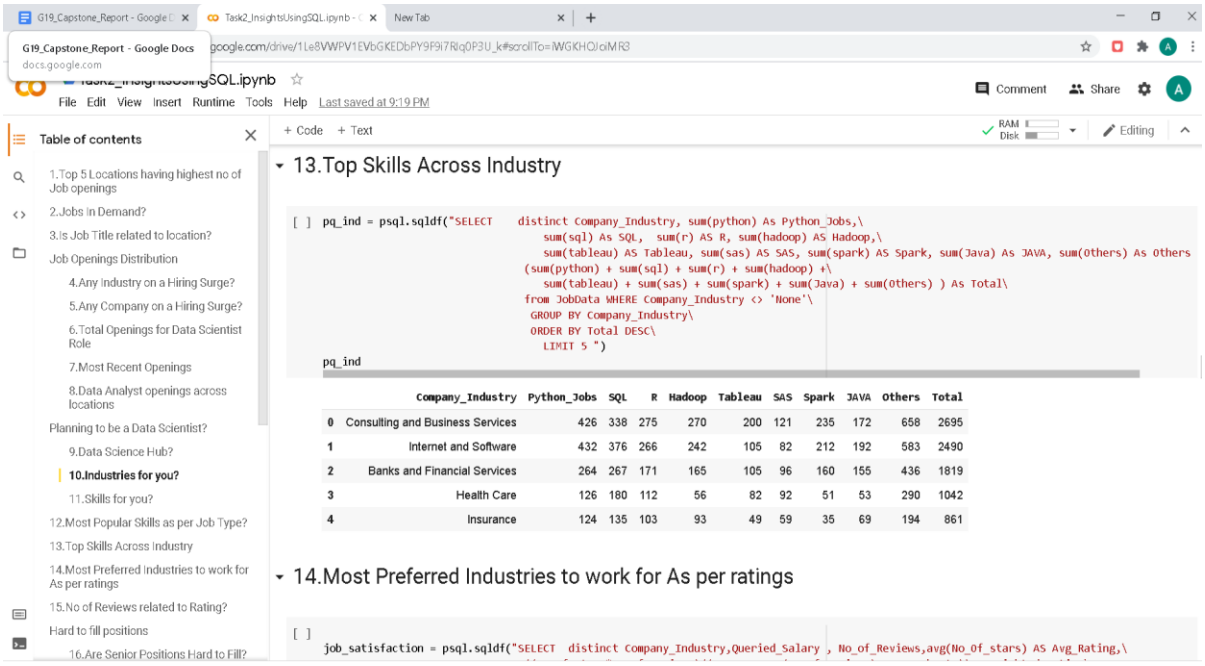
Planning to be a Data Scientist?

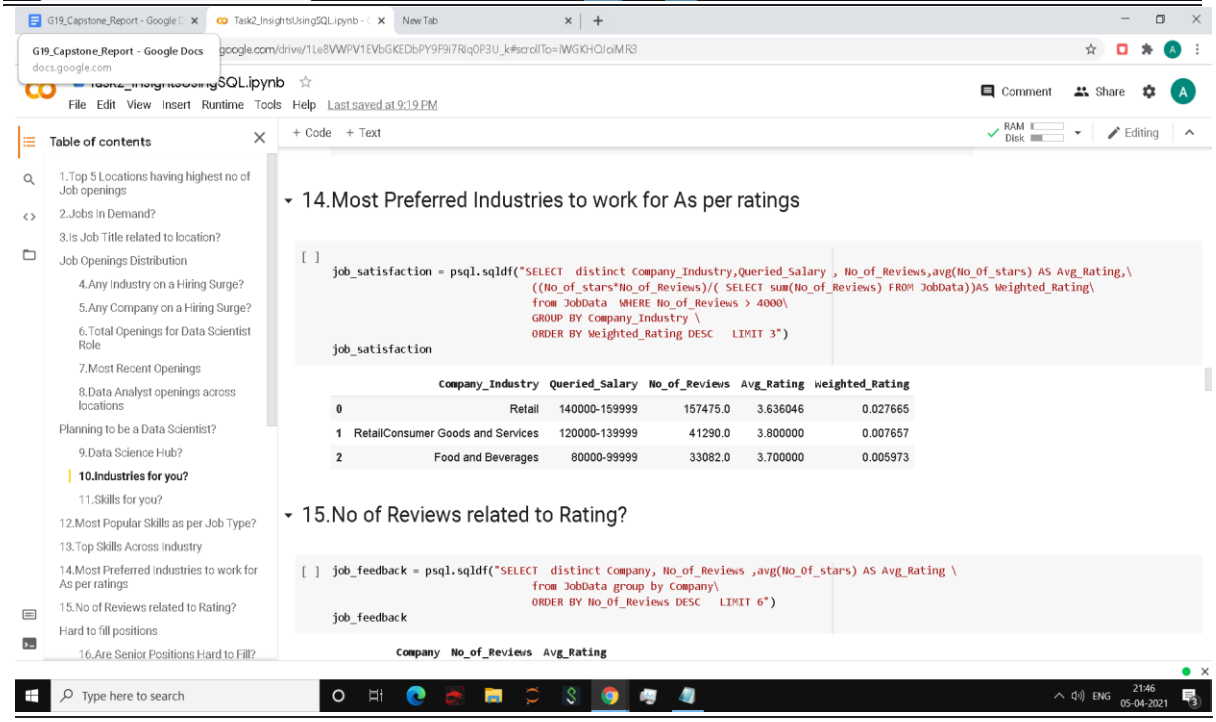
9.Data Science Hub?

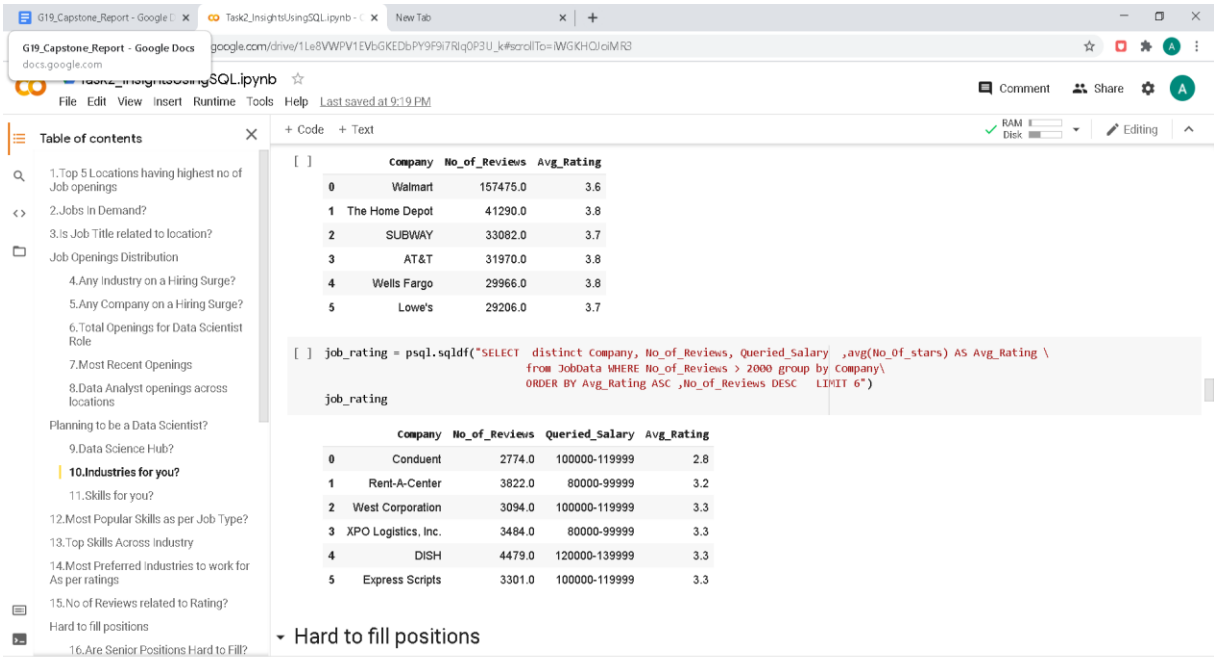
```
job_ds_loc = psql.sqldf("SELECT     Location, count(Job_Title) As No_Of_DataScience_Jobs,\
                        ((Count(Job_Title)* 100 / (SELECT     count(Job_Title) As No_Of_DataScience_Jobs from JobData WHERE Job_
                        || '%') AS PCT\
                        from JobData\
                        WHERE Job_Title like '%Data Scientist%'\
                        GROUP BY Location\
```

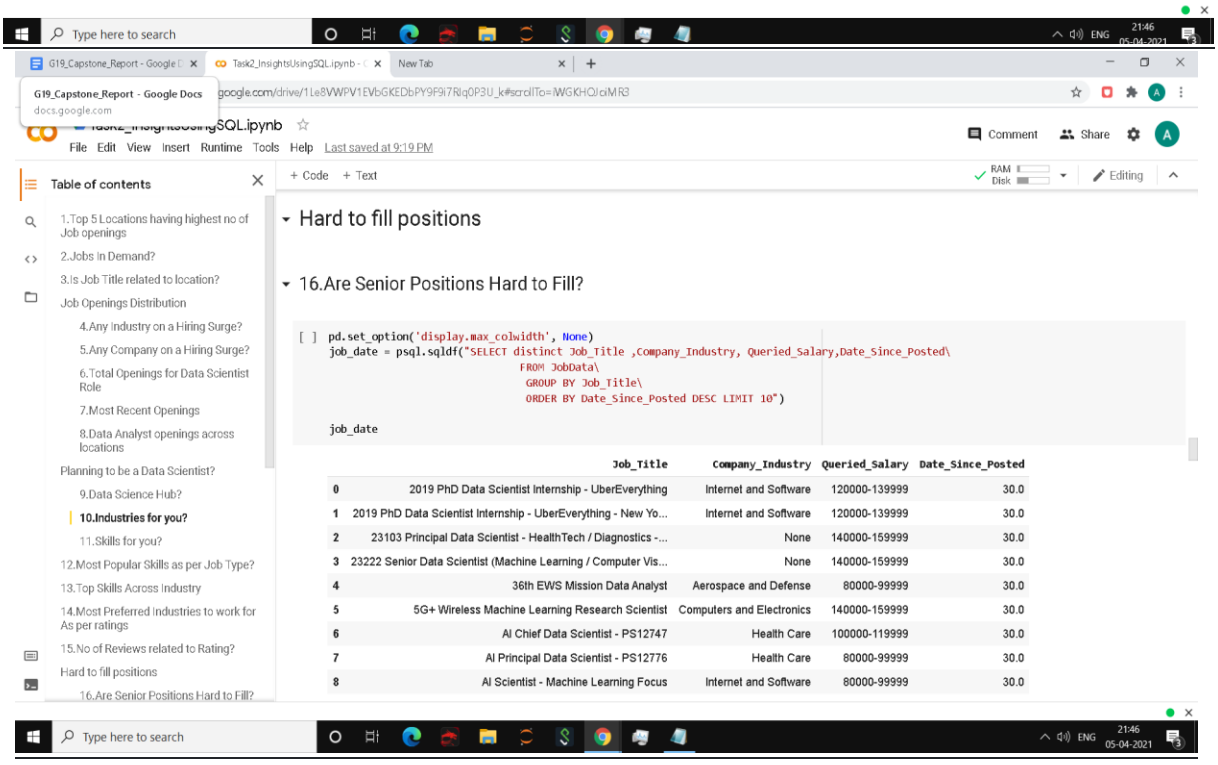

G19_Capstone_Report - Google Docs
docs.google.com
Task2_InsightsUsingSQL.ipynb
File Edit View Insert Runtime Tools Help Last saved at 9:19 PM
Comment
Share
Table of contents
1.Top 5 Locations having highest no of Job openings
2.Jobs In Demand?
3.Is Job Title related to location?
Job Openings Distribution
4.Any Industry on a Hiring Surge?
5.Any Company on a Hiring Surge?
6.Total Openings for Data Scientist Role
7.Most Recent Openings
8.Data Analyst openings across locations
Planning to be a Data Scientist?
9.Data Science Hub?
10.Industries for you?
11.Skills for you?
12.Most Popular Skills as per Job Type?
13.Top Skills Across Industry
14.Most Preferred Industries to work for As per ratings
15.No of Reviews related to Rating?
Hard to fill positions
16.Are Senior Positions Hard to Fill?
+ Code + Text
Editing
from JobData\
WHERE Job_Title like '%Data Scientist%'\
GROUP BY Location\
ORDER BY No_Of_DataScience_Jobs DESC\
LIMIT 5 ")
job_ds_loc
Location No_Of_DataScience_Jobs PCT
0 CA 640 26%
1 NY 228 9%
2 VA 177 7%
3 TX 133 5%
4 MA 121 5%
fig = plt.figure(figsize = (15, 5))
plt.xlabel("Location")
plt.ylabel("No Of Data Science Jobs")
plt.title("Data Science Hub?")
plt.bar(job_ds_loc['Location'], job_ds_loc['No_Of_DataScience_Jobs'])
<BarContainer object of 5 artists>
Data Science Hub?
Jobs
600
500
Type here to search
ENG
21:29
05-04-2021
Task2_InsightsUsingSQL.ipynb
File Edit View Insert Runtime Tools Help Last saved at 9:19 PM
Comment
Share
Table of contents
1.Top 5 Locations having highest no of Job openings
2.Jobs In Demand?
3.Is Job Title related to location?
Job Openings Distribution
4.Any Industry on a Hiring Surge?
5.Any Company on a Hiring Surge?
6.Total Openings for Data Scientist Role
7.Most Recent Openings
8.Data Analyst openings across locations
Planning to be a Data Scientist?
9.Data Science Hub?
10.Industries for you?
11.Skills for you?
12.Most Popular Skills as per Job Type?
13.Top Skills Across Industry
14.Most Preferred Industries to work for As per ratings
15.No of Reviews related to Rating?
Hard to fill positions
16.Are Senior Positions Hard to Fill?
+ Code + Text
Editing
<BarContainer object of 5 artists>
Data Science Hub?
No Of Data Science Jobs
600
500
400
300
200
100
0
CA
NY
VA
TX
MA
Location
10.Industries for you?
job_ds_ind = psql.sqldf("SELECT Company_Industry, count(Job_Title) As No_Of_DataScience_Jobs,\
((Count(Job_Title)* 100 / (SELECT count(Job_Title) As No_Of_DataScience_Jobs from JobData WHERE Job_
|| '%') AS PCT\
from JobData\
WHERE Job_Title like '%Data Scientist%' AND Company_Industry <>'None'\
GROUP BY Company_Industry\
ORDER BY No_Of_DataScience_Jobs DESC\
LIMIT 5 ")
Type here to search
ENG
21:29
05-04-2021


Task2_InsightsUsingSQL.ipynb
File Edit View Insert Runtime Tools Help Last saved at 9:19 PM
Comment Share
+ Code + Text
RAM Disk
Editing
Table of contents
1.Top 5 Locations having highest no of Job openings
2.Jobs In Demand?
3.Is Job Title related to location?
Job Openings Distribution
4.Any Industry on a Hiring Surge?
5.Any Company on a Hiring Surge?
6.Total Openings for Data Scientist Role
7.Most Recent Openings
8.Data Analyst openings across locations
Planning to be a Data Scientist?
9.Data Science Hub?
10.Industries for you?
11.Skills for you?
12.Most Popular Skills as per Job Type?
13.Top Skills Across Industry
14.Most Preferred Industries to work for As per ratings
15.No of Reviews related to Rating?
Hard to fill positions
16.Are Senior Positions Hard to Fill?
0 Consulting and Business Services 372 22%
1 Internet and Software 298 18%
2 Banks and Financial Services 163 10%
3 Health Care 113 6%
4 Insurance 84 5%
11.Skills for you?
query = psql.sqldf("SELECT * FROM JobData where 1=2")
query
Unnamed: 0 Job_Title Link Queried_Salary Job_Type Skill No_of_Skills Company No_of_Reviews No_of_Stars Date_Since_Posted Descripti
pd.set_option('display.max_colwidth', None)
sq = psql.sqldf("SELECT distinct Skill , count(*) AS No_Of_Count FROM JobData\
WHERE Job_Title like '%Data Scientist%' GROUP BY Skill\
ORDER BY No_Of_Count DESC LIMIT 5 ")
sq
Skill No_Of_Count
0 None 41
Type here to search
ENG 21:44 05-04-2021
G19_Capstone_Report - Google Docs
docs.google.com
Skill No_Of_Count
0 None 41
1 ['TensorFlow', 'Linux', 'Machine Learning', 'Google Cloud Platform', 'Azure', 'Solr', 'SQL', 'Natural Language Processing', 'OOP', 'Threading', 'AI', 'Image Processing', 'Git', 'Python', 'Elasticsearch', 'AWS'] 17
2 ['Machine Learning'] 14
3 ['TensorFlow', 'Linux', 'Machine Learning', 'Google Cloud Platform', 'Azure', 'Solr', 'SQL', 'Natural Language Processing', 'Threading', 'AI', 'Image Processing', 'Git', 'Python', 'Elasticsearch', 'AWS'] 14
4 ['Machine Learning', 'Python'] 13
12.Most Popular Skills as per Job Type?
pq = psql.sqldf("SELECT distinct Job_Type, sum(python) As Python_Jobs,\
sum(sql) As SQL, sum(r) AS R, sum(hadoop) AS Hadoop,\
sum(tableau) AS Tableau, sum(sas) AS SAS, sum(spark) AS Spark, sum(Java) As JAVA, sum(Others) As Others
from JobData\
GROUP BY Job_Type\
LIMIT 5 ")
pq
Job_Type Python_Jobs SQL R Hadoop Tableau SAS Spark JAVA Others
0 data_analyst 512 1044 456 126 538 310 67 101 1462
1 data_engineer 901 867 228 761 214 67 719 704 1338
2 data_scientist 1912 1193 1550 827 484 564 745 675 2352
Type here to search
ENG 21:45 05-04-2021


G19_Capstone_Report - Google Docs
docs.google.com
File Edit View Insert Runtime Tools Help Last saved at 9:19 PM
Comment
Share
+ Code + Text
RAM
Disk
Editing
Table of contents
1.Top 5 Locations having highest no of Job openings
2.Jobs In Demand?
3.Is Job Title related to location?
Job Openings Distribution
4.Any Industry on a Hiring Surge?
5.Any Company on a Hiring Surge?
6.Total Openings for Data Scientist Role
7.Most Recent Openings
8.Data Analyst openings across locations
Planning to be a Data Scientist?
9.Data Science Hub?
10.Industries for you?
11.Skills for you?
12.Most Popular Skills as per Job Type?
13.Top Skills Across Industry
14.Most Preferred Industries to work for As per ratings
15.No of Reviews related to Rating?
Hard to fill positions
16.Are Senior Positions Hard to Fill?
13.Top Skills Across Industry
[ ] pq_ind = psql.sqldf("SELECT distinct Company_Industry, sum(python) As Python_Jobs,\
sum(sql) As SQL, sum(r) AS R, sum(hadoop) AS Hadoop,\
sum(tableau) AS Tableau, sum(sas) AS SAS, sum(spark) AS Spark, sum(Java) As JAVA, sum(Others) As Others
(sum(python) + sum(sql) + sum(r) + sum(hadoop) +\
sum(tableau) + sum(sas) + sum(spark) + sum(Java) + sum(Others)) As Total\
from JobData WHERE Company_Industry <> 'None'\
GROUP BY Company_Industry\
ORDER BY Total DESC\
LIMIT 5 ")
pq_ind
Company_Industry Python_Jobs SQL R Hadoop Tableau SAS Spark JAVA Others Total
0 Consulting and Business Services 426 338 275 270 200 121 235 172 658 2695
1 Internet and Software 432 376 266 242 105 82 212 192 583 2490
2 Banks and Financial Services 264 267 171 165 105 96 160 155 436 1819
3 Health Care 126 180 112 56 82 92 51 53 290 1042
4 Insurance 124 135 103 93 49 59 35 69 194 861
14.Most Preferred Industries to work for As per ratings
[ ] job_satisfaction = psql.sqldf("SELECT distinct Company_Industry,Queried_Salary , No_of_Reviews,avg(No_Of_stars) AS Avg_Rating,\
Type here to search
ENG
21:45
05-04-2021


G19_Capstone_Report - Google Docs
docs.google.com
File Edit View Insert Runtime Tools Help Last saved at 9:19 PM
Comment
Share
+ Code + Text
RAM
Disk
Editing
Table of contents
1.Top 5 Locations having highest no of Job openings
2.Jobs In Demand?
3.Is Job Title related to location?
Job Openings Distribution
4.Any Industry on a Hiring Surge?
5.Any Company on a Hiring Surge?
6.Total Openings for Data Scientist Role
7.Most Recent Openings
8.Data Analyst openings across locations
Planning to be a Data Scientist?
9.Data Science Hub?
10.Industries for you?
11.Skills for you?
12.Most Popular Skills as per Job Type?
13.Top Skills Across Industry
14.Most Preferred Industries to work for As per ratings
15.No of Reviews related to Rating?
Hard to fill positions
16.Are Senior Positions Hard to Fill?
14.Most Preferred Industries to work for As per ratings
[ ] job_satisfaction = psql.sqldf("SELECT distinct Company_Industry,Queried_Salary , No_of_Reviews,avg(No_Of_stars) AS Avg_Rating,\
((No_of_stars*No_of_Reviews)/(SELECT sum(No_of_Reviews) FROM JobData))AS Weighted_Rating\
from JobData WHERE No_of_Reviews > 4000\
GROUP BY Company_Industry \
ORDER BY Weighted_Rating DESC LIMIT 3")
job_satisfaction
Company_Industry Queried_Salary No_of_Reviews Avg_Rating Weighted_Rating
0 Retail 140000-159999 157475.0 3.636046 0.027665
1 RetailConsumer Goods and Services 120000-139999 41290.0 3.800000 0.007657
2 Food and Beverages 80000-99999 33082.0 3.700000 0.005973
15.No of Reviews related to Rating?
[ ] job_feedback = psql.sqldf("SELECT distinct Company, No_of_Reviews ,avg(No_Of_stars) AS Avg_Rating \
from JobData group by Company\
ORDER BY No_Of_Reviews DESC LIMIT 6")
job_feedback
Company No_of_Reviews Avg_Rating
Type here to search
ENG
21:46
05-04-2021


G19_Capstone_Report - Google Docs
docs.google.com
File Edit View Insert Runtime Tools Help Last saved at 9:19 PM
Comment Share
+ Code + Text
RAM Disk
Editing
Table of contents
1.Top 5 Locations having highest no of Job openings
2.Jobs In Demand?
3.Is Job Title related to location?
Job Openings Distribution
4.Any Industry on a Hiring Surge?
5.Any Company on a Hiring Surge?
6.Total Openings for Data Scientist Role
7.Most Recent Openings
8.Data Analyst openings across locations
Planning to be a Data Scientist?
9.Data Science Hub?
10.Industries for you?
11.Skills for you?
12.Most Popular Skills as per Job Type?
13.Top Skills Across Industry
14.Most Preferred Industries to work for As per ratings
15.No of Reviews related to Rating?
Hard to fill positions
16.Are Senior Positions Hard to Fill?
[ ]
Company No_of_Reviews Avg_Rating
0 Walmart 157475.0 3.6
1 The Home Depot 41290.0 3.8
2 SUBWAY 33082.0 3.7
3 AT&T 31970.0 3.8
4 Wells Fargo 29966.0 3.8
5 Lowe's 29206.0 3.7
[ ] job_rating = psql.sqldf("SELECT distinct Company, No_of_Reviews, Queried_Salary ,avg(No_Of_stars) AS Avg_Rating \
from JobData WHERE No_of_Reviews > 2000 group by Company\
ORDER BY Avg_Rating ASC ,No_of_Reviews DESC LIMIT 6")
job_rating
Company No_of_Reviews Queried_Salary Avg_Rating
0 Conduent 2774.0 100000-119999 2.8
1 Rent-A-Center 3822.0 80000-99999 3.2
2 West Corporation 3094.0 100000-119999 3.3
3 XPO Logistics, Inc. 3484.0 80000-99999 3.3
4 DISH 4479.0 120000-139999 3.3
5 Express Scripts 3301.0 100000-119999 3.3
Hard to fill positions


G19_Capstone_Report - Google Docs
docs.google.com
File Edit View Insert Runtime Tools Help Last saved at 9:19 PM
Comment Share
+ Code + Text
RAM Disk
Editing
Table of contents
1.Top 5 Locations having highest no of Job openings
2.Jobs In Demand?
3.Is Job Title related to location?
Job Openings Distribution
4.Any Industry on a Hiring Surge?
5.Any Company on a Hiring Surge?
6.Total Openings for Data Scientist Role
7.Most Recent Openings
8.Data Analyst openings across locations
Planning to be a Data Scientist?
9.Data Science Hub?
10.Industries for you?
11.Skills for you?
12.Most Popular Skills as per Job Type?
13.Top Skills Across Industry
14.Most Preferred Industries to work for As per ratings
15.No of Reviews related to Rating?
Hard to fill positions
16.Are Senior Positions Hard to Fill?
Hard to fill positions
16.Are Senior Positions Hard to Fill?
[ ] pd.set_option('display.max_colwidth', None)
job_date = psql.sqldf("SELECT distinct Job_Title ,Company_Industry, Queried_Salary,Date_Since_Posted\
FROM JobData\
GROUP BY Job_Title\
ORDER BY Date_Since_Posted DESC LIMIT 10")
job_date
Job_Title Company_Industry Queried_Salary Date_Since_Posted
0 2019 PhD Data Scientist Internship - UberEverything Internet and Software 120000-139999 30.0
1 2019 PhD Data Scientist Internship - UberEverything - New Yo... Internet and Software 120000-139999 30.0
2 23103 Principal Data Scientist - HealthTech / Diagnostics -... None 140000-159999 30.0
3 23222 Senior Data Scientist (Machine Learning / Computer Vis... None 140000-159999 30.0
4 36th EWS Mission Data Analyst Aerospace and Defense 80000-99999 30.0
5 5G+ Wireless Machine Learning Research Scientist Computers and Electronics 140000-159999 30.0
6 AI Chief Data Scientist - PS12747 Health Care 100000-119999 30.0
7 AI Principal Data Scientist - PS12776 Health Care 80000-99999 30.0
8 AI Scientist - Machine Learning Focus Internet and Software 80000-99999 30.0

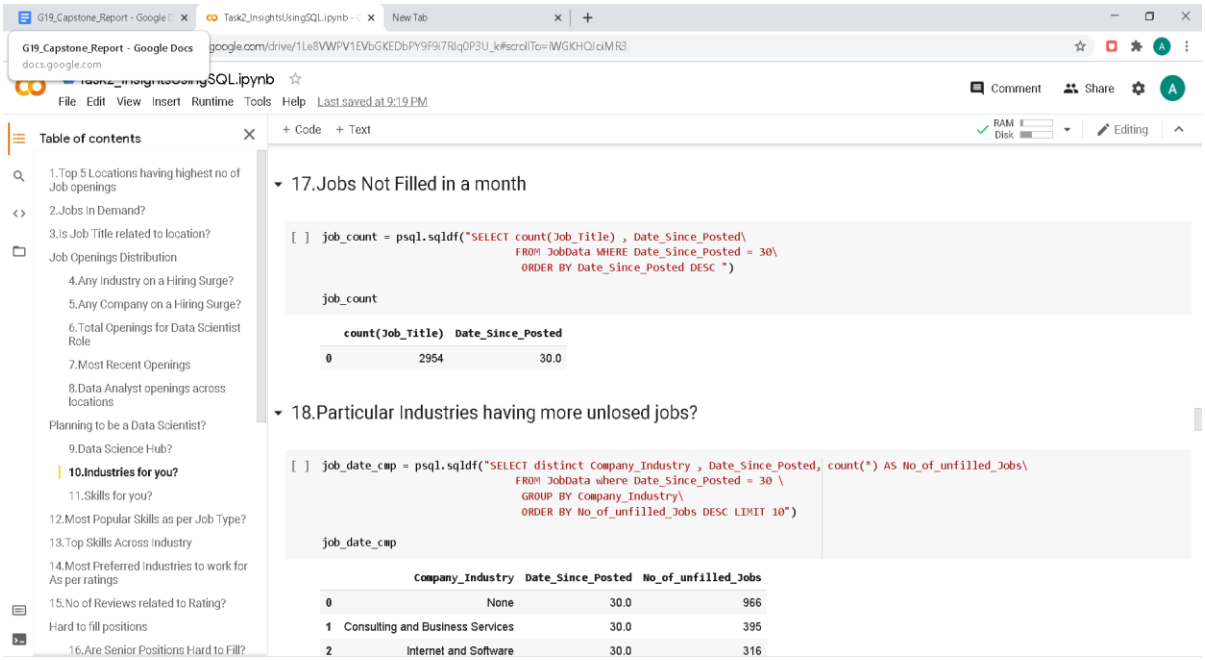


Table of contents

- 1.Top 5 Locations having highest no of Job openings
- 2.Jobs In Demand?
- 3.Is Job Title related to location?
- Job Openings Distribution
  - 4.Any Industry on a Hiring Surge?
  - 5.Any Company on a Hiring Surge?
  - 6.Total Openings for Data Scientist Role
  - 7.Most Recent Openings
  - 8.Data Analyst openings across locations
- Planning to be a Data Scientist?
  - 9.Data Science Hub?
  - **10.Industries for you?**
  - 11.Skills for you?
- 12.Most Popular Skills as per Job Type?
- 13.Top Skills Across Industry
- 14.Most Preferred Industries to work for As per ratings
- 15.No of Reviews related to Rating?
- Hard to fill positions
  - 16.Are Senior Positions Hard to Fill?

17.Jobs Not Filled in a month

```
job_count = psql.sqldf("SELECT count(Job_Title) , Date_Since_Posted\
                        FROM JobData WHERE Date_Since_Posted = 30\
                        ORDER BY Date_Since_Posted DESC ")

job_count
```

	count(Job_Title)	Date_Since_Posted
0	2954	30.0

18.Particular Industries having more unlosed jobs?

```
job_date_cmp = psql.sqldf("SELECT distinct Company_Industry , Date_Since_Posted, count(*) AS No_of_unfilled_Jobs\
                           FROM JobData where Date_Since_Posted = 30 \
                           GROUP BY Company_Industry\
                           ORDER BY No_of_unfilled_Jobs DESC LIMIT 10")

job_date_cmp
```

	Company_Industry	Date_Since_Posted	No_of_unfilled_Jobs
0	None	30.0	966
1	Consulting and Business Services	30.0	395
2	Internet and Software	30.0	316

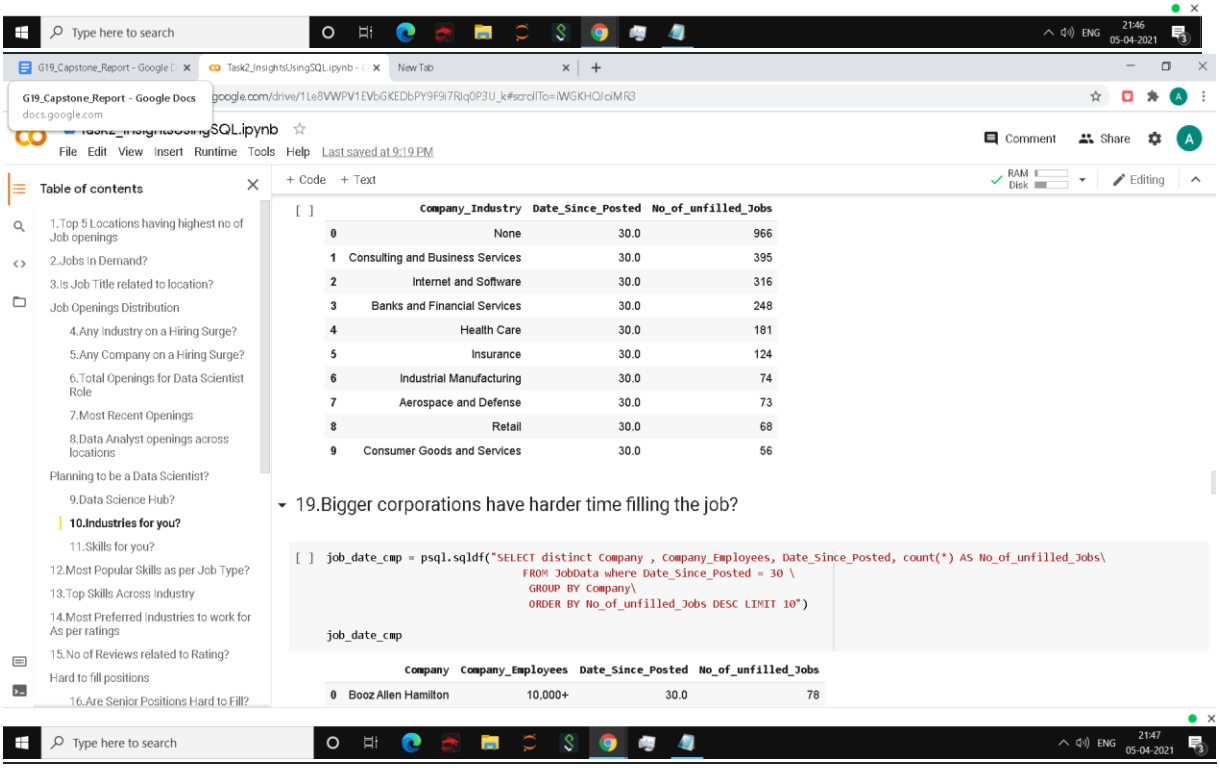


Table of contents

- 1.Top 5 Locations having highest no of Job openings
- 2.Jobs In Demand?
- 3.Is Job Title related to location?
- Job Openings Distribution
  - 4.Any Industry on a Hiring Surge?
  - 5.Any Company on a Hiring Surge?
  - 6.Total Openings for Data Scientist Role
  - 7.Most Recent Openings
  - 8.Data Analyst openings across locations
- Planning to be a Data Scientist?
  - 9.Data Science Hub?
  - **10.Industries for you?**
  - 11.Skills for you?
- 12.Most Popular Skills as per Job Type?
- 13.Top Skills Across Industry
- 14.Most Preferred Industries to work for As per ratings
- 15.No of Reviews related to Rating?
- Hard to fill positions
  - 16.Are Senior Positions Hard to Fill?

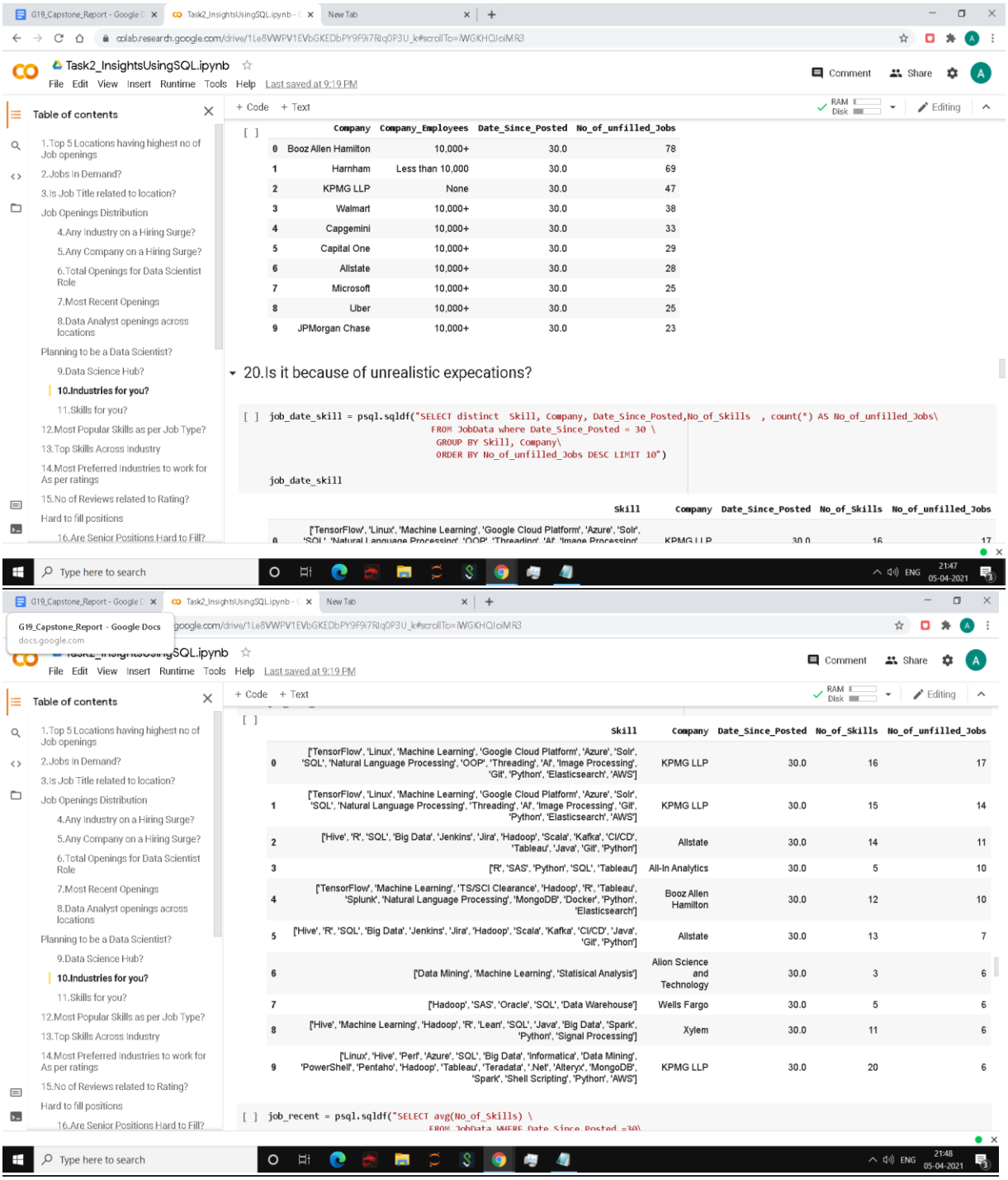
```
[ ]
```

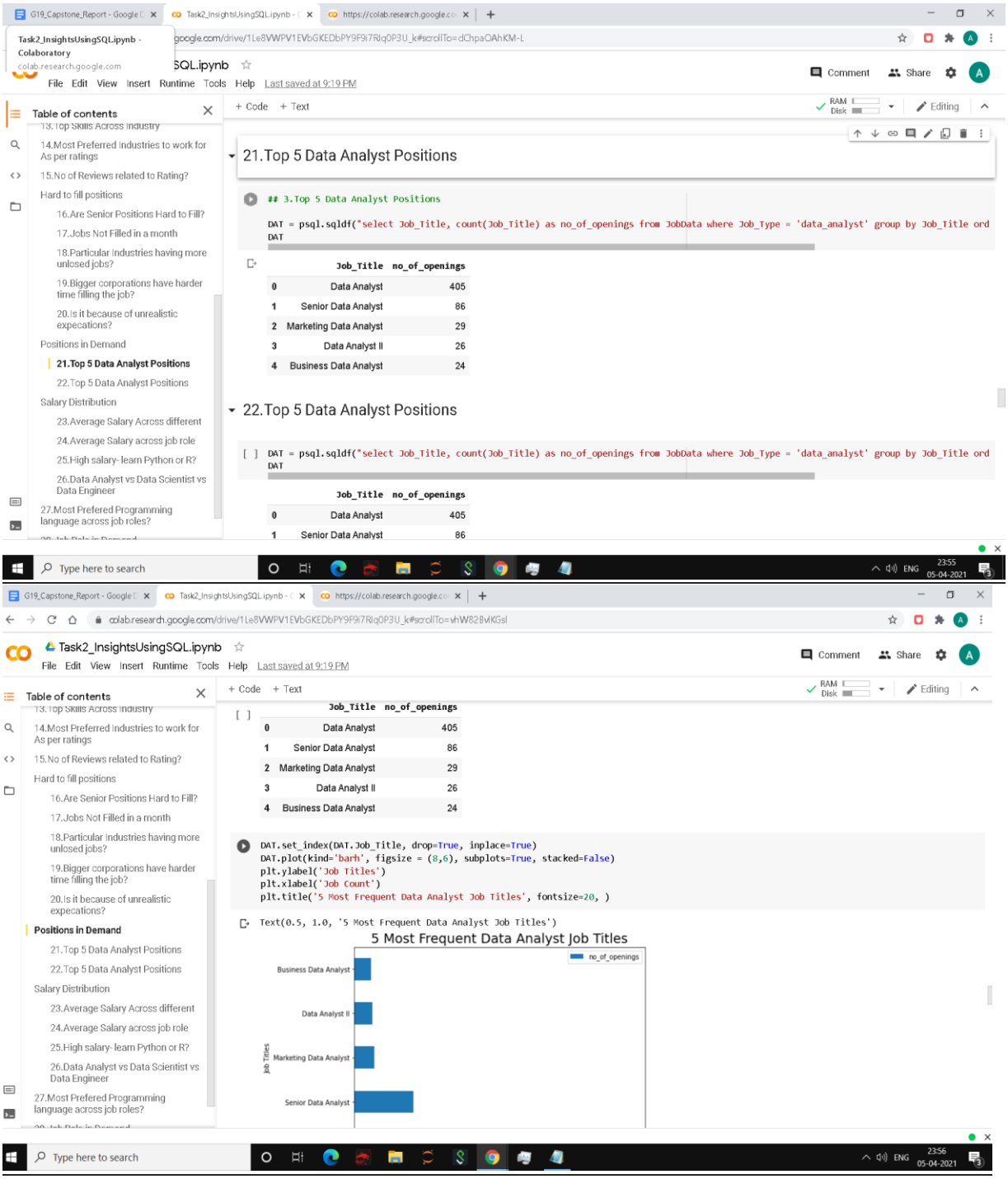
	Company_Industry	Date_Since_Posted	No_of_unfilled_Jobs
0	None	30.0	966
1	Consulting and Business Services	30.0	395
2	Internet and Software	30.0	316
3	Banks and Financial Services	30.0	248
4	Health Care	30.0	181
5	Insurance	30.0	124
6	Industrial Manufacturing	30.0	74
7	Aerospace and Defense	30.0	73
8	Retail	30.0	68
9	Consumer Goods and Services	30.0	56

19.Bigger corporations have harder time filling the job?

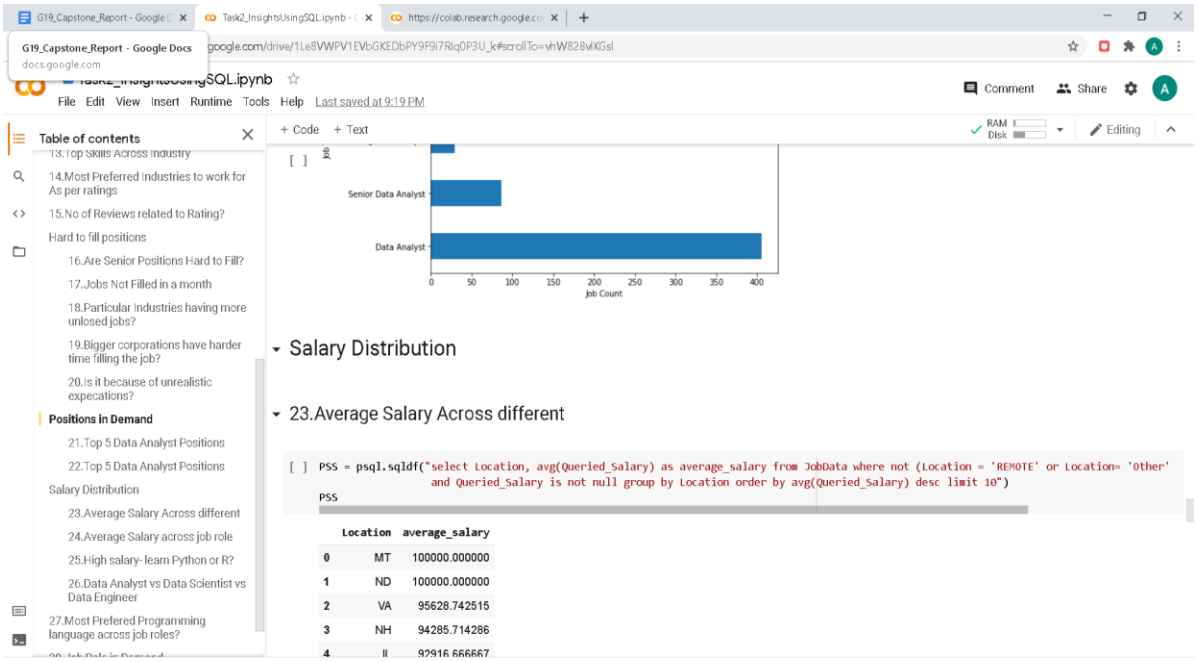
```
job_date_cmp = psql.sqldf("SELECT distinct Company , Company_Employees, Date_Since_Posted, count(*) AS No_of_unfilled_Jobs\
                           FROM JobData where Date_Since_Posted = 30 \
                           GROUP BY Company\
                           ORDER BY No_of_unfilled_Jobs DESC LIMIT 10")

job_date_cmp
```

	Company	Company_Employees	Date_Since_Posted	No_of_unfilled_Jobs
0	Booz Allen Hamilton	10,000+	30.0	78


Task2_InsightsUsingSQL.ipynb
File Edit View Insert Runtime Tools Help Last saved at 9:19 PM
Comment Share
Table of contents
1.Top 5 Locations having highest no of Job openings
2.Jobs In Demand?
3.Is Job Title related to location?
Job Openings Distribution
4.Any Industry on a Hiring Surge?
5.Any Company on a Hiring Surge?
6.Total Openings for Data Scientist Role
7.Most Recent Openings
8.Data Analyst openings across locations
Planning to be a Data Scientist?
9.Data Science Hub?
10.Industries for you?
11.Skills for you?
12.Most Popular Skills as per Job Type?
13.Top Skills Across Industry
14.Most Preferred Industries to work for As per ratings
15.No of Reviews related to Rating?
Hard to fill positions
16.Are Senior Positions Hard to Fill?
+ Code + Text
RAM Disk
Editing
Company Company_Employees Date_Since_Posted No_of_unfilled_Jobs
0 Booz Allen Hamilton 10,000+ 30.0 78
1 Harnham Less than 10,000 30.0 69
2 KPMG LLP None 30.0 47
3 Walmart 10,000+ 30.0 38
4 Capgemini 10,000+ 30.0 33
5 Capital One 10,000+ 30.0 29
6 Allstate 10,000+ 30.0 28
7 Microsoft 10,000+ 30.0 25
8 Uber 10,000+ 30.0 25
9 JPMorgan Chase 10,000+ 30.0 23
20.Is it because of unrealistic expecations?
job_date_skill = psql.sqldf("SELECT distinct Skill, Company, Date_Since_Posted,No_of_Skills , count(*) AS No_of_unfilled_Jobs\
FROM JobData where Date_Since_Posted = 30 \
GROUP BY Skill, Company\
ORDER BY No_of_unfilled_Jobs DESC LIMIT 10")
job_date_skill
Skill Company Date_Since_Posted No_of_Skills No_of_unfilled_Jobs
0 ['TensorFlow', 'Linux', 'Machine Learning', 'Google Cloud Platform', 'Azure', 'Solr', 'SQL', 'Natural Language Processing', 'OOP', 'Threading', 'AI', 'Image Processing', KPMG LLP 30.0 16 17
Type here to search
ENG 21:47 05-04-2021
G19_Capstone_Report - Google Docs
docs.google.com
Skill Company Date_Since_Posted No_of_Skills No_of_unfilled_Jobs
0 ['TensorFlow', 'Linux', 'Machine Learning', 'Google Cloud Platform', 'Azure', 'Solr', 'SQL', 'Natural Language Processing', 'OOP', 'Threading', 'AI', 'Image Processing', 'Git', 'Python', 'Elasticsearch', 'AWS'] KPMG LLP 30.0 16 17
1 ['TensorFlow', 'Linux', 'Machine Learning', 'Google Cloud Platform', 'Azure', 'Solr', 'SQL', 'Natural Language Processing', 'Threading', 'AI', 'Image Processing', 'Git', 'Python', 'Elasticsearch', 'AWS'] KPMG LLP 30.0 15 14
2 ['Hive', 'R', 'SQL', 'Big Data', 'Jenkins', 'Jira', 'Hadoop', 'Scala', 'Kafka', 'CI/CD', 'Tableau', 'Java', 'Git', 'Python'] Allstate 30.0 14 11
3 ['R', 'SAS', 'Python', 'SQL', 'Tableau'] All-In Analytics 30.0 5 10
4 ['TensorFlow', 'Machine Learning', 'TS/SCI Clearance', 'Hadoop', 'R', 'Tableau', 'Splunk', 'Natural Language Processing', 'MongoDB', 'Docker', 'Python', 'Elasticsearch'] Booz Allen Hamilton 30.0 12 10
5 ['Hive', 'R', 'SQL', 'Big Data', 'Jenkins', 'Jira', 'Hadoop', 'Scala', 'Kafka', 'CI/CD', 'Java', 'Git', 'Python'] Allstate 30.0 13 7
6 ['Data Mining', 'Machine Learning', 'Statisical Analysis'] Alion Science and Technology 30.0 3 6
7 ['Hadoop', 'SAS', 'Oracle', 'SQL', 'Data Warehouse'] Wells Fargo 30.0 5 6
8 ['Hive', 'Machine Learning', 'Hadoop', 'R', 'Lean', 'SQL', 'Java', 'Big Data', 'Spark', 'Python', 'Signal Processing'] Xylem 30.0 11 6
9 ['Linux', 'Hive', 'Perl', 'Azure', 'SQL', 'Big Data', 'Informatica', 'Data Mining', 'PowerShell', 'Pentaho', 'Hadoop', 'Tableau', 'Teradata', '.Net', 'Alteryx', 'MongoDB', 'Spark', 'Shell Scripting', 'Python', 'AWS'] KPMG LLP 30.0 20 6
job_recent = psql.sqldf("SELECT avg(No_of_Skills) \
Type here to search
ENG 21:48 05-04-2021


Task2_InsightsUsingSQL.ipynb - Colaboratory
colab.research.google.com
Task2_InsightsUsingSQL.ipynb
File Edit View Insert Runtime Tools Help Last saved at 9:19 PM
Comment Share
+ Code + Text
RAM Disk
Editing
Table of contents
13. Top Skills Across Industry
14.Most Preferred Industries to work for As per ratings
15.No of Reviews related to Rating?
Hard to fill positions
16.Are Senior Positions Hard to Fill?
17.Jobs Not Filled in a month
18.Particular Industries having more unlosed jobs?
19.Bigger corporations have harder time filling the job?
20.Is it because of unrealistic expecations?
Positions in Demand
21.Top 5 Data Analyst Positions
22.Top 5 Data Analyst Positions
Salary Distribution
23.Average Salary Across different
24.Average Salary across job role
25.High salary- learn Python or R?
26.Data Analyst vs Data Scientist vs Data Engineer
27.Most Prefered Programming language across job roles?
21.Top 5 Data Analyst Positions
3.Top 5 Data Analyst Positions
DAT = psql.sqldf("select Job_Title, count(Job_Title) as no_of_openings from JobData where Job_Type = 'data_analyst' group by Job_Title ord
DAT
Job_Title no_of_openings
0 Data Analyst 405
1 Senior Data Analyst 86
2 Marketing Data Analyst 29
3 Data Analyst II 26
4 Business Data Analyst 24
22.Top 5 Data Analyst Positions
DAT = psql.sqldf("select Job_Title, count(Job_Title) as no_of_openings from JobData where Job_Type = 'data_analyst' group by Job_Title ord
DAT
Job_Title no_of_openings
0 Data Analyst 405
1 Senior Data Analyst 86
Type here to search
ENG 23:55 05-04-2021

colab.research.google.com/drive/1Le8VWPV1EVbGKEDbPY9F9i7RIq0P3U_k#scrollTo=vhW82BvKGsI
Task2_InsightsUsingSQL.ipynb
File Edit View Insert Runtime Tools Help Last saved at 9:19 PM
Comment Share
+ Code + Text
RAM Disk
Editing
Table of contents
13. Top Skills Across Industry
14.Most Preferred Industries to work for As per ratings
15.No of Reviews related to Rating?
Hard to fill positions
16.Are Senior Positions Hard to Fill?
17.Jobs Not Filled in a month
18.Particular Industries having more unlosed jobs?
19.Bigger corporations have harder time filling the job?
20.Is it because of unrealistic expecations?
Positions in Demand
21.Top 5 Data Analyst Positions
22.Top 5 Data Analyst Positions
Salary Distribution
23.Average Salary Across different
24.Average Salary across job role
25.High salary- learn Python or R?
26.Data Analyst vs Data Scientist vs Data Engineer
27.Most Prefered Programming language across job roles?
Job_Title no_of_openings
0 Data Analyst 405
1 Senior Data Analyst 86
2 Marketing Data Analyst 29
3 Data Analyst II 26
4 Business Data Analyst 24
DAT.set_index(DAT.Job_Title, drop=True, inplace=True)
DAT.plot(kind='barh', figsize = (8,6), subplots=True, stacked=False)
plt.ylabel('Job Titles')
plt.xlabel('Job Count')
plt.title('5 Most Frequent Data Analyst Job Titles', fontsize=20,)
Text(0.5, 1.0, '5 Most Frequent Data Analyst Job Titles')
5 Most Frequent Data Analyst Job Titles
no_of_openings
Business Data Analyst
Data Analyst II
Marketing Data Analyst
Senior Data Analyst
Job Titles
Type here to search
ENG 23:56 05-04-2021

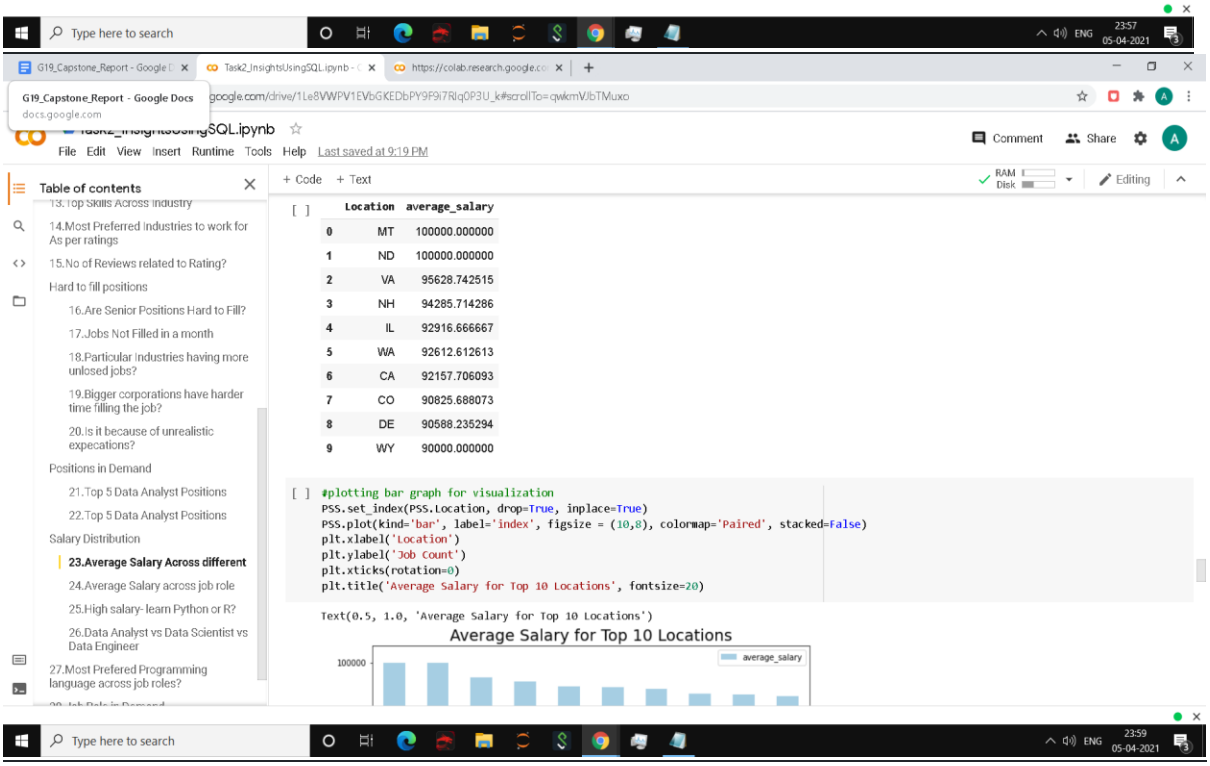


Salary Distribution

23.Average Salary Across different

```
PSS = psql.sqldf("select Location, avg(Queried_Salary) as average_salary from JobData where not (Location = 'REMOTE' or Location= 'Other'
                  and Queried_Salary is not null group by Location order by avg(Queried_Salary) desc limit 10")
PSS
```

	Location	average_salary
0	MT	100000.000000
1	ND	100000.000000
2	VA	95628.742515
3	NH	94285.714286
4	IL	92916.666667



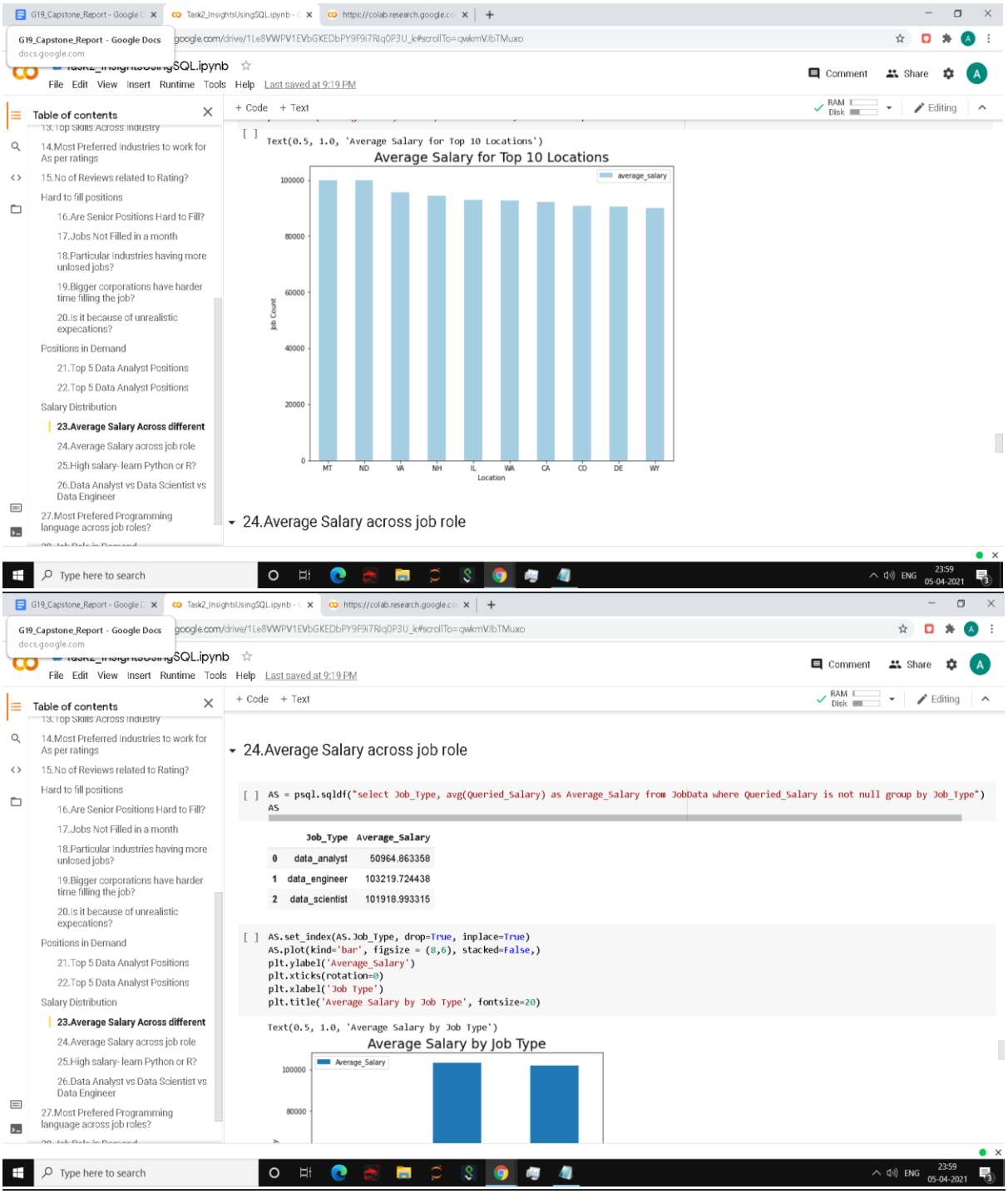


Table of contents
13. Top Skills Across Industry
14.Most Preferred Industries to work for As per ratings
15.No of Reviews related to Rating?
Hard to fill positions
16.Are Senior Positions Hard to Fill?
17.Jobs Not Filled in a month
18.Particular Industries having more unlosed jobs?
19.Bigger corporations have harder time filling the job?
20.Is it because of unrealistic expecations?
Positions in Demand
21.Top 5 Data Analyst Positions
22.Top 5 Data Analyst Positions
Salary Distribution
23.Average Salary Across different
24.Average Salary across job role
25.High salary- learn Python or R?
26.Data Analyst vs Data Scientist vs Data Engineer
27.Most Prefered Programming language across job roles?
Text(0.5, 1.0, 'Average Salary for Top 10 Locations')
Average Salary for Top 10 Locations
average_salary
Job Count
Location
MT
ND
VA
NH
IL
WA
CA
CO
DE
WY
24.Average Salary across job role
Type here to search
Table of contents
24.Average Salary across job role
AS = psql.sqldf("select Job_Type, avg(Queried_Salary) as Average_Salary from JobData where Queried_Salary is not null group by Job_Type")
AS
Job_Type Average_Salary
0 data_analyst 50964.863358
1 data_engineer 103219.724438
2 data_scientist 101918.993315
AS.set_index(AS.Job_Type, drop=True, inplace=True)
AS.plot(kind='bar', figsize = (8,6), stacked=False,)
plt.ylabel('Average_Salary')
plt.xticks(rotation=0)
plt.xlabel('Job Type')
plt.title('Average Salary by Job Type', fontsize=20)
Text(0.5, 1.0, 'Average Salary by Job Type')
Average Salary by Job Type
Average_Salary
Type here to search

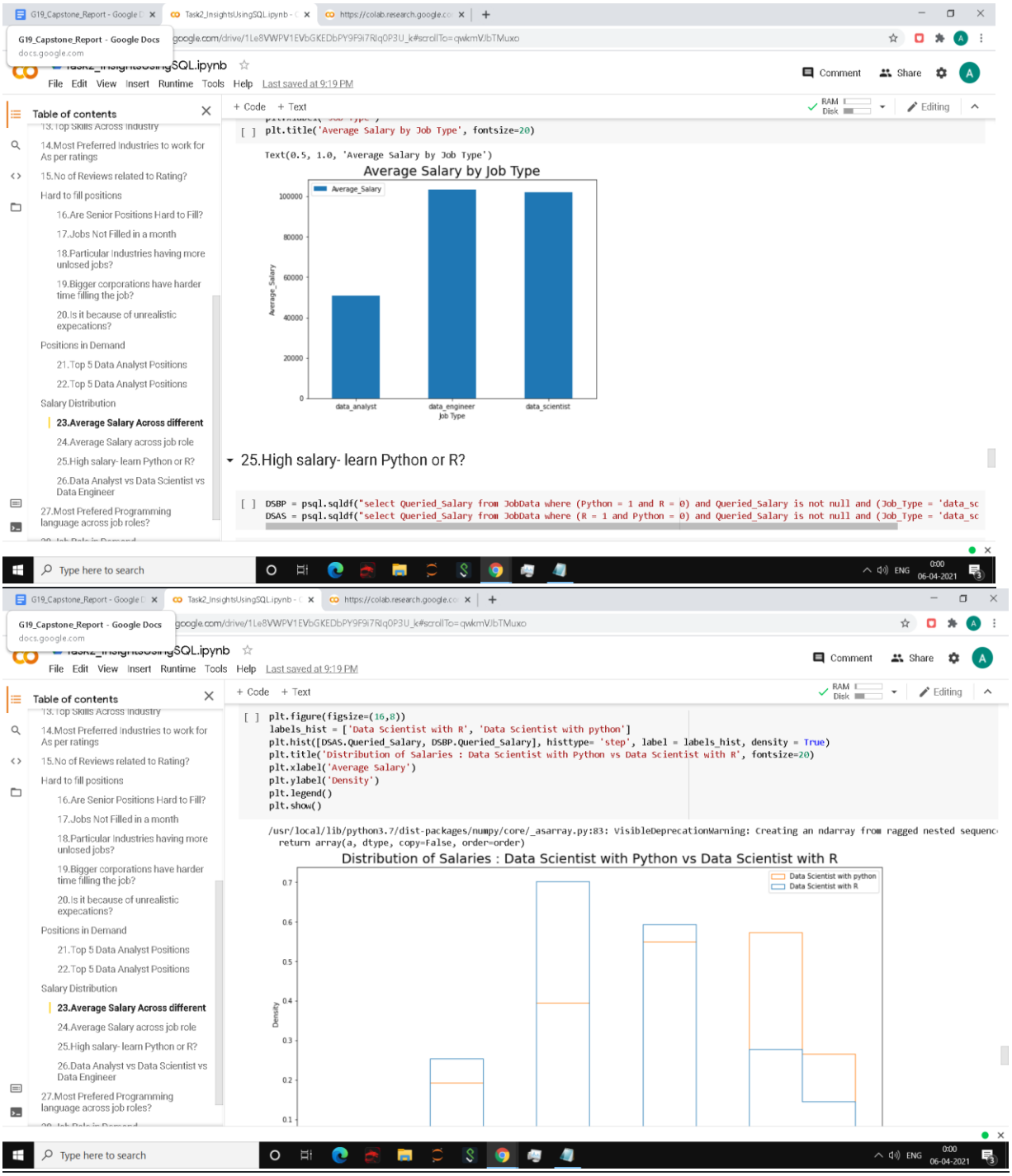
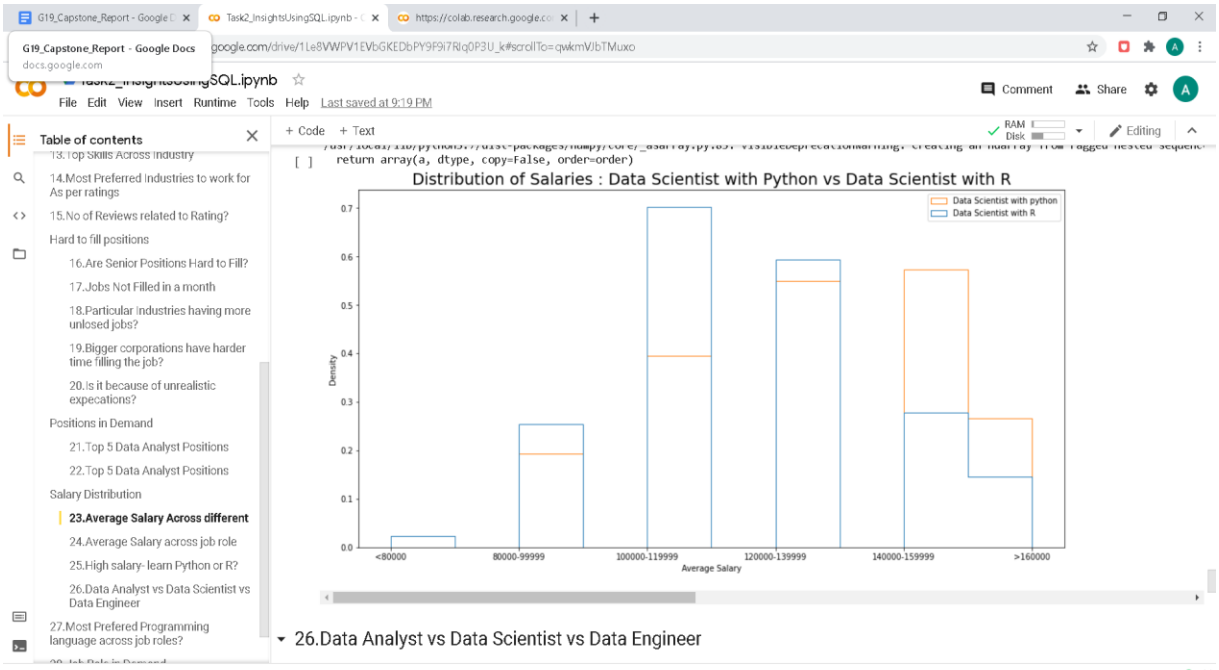
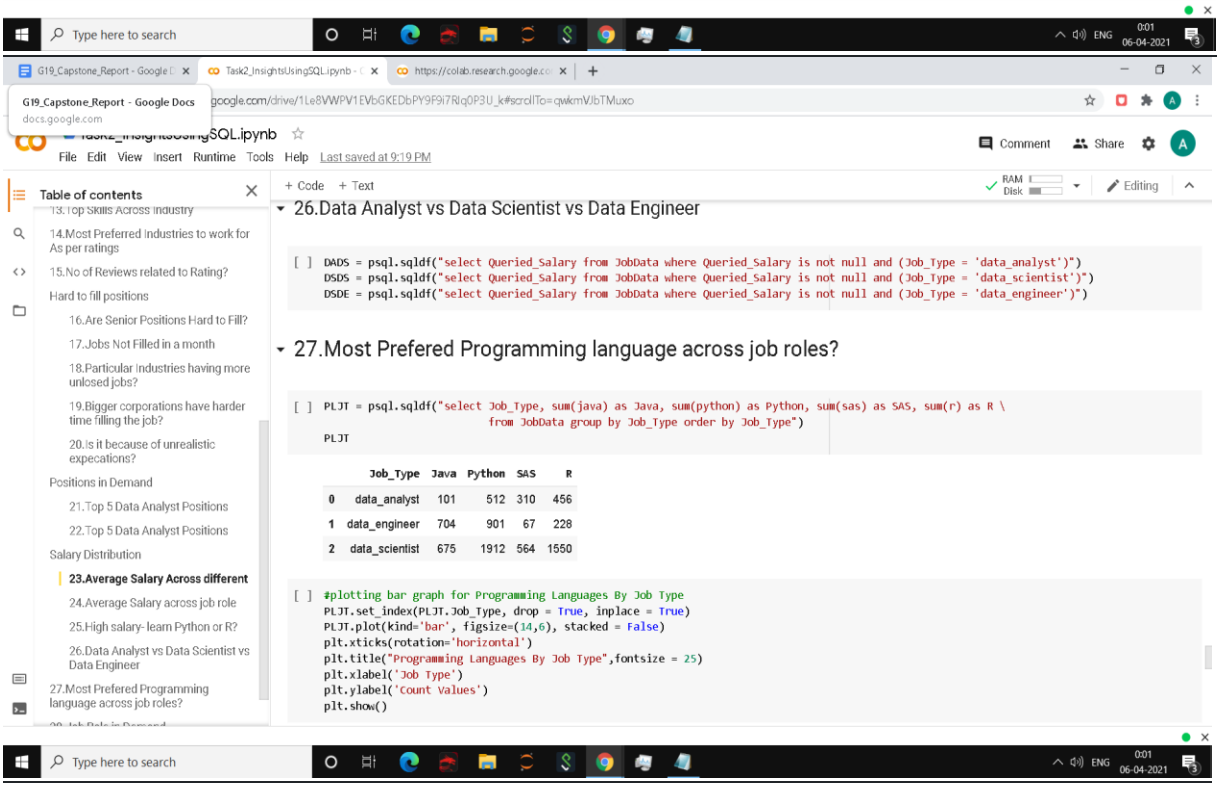


Table of contents
13. Top Skills Across Industry
14.Most Preferred Industries to work for As per ratings
15.No of Reviews related to Rating?
Hard to fill positions
16.Are Senior Positions Hard to Fill?
17.Jobs Not Filled in a month
18.Particular Industries having more unlosed jobs?
19.Bigger corporations have harder time filling the job?
20.Is it because of unrealistic expecations?
Positions in Demand
21.Top 5 Data Analyst Positions
22.Top 5 Data Analyst Positions
Salary Distribution
23.Average Salary Across different
24.Average Salary across job role
25.High salary- learn Python or R?
26.Data Analyst vs Data Scientist vs Data Engineer
27.Most Prefered Programming language across job roles?
[ ] plt.title('Average Salary by Job Type', fontsize=20)
Text(0.5, 1.0, 'Average Salary by Job Type')
Average Salary by Job Type
Average_Salary
data_analyst
data_engineer
data_scientist
Job Type
25.High salary- learn Python or R?
[ ] DSBP = psql.sqldf("select Queried_Salary from JobData where (Python = 1 and R = 0) and Queried_Salary is not null and (Job_Type = 'data_sc
DSAS = psql.sqldf("select Queried_Salary from JobData where (R = 1 and Python = 0) and Queried_Salary is not null and (Job_Type = 'data_sc
[ ] plt.figure(figsize=(16,8))
labels_hist = ['Data Scientist with R', 'Data Scientist with python']
plt.hist([DSAS.Queried_Salary, DSBP.Queried_Salary], histtype= 'step', label = labels_hist, density = True)
plt.title('Distribution of Salaries : Data Scientist with Python vs Data Scientist with R', fontsize=20)
plt.xlabel('Average Salary')
plt.ylabel('Density')
plt.legend()
plt.show()
/usr/local/lib/python3.7/dist-packages/numpy/core/_asarray.py:83: VisibleDeprecationWarning: Creating an ndarray from ragged nested sequenc
return array(a, dtype, copy=False, order=order)
Distribution of Salaries : Data Scientist with Python vs Data Scientist with R
Data Scientist with python
Data Scientist with R
Density



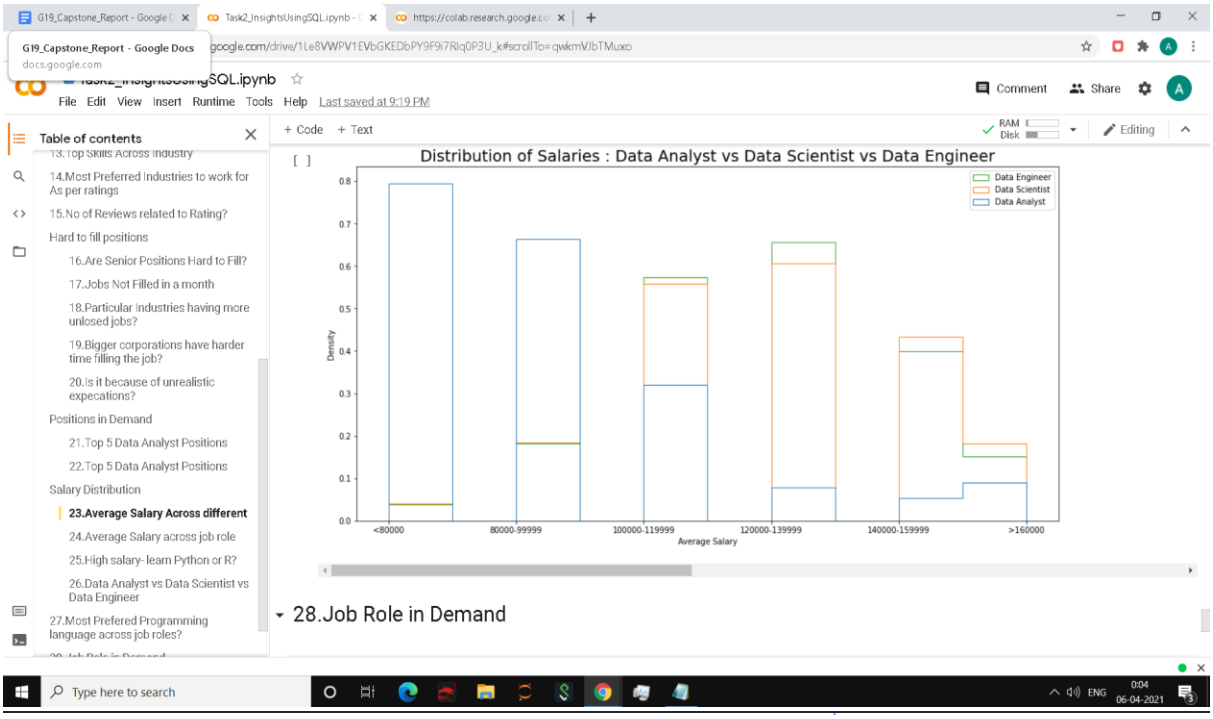


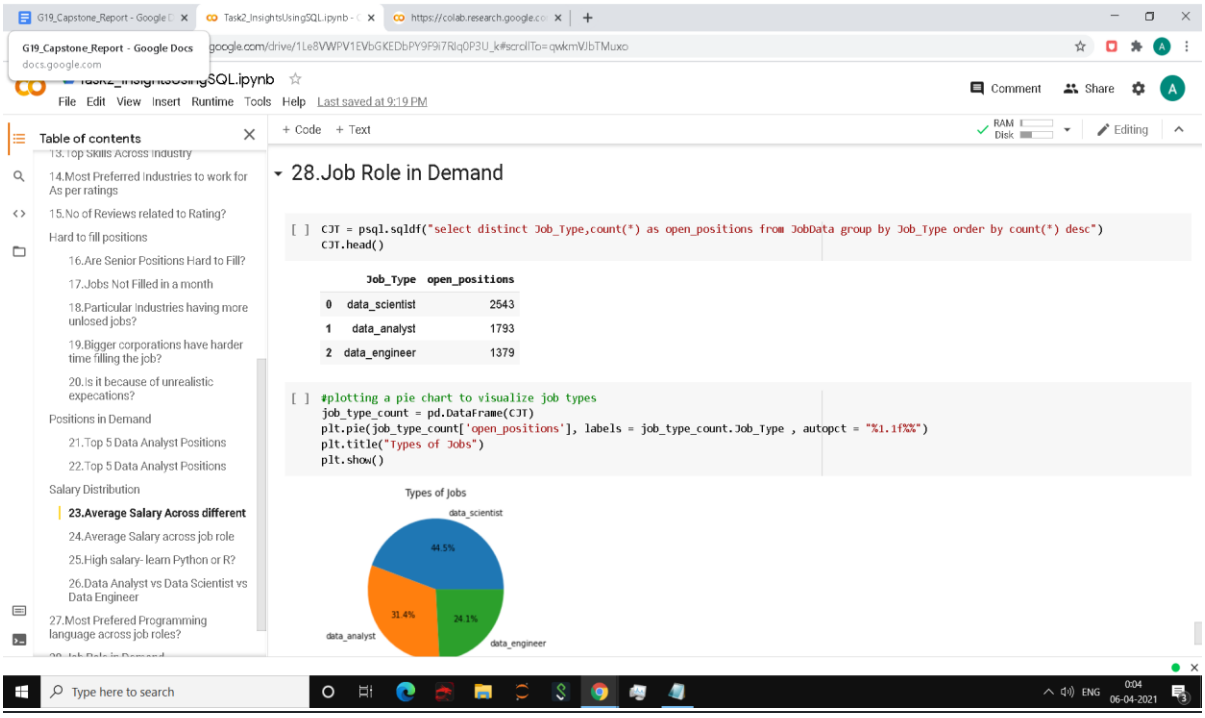

```
DADS = psql.sqldf("select Queried_Salary from JobData where Queried_Salary is not null and (Job_Type = 'data_analyst')")
DSDS = psql.sqldf("select Queried_Salary from JobData where Queried_Salary is not null and (Job_Type = 'data_scientist')")
DSDE = psql.sqldf("select Queried_Salary from JobData where Queried_Salary is not null and (Job_Type = 'data_engineer')")
```

```
PLJT = psql.sqldf("select Job_Type, sum(java) as Java, sum(python) as Python, sum(sas) as SAS, sum(r) as R \
                   from JobData group by Job_Type order by Job_Type")
PLJT
```

	Job_Type	Java	Python	SAS	R
0	data_analyst	101	512	310	456
1	data_engineer	704	901	67	228
2	data_scientist	675	1912	564	1550

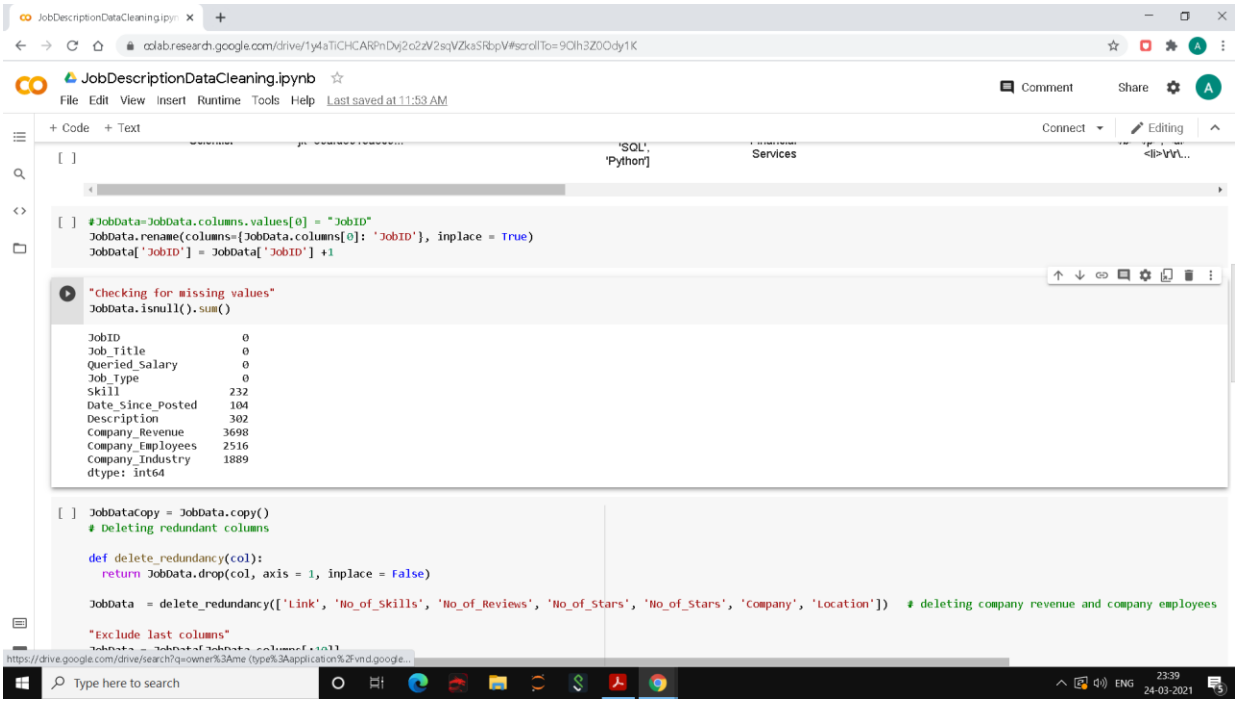
```
#plotting bar graph for Programming Languages By Job Type
PLJT.set_index(PLJT.Job_Type, drop = True, inplace = True)
PLJT.plot(kind='bar', figsize=(14,6), stacked = False)
plt.xticks(rotation='horizontal')
plt.title("Programming Languages By Job Type",fontsize = 25)
plt.xlabel('Job Type')
plt.ylabel('Count Values')
plt.show()
```


Distribution of Salaries : Data Analyst vs Data Scientist vs Data Engineer
Data Engineer
Data Scientist
Data Analyst
Density
Average Salary
<80000
80000-99999
100000-119999
120000-139999
140000-159999
>160000
28.Job Role in Demand

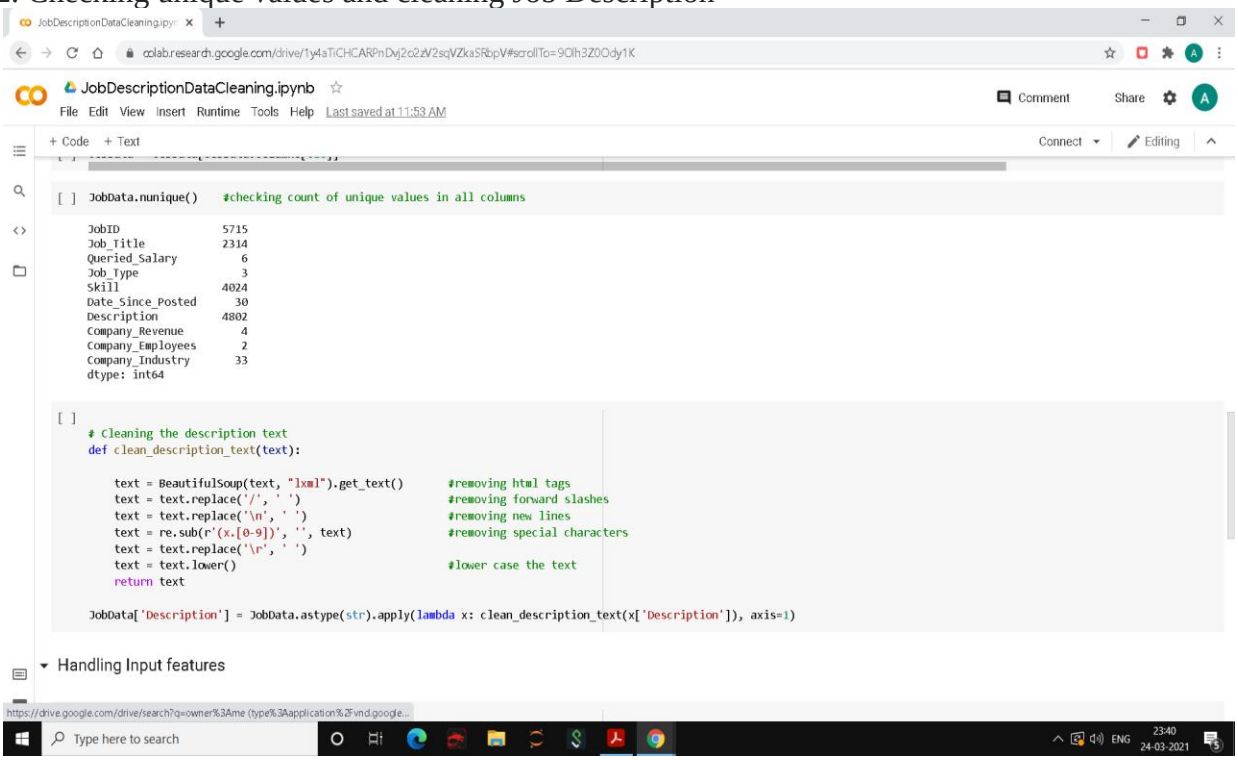

28.Job Role in Demand
CJT = psql.sqldf("select distinct Job_Type,count(*) as open_positions from JobData group by Job_Type order by count(*) desc")
CJT.head()
Job_Type open_positions
0 data_scientist 2543
1 data_analyst 1793
2 data_engineer 1379
#plotting a pie chart to visualize job types
job_type_count = pd.DataFrame(CJT)
plt.pie(job_type_count['open_positions'], labels = job_type_count.Job_Type , autopct = "%1.1f%%")
plt.title("Types of Jobs")
plt.show()
Types of Jobs
data_scientist
44.5%
31.4%
24.1%
data_analyst
data_engineer

Data Cleaning

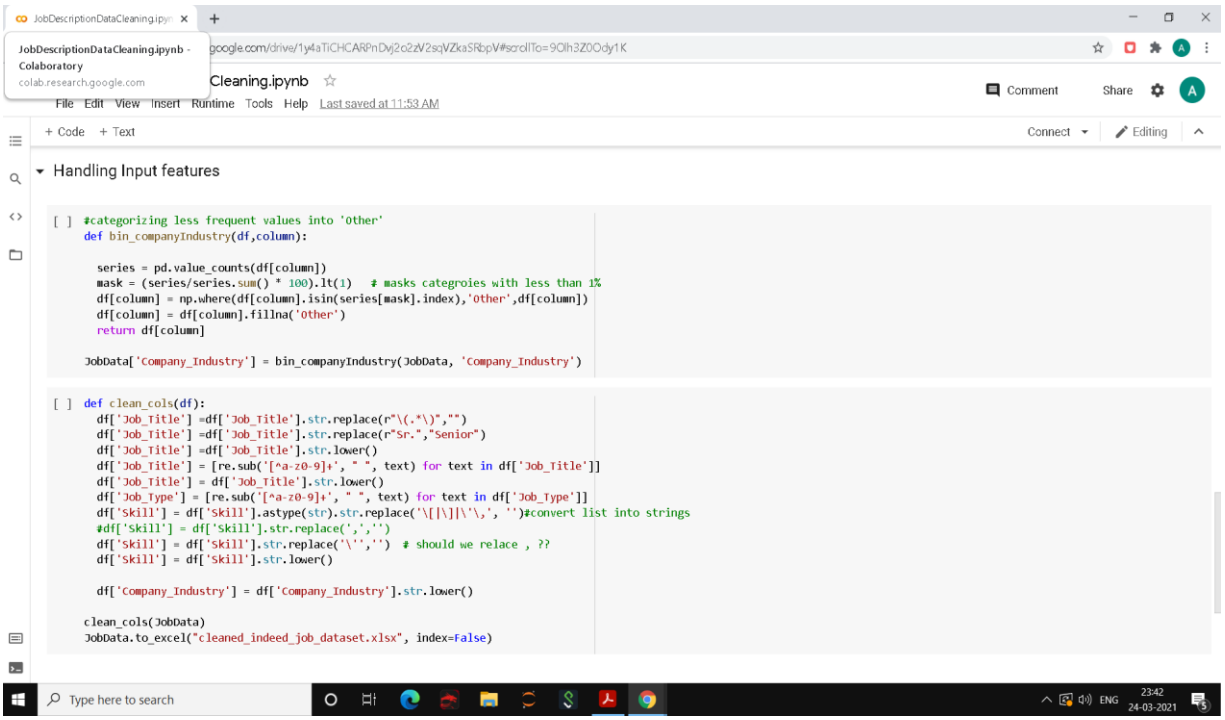
1. Removing Redundancy



2. Checking unique values and cleaning Job Description



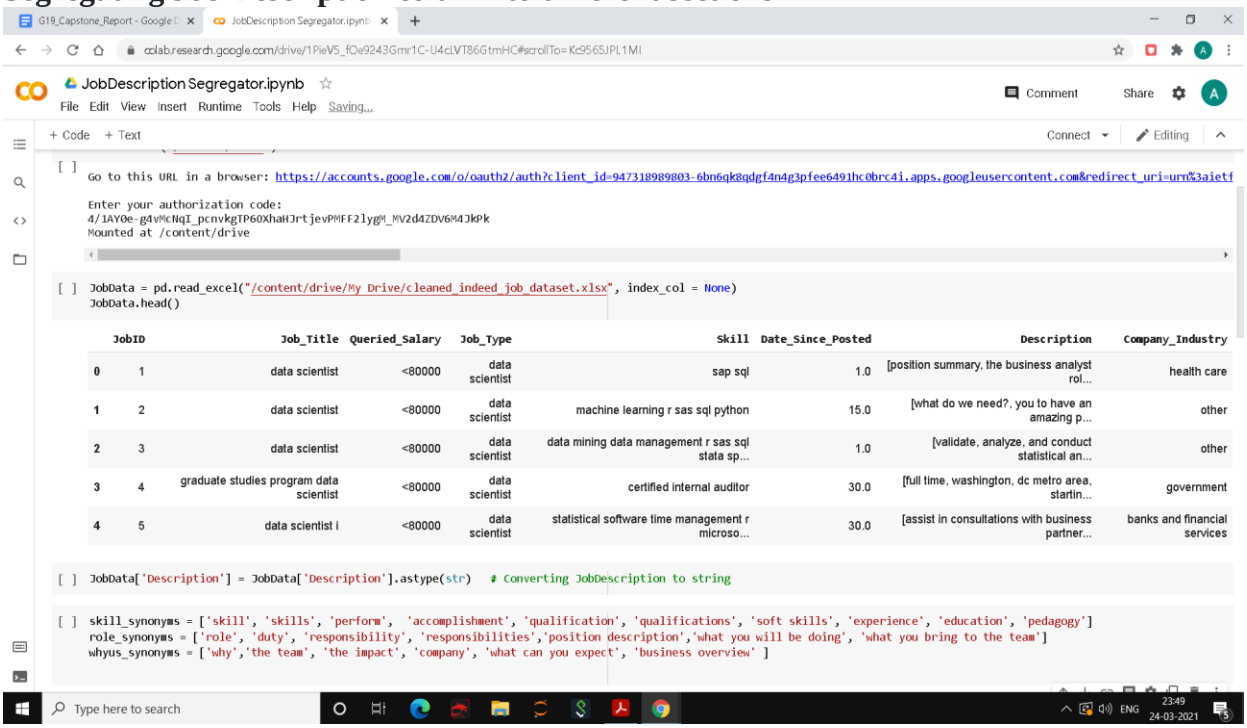
4. Cleaning Input columns and saving the clean dataset to an excel.



```
Handling Input features

[ ] #categorizing less frequent values into 'Other'
    def bin_companyIndustry(df,column):

      series = pd.value_counts(df[column])
      mask = (series/series.sum() * 100).lt(1)   # masks categroies with less than 1%
      df[column] = np.where(df[column].isin(series[mask].index),'Other',df[column])
      df[column] = df[column].fillna('Other')
      return df[column]

    JobData['Company_Industry'] = bin_companyIndustry(JobData, 'Company_Industry')

[ ] def clean_cols(df):
      df['Job_Title'] =df['Job_Title'].str.replace(r"\(.*\)","")
      df['Job_Title'] =df['Job_Title'].str.replace(r"Sr.","Senior")
      df['Job_Title'] =df['Job_Title'].str.lower()
      df['Job_Title'] = [re.sub('[^a-z0-9]+', " ", text) for text in df['Job_Title']]
      df['Job_Title'] = df['Job_Title'].str.lower()
      df['Job_Type'] = [re.sub('[^a-z0-9]+', " ", text) for text in df['Job_Type']]
      df['Skill'] = df['Skill'].astype(str).str.replace('\[|\]|\'\,', '')#convert list into strings
      #df['Skill'] = df['Skill'].str.replace(',','')
      df['Skill'] = df['Skill'].str.replace('\'','')  # should we relace , ??
      df['Skill'] = df['Skill'].str.lower()

      df['Company_Industry'] = df['Company_Industry'].str.lower()

    clean_cols(JobData)
    JobData.to_excel("cleaned_indeed_job_dataset.xlsx", index=False)
```

Feature Engineering Code Screenshots

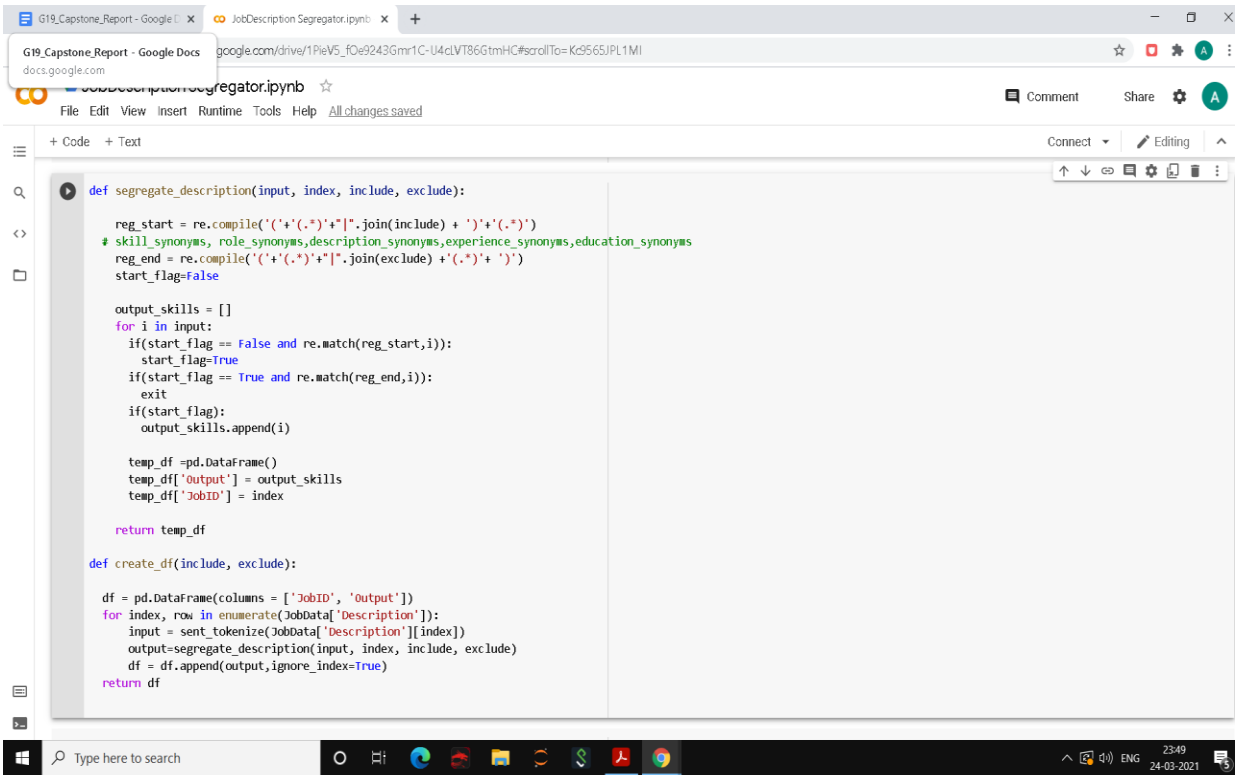
Segregating Job Description column to different sections



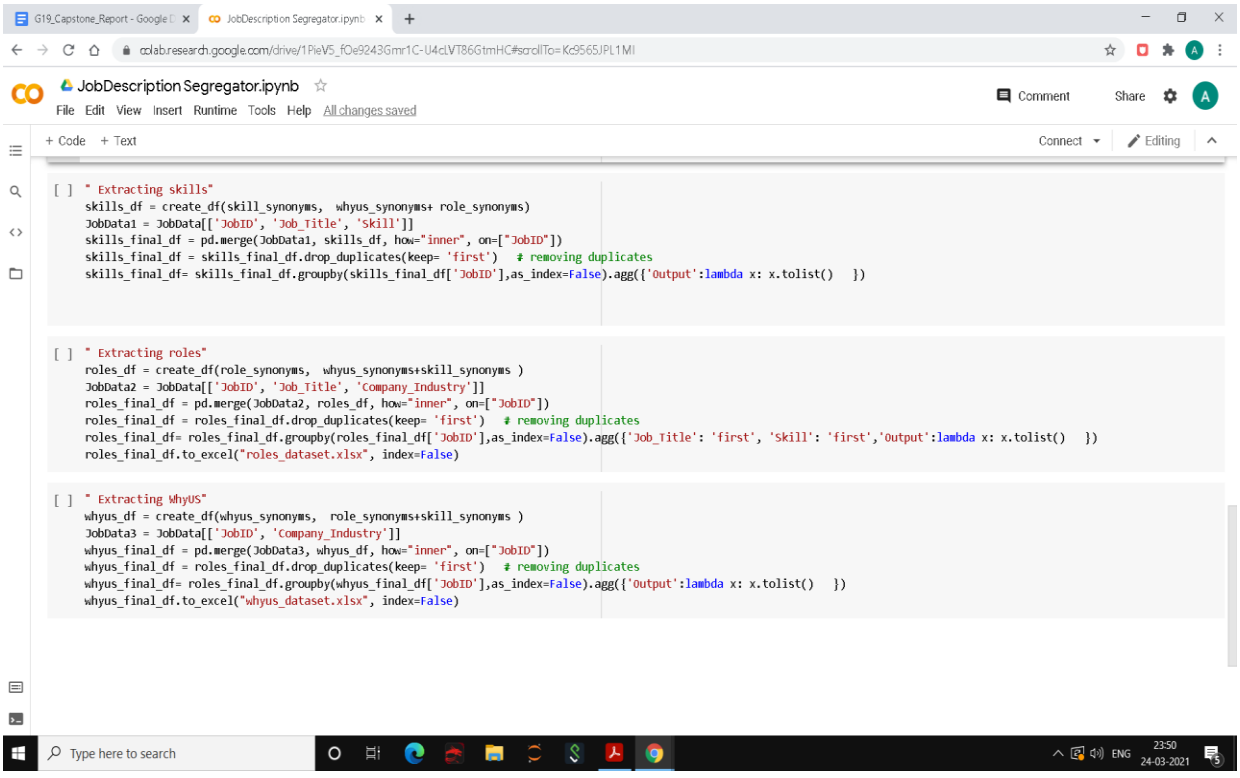
```
Go to this URL in a browser: https://accounts.google.com/o/oauth2/auth?client_id=947318989803-6bn6qk8qdgf4n4g3pfee6491hc0brc4i.apps.googleusercontent.com&redirect_uri=urn%3aietf

Enter your authorization code:
4/1AY0e-g4vMcNqI_pcnvkgTP6OXhaHJrtjevPMFF2lygM_MV2d4ZDV6M4JkPk
Mounted at /content/drive

[ ] JobData = pd.read_excel("/content/drive/My Drive/cleaned_indeed_job_dataset.xlsx", index_col = None)
    JobData.head()
```

	JobID	Job_Title	Queried_Salary	Job_Type	Skill	Date_Since_Posted	Description	Company_Industry
0	1	data scientist	<80000	data scientist	sap sql	1.0	[position summary, the business analyst rol...	health care
1	2	data scientist	<80000	data scientist	machine learning r sas sql python	15.0	[what do we need?, you to have an amazing p...	other
2	3	data scientist	<80000	data scientist	data mining data management r sas sql stata sp...	1.0	[validate, analyze, and conduct statistical an...	other
3	4	graduate studies program data scientist	<80000	data scientist	certified internal auditor	30.0	[full time, washington, dc metro area, startin...	government
4	5	data scientist i	<80000	data scientist	statistical software time management r microso...	30.0	[assist in consultations with business partner...	banks and financial services

```
[ ] JobData['Description'] = JobData['Description'].astype(str)   # Converting JobDescription to string

[ ] skill_synonyms = ['skill', 'skills', 'perform',  'accomplishment', 'qualification', 'qualifications', 'soft skills', 'experience', 'education', 'pedagogy']
    role_synonyms = ['role', 'duty', 'responsibility', 'responsibilities','position description','what you will be doing', 'what you bring to the team']
    whyus_synonyms = ['why','the team', 'the impact', 'company', 'what can you expect', 'business overview' ]
```



```
def segregate_description(input, index, include, exclude):

    reg_start = re.compile('('+'(.*)'+"|".join(include) + ')'+'(.*)')
  # skill_synonyms, role_synonyms,description_synonyms,experience_synonyms,education_synonyms
    reg_end = re.compile('('+'(.*)'+"|".join(exclude) +'(.*)'+ ')')
    start_flag=False

    output_skills = []
    for i in input:
      if(start_flag == False and re.match(reg_start,i)):
        start_flag=True
      if(start_flag == True and re.match(reg_end,i)):
        exit
      if(start_flag):
        output_skills.append(i)

      temp_df =pd.DataFrame()
      temp_df['Output'] = output_skills
      temp_df['JobID'] = index

    return temp_df

def create_df(include, exclude):

  df = pd.DataFrame(columns = ['JobID', 'Output'])
  for index, row in enumerate(JobData['Description']):
      input = sent_tokenize(JobData['Description'][index])
      output=segregate_description(input, index, include, exclude)
      df = df.append(output,ignore_index=True)
  return df
```



```
[ ] " Extracting skills"
    skills_df = create_df(skill_synonyms,  whyus_synonyms+ role_synonyms)
    JobData1 = JobData[['JobID', 'Job_Title', 'Skill']]
    skills_final_df = pd.merge(JobData1, skills_df, how="inner", on=["JobID"])
    skills_final_df = skills_final_df.drop_duplicates(keep= 'first')  # removing duplicates
    skills_final_df= skills_final_df.groupby(skills_final_df['JobID'],as_index=False).agg({'Output':lambda x: x.tolist()  })
```

```
[ ] " Extracting roles"
    roles_df = create_df(role_synonyms,  whyus_synonyms+skill_synonyms )
    JobData2 = JobData[['JobID', 'Job_Title', 'Company_Industry']]
    roles_final_df = pd.merge(JobData2, roles_df, how="inner", on=["JobID"])
    roles_final_df = roles_final_df.drop_duplicates(keep= 'first')  # removing duplicates
    roles_final_df= roles_final_df.groupby(roles_final_df['JobID'],as_index=False).agg({'Job_Title': 'first', 'Skill': 'first','Output':lambda x: x.tolist()  })
    roles_final_df.to_excel("roles_dataset.xlsx", index=False)
```

```
[ ] " Extracting WhyUS"
    whyus_df = create_df(whyus_synonyms,  role_synonyms+skill_synonyms )
    JobData3 = JobData[['JobID', 'Company_Industry']]
    whyus_final_df = pd.merge(JobData3, whyus_df, how="inner", on=["JobID"])
    whyus_final_df = roles_final_df.drop_duplicates(keep= 'first')  # removing duplicates
    whyus_final_df= roles_final_df.groupby(whyus_final_df['JobID'],as_index=False).agg({'Output':lambda x: x.tolist()  })
    whyus_final_df.to_excel("whyus_dataset.xlsx", index=False)
```

Transformer

```
[ ]
df = pd.read_excel('/content/drive/My Drive/JobDescriptionProject/skill_dataset.xlsx')
df[['JobID','Job_Title','Skill','SkillOutput']].head(n=2)
```

```
[ ] TRAIN_SIZE      = 0.8
def split_data(df, S=TRAIN_SIZE):
    print(TRAIN_SIZE)

    # Split into training and validation sets
    train_size = int(S * len(df))


    train_data = df[:train_size]
    val_data = df[train_size:]

    return train_data, val_data
```

```
[ ] train_data, test_data = split_data(df, TRAIN_SIZE)

    f'There are {len(train_data) :,} samples for training, and {len(val_data) :,} samples for validation testing'
```

```
0.8
'There are 3,337 samples for training, and 835 samples for validation testing'
```

```
train_data.to_csv('/content/drive/My Drive/skill_dataset_v1.txt', header=True, index=False, sep='\t', mode='w')
test_data.to_csv('/content/drive/My Drive/skill_dataset_v1.txt', header=True, index=False, sep='\t', mode='w')
```

Training the GPT-2

```
[ ] gpt2.generate(sess,
                  prefix="Associate data scientist AI Machine Learning Network R SAS C/C++ Java SPSS Data SciencePython",
                  length=150,
                  temperature=0.7,
                  nsamples=2
                  )
```

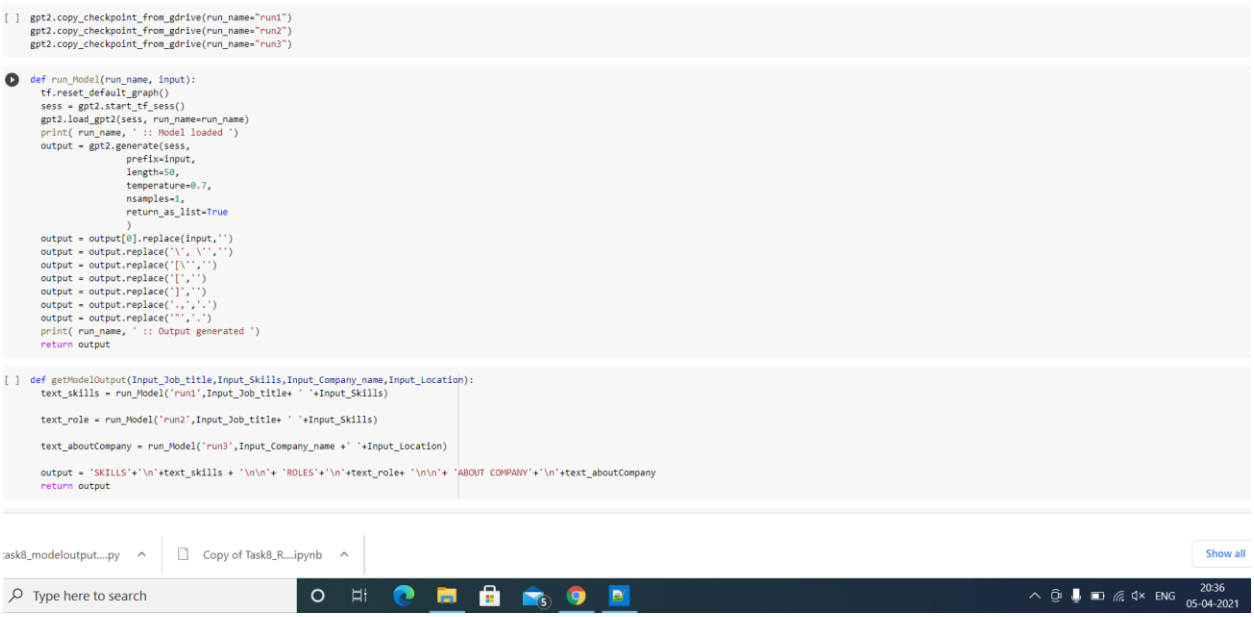
```
Associate data scientist AI Machine Learning Network R SAS C/C++ Java SPSS Data SciencePython   ['[   design, develop, and launch efficient and s
====================
Associate data scientist AI Machine Learning Network R SAS C/C++ Java SPSS Data SciencePython   ['[   apply mathematical methods and techniques t
====================
```

```
[ ] gpt2.generate(sess,
                  prefix="data scientist TensorFlow Machine Learning Hadoop Scala Kafka HBase Big Data Java Software Development Python Elasticsearc
                  length=150,
                  temperature=0.7,
                  nsamples=2
                  )
```

```
data scientist TensorFlow Machine Learning Hadoop Scala Kafka HBase Big Data Java Software Development Python Elasticsearch data science        [
====================
data scientist TensorFlow Machine Learning Hadoop Scala Kafka HBase Big Data Java Software Development Python Elasticsearch     ['[   excellent w
====================
```

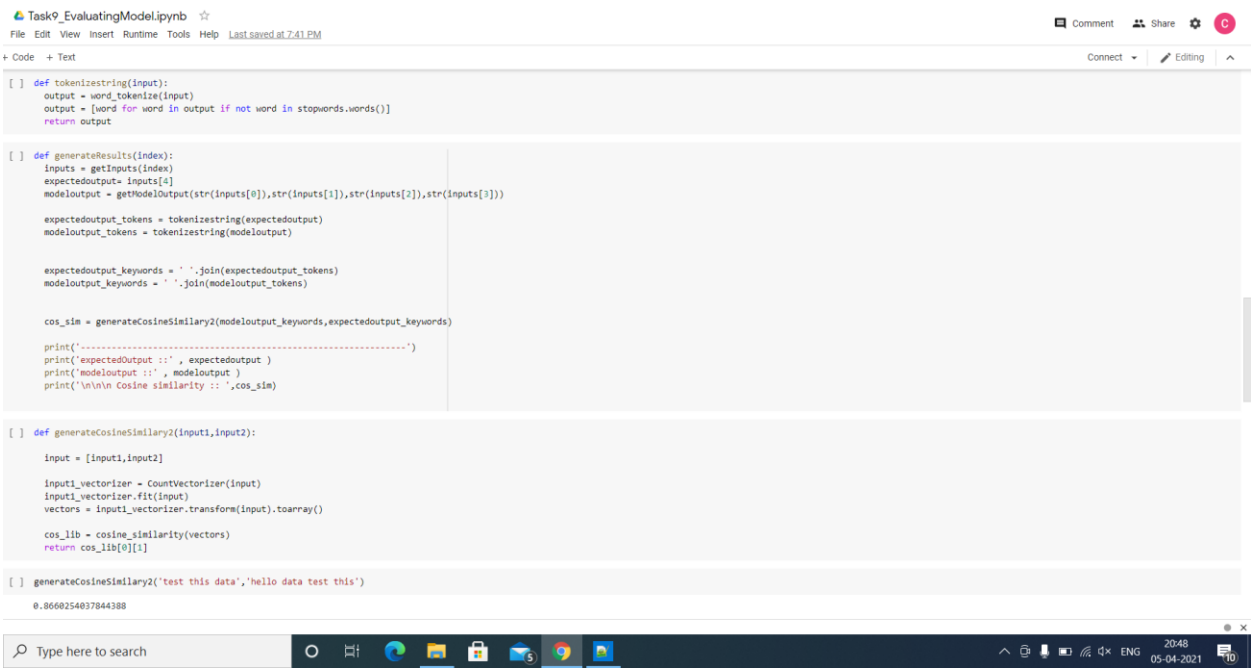
```
[ ]
```

Combining Model Output

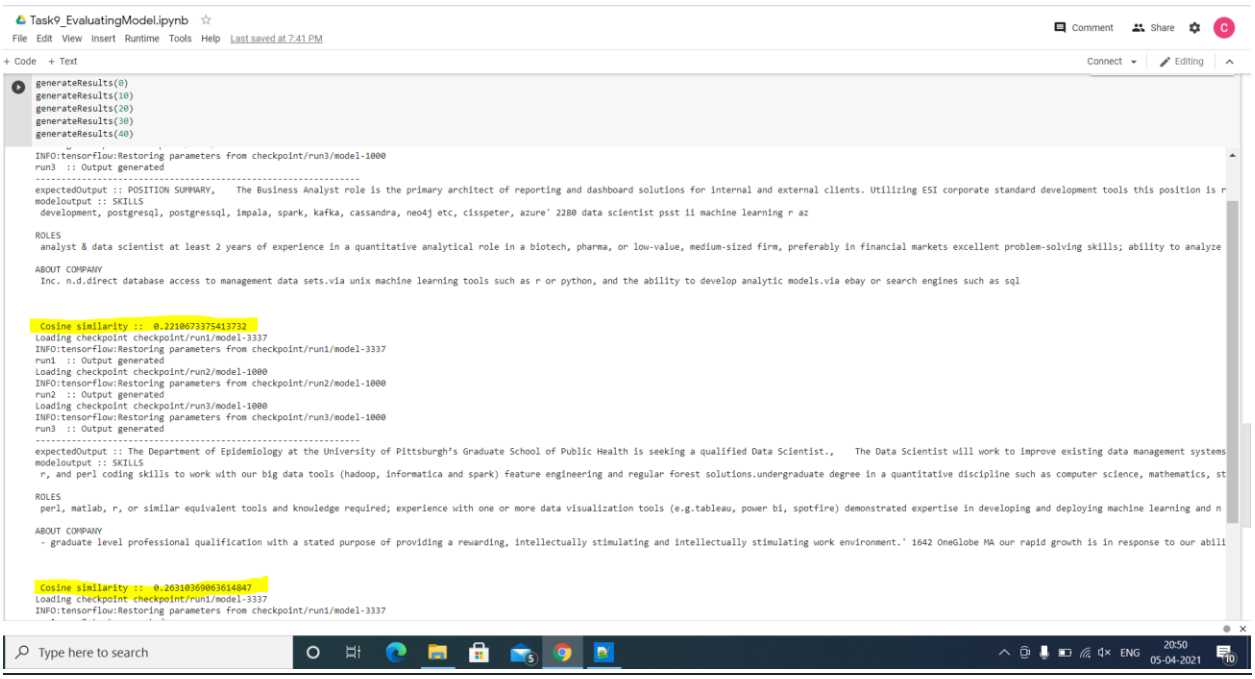


```
[ ] gpt2.copy_checkpoint_from_gdrive(run_name="run1")
    gpt2.copy_checkpoint_from_gdrive(run_name="run2")
    gpt2.copy_checkpoint_from_gdrive(run_name="run3")

def run_Model(run_name, input):
    tf.reset_default_graph()
    sess = gpt2.start_tf_sess()
    gpt2.load_gpt2(sess, run_name=run_name)
    print( run_name, ' :: Model loaded ')
    output = gpt2.generate(sess,
                   prefix=input,
                   length=50,
                   temperature=0.7,
                   nsamples=1,
                   return_as_list=True
                   )
    output = output[0].replace(input,'')
    output = output.replace('\', \'','')
    output = output.replace('[\'','')
    output = output.replace('[','')
    output = output.replace(']','')
    output = output.replace('.,','.')
    output = output.replace('"','.')
    print( run_name, ' :: Output generated ')
    return output

[ ] def getModelOutput(Input_Job_title,Input_Skills,Input_Company_name,Input_Location):
      text_skills = run_Model('run1',Input_Job_title+ ' '+Input_Skills)

      text_role = run_Model('run2',Input_Job_title+ ' '+Input_Skills)

      text_aboutCompany = run_Model('run3',Input_Company_name +' '+Input_Location)

      output = 'SKILLS'+'\n'+text_skills + '\n\n'+ 'ROLES'+'\n'+text_role+ '\n\n'+ 'ABOUT COMPANY'+'\n'+text_aboutCompany
      return output
```


Cosine Similarity for Predicted Output and Expected Output




```
[ ] def tokenizestring(input):
      output = word_tokenize(input)
      output = [word for word in output if not word in stopwords.words()]
      return output

[ ] def generateResults(index):
      inputs = getInputs(index)
      expectedoutput= inputs[4]
      modeloutput = getModelOutput(str(inputs[0]),str(inputs[1]),str(inputs[2]),str(inputs[3]))

      expectedoutput_tokens = tokenizestring(expectedoutput)
      modeloutput_tokens = tokenizestring(modeloutput)


      expectedoutput_keywords = ' '.join(expectedoutput_tokens)
      modeloutput_keywords = ' '.join(modeloutput_tokens)


      cos_sim = generateCosineSimilary2(modeloutput_keywords,expectedoutput_keywords)

      print('-----------------------------------------------------------------')
      print('expectedOutput ::' , expectedoutput )
      print('modeloutput ::' , modeloutput )
      print('\n\n\n Cosine similarity :: ',cos_sim)

[ ] def generateCosineSimilary2(input1,input2):

      input = [input1,input2]

      input1_vectorizer = CountVectorizer(input)
      input1_vectorizer.fit(input)
      vectors = input1_vectorizer.transform(input).toarray()

      cos_lib = cosine_similarity(vectors)
      return cos_lib[0][1]

[ ] generateCosineSimilary2('test this data','hello data test this')

    0.8660254037844388
```


Task9_EvaluatingModel.ipynb
File Edit View Insert Runtime Tools Help Last saved at 7:41 PM
Comment Share
+ Code + Text
Connect Editing
generateResults(0)
generateResults(10)
generateResults(20)
generateResults(30)
generateResults(40)
INFO:tensorflow:Restoring parameters from checkpoint/run3/model-1000
run3 :: Output generated
expectedOutput :: POSITION SUMMARY, The Business Analyst role is the primary architect of reporting and dashboard solutions for internal and external clients. Utilizing ESI corporate standard development tools this position is r
modeloutput :: SKILLS
development, postgresql, postgressql, impala, spark, kafka, cassandra, neo4j etc, cisspeter, azure' 2280 data scientist psst ii machine learning r az
ROLES
analyst & data scientist at least 2 years of experience in a quantitative analytical role in a biotech, pharma, or low-value, medium-sized firm, preferably in financial markets excellent problem-solving skills; ability to analyze
ABOUT COMPANY
Inc. n.d.direct database access to management data sets.via unix machine learning tools such as r or python, and the ability to develop analytic models.via ebay or search engines such as sql
Cosine similarity :: 0.2210673375413732
Loading checkpoint checkpoint/run1/model-3337
INFO:tensorflow:Restoring parameters from checkpoint/run1/model-3337
run1 :: Output generated
Loading checkpoint checkpoint/run2/model-1000
INFO:tensorflow:Restoring parameters from checkpoint/run2/model-1000
run2 :: Output generated
Loading checkpoint checkpoint/run3/model-1000
INFO:tensorflow:Restoring parameters from checkpoint/run3/model-1000
run3 :: Output generated
expectedOutput :: The Department of Epidemiology at the University of Pittsburgh's Graduate School of Public Health is seeking a qualified Data Scientist., The Data Scientist will work to improve existing data management systems
modeloutput :: SKILLS
r, and perl coding skills to work with our big data tools (hadoop, informatica and spark) feature engineering and regular forest solutions.undergraduate degree in a quantitative discipline such as computer science, mathematics, st
ROLES
perl, matlab, r, or similar equivalent tools and knowledge required; experience with one or more data visualization tools (e.g.tableau, power bi, spotfire) demonstrated expertise in developing and deploying machine learning and n
ABOUT COMPANY
- graduate level professional qualification with a stated purpose of providing a rewarding, intellectually stimulating and intellectually stimulating work environment.' 1642 OneGlobe MA our rapid growth is in response to our abili
Cosine similarity :: 0.26310369063614847
Loading checkpoint checkpoint/run1/model-3337
INFO:tensorflow:Restoring parameters from checkpoint/run1/model-3337
Type here to search
ENG
20:50
05-04-2021

Interpretation

- **Metrics used**

  Cosine Similarity Matrix:

Cosine similarity is a metric used to determine how similar two entities are irrespective of their size. Mathematically, it measures the cosine of the angle between two vectors. Mathematically, if ‘a’ and ‘b’ are two vectors, the cosine equation gives the angle between the two. To compute the cosine similarity, you need the word count of the words in each document. We can either use either CountVectorizer or the TfidfVectorizer from scikit-learn for word count.

- **Project output in terms of cosine metrics**

Inputs : We passed some inputs from the cleaned dataset [input 1] and compared model output to the description already present for that row [input 2].Steps - We removed stop words from both the inputs and used count vectorizer for word count. Used cosine_similarity from sklearn for calculating the similarity.

Output - On an avg, the similarity was 0.25

Scope of improvements

Considering the time constraints of this project, a couple of models have been tried and few hyper parameters were explored. However, there is scope to improve further by exploring more models and as well as getting better data. At present, we tried with several clustering algorithms, such as K-means clustering and DBSCAN Clustering. K means clustering was working good in understanding the contexts, but it was adding a lot of noise as well in the result. Clustering and Topic Modelling can be explored further to check if they improve our model.

On similar lines several new investigations may be carried out in future and some of the idea's worth exploring are:

1. Extracting various sections of Job Description with POS tagging.
2. For creating a better synonyms list, machine learning algorithms, like Topic Modelling, Clustering can be tried which will help in generating better data to feed to the model.
3. If these do not work, then other text generation models can be checked to see if they provide better results.

Bibliography / References

1. https://science.jrank.org/computer-science/Natural_Language_Processing.html
2. https://medium.com/acing-ai/what-is-cosine-similarity-matrix-f0819e674ad1
3. Data Mining Medical Records With Machine Learning - 5 Current Applications
4. https://towardsdatascience.com/how-to-use-nlp-in-python-a-practical-step-by-step-example-bd82ca2d2e1e
5. https://machinelearningmastery.com/text-generation-lstm-recurrent-neural-networks-python-keras/
6. https://towardsdatascience.com/how-our-device-thinks-e1f5ab15071e
7. https://towardsdatascience.com/how-our-device-thinks-e1f5ab15071e

Duly Completed Checklist for Final Report

a)	Is the Cover page in proper format?	Y
b)	Is the Title page in proper format?	Y
c)	Is the Certificate from the Mentor in proper format and signed?	Y
d)	Is Abstract included in the Report? Is it properly written?	Y
e)	Does the Table of Contents page include chapter page numbers?	Y
f)	Does the Report contain a summary of the literature survey?	Y
	i. Are the Pages numbered properly?	Y
	ii. Are the Figures numbered properly?	Y
	iii. Are the Tables numbered properly?	Y
	iv. Are the Captions for the Figures and Tables proper?	Y
	v. Are the Appendices numbered?	Y
g)	Does the Report have Conclusion / Recommendations of the work?	Y
h)	Are References/Bibliography given in the Report?	Y
i)	Have the References been cited in the Report?	Y
j)	Is the citation of References / Bibliography in proper format?	Y