# RESUME PARSER

## ( END CAPSTONE PROJECT )

SUBMITTED BY: T.SASHIDHAR

COHORT: 20

INTEGRATED DATA SCIENCE

# TABLE OF CONTENTS

# 1. INTRODUCTION

## 1.1 Project Objective

The objective of this project is to develop an automated system for extracting, cleaning, storing, and analyzing resume data. The system leverages Python for data extraction and statistical analysis, SQL for storage and transaction management, Excel VBA for data cleaning, and Power BI for visualizations. The goal is to enhance the resume management and provide data-driveninsights that aid recruitment decision-making.

## 1.2 Project Deliverables

The deliverables for this project include:

- A Python script for extracting resume data from PDF and DOCX files.
- An Excel VBA macro for cleaning and standardizing the extracted resume data.
- SQL transactions, triggers, and stored procedures for data management.
- A Python class for statistical analysis of resume data.
- An XG Boost machine learning model to predict the primary skill from a resume.
- A Power BI dashboard that visualizes key resume metrics.

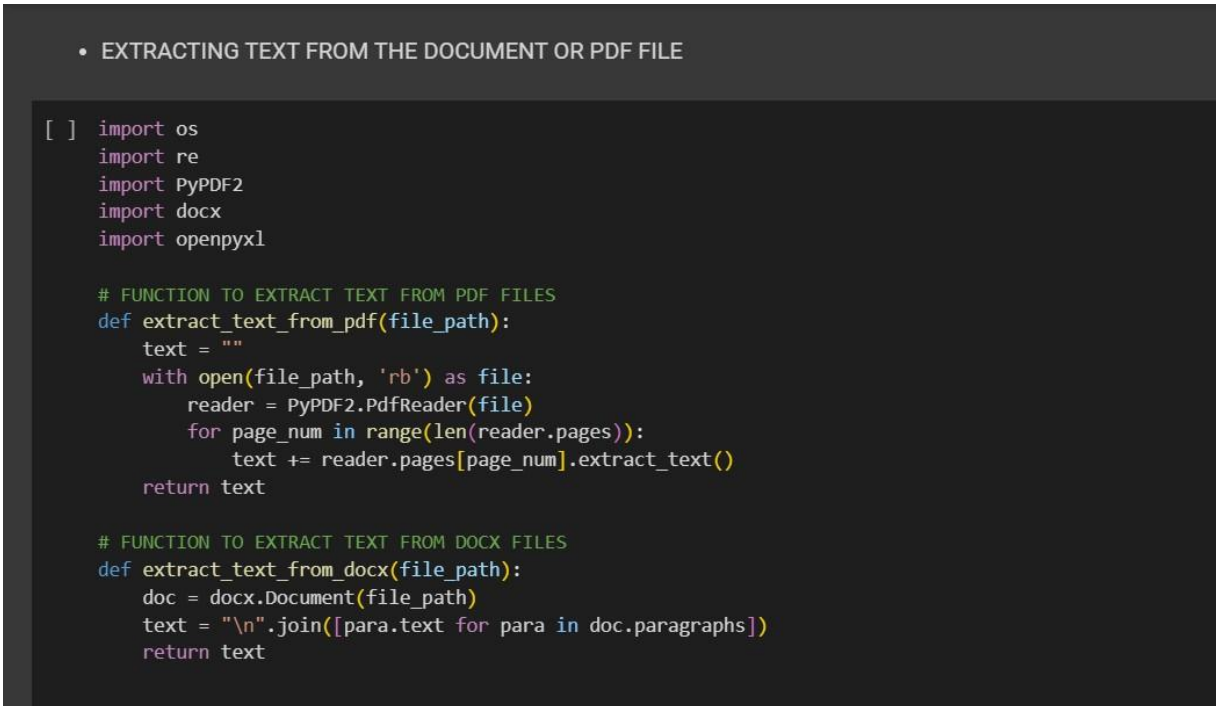
## 2.2 Tools and Technologies

- **Python Libraries**: PyPDF2, python-docx, Seaborn, NumPy, XGBoost, pandas, etc.
- **SQL**: SQL SMS for storage, transactions, triggers, and stored procedures.
- **Excel**: VBA for cleaning, Power Query for standardization.
- **Power BI**: Visualization and DAX calculations.

# 3. Data Extraction

## 3.1 Python Script for Extracting Resume Data

A Python script is written using PyPDF2 and python-docx to extract text from resumes. The extracted data is parsed into structured fields such as Name, Skills, Experience, and Education.

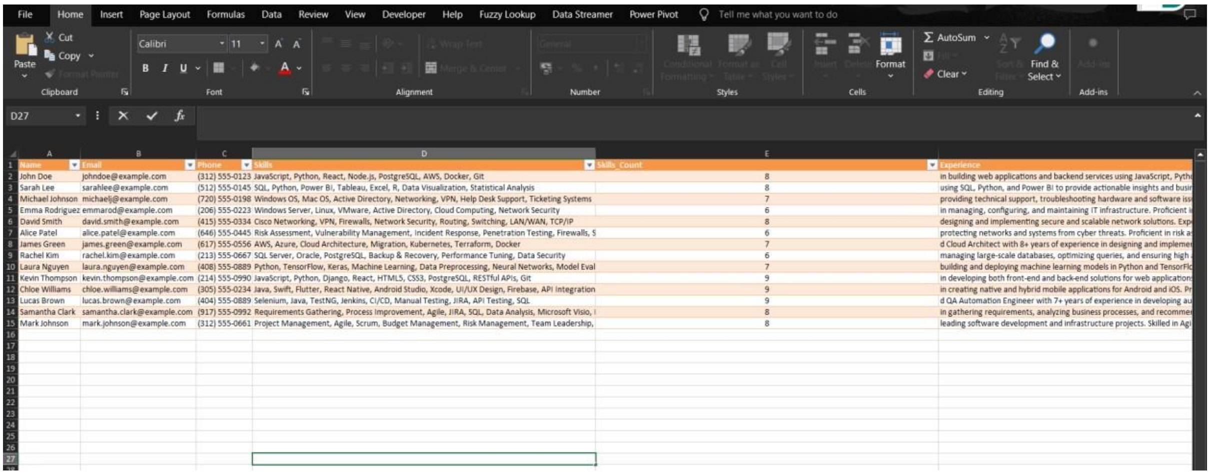


- EXTRACTING TEXT FROM THE DOCUMENT OR PDF FILE

```
[ ]  import os
     import re
     import PyPDF2
     import docx
     import openpyxl

     # FUNCTION TO EXTRACT TEXT FROM PDF FILES
     def extract_text_from_pdf(file_path):
         text = ""
         with open(file_path, 'rb') as file:
             reader = PyPDF2.PdfReader(file)
             for page_num in range(len(reader.pages)):
                 text += reader.pages[page_num].extract_text()
         return text

     # FUNCTION TO EXTRACT TEXT FROM DOCX FILES
     def extract_text_from_docx(file_path):
         doc = docx.Document(file_path)
         text = "\n".join([para.text for para in doc.paragraphs])
         return text
```

## 3.2 Mapping Extracted Data to Excel

The extracted data is then mapped to predefined columns (e.g., Name, Skills, Experience, Education) and stored in an Excel sheet.



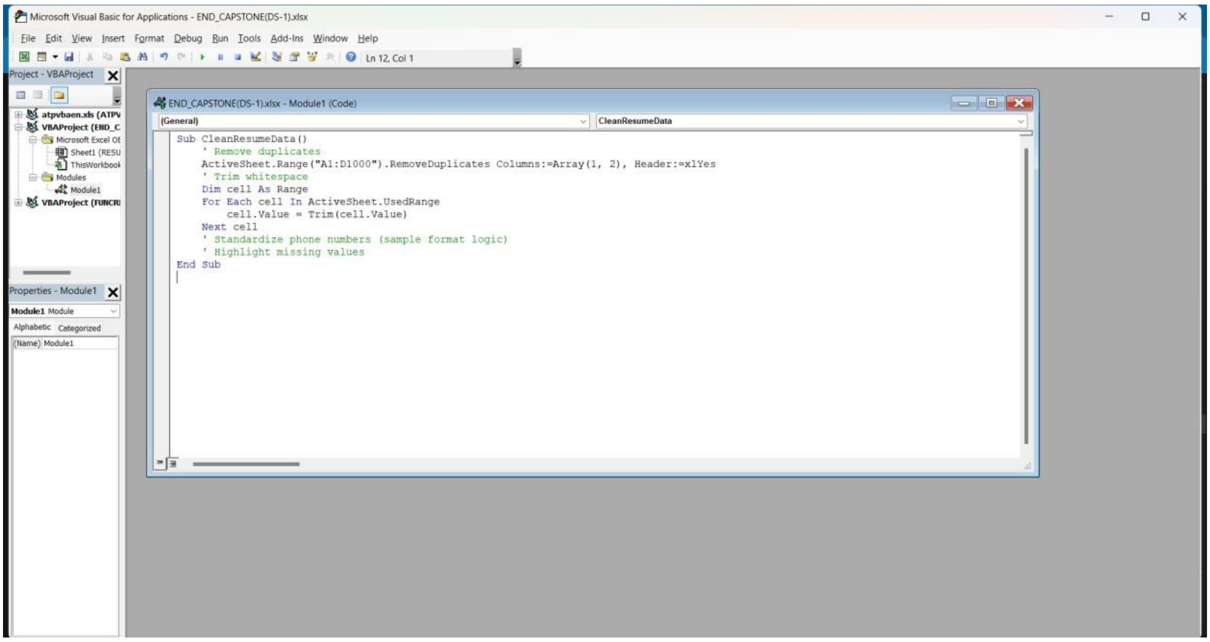
| Name | Email | Phone | Skills | Skills_Count | Experience |
|---|---|---|---|---|---|
| John Doe | johndoe@example.com | (312) 555-0123 | JavaScript, Python, React, Node.js, PostgreSQL, AWS, Docker, Git | 8 | in building web applications and backend services using JavaScript, Pytho |
| Sarah Lee | sarahlee@example.com | (512) 555-0145 | SQL, Python, Power BI, Tableau, Excel, R, Data Visualization, Statistical Analysis | 8 | using SQL, Python, and Power BI to provide actionable insights and busin |
| Michael Johnson | michaelj@example.com | (720) 555-0198 | Windows OS, Mac OS, Active Directory, Networking, VPN, Help Desk Support, Ticketing Systems | 7 | providing technical support, troubleshooting hardware and software iss |
| Emma Rodriguez | emmarod@example.com | (206) 555-0223 | Windows Server, Linux, VMware, Active Directory, Cloud Computing, Network Security | 6 | in managing, configuring, and maintaining IT infrastructure. Proficient i |
| David Smith | david.smith@example.com | (415) 555-0334 | Cisco Networking, VPN, Firewalls, Network Security, Routing, Switching, LAN/WAN, TCP/IP | 8 | designing and implementing secure and scalable network solutions. Exp |
| Alice Patel | alice.patel@example.com | (646) 555-0445 | Risk Assessment, Vulnerability Management, Incident Response, Penetration Testing, Firewalls, S | 6 | protecting networks and systems from cyber threats. Proficient in risk a |
| James Green | james.green@example.com | (617) 555-0556 | AWS, Azure, Cloud Architecture, Migration, Kubernetes, Terraform, Docker | 7 | d Cloud Architect with 8+ years of experience in designing and impleme |
| Rachel Kim | rachel.kim@example.com | (213) 555-0667 | SQL Server, Oracle, PostgreSQL, Backup & Recovery, Performance Tuning, Data Security | 6 | managing large-scale databases, optimizing queries, and ensuring high |
| Laura Nguyen | laura.nguyen@example.com | (408) 555-0889 | Python, TensorFlow, Keras, Machine Learning, Data Preprocessing, Neural Networks, Model Eval | 7 | building and deploying machine learning models in Python and TensorFlo |
| Kevin Thompson | kevin.thompson@example.com | (214) 555-0990 | JavaScript, Python, Django, React, HTML5, CSS3, PostgreSQL, RESTful APIs, Git | 9 | in developing both front-end and back-end solutions for web application |
| Chloe Williams | chloe.williams@example.com | (305) 555-0234 | Java, Swift, Flutter, React Native, Android Studio, Xcode, UI/UX Design, Firebase, API Integration | 9 | in creating native and hybrid mobile applications for Android and iOS. Pr |
| Lucas Brown | lucas.brown@example.com | (404) 555-0889 | Selenium, Java, TestNG, Jenkins, CI/CD, Manual Testing, JIRA, API Testing, SQL | 9 | d QA Automation Engineer with 7+ years of experience in developing au |
| Samantha Clark | samantha.clark@example.com | (917) 555-0992 | Requirements Gathering, Process Improvement, Agile, JIRA, SQL, Data Analysis, Microsoft Visio, | 8 | in gathering requirements, analyzing business processes, and recomme |
| Mark Johnson | mark.johnson@example.com | (312) 555-0661 | Project Management, Agile, Scrum, Budget Management, Risk Management, Team Leadership, | 8 | leading software development and infrastructure projects. Skilled in Agi |

# 4. Data Cleaning

## 4.1 Excel VBA Macros for Data Cleaning

A VBA macro automates the cleaning process by:

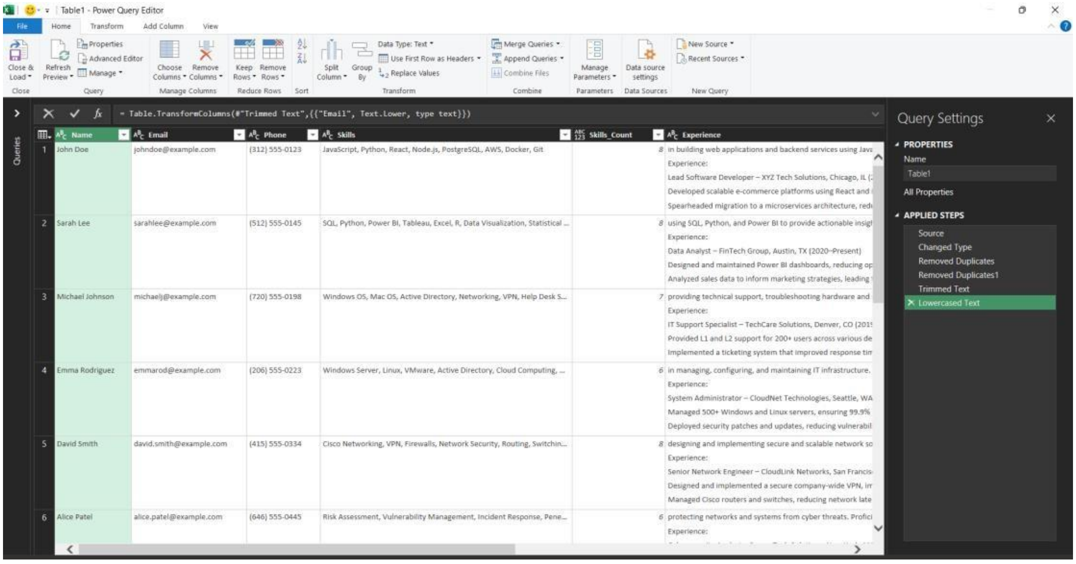
- Removing duplicate entries.
- Trimming leading and trailing whitespaces from text fields.
- Standardizing phone numbers and email formats.
- Highlighting missing data.



```vba
Sub CleanResumeData()
    ' Remove duplicates
    ActiveSheet.Range("A1:D1000").RemoveDuplicates Columns:=Array(1, 2), Header:=xlYes
    ' Trim whitespace
    Dim cell As Range
    For Each cell In ActiveSheet.UsedRange
        cell.Value = Trim(cell.Value)
    Next cell
    ' Standardize phone numbers (sample format logic)
    ' Highlight missing values
End Sub
```

## 4.2 Power Query for Data Standardization

Power Query is used to further standardize the resume data, ensuring consistent formats and filtering out incorrect entries.



```
= Table.TransformColumns(#"Trimmed Text",{{"Email", Text.Lower, type text}})
```

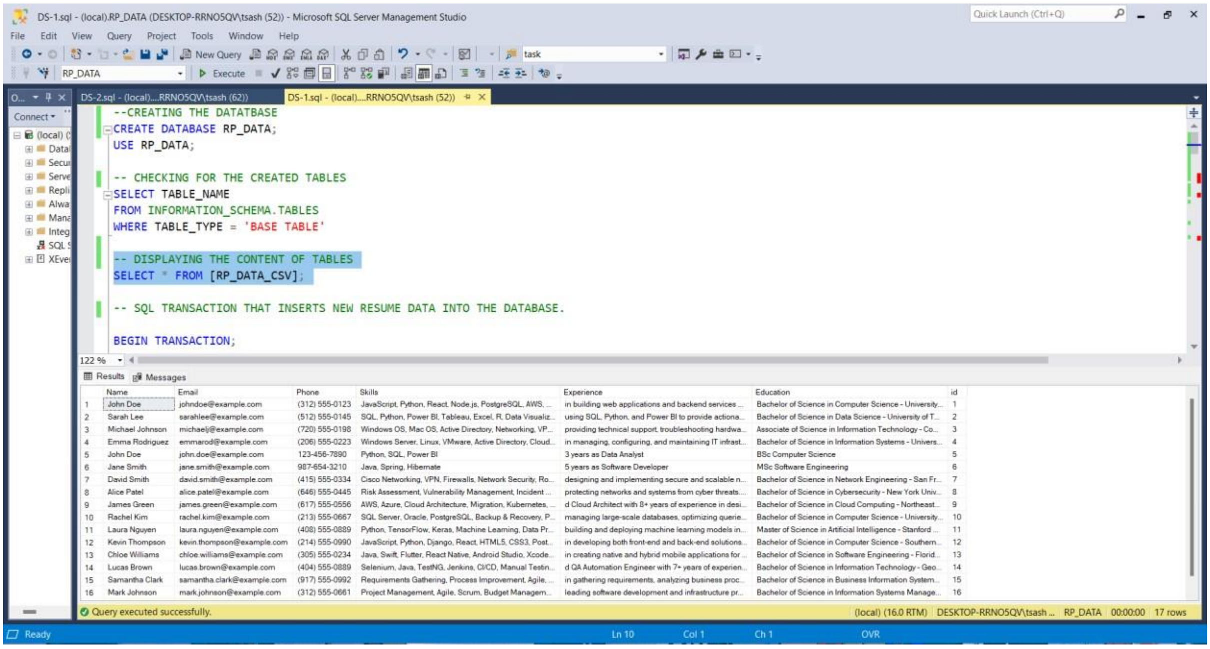
| | Name | Email | Phone | Skills | Skills_Count | Experience |
|---|---|---|---|---|---|---|
| 1 | John Doe | johndoe@example.com | (312) 555-0123 | JavaScript, Python, React, Node.js, PostgreSQL, AWS, Docker, Git | 8 | in building web applications and backend services using Java<br>Experience:<br>Lead Software Developer – XYZ Tech Solutions, Chicago, IL (<br>Developed scalable e-commerce platforms using React and<br>Spearheaded migration to a microservices architecture, red |
| 2 | Sarah Lee | sarahlee@example.com | (512) 555-0145 | SQL, Python, Power BI, Tableau, Excel, R, Data Visualization, Statistical ... | 8 | using SQL, Python, and Power BI to provide actionable insig<br>Experience:<br>Data Analyst – FinTech Group, Austin, TX (2020–Present)<br>Designed and maintained Power BI dashboards, reducing op<br>Analyzed sales data to inform marketing strategies, leading |
| 3 | Michael Johnson | michaelj@example.com | (720) 555-0198 | Windows OS, Mac OS, Active Directory, Networking, VPN, Help Desk S... | 7 | providing technical support, troubleshooting hardware and<br>Experience:<br>IT Support Specialist – TechCare Solutions, Denver, CO (201<br>Provided L1 and L2 support for 200+ users across various de<br>Implemented a ticketing system that improved response tin |
| 4 | Emma Rodriguez | emmarod@example.com | (206) 555-0223 | Windows Server, Linux, VMware, Active Directory, Cloud Computing, ... | 6 | in managing, configuring, and maintaining IT infrastructure.<br>Experience:<br>System Administrator – CloudNet Technologies, Seattle, WA<br>Managed 500+ Windows and Linux servers, ensuring 99.9%<br>Deployed security patches and updates, reducing vulnerabil |
| 5 | David Smith | david.smith@example.com | (415) 555-0334 | Cisco Networking, VPN, Firewalls, Network Security, Routing, Switchin... | 8 | designing and implementing secure and scalable network so<br>Experience:<br>Senior Network Engineer – CloudLink Networks, San Francis<br>Designed and implemented a secure company-wide VPN, im<br>Managed Cisco routers and switches, reducing network late |
| 6 | Alice Patel | alice.patel@example.com | (646) 555-0445 | Risk Assessment, Vulnerability Management, Incident Response, Pene... | 6 | protecting networks and systems from cyber threats. Profici<br>Experience: |

Name: Table1

Applied steps:
Source
Changed Type
Removed Duplicates
Removed Duplicates1
Trimmed Text
Lowercased Text

## **5.** Data Storage and Transaction Management

A SQL transaction is created to ensure data integrity. If an error occurs during insertion, the transaction is rolled back, ensuring that either all records are committed or none are.



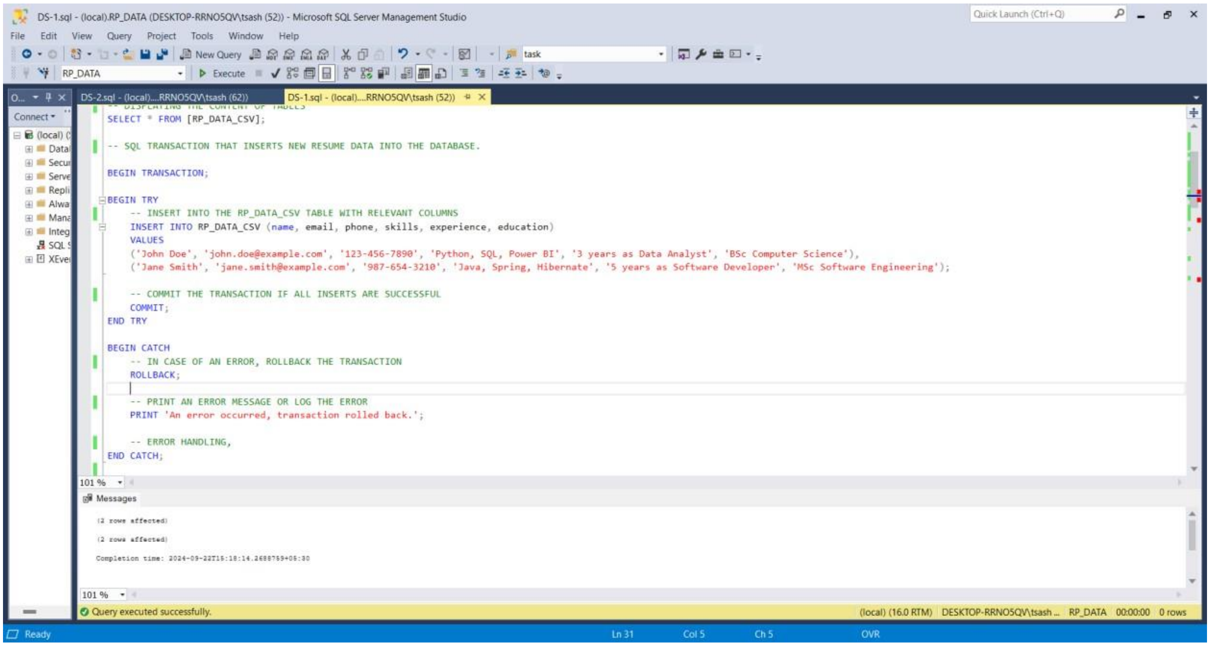
DS-1.sql - (local).RP_DATA (DESKTOP-RRNO5QV\tsash (52)) - Microsoft SQL Server Management Studio
File Edit View Query Project Tools Window Help
New Query
Execute
RP_DATA
task
DS-2.sql - (local)....RRNO5QV\tsash (62))
DS-1.sql - (local)....RRNO5QV\tsash (52))
122 %
Results
Messages
Ready
Ln 10 Col 1 Ch 1 OVR

```sql
--CREATING THE DATATBASE
CREATE DATABASE RP_DATA;
USE RP_DATA;

-- CHECKING FOR THE CREATED TABLES
SELECT TABLE_NAME
FROM INFORMATION_SCHEMA.TABLES
WHERE TABLE_TYPE = 'BASE TABLE'

-- DISPLAYING THE CONTENT OF TABLES
SELECT * FROM [RP_DATA_CSV];

-- SQL TRANSACTION THAT INSERTS NEW RESUME DATA INTO THE DATABASE.

BEGIN TRANSACTION;
```

| | Name | Email | Phone | Skills | Experience | Education | id |
|---|---|---|---|---|---|---|---|
| 1 | John Doe | johndoe@example.com | (312) 555-0123 | JavaScript, Python, React, Node.js, PostgreSQL, AWS, ... | in building web applications and backend services ... | Bachelor of Science in Computer Science - University... | 1 |
| 2 | Sarah Lee | sarahlee@example.com | (512) 555-0145 | SQL, Python, Power BI, Tableau, Excel, R, Data Visualiz... | using SQL, Python, and Power BI to provide actiona... | Bachelor of Science in Data Science - University of T... | 2 |
| 3 | Michael Johnson | michaelj@example.com | (720) 555-0198 | Windows OS, Mac OS, Active Directory, Networking, VP... | providing technical support, troubleshooting hardwa... | Associate of Science in Information Technology - Co... | 3 |
| 4 | Emma Rodriguez | emmarod@example.com | (206) 555-0223 | Windows Server, Linux, VMware, Active Directory, Cloud... | in managing, configuring, and maintaining IT infrast... | Bachelor of Science in Information Systems - Univers... | 4 |
| 5 | John Doe | john.doe@example.com | 123-456-7890 | Python, SQL, Power BI | 3 years as Data Analyst | BSc Computer Science | 5 |
| 6 | Jane Smith | jane.smith@example.com | 987-654-3210 | Java, Spring, Hibernate | 5 years as Software Developer | MSc Software Engineering | 6 |
| 7 | David Smith | david.smith@example.com | (415) 555-0334 | Cisco Networking, VPN, Firewalls, Network Security, Ro... | designing and implementing secure and scalable n... | Bachelor of Science in Network Engineering - San Fr... | 7 |
| 8 | Alice Patel | alice.patel@example.com | (646) 555-0445 | Risk Assessment, Vulnerability Management, Incident ... | protecting networks and systems from cyber threats.... | Bachelor of Science in Cybersecurity - New York Univ... | 8 |
| 9 | James Green | james.green@example.com | (617) 555-0556 | AWS, Azure, Cloud Architecture, Migration, Kubernetes, ... | d Cloud Architect with 8+ years of experience in desi... | Bachelor of Science in Cloud Computing - Northeast... | 9 |
| 10 | Rachel Kim | rachel.kim@example.com | (213) 555-0667 | SQL Server, Oracle, PostgreSQL, Backup & Recovery, P... | managing large-scale databases, optimizing querie... | Bachelor of Science in Computer Science - University... | 10 |
| 11 | Laura Nguyen | laura.nguyen@example.com | (408) 555-0889 | Python, TensorFlow, Keras, Machine Learning, Data Pr... | building and deploying machine learning models in... | Master of Science in Artificial Intelligence - Stanford ... | 11 |
| 12 | Kevin Thompson | kevin.thompson@example.com | (214) 555-0990 | JavaScript, Python, Django, React, HTML5, CSS3, Post... | in developing both front-end and back-end solutions... | Bachelor of Science in Computer Science - Southern... | 12 |
| 13 | Chloe Williams | chloe.williams@example.com | (305) 555-0234 | Java, Swift, Flutter, React Native, Android Studio, Xcode... | in creating native and hybrid mobile applications for ... | Bachelor of Science in Software Engineering - Florid... | 13 |
| 14 | Lucas Brown | lucas.brown@example.com | (404) 555-0889 | Selenium, Java, TestNG, Jenkins, CI/CD, Manual Testin... | d QA Automation Engineer with 7+ years of experien... | Bachelor of Science in Information Technology - Geo... | 14 |
| 15 | Samantha Clark | samantha.clark@example.com | (917) 555-0992 | Requirements Gathering, Process Improvement, Agile, ... | in gathering requirements, analyzing business proc... | Bachelor of Science in Business Information System... | 15 |
| 16 | Mark Johnson | mark.johnson@example.com | (312) 555-0661 | Project Management, Agile, Scrum, Budget Managem... | leading software development and infrastructure pr... | Bachelor of Science in Information Systems Manage... | 16 |

```
Query executed successfully.   (local) (16.0 RTM)   DESKTOP-RRNO5QV\tsash ...   RP_DATA   00:00:00   17 rows
```

CREATE A SQL TRIGGER THAT AUTOMATICALLY LOGS THE INSERTION OF NEW RESUME DATA INTO A SEPARATE LOG TABLE



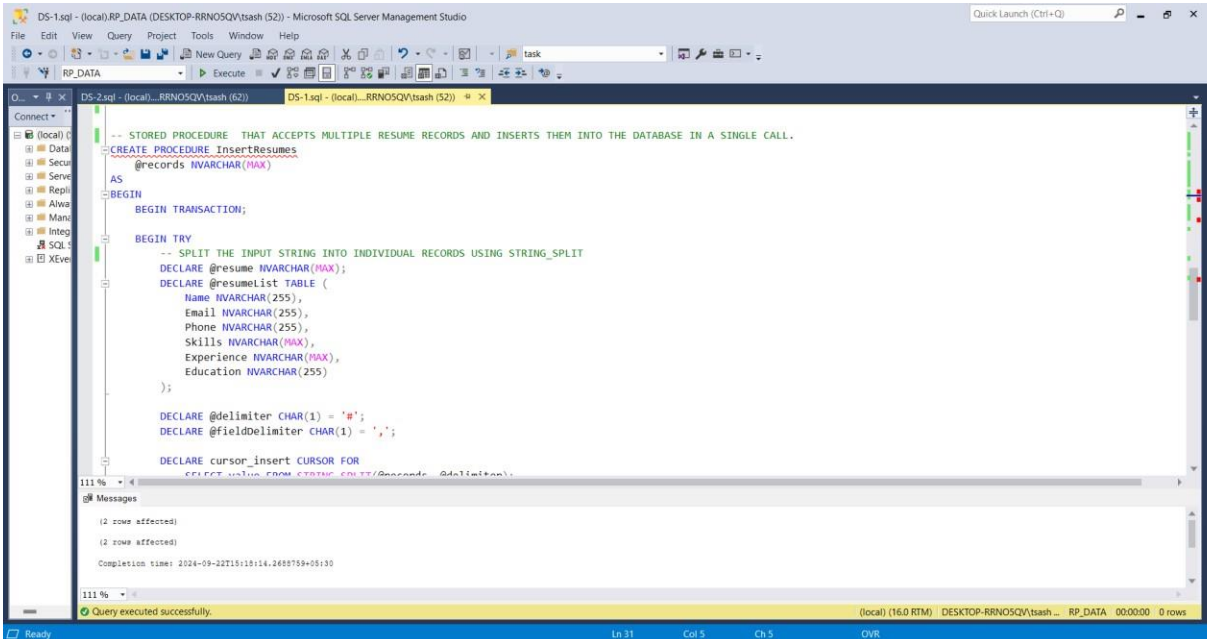
DS-1.sql - (local).RP_DATA (DESKTOP-RRNO5QV\tsash (52)) - Microsoft SQL Server Management Studio
File Edit View Query Project Tools Window Help
New Query
Execute
RP_DATA
task
DS-2.sql - (local)....RRNO5QV\tsash (62))
DS-1.sql - (local)....RRNO5QV\tsash (52))
101 %
Messages
Ready
Ln 31 Col 5 Ch 5 OVR

```sql
SELECT * FROM [RP_DATA_CSV];

-- SQL TRANSACTION THAT INSERTS NEW RESUME DATA INTO THE DATABASE.

BEGIN TRANSACTION;

BEGIN TRY
    -- INSERT INTO THE RP_DATA_CSV TABLE WITH RELEVANT COLUMNS
    INSERT INTO RP_DATA_CSV (name, email, phone, skills, experience, education)
    VALUES
    ('John Doe', 'john.doe@example.com', '123-456-7890', 'Python, SQL, Power BI', '3 years as Data Analyst', 'BSc Computer Science'),
    ('Jane Smith', 'jane.smith@example.com', '987-654-3210', 'Java, Spring, Hibernate', '5 years as Software Developer', 'MSc Software Engineering');

    -- COMMIT THE TRANSACTION IF ALL INSERTS ARE SUCCESSFUL
    COMMIT;
END TRY

BEGIN CATCH
    -- IN CASE OF AN ERROR, ROLLBACK THE TRANSACTION
    ROLLBACK;

    -- PRINT AN ERROR MESSAGE OR LOG THE ERROR
    PRINT 'An error occurred, transaction rolled back.';

    -- ERROR HANDLING,
END CATCH;
```

```
(2 rows affected)

(2 rows affected)

Completion time: 2024-09-22T15:18:14.2688759+05:30
```

```
Query executed successfully.   (local) (16.0 RTM)   DESKTOP-RRNO5QV\tsash ...   RP_DATA   00:00:00   0 rows
```

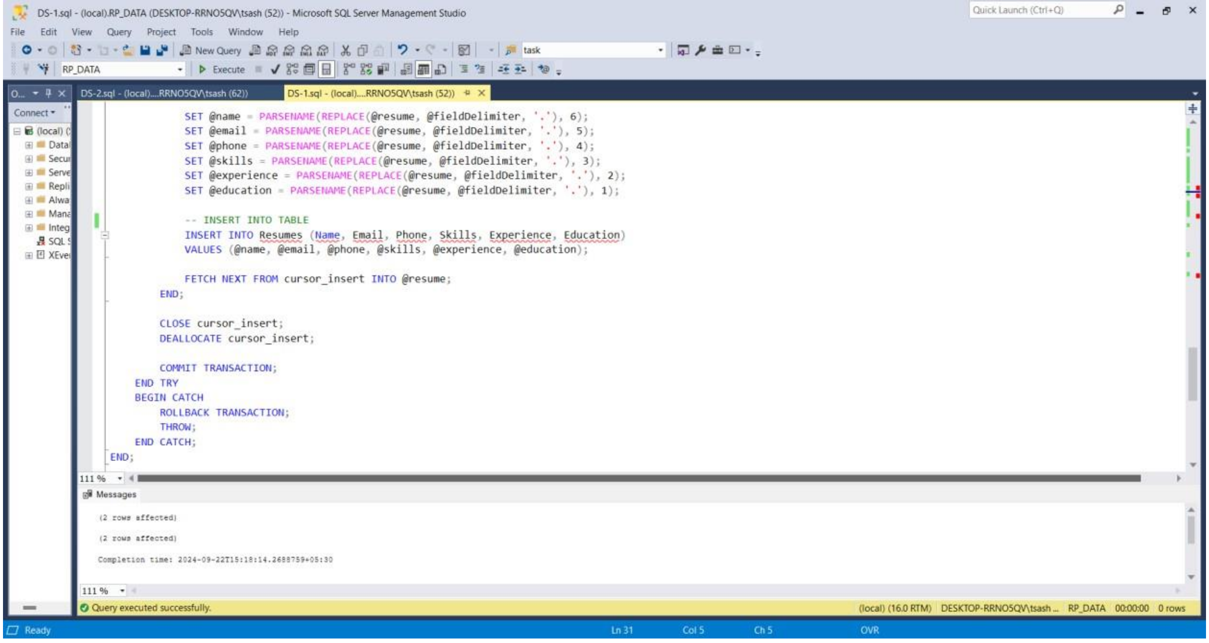
STORED PROCEDURE THAT ACCEPTS MULTIPLE RESUME RECORDS AND INSERTS THEM INTO THE DATABASE IN A SINGLE CALL.



```
-- STORED PROCEDURE  THAT ACCEPTS MULTIPLE RESUME RECORDS AND INSERTS THEM INTO THE DATABASE IN A SINGLE CALL.
CREATE PROCEDURE InsertResumes
    @records NVARCHAR(MAX)
AS
BEGIN
    BEGIN TRANSACTION;

    BEGIN TRY
        -- SPLIT THE INPUT STRING INTO INDIVIDUAL RECORDS USING STRING_SPLIT
        DECLARE @resume NVARCHAR(MAX);
        DECLARE @resumeList TABLE (
            Name NVARCHAR(255),
            Email NVARCHAR(255),
            Phone NVARCHAR(255),
            Skills NVARCHAR(MAX),
            Experience NVARCHAR(MAX),
            Education NVARCHAR(255)
        );

        DECLARE @delimiter CHAR(1) = '#';
        DECLARE @fieldDelimiter CHAR(1) = ',';

        DECLARE cursor_insert CURSOR FOR
            SELECT value FROM STRING_SPLIT(@records, @delimiter);
```

```
(2 rows affected)

(2 rows affected)

Completion time: 2024-09-22T15:18:14.2688759+05:30
```

Query executed successfully.



```
            SET @name = PARSENAME(REPLACE(@resume, @fieldDelimiter, '.'), 6);
            SET @email = PARSENAME(REPLACE(@resume, @fieldDelimiter, '.'), 5);
            SET @phone = PARSENAME(REPLACE(@resume, @fieldDelimiter, '.'), 4);
            SET @skills = PARSENAME(REPLACE(@resume, @fieldDelimiter, '.'), 3);
            SET @experience = PARSENAME(REPLACE(@resume, @fieldDelimiter, '.'), 2);
            SET @education = PARSENAME(REPLACE(@resume, @fieldDelimiter, '.'), 1);

            -- INSERT INTO TABLE
            INSERT INTO Resumes (Name, Email, Phone, Skills, Experience, Education)
            VALUES (@name, @email, @phone, @skills, @experience, @education);

            FETCH NEXT FROM cursor_insert INTO @resume;
        END;

        CLOSE cursor_insert;
        DEALLOCATE cursor_insert;

        COMMIT TRANSACTION;
    END TRY
    BEGIN CATCH
        ROLLBACK TRANSACTION;
        THROW;
    END CATCH;
END;
```

```
(2 rows affected)

(2 rows affected)

Completion time: 2024-09-22T15:18:14.2688759+05:30
```

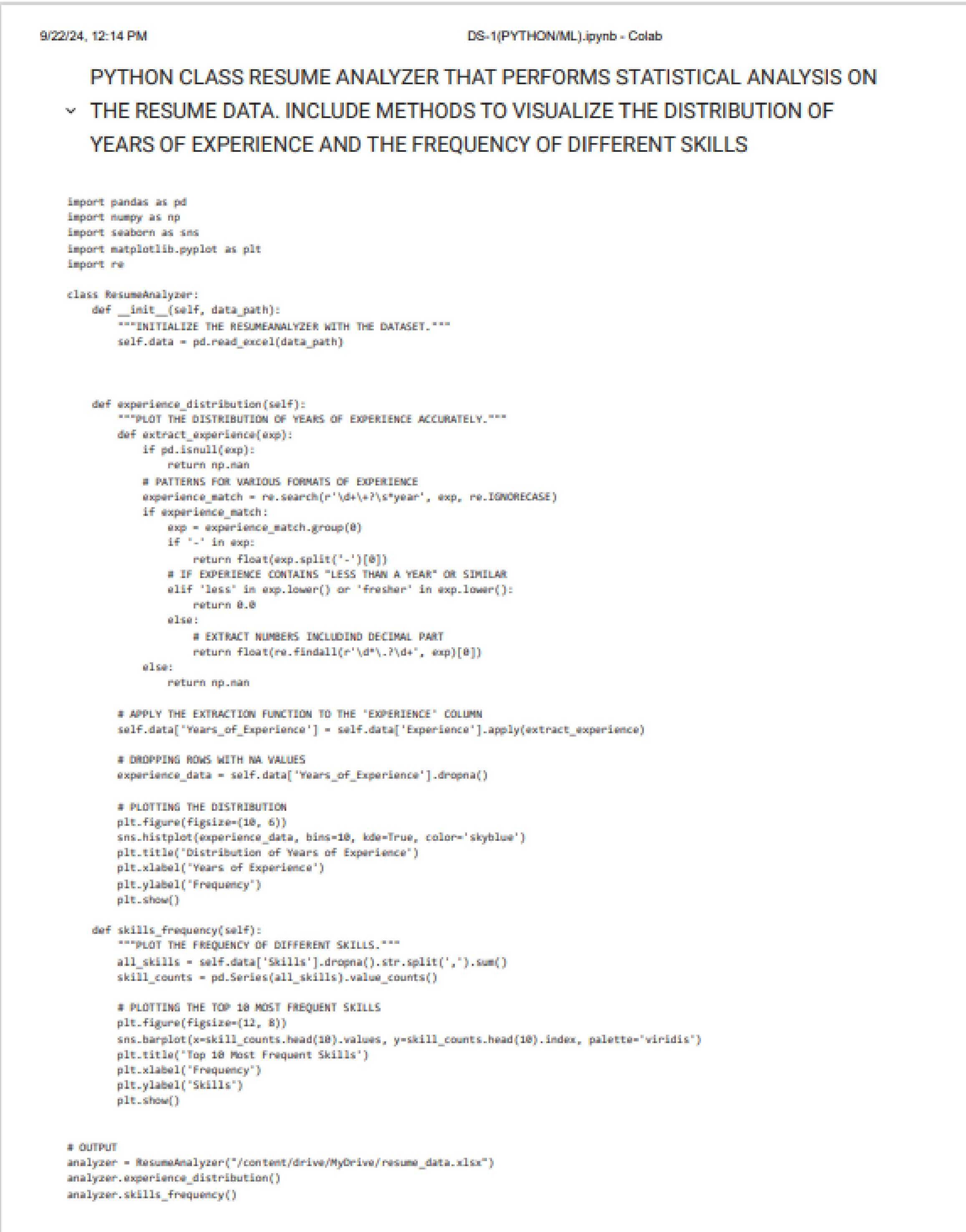
Query executed successfully.

## 6. Statistical Analysis and Machine Learning

### 6.1 Python Resume Analyzer Class for Statistical Analysis

The Resume Analyzer class uses Seaborn and NumPy to perform statistical analysis on the resume data, such as analyzing the distribution of experience and the frequency of skills.

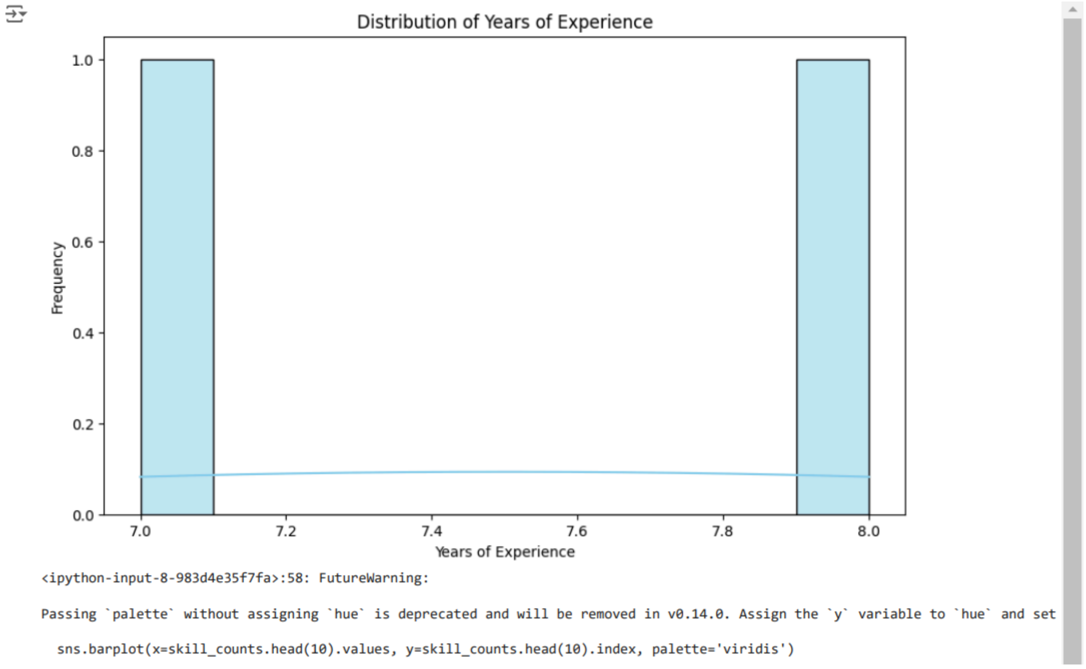
## Source Code:



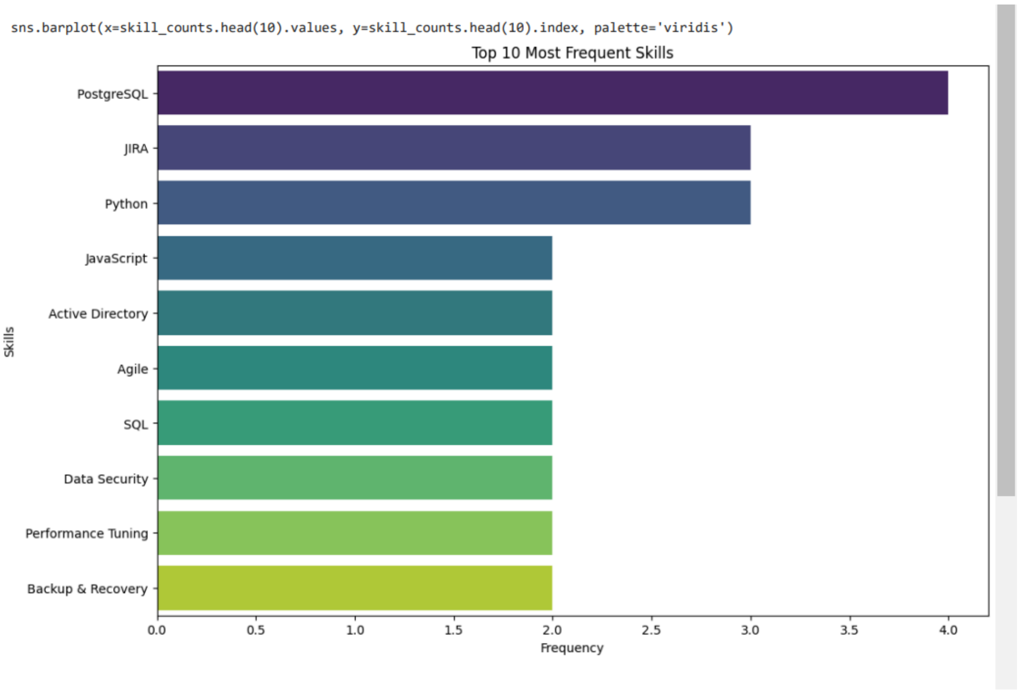
### PYTHON CLASS RESUME ANALYZER THAT PERFORMS STATISTICAL ANALYSIS ON THE RESUME DATA. INCLUDE METHODS TO VISUALIZE THE DISTRIBUTION OF YEARS OF EXPERIENCE AND THE FREQUENCY OF DIFFERENT SKILLS

```python
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import re

class ResumeAnalyzer:
    def __init__(self, data_path):
        """INITIALIZE THE RESUMEANALYZER WITH THE DATASET."""
        self.data = pd.read_excel(data_path)



    def experience_distribution(self):
        """PLOT THE DISTRIBUTION OF YEARS OF EXPERIENCE ACCURATELY."""
        def extract_experience(exp):
            if pd.isnull(exp):
                return np.nan
            # PATTERNS FOR VARIOUS FORMATS OF EXPERIENCE
            experience_match = re.search(r'\d+\+?\s*year', exp, re.IGNORECASE)
            if experience_match:
                exp = experience_match.group(0)
                if '-' in exp:
                    return float(exp.split('-')[0])
                # IF EXPERIENCE CONTAINS "LESS THAN A YEAR" OR SIMILAR
                elif 'less' in exp.lower() or 'fresher' in exp.lower():
                    return 0.0
                else:
                    # EXTRACT NUMBERS INCLUDIND DECIMAL PART
                    return float(re.findall(r'\d*\.?\d+', exp)[0])
            else:
                return np.nan

        # APPLY THE EXTRACTION FUNCTION TO THE 'EXPERIENCE' COLUMN
        self.data['Years_of_Experience'] = self.data['Experience'].apply(extract_experience)

        # DROPPING ROWS WITH NA VALUES
        experience_data = self.data['Years_of_Experience'].dropna()

        # PLOTTING THE DISTRIBUTION
        plt.figure(figsize=(10, 6))
        sns.histplot(experience_data, bins=10, kde=True, color='skyblue')
        plt.title('Distribution of Years of Experience')
        plt.xlabel('Years of Experience')
        plt.ylabel('Frequency')
        plt.show()

    def skills_frequency(self):
        """PLOT THE FREQUENCY OF DIFFERENT SKILLS."""
        all_skills = self.data['Skills'].dropna().str.split(',').sum()
        skill_counts = pd.Series(all_skills).value_counts()

        # PLOTTING THE TOP 10 MOST FREQUENT SKILLS
        plt.figure(figsize=(12, 8))
        sns.barplot(x=skill_counts.head(10).values, y=skill_counts.head(10).index, palette='viridis')
        plt.title('Top 10 Most Frequent Skills')
        plt.xlabel('Frequency')
        plt.ylabel('Skills')
        plt.show()


# OUTPUT
analyzer = ResumeAnalyzer("/content/drive/MyDrive/resume_data.xlsx")
analyzer.experience_distribution()
analyzer.skills_frequency()
```

DISTRIBUTION OF YEARS OF EXPERIENCE AND THE FREQUENCY OF DIFFERENT SKILLS



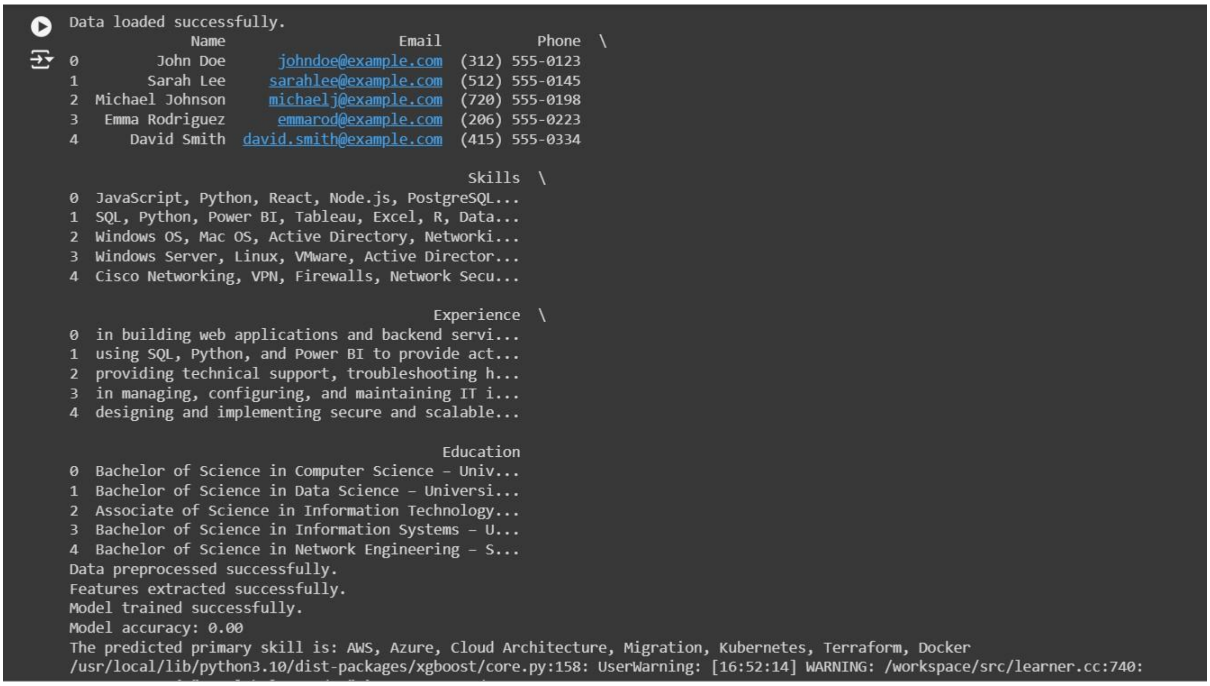
```
<ipython-input-8-983d4e35f7fa>:58: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and set

  sns.barplot(x=skill_counts.head(10).values, y=skill_counts.head(10).index, palette='viridis')
```



```
sns.barplot(x=skill_counts.head(10).values, y=skill_counts.head(10).index, palette='viridis')
```

## 6.2 XGBoost Model for Skill Prediction

An XGBoost model is trained to predict the primary skill of a candidate based on resume content. The model is encapsulated in a Python class.



```
Data loaded successfully.
             Name                   Email           Phone  \
0        John Doe      johndoe@example.com  (312) 555-0123
1       Sarah Lee     sarahlee@example.com  (512) 555-0145
2  Michael Johnson     michaelj@example.com  (720) 555-0198
3   Emma Rodriguez      emmarod@example.com  (206) 555-0223
4     David Smith  david.smith@example.com  (415) 555-0334

                                      Skills  \
0  JavaScript, Python, React, Node.js, PostgreSQL...
1  SQL, Python, Power BI, Tableau, Excel, R, Data...
2  Windows OS, Mac OS, Active Directory, Networki...
3  Windows Server, Linux, VMware, Active Director...
4  Cisco Networking, VPN, Firewalls, Network Secu...

                                       Experience  \
0  in building web applications and backend servi...
1  using SQL, Python, and Power BI to provide act...
2  providing technical support, troubleshooting h...
3  in managing, configuring, and maintaining IT i...
4  designing and implementing secure and scalable...

                                        Education
0  Bachelor of Science in Computer Science - Univ...
1  Bachelor of Science in Data Science - Universi...
2  Associate of Science in Information Technology...
3  Bachelor of Science in Information Systems - U...
4  Bachelor of Science in Network Engineering - S...
Data preprocessed successfully.
Features extracted successfully.
Model trained successfully.
Model accuracy: 0.00
The predicted primary skill is: AWS, Azure, Cloud Architecture, Migration, Kubernetes, Terraform, Docker
/usr/local/lib/python3.10/dist-packages/xgboost/core.py:158: UserWarning: [16:52:14] WARNING: /workspace/src/learner.cc:740:
```

# 7. Data Visualization

## 7.1 Power BI Dashboard Overview

A Power BI dashboard visualizes key metrics derived from the resume data, including:

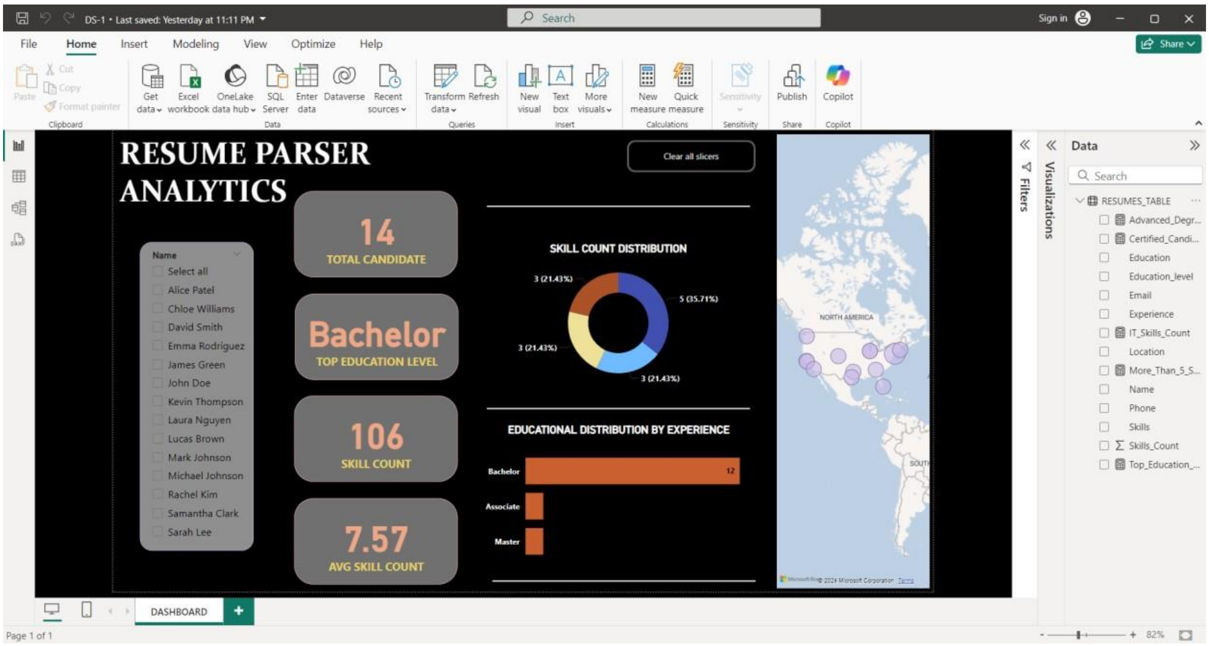
- **Education Level Distribution**: Displays the highest education level of candidates.
- **Geographic Distribution**: Visualizes where candidates are located.

## 7.2 Key Metrics and Insights

- **Average Number of Skills per Candidate**: Helps identify candidates with diverse skills.

## 7.3 DAX Formulas for Calculations

Key metrics are calculated using DAX formulas, such as the average number of skills per candidate and the submission trends.



RESUME PARSER ANALYTICS

14 TOTAL CANDIDATE

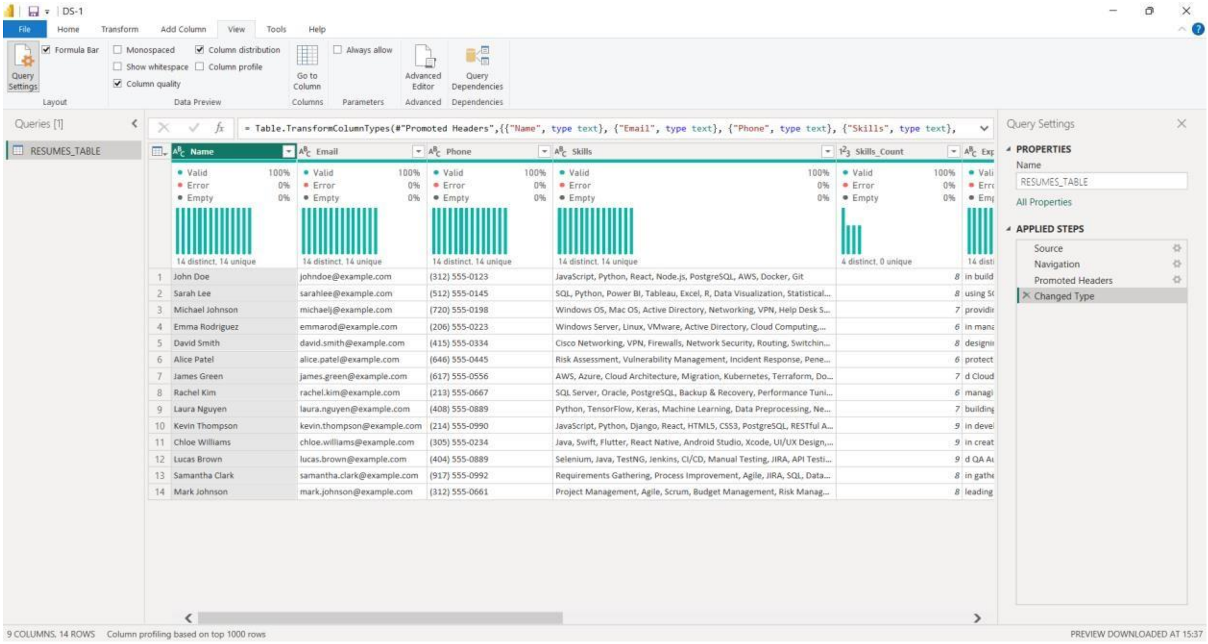
Bachelor TOP EDUCATION LEVEL

106 SKILL COUNT

7.57 AVG SKILL COUNT

Name
- ☐ Select all
- ☐ Alice Patel
- ☐ Chloe Williams
- ☐ David Smith
- ☐ Emma Rodriguez
- ☐ James Green
- ☐ John Doe
- ☐ Kevin Thompson
- ☐ Laura Nguyen
- ☐ Lucas Brown
- ☐ Mark Johnson
- ☐ Michael Johnson
- ☐ Rachel Kim
- ☐ Samantha Clark
- ☐ Sarah Lee

RESUMES_TABLE
- ☐ Advanced_Degr...
- ☐ Certified_Candi...
- ☐ Education
- ☐ Education_level
- ☐ Email
- ☐ Experience
- ☐ IT_Skills_Count
- ☐ Location
- ☐ More_Than_5_S...
- ☐ Name
- ☐ Phone
- ☐ Skills
- ☐ Σ Skills_Count
- ☐ Top_Education_...



☑ Formula Bar
☐ Monospaced ☑ Column distribution ☐ Always allow
☐ Show whitespace ☐ Column profile
☑ Column quality

```
= Table.TransformColumnTypes(#"Promoted Headers",{{"Name", type text}, {"Email", type text}, {"Phone", type text}, {"Skills", type text},
```

| | Name | Email | Phone | Skills | Skills_Count | Exp |
|---|---|---|---|---|---|---|
| | • Valid 100%<br>• Error 0%<br>• Empty 0% | • Valid 100%<br>• Error 0%<br>• Empty 0% | • Valid 100%<br>• Error 0%<br>• Empty 0% | • Valid 100%<br>• Error 0%<br>• Empty 0% | • Valid 100%<br>• Error 0%<br>• Empty 0% | • Vali<br>• Erro<br>• Emp |
| | 14 distinct, 14 unique | 14 distinct, 14 unique | 14 distinct, 14 unique | 14 distinct, 14 unique | 4 distinct, 0 unique | 14 disti |
| 1 | John Doe | johndoe@example.com | (312) 555-0123 | JavaScript, Python, React, Node.js, PostgreSQL, AWS, Docker, Git | 8 | in build |
| 2 | Sarah Lee | sarahlee@example.com | (512) 555-0145 | SQL, Python, Power BI, Tableau, Excel, R, Data Visualization, Statistical... | 8 | using S( |
| 3 | Michael Johnson | michaelj@example.com | (720) 555-0198 | Windows OS, Mac OS, Active Directory, Networking, VPN, Help Desk S... | 7 | providir |
| 4 | Emma Rodriguez | emmarod@example.com | (206) 555-0223 | Windows Server, Linux, VMware, Active Directory, Cloud Computing,... | 6 | in mana |
| 5 | David Smith | david.smith@example.com | (415) 555-0334 | Cisco Networking, VPN, Firewalls, Network Security, Routing, Switchin... | 8 | designi |
| 6 | Alice Patel | alice.patel@example.com | (646) 555-0445 | Risk Assessment, Vulnerability Management, Incident Response, Pene... | 6 | protect |
| 7 | James Green | james.green@example.com | (617) 555-0556 | AWS, Azure, Cloud Architecture, Migration, Kubernetes, Terraform, Do... | 7 | d Cloud |
| 8 | Rachel Kim | rachel.kim@example.com | (213) 555-0667 | SQL Server, Oracle, PostgreSQL, Backup & Recovery, Performance Tuni... | 6 | managi |
| 9 | Laura Nguyen | laura.nguyen@example.com | (408) 555-0889 | Python, TensorFlow, Keras, Machine Learning, Data Preprocessing, Ne... | 7 | building |
| 10 | Kevin Thompson | kevin.thompson@example.com | (214) 555-0990 | JavaScript, Python, Django, React, HTML5, CSS3, PostgreSQL, RESTful A... | 9 | in devel |
| 11 | Chloe Williams | chloe.williams@example.com | (305) 555-0234 | Java, Swift, Flutter, React Native, Android Studio, Xcode, UI/UX Design,... | 9 | in creat |
| 12 | Lucas Brown | lucas.brown@example.com | (404) 555-0889 | Selenium, Java, TestNG, Jenkins, CI/CD, Manual Testing, JIRA, API Testi... | 9 | d QA Au |
| 13 | Samantha Clark | samantha.clark@example.com | (917) 555-0992 | Requirements Gathering, Process Improvement, Agile, JIRA, SQL, Data... | 8 | in gathe |
| 14 | Mark Johnson | mark.johnson@example.com | (312) 555-0661 | Project Management, Agile, Scrum, Budget Management, Risk Manag... | 8 | leading |

Query Settings

PROPERTIES

Name: RESUMES_TABLE

All Properties

APPLIED STEPS
- Source
- Navigation
- Promoted Headers
- Changed Type

9 COLUMNS, 14 ROWS Column profiling based on top 1000 rows

PREVIEW DOWNLOADED AT 15:37

# 8. Conclusion

## 8.1 Summary of Achievements

The project successfully automates the extraction, cleaning, storage, and analysis of resume data using a combination of Python, SQL, Excel VBA, and Power BI.

### 8.2 Key Insights Derived

- **Skills**: Identified multi-skilled candidates.
- **Geographic Insights**: Helped understand where candidates are concentrated.
- **Trends**: Resume submission patterns inform recruitment strategies.

### 8.3 Future Enhancements

Future improvements could include expanding the machine learning model to predict other attributes such as job fit, and integrating external APIs to further enrich the resume data.