Exploring the Data Analyst Job Market: Insights and Trends

Prepared by: Marcus Nogueira

Section 1

Introduction:

This dataset comprises job listings for data analyst positions scraped from LinkedIn. It includes information such as job title, company name, job description, location (onsite/remote), salary range, company location, job criteria, posting date, and a URL link to the job. This dataset includes data from Africa,Canada and the USA. It's as of November 2022.

Figure 1: Variables and Definitions

Variable Name	Description
title	Job Title
company	Name of the company
description	Description of the job and company
onsite_remote	Location where the employee will be working from (Onsite/Remote)
salary	Salary for the job (may be yearly or hourly, usually in a range from min to max)
location	Location of the company with the job opening
criteria	Seniority Level, Employment Type, Industry and Job Function
posted_date	The date the job was posted
link	URL link to the job posting

Section 2

Key Objective and Outcome:

The overarching motivation behind this data analysis is to gain comprehensive insights into the data analyst job market landscape. By addressing fundamental questions such as the top skills sought by employers, the distribution of onsite versus remote job opportunities, industry-specific job trends, salary ranges based on seniority level, and potential salary discrepancies between different work arrangements, we aim to provide valuable guidance to both job seekers and employers. This analysis seeks to illuminate key trends, challenges, and opportunities in the data analyst field, empowering individuals to make informed decisions regarding skill development, job searching, hiring practices, and compensation negotiations. Ultimately, the goal is to enhance understanding of the evolving dynamics of the data analyst profession and facilitate strategic decision-making in the context of this rapidly evolving industry.

Section 3

Data Cleaning:

- Combined 3 separate files
- Extracted information from the criteria column
- Checked job descriptions for all skills
- Cleaned the salary column: extra characters, selected lower bound to be conservative
- Converted hourly salary figures
- Challenges: No reliable way of scraping years of experience from the job description, location column not uniform,

3.1 Combining datasets

I compiled data from three distinct regions: USA, Canada and Africa. Each region's data was stored in separate CSV files, which we merged into a single master dataset name 'MainDataSet'.

	title	company	description	onsite_remote	salary	location	criteria	posted_date	link
0	Data Analyst - Recent Graduate	PayPal	At PayPal (NASDAQ: PYPL), we believe that ever...	onsite	NaN	Buffalo-Niagara Falls Area	[{'Seniority level': 'Not Applicable'}, {'Empl...	2022-11-22	https://www.linkedin.com/jobs/view/data-analys...
1	Data Analyst - Recent Graduate	PayPal	At PayPal (NASDAQ: PYPL), we believe that ever...	onsite	NaN	San Jose, CA	[{'Seniority level': 'Not Applicable'}, {'Empl...	2022-11-22	https://www.linkedin.com/jobs/view/data-analys...
2	Data Analyst	PayPal	At PayPal (NASDAQ: PYPL), we believe that ever...	onsite	NaN	Texas, United States	[{'Seniority level': 'Not Applicable'}, {'Empl...	2022-11-17	https://www.linkedin.com/jobs/view/data-analys...
3	Data Analyst	PayPal	At PayPal (NASDAQ: PYPL), we believe that ever...	onsite	NaN	Illinois, United States	[{'Seniority level': 'Not Applicable'}, {'Empl...	2022-11-17	https://www.linkedin.com/jobs/view/data-analys...
4	Entry-Level Data Analyst	The Federal Savings Bank	The Federal Savings Bank, a national bank and ...	onsite	NaN	Chicago, IL	[{'Seniority level': 'Entry level'}, {'Employm...	2022-11-17	https://www.linkedin.com/jobs/view/entry-level...

The 'criteria' column encompassed valuable information including seniority level, employment type (onsite/remote/hybrid), job function, and industry affiliation for each job listing. However, this data was initially stored as a list of dictionaries. To facilitate analysis, we meticulously parsed each detail into its own dedicated column.

```python
def extract_criteria(criteria):
    seniority_level = None
    job_function = None
    industries = None

    for detail in eval(criteria):
        if 'Seniority level' in detail:
            seniority_level = detail['Seniority level']
        if 'Job function' in detail:
            job_function = detail['Job function']
        if 'Industries' in detail:
            industries = detail['Industries']

    return seniority_level, job_function, industries

MainDataSet[['Seniority level', 'Job function', 'Industries']] = MainDataSet['criteria'].apply(extract_criteria).apply(pd.Series)

MainDataSet.drop('criteria', axis=1, inplace=True)
```

	title	company	description	onsite_remote	location	posted_date	Seniority level	Job function	Industries
0	Data Analyst - Recent Graduate	PayPal	At PayPal (NASDAQ: PYPL), we believe that ever...	onsite	Buffalo-Niagara Falls Area	2022-11-22	Not Applicable	Information Technology	Software Development, Technology, Information ...
1	Data Analyst - Recent Graduate	PayPal	At PayPal (NASDAQ: PYPL), we believe that ever...	onsite	San Jose, CA	2022-11-22	Not Applicable	Information Technology	Software Development, Technology, Information ...
2	Data Analyst	PayPal	At PayPal (NASDAQ: PYPL), we believe that ever...	onsite	Texas, United States	2022-11-17	Not Applicable	Information Technology	Software Development, Technology, Information ...
3	Data Analyst	PayPal	At PayPal (NASDAQ: PYPL), we believe that ever...	onsite	Illinois, United States	2022-11-17	Not Applicable	Information Technology	Software Development, Technology, Information ...
4	Entry-Level Data Analyst	The Federal Savings Bank	The Federal Savings Bank, a national bank and ...	onsite	Chicago, IL	2022-11-17	Entry level	Information Technology	Savings Institutions

3.2. Extracting skills

I proceeded by compiling a comprehensive list of the most prevalent tools and skill sets typically demanded in data analyst positions. Then, we parsed each job description to ascertain the presence of these requisites. To facilitate systematic analysis, we introduced binary columns corresponding to each skill, indicating its presence or absence within the job descriptions

Code:

```
# Scraping skills from the job description
top_skills = ['sql', ' r ', 'python', 'matlab', 'excel', 'tableau', 'power bi', 'java', 'data visualization',
              'data mining', 'statistics', 'machine learning', 'data cleaning', 'data manipulation', 'spark',
              'hadoop', 'google analytics', 'aws', 'machine learning', 'qlik', 'data modeling']
for i in top_skills:
    skill = []
    for desc in MainDataSet['description']:
        if i in desc.lower():
            skill.append(1)
        else:
            skill.append(0)
    MainDataSet[i] = skill

MainDataSet.head()
```

Result:

location	posted_date	Seniority level	Job function	Industries	sql	...	statistics	machine learning	data cleaning	data manipulation	spark	hadoop	google analytics	aws	qlik	data modeling
Buffalo-Niagara Falls Area	2022-11-22	Not Applicable	Information Technology	Software Development, Technology, Information ...	1	...	1	0	0	0	0	1	0	0	0	0
San Jose, CA	2022-11-22	Not Applicable	Information Technology	Software Development, Technology, Information ...	1	...	1	0	0	0	0	1	0	0	0	0
Texas, United States	2022-11-17	Not Applicable	Information Technology	Software Development, Technology, Information ...	1	...	0	0	0	0	0	1	0	0	0	0
Illinois, United States	2022-11-17	Not Applicable	Information Technology	Software Development, Technology, Information ...	1	...	0	0	0	0	0	1	0	0	0	0
Chicago, IL	2022-11-17	Entry level	Information Technology	Savings Institutions	1	...	0	0	0	0	0	0	0	0	0	0

3.2. Cleaning and Converting Salary data

	index	title	company	description	onsite_remote	salary	location	criteria	posted_date	
	8	Data Analyst	London Approach	The ideal candidate for the Data Analyst/Finan...	onsite	100, 000.00\r\n – \r\n 1...	Franklin, TN	[{'Seniority level': 'Associate'}, {'Employmen...	2022-11-21	https://ww
	24	Data Analyst	Eva Garland Consulting, LLC	Reporting to the Director of Operations, the D...	onsite	50, 000.00\r\n – \r\n 55...	Raleigh, NC	[{'Seniority level': 'Entry level'}, {'Employm...	2022-11-22	https://ww
	35	Data Analyst	London Approach	The ideal candidate for the Data Analyst/Finan...	onsite	100, 000.00\r\n – \r\n 1...	Franklin, TN	[{'Seniority level': 'Associate'}, {'Employmen...	2022-11-21	https://ww
	61	Data Analyst	London Approach	The ideal candidate for the Data Analyst/Finan...	onsite	100, 000.00\r\n – \r\n 1...	Franklin, TN	[{'Seniority level': 'Associate'}, {'Employmen...	2022-11-21	https://ww
	88	Data Analyst	London Approach	The ideal candidate for the Data Analyst/Finan...	onsite	100, 000.00\r\n – \r\n 1...	Franklin, TN	[{'Seniority level': 'Associate'}, {'Employmen...	2022-11-21	https://ww
	...	...	...	...	...	...	...	...	...	
	2834	Junior Data Analyst	Insight Global	This is a One Year Contract to Hire, 40 hours ...	hybrid	30.00\r\n – \r\n 33.00	Denver Metropolitan Area	[{'Seniority level': 'Entry level'}, {'Employm...	2022-11-16	https://ww

The salary column initially contained strings with excessive characters, rendering the data unfit for quantitative analysis. To rectify this, we employed a two-step approach. Firstly, I split each string at the first instance of the character "\" and eliminated commas to extract the lower bound of the salary estimate. Secondly, we addressed the issue of hourly rates by converting them into yearly figures. This involved multiplying each two-digit figure by 2080, which represents the average number of hours worked in a year in the US.

Code:

```
sal = []
sal_converted = []
for salary in usa['salary']:
    value = int(salary.split(".")[0].replace('$', '').replace(',', ''))
    sal.append(value)

sal_converted = []
for number in sal:
    if len(str(number)) == 2:
        sal_converted.append(number*2080)
    else:
        sal_converted.append(number)

usa['salary_converted'] = sal_converted
```

data anipulation	salary_cleaned	salary_converted
0	100000	100000
0	50000	50000
0	100000	100000
0	100000	100000
0	100000	100000

3.3. Challenges with data cleaning

Parsing the minimum years of work experience required from job descriptions posed a challenge. Our approach involved scanning each job description for the presence of the terms 'years' or 'yrs' and extracting any integers nearby. To accomplish this, I utilized the 're' Python library, which offers robust support for regular expressions (using the match function)

Code:

```
def find_years(string):
    # Define the pattern to match integers followed by 'year' or 'yrs'
    pattern = r'\b\d+\s*(?:years?|yrs?)\b'

    # Find all matches in the text
    matches = re.findall(pattern, string)

    # Convert matched strings to integers and return
    return [int(match.split()[0]) for match in matches]

yrs = []
for text in MainDataSet['description']:  # Changes string to text
    y = find_years(text)
    if y:  # Check if y is not empty
        yrs.append(min(y))
    else:
        yrs.append(np.nan)

MainDataSet['years_of_experience'] = yrs
```

while our initial attempt to extract the minimum years of work experience from job descriptions yielded some success, it proved unreliable across all instances. Many job descriptions either lacked this information entirely or contained misleading numerical references, such as the company's years in operation. Despite modifying the code to return the minimum value from a list of integers found near the terms 'years' or 'yrs', we still encountered inaccurate outputs, including absurdly high figures like 175 years. Consequently, I made the decision to exclude the years of experience column from our analysis due to its inconsistency and potential for misinterpretation.

Section 4

A. Data Analysis Questions:

Guiding Question: What factors drive success for data analysts in terms of skills, location, industry, seniority, and compensation?

1. What are the top skills employers seek for data analysts?

- This question is fundamental as it sheds light on the most sought-after competencies in the field of data analysis, guiding both job seekers and educators. We aim to identify the essential skills by analyzing job postings and extracting recurring skill requirements.

2. What is the onsite vs. remote job distribution by location?

- Understanding the distribution of onsite and remote job opportunities across different locations provides insights into evolving work arrangements, particularly post-pandemic. This analysis aids in comprehending the shifting dynamics of work preferences and employer policies.

3. What are the most sought-after skills for data analysts at different seniority levels?

- Examining job distribution by seniority and required skills enables us to customize skill development and training initiatives to match precise industry needs. By pinpointing trends unique to each role, we can effectively equip aspiring data analysts for their preferred sectors.

4. What is the salary range for data analysts based on seniority level?

- Salary range analysis based on seniority and experience provides valuable insights into the earning potential of data analysts. Understanding the compensation landscape enables

professionals to negotiate competitive salaries and organizations to benchmark their pay scales effectively.

5. Are there salary discrepancies between remote and onsite jobs for data analysts?

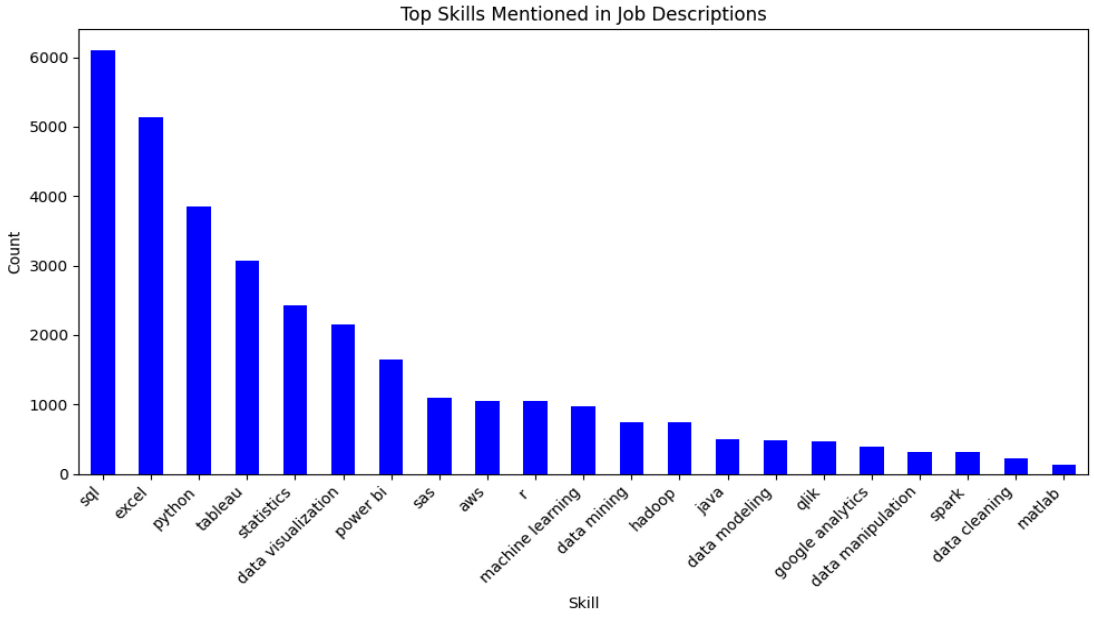
- Exploring potential salary differences between remote and onsite positions offers insights into the financial implications of different work arrangements. Understanding these disparities informs both job seekers and employers in negotiating fair compensation packages and evaluating the cost-benefit analysis of remote work policies.

B. Explanation of Methods:

1. Top Skills Analysis:

Our approach to analyzing the top skills sought by employers for data analysts was systematic and structured. We began by compiling a list of essential skills based on industry knowledge and market trends. Next, we iteratively examined the job descriptions in our dataset to determine the presence of these skills. Using a for loop, we systematically searched each job description, employing case-insensitive matching to ensure comprehensive coverage. For each skill, we created binary columns indicating its presence or absence in the job description. This step-by-step process enabled us to quantify the demand for each skill and identify the most sought-after competencies. The resulting visualization provided a clear overview of the top skills desired by employers, with SQL, Excel, and Python emerging as the most prominent.

Figure 2: Top Skills Mentioned in Job Descriptions



2. Onsite vs. Remote Distribution:

Africa

In order to conduct our analysis effectively, we segmented our data into separate datasets for the USA, Africa, and Canada, applying a consistent methodology with slight variations across each dataset. Initially, we devised two functions to discern whether a job is onsite/hybrid or remote, which we then utilized to generate binary columns facilitating our analysis and visualization tasks. Extracting city and country information from the location column via splitting, we encountered instances of missing country data, which we addressed by manually mapping cities to their respective countries.

We then aggregated the data by country, computing the total number of jobs available in both onsite and remote capacities. Our visualization efforts included exploring the utilization of a

heatmap, accomplished using a shapefile to depict job density by country. Notably, our findings revealed a significant concentration of jobs in South Africa, a trend that aligns with industry insights. Despite the noticeable disparity in job distribution across Africa in the data analyst field, our results corroborate expert opinions, attributing South Africa's prominence to its robust infrastructure, stable political climate, and strong educational system. Additionally, the presence of key tech hubs, governmental backing for innovation, and an expanding pool of skilled professionals further solidify its allure to tech companies.

Figure 3: Distribution Jobs by Country in Africa for Onsite Data Analyst

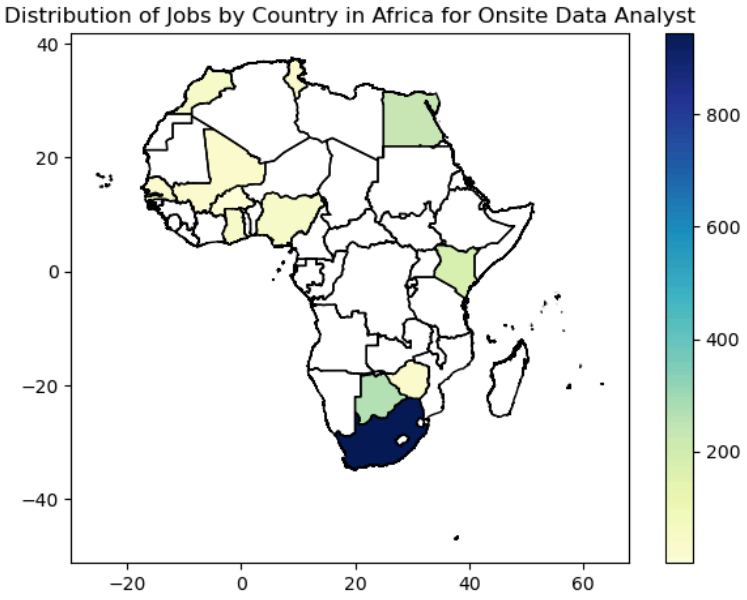
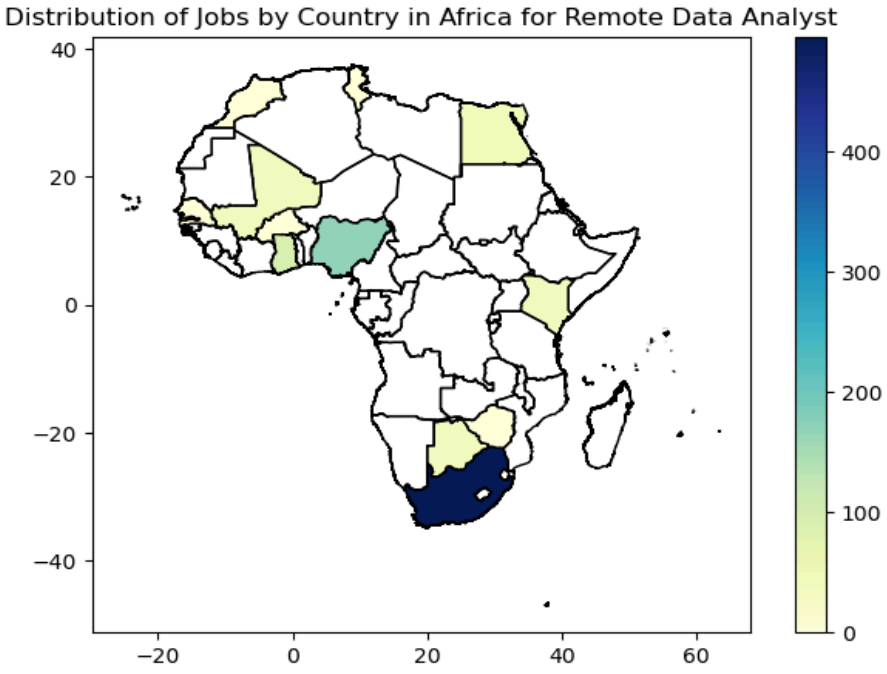


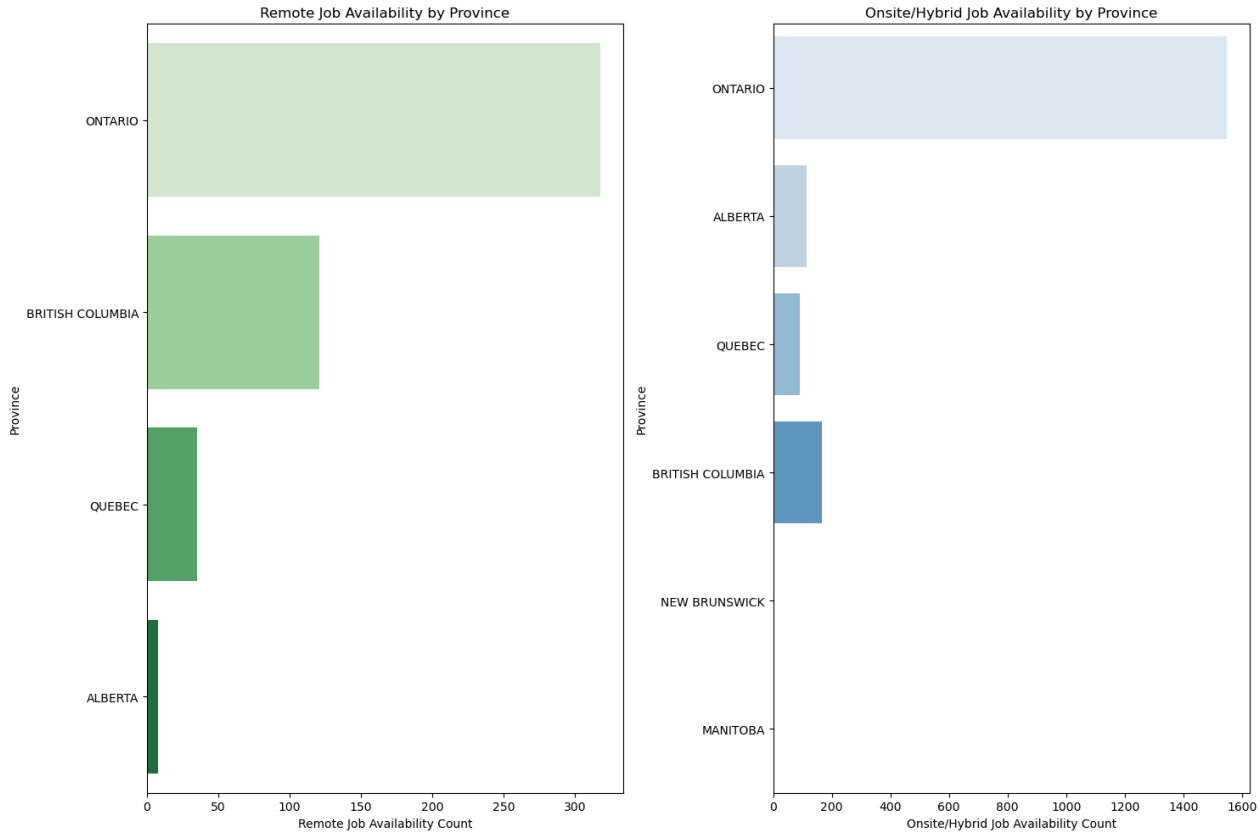
Figure 4: Distribution Jobs by Country in Africa for Remote Data Analyst



Canada

We followed similar data preparation procedures as in the previous datasets, focusing on extracting province and city details instead of country information. By doing so, we organized the data by province and computed the total number of onsite/hybrid and remote jobs. Notably, our analysis of Canada highlighted Ontario as the dominant province for job postings, indicating a significantly higher demand for onsite roles compared to remote positions across the country. This dominance of Ontario in job postings could be attributed to several factors, including its status as Canada's economic and technological hub, with major cities like Toronto and Ottawa hosting numerous tech companies, research institutions, and educational facilities. Additionally, Ontario's strong economy, diverse industries, and skilled workforce likely contribute to its appeal for job seekers and employers alike.

Figure 5: Distribution of Remote vs Onsite Jobs by Province

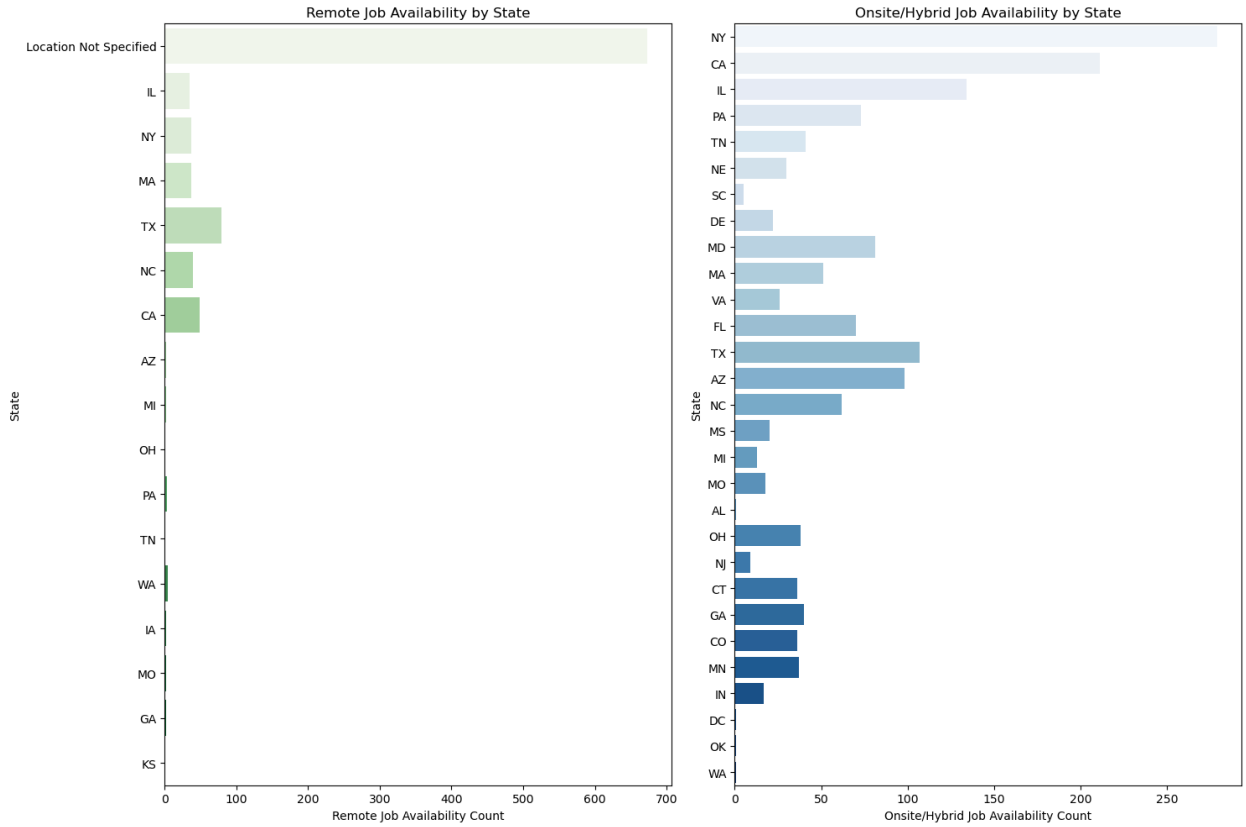


USA

For this dataset, we followed the same systematic approach as with the other two datasets, but instead of extracting country or city information, we extracted the state from the location column. As expected, some data entries lacked state information, prompting us to manually assign these outliers to their appropriate states. We then organized the data by state and computed the total number of onsite/hybrid and remote jobs. Interestingly, our analysis revealed that California, New York, Texas, and Illinois emerged as the top states for data analyst opportunities across all seniority levels, indicative of their robust tech ecosystems, thriving job

markets, and concentrations of tech companies, universities, and innovation hubs. These states, being home to renowned tech hubs such as Silicon Valley, New York City, and Austin, offer an array of opportunities for data analysts due to their vibrant tech scenes, strong infrastructure, and access to a skilled workforce, making them highly sought-after destinations for tech professionals.

Figure 6: Distribution of Remote vs Onsite Jobs by State



3. Industry-based Job Distribution and Skills:

In our data cleaning process, we successfully extracted the Seniority level, a column that provided significant insights throughout our analysis. Notably, we observed that SQL emerged as the most in-demand skill overall, closely followed by Excel. However, upon delving into Mid Senior level roles, SQL surpassed Excel in demand, unlike in Associate and Entry Level roles.

Understanding these employer preferences was crucial for us as aspiring data analysts, aiming to refine and enhance our skill sets. One surprising revelation was the limited demand for R in roles beyond entry level. Our analysis of the top 10 skills for each Seniority level reaffirms that Excel's widespread familiarity makes it more sought after for Associate roles, while R's specialized nature restricts its demand beyond entry-level positions, primarily valued for statistical analysis and data science tasks. These conclusions are grounded in our data-driven approach, emphasizing empirical evidence over preconceptions.

Additionally, we encountered a scenario where the amount of Onsite Jobs that only had "United States" as the location initially misled us. It was believed that these jobs would become more region-specific further in the application process; however, we were not privy to this information. As a group, we decided it would be best to drop the rows where jobs were onsite and the location extracted was "United States". This decision led to a significant decrease in the availability of values but allowed us to paint a more accurate picture of the data.

Figure 7: Top 10 Skills for Mid-Senior Level Roles

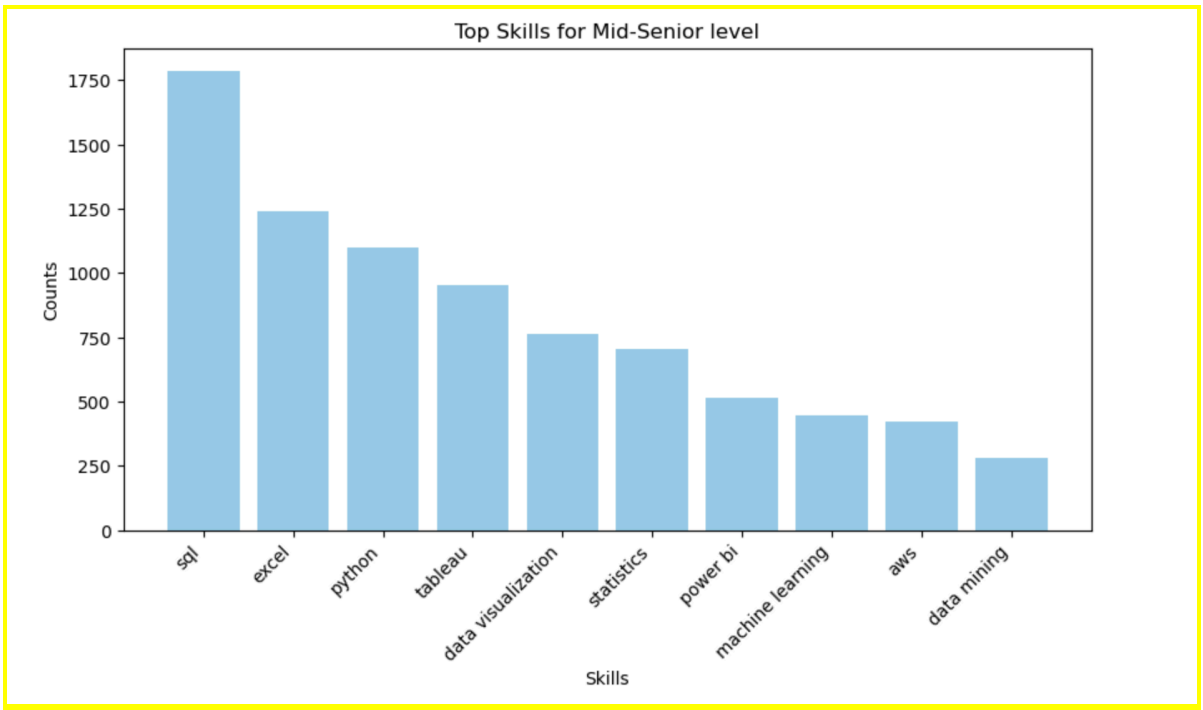


Figure 8: Top 10 Skills for Associate Level Roles

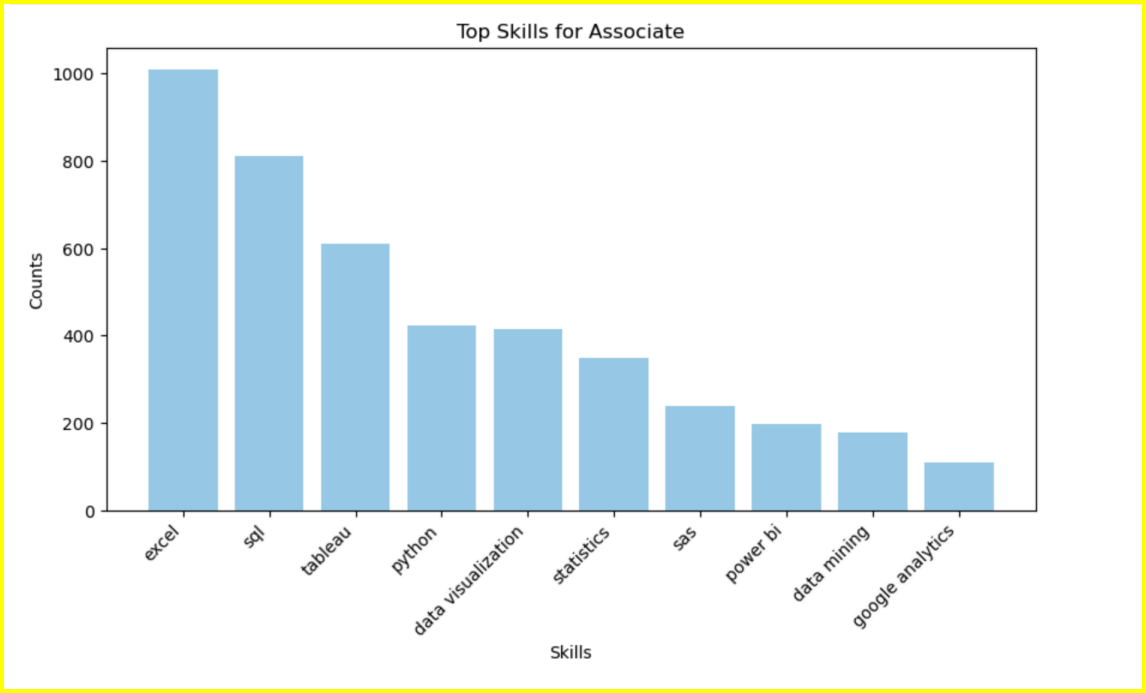
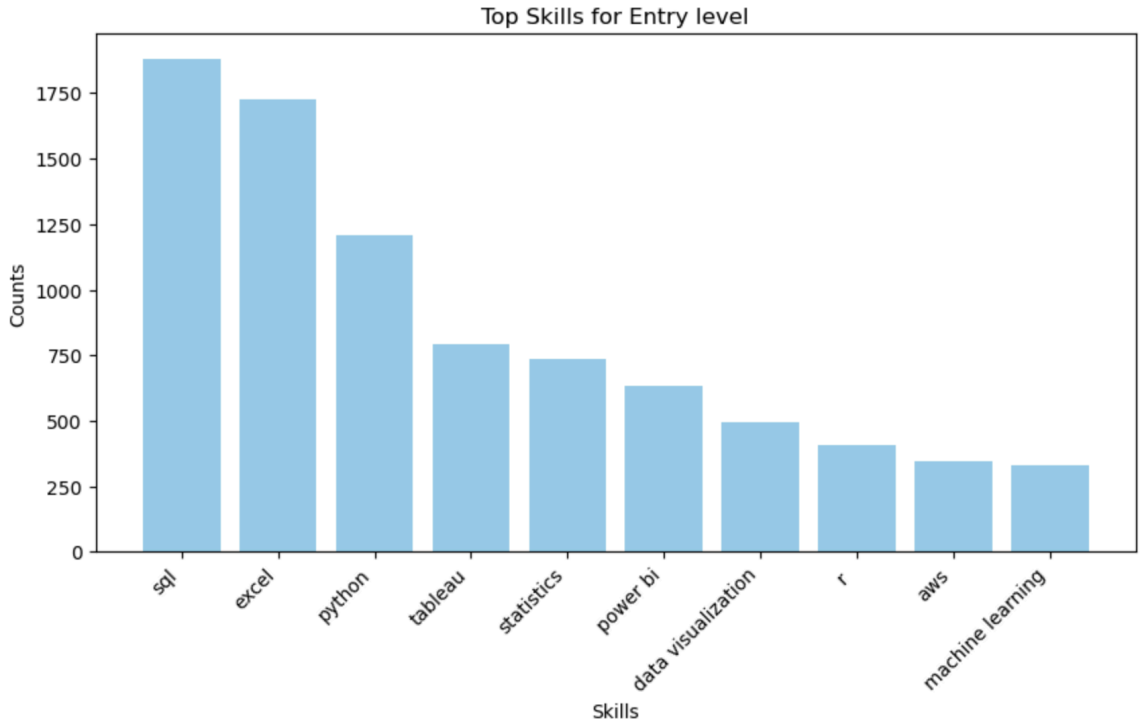


Figure 9: Top 10 Skills for Entry Level Roles



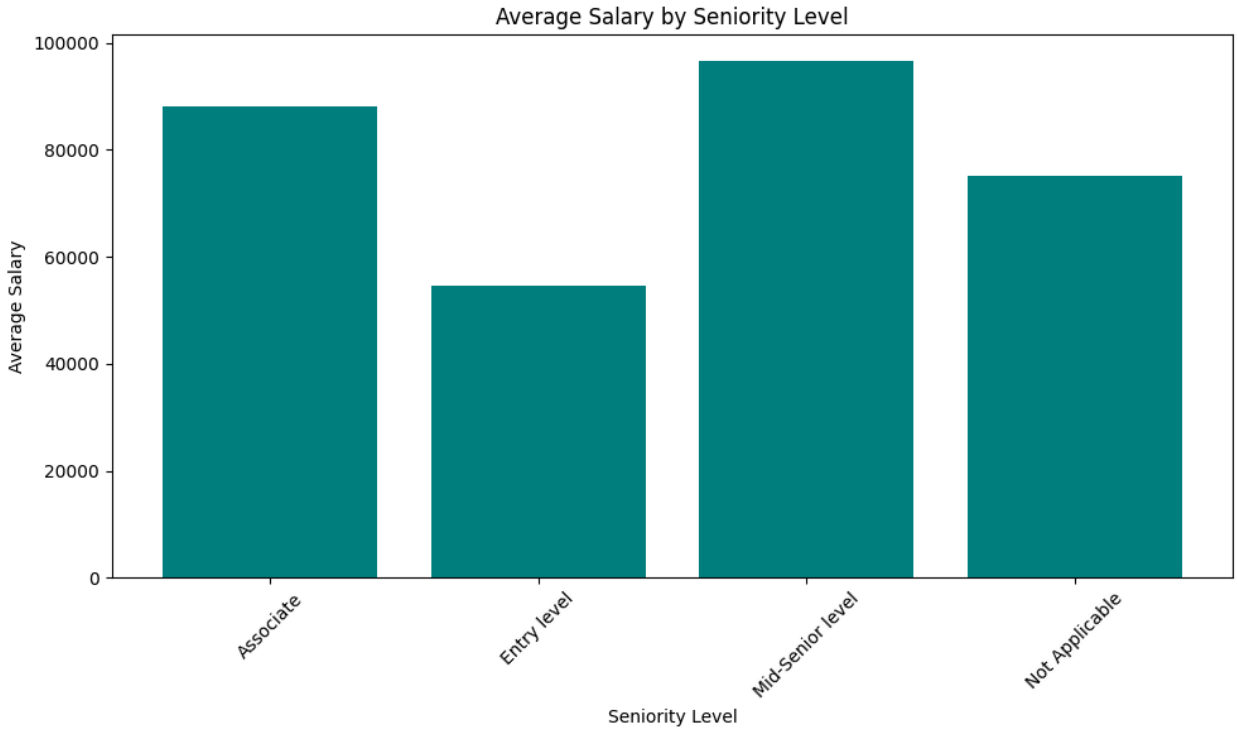
4. Salary Range Analysis:

Our analysis aimed to explore the relationship between seniority level and average salary. Notably, we focused solely on the USA dataset for salary analysis due to the absence of significant salary information in the Africa and Canada datasets, prompting us to exclude them from this analysis. Initially, all relevant criteria, including Industry, Job Function, and Employment Type, were consolidated within the "Criteria" column. However, unraveling this information posed challenges as not all keys within the column had corresponding values.

After meticulous examination, we successfully parsed the "Criteria" column, segregating each criterion into its own distinct column, thereby facilitating our analysis. We hypothesized that Mid-Senior Level roles would command higher salaries compared to entry-level positions. However, we were surprised to find that roles not specifying a seniority level exhibited noteworthy average salary levels, signifying that factors beyond traditional hierarchical structures influenced compensation.

This discovery underscores the evolving dynamics of the job market, where qualifications and skill sets often outweigh conventional measures of experience. It suggests a paradigm shift towards skill-based hiring practices, where candidates are evaluated based on their abilities and contributions rather than strictly adhering to traditional seniority levels. Thus, our analysis underscores the increasing importance of skill acquisition and expertise in shaping employment opportunities and compensation levels in the data analyst domain, particularly within the USA.

Figure 10: Average Salary by Seniority Level

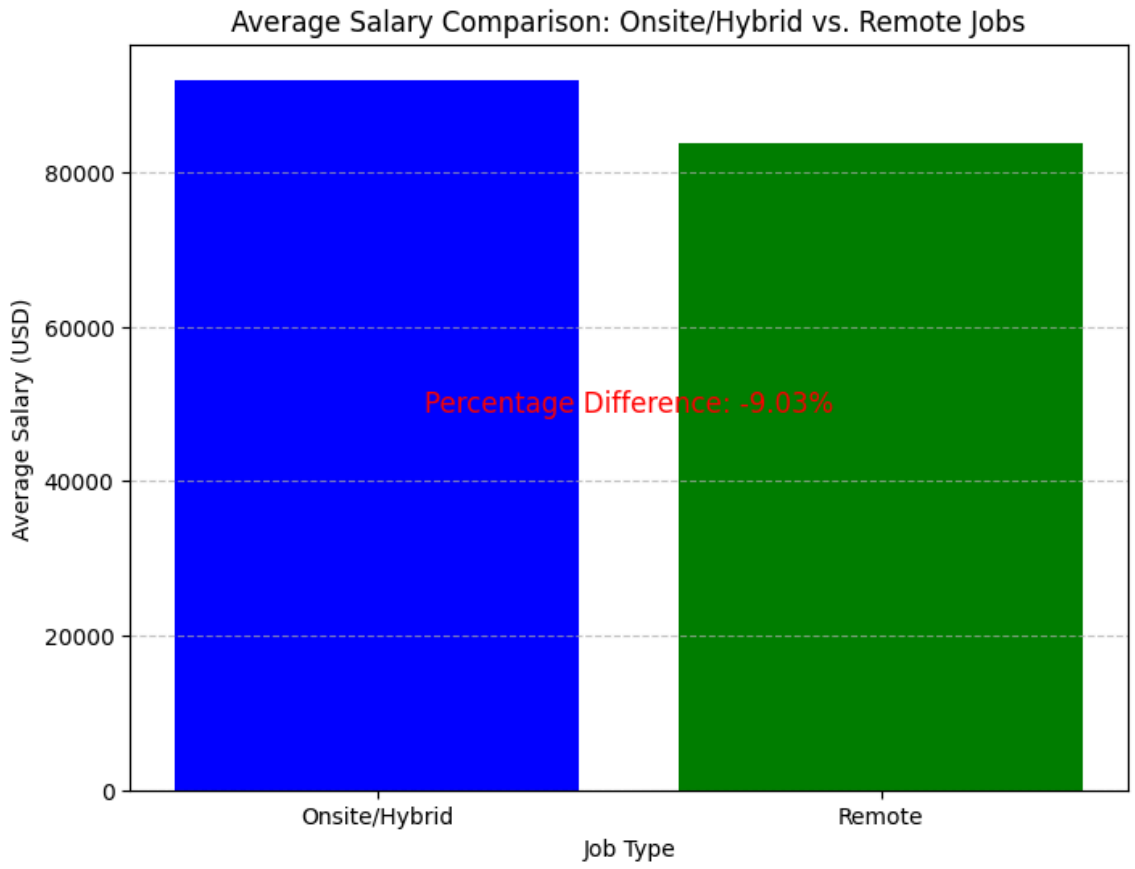


5. Salary Discrepancies Analysis:

With the goal of examining salary differentials between onsite and remote positions in the United States, we initially grouped the USA dataset by the 'onsite_hybrid' and 'remote' columns. This allowed us to compute the mean salary using the 'salary_converted' column for each category, resulting in the generation of the 'salary_comparison' DataFrame. Our analysis highlights that while onsite/hybrid roles generally offer higher compensation, with an approximate 9% disparity observed in our study, the slight discrepancy in salary is noteworthy. Moreover, we encourage contemplation of the potential impact of onsite work expenses on this variance, encompassing factors like commuting expenses and the opportunity cost associated with onsite employment. This sheds light on the evolving landscape of remote work in the USA,

where the argument for remote roles is gaining traction, particularly given the potential for enhanced productivity in some scenarios. As students, this resonates strongly, especially considering the commuting challenges many of us face. It initiates an intriguing discussion about the trade-offs between flexibility and monetary compensation, with the initial answer revealing a 9% differential in favor of onsite roles.

Figure 11: Average Salary Comparison for Onsite vs Remote Jobs in USA



Answering our Driving Question Directly:

Through our comprehensive analysis, we've illuminated crucial insights into the multifaceted landscape of factors influencing success for data analysts. We've identified the top skills sought by employers, shedding light on competencies crucial for professional advancement. Our examination of job distribution by location and seniority has provided nuanced perspectives on evolving work arrangements and industry demands. Moreover, our exploration of salary ranges and discrepancies has highlighted the intricate interplay between compensation, skill level, and job preferences. By synthesizing these findings, we've addressed our guiding question, elucidating how a combination of skills, industry focus, location preferences, seniority levels, and compensation considerations collectively drive success for data analysts in today's dynamic job market.

Section 5

Model and Results:

After discussing the possible methods that could potentially be used to create a prediction model, we decided to use the Least Squares method. To aid in our analysis we employed the statsmodel library. The dependent variable that aligned best with our analysis was salary. The dependent variables we wanted to include into our model were skills associated with job listings seeking data analysts.

Since many skills are sought after while searching for potential candidates, our model encompassed as many of those skills that were significant enough to keep. Our model yielded an R - squared of 0.499, adjusted R - squared of 0.491, and F - statistic of 60.74. The final regression equation was derived as follows:

$$\textbf{\textit{Salary}} = 8.42 + 3113.25X_{sql} + 4668.39X_{python} + 2.43X_{matlab} - 1.61X_{excel} + 1.51X_{tableau} + 1.41X_{power\ bi} + 1.68X_{java} + 4.53X_{data\ modeling} - 1.27X_{spark} + 2.24X_{hadoop} - 3.11X_{google\ analytics} - 1.22X_{aws} + 9105.28X_{machine\ learning} - 4.05X_{qlik}$$

Below summarizes the results of the predictors that made it into the final regression equation and their relationship to Salary.

Sql	Per skill count of sql, salary increases by $3113.25, assuming every other variable is held constant.
Python	Per skill count of python, salary increases by $4668.39, assuming every other variable is held constant.
Matlab	Per skill count of matlab, salary increases by $2.43, assuming every other variable is held constant.
Excel	Per skill count of excel, salary decreases by $1.61, assuming every other variable is held constant.

Tableau	Per skill count of tableau, salary increases by $1.51, assuming every other variable is held constant.
Power bi	Per skill count of power bi, salary increases by $1.41, assuming every other variable is held constant.
Java	Per skill count of java, salary increases by $1.68, assuming every other variable is held constant.
Data modeling	Per skill count of data modeling, salary increases by $4.53, assuming every other variable is held constant.
Spark	Per skill count of spark, salary decreases by $1.27, assuming every other variable is held constant.
Hadoop	Per skill count of hadoop, salary increases by $2.24, assuming every other variable is held constant.
Google analytics	Per skill count of google analytics, salary decreases by $3.11, assuming every other variable is held constant.
Aws	Per skill count of aws, salary decreases by $1.22, assuming every other variable is held constant.
Machine learning	Per skill count of machine learning, salary increases by $9105.28, assuming every other variable is held constant.
Qlik	Per skill count of qlik, salary decreases by $4.05, assuming every other variable is held constant.