**Overview**

Welcome to my analysis of the data job market, focusing on data analyst roles. This project was created out of a desire to navigate and understand the job market more effectively. It delves into the top-paying and in-demand skills to help find optimal job opportunities for data analysts.

The data sourced from [Luke Barousse's Python Course](https://lukebarousse.com/python) which provides a foundation for my analysis, containing detailed information on job titles, salaries, locations, and essential skills. Through a series of Python scripts, I explore key questions such as the most demanded skills, salary trends, and the intersection of demand and salary in data analytics.

**The Questions**

Below are the questions I want to answer in my project:

1. What are the skills most in demand for the top 3 most popular data roles?
2. How are in-demand skills trending for Data Analysts?
3. How well do jobs and skills pay for Data Analysts?
4. What are the optimal skills for data analysts to learn? (High Demand AND High Paying)

**Tools I Used**

For my deep dive into the data analyst job market, I harnessed the power of several key tools:

* **Python:** The backbone of my analysis, allowing me to analyze the data and find critical insights.I also used the following Python libraries:
  + **Pandas Library:** This was used to analyze the data.
  + **Matplotlib Library:** I visualized the data.
  + **Seaborn Library:** Helped me create more advanced visuals.
* **Jupyter Notebooks:** The tool I used to run my Python scripts which let me easily include my notes and analysis.
* **Visual Studio Code:** My go-to for executing my Python scripts.
* **Git & GitHub:** Essential for version control and sharing my Python code and analysis, ensuring collaboration and project tracking.

**Data Preparation and Cleanup**

This section outlines the steps taken to prepare the data for analysis, ensuring accuracy and usability.

**Import & Clean Up Data**

I start by importing necessary libraries and loading the dataset, followed by initial data cleaning tasks to ensure data quality.

# Importing Libraries

import ast

import pandas as pd

import seaborn as sns

from datasets import load\_dataset

import matplotlib.pyplot as plt

# Loading Data

dataset = load\_dataset('lukebarousse/data\_jobs')

df = dataset['train'].to\_pandas()

# Data Cleanup

df['job\_posted\_date'] = pd.to\_datetime(df['job\_posted\_date'])

df['job\_skills'] = df['job\_skills'].apply(lambda x: ast.literal\_eval(x) if pd.notna(x) else x)

**Filter India Jobs**

To focus my analysis on the India job market, I apply filters to the dataset, narrowing down to roles based in the India.

df\_US = df[df['job\_country'] == 'India']

**The Analysis**

Each Jupyter notebook for this project aimed at investigating specific aspects of the data job market. Here’s how I approached each question:

**1. What are the most demanded skills for the top 3 most popular data roles?**

To find the most demanded skills for the top 3 most popular data roles. I filtered out those positions by which ones were the most popular, and got the top 5 skills for these top 3 roles. This query highlights the most popular job titles and their top skills, showing which skills I should pay attention to depending on the role I'm targeting.

**Visualize Data**

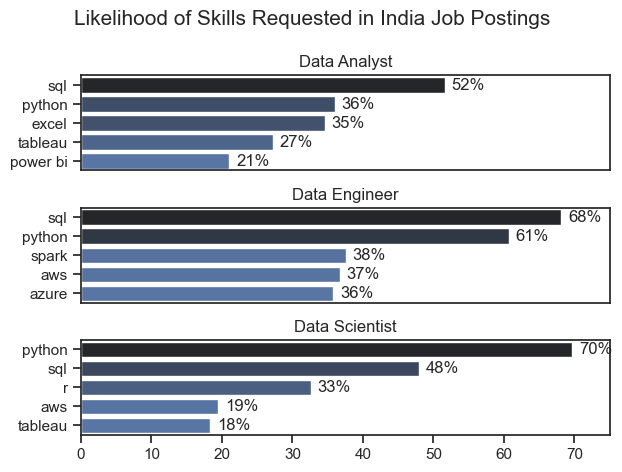
fig, ax = plt.subplots(len(job\_titles), 1)

for i, job\_title in enumerate(job\_titles):

df\_plot = df\_skills\_perc[df\_skills\_perc['job\_title\_short'] == job\_title].head(5)[::-1]

sns.barplot(data=df\_plot, x='skill\_percent', y='job\_skills', ax=ax[i], hue='skill\_count', palette='dark:b\_r')

plt.show()

**Results**

*Bar graph visualizing the salary for the top 3 data roles and their top 5 skills associated with each.*

**Insights:**

* SQL is the most requested skill for Data Analysts and Data Engineer with it in over half the job postings for both roles. For Data Scientist, Python is the most sought-after skill, appearing in 70% of job postings.
* Data Engineers require more specialized technical skills (AWS, Azure, Spark) compared to Data Analysts and Data Scientists who are expected to be proficient in more general data management and analysis tools (sql, Tableau).
* Python is a versatile skill, highly demanded across all three roles, but most prominently for Data Scientists (70%) and Data Engineers (61%).

**2. How are in-demand skills trending for Data Analysts?**

To find how skills are trending in 2023 for Data Analysts, I filtered data analyst positions and grouped the skills by the month of the job postings. This got me the top 5 skills of data analysts by month, showing how popular skills were throughout 2023.

**Visualize Data**

from matplotlib.ticker import PercentFormatter

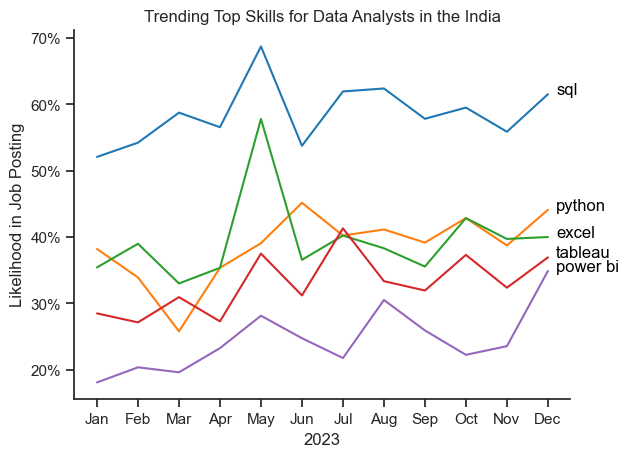
df\_plot = df\_DA\_US\_percent.iloc[:, :5]

sns.lineplot(data=df\_plot, dashes=False, legend='full', palette='tab10')

plt.gca().yaxis.set\_major\_formatter(PercentFormatter(decimals=0))

plt.show()

**Results**

  
*Bar graph visualizing the trending top skills for data analysts in the India in 2023.*

**Insights:**

* SQL remains the most consistently demanded skill throughout the year, although it shows a gradual decrease in demand.
* Excel experienced a significant increase in demand starting around May,
* Both Python and Tableau show relatively stable demand throughout the year with some fluctuations but remain essential skills for data analysts. Power BI, while less demanded compared to the others, shows a slight upward trend towards the year's end.

**3. How well do jobs and skills pay for Data Analysts?**

To identify the highest-paying roles and skills, I only got jobs in the India and looked at their median salary. But first I looked at the salary distributions of common data jobs like Data Scientist, Data Engineer, and Data Analyst, to get an idea of which jobs are paid the most.

**Visualize Data**

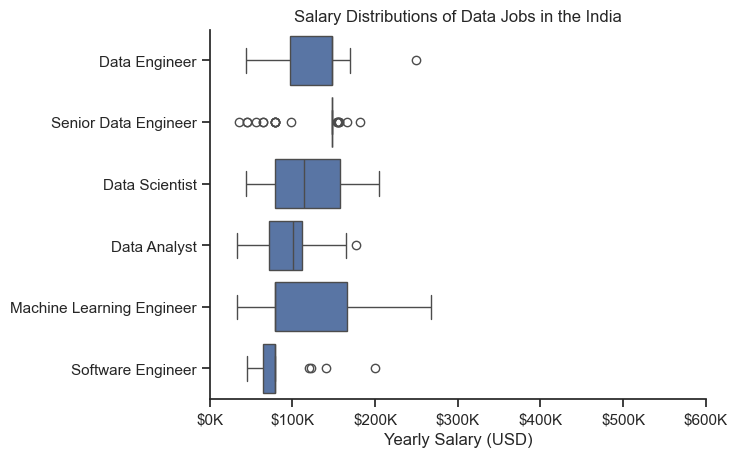
sns.boxplot(data=df\_US\_top6, x='salary\_year\_avg', y='job\_title\_short', order=job\_order)

ticks\_x = plt.FuncFormatter(lambda y, pos: f'${int(y/1000)}K')

plt.gca().xaxis.set\_major\_formatter(ticks\_x)

plt.show()

**Results**

  
*Box plot visualizing the salary distributions for the top 6 data job titles.*

**Insights**

* **Salary distributions vary significantly** across different data roles, with Machine Learning Engineers and Data Scientists showing higher earning potential compared to Data Analysts and Software Engineers.
* **Machine Learning Engineers** exhibit a broad salary range and high upper outliers, indicating strong demand and premium compensation for specialized AI/ML skills.
* **Senior Data Engineer** roles, while showing a narrower box, have a large number of outliers on the higher end, suggesting that exceptional expertise or niche roles command higher salaries.
* **Data Analysts** have a more compact and consistent salary distribution, reflecting more standardized pay for this role with fewer high-paying anomalies.
* **Software Engineers** in this context appear to earn lower salaries compared to data-focused roles, with a tighter range and more low-end outliers.

**Highest Paid & Most Demanded Skills for Data Analysts**

Next, I narrowed my analysis and focused only on data analyst roles. I looked at the highest-paid skills and the most in-demand skills. I used two bar charts to showcase these.

**Visualize Data**

fig, ax = plt.subplots(2, 1)

# Top 10 Highest Paid Skills for Data Analysts

sns.barplot(data=df\_DA\_top\_pay, x='median', y=df\_DA\_top\_pay.index, hue='median', ax=ax[0], palette='dark:b\_r')

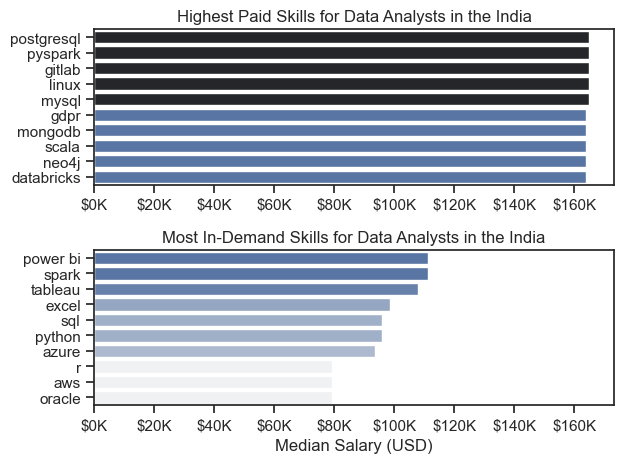
# Top 10 Most In-Demand Skills for Data Analystsr')

sns.barplot(data=df\_DA\_skills, x='median', y=df\_DA\_skills.index, hue='median', ax=ax[1], palette='light:b')

plt.show()

**Results**

Here's the breakdown of the highest-paid & most in-demand skills for data analysts in the India:

 *Two separate bar graphs visualizing the highest paid skills and most in-demand skills for data analysts in the India.*

**Insights:**

 The top chart reveals that **specialized backend and data engineering skills** such as PostgreSQL, PySpark, and GitLab command the **highest median salaries** for Data Analysts in India, often approaching or exceeding **$150K–$160K**.

 In contrast, the bottom chart highlights that tools like **Power BI, Spark, Tableau, Excel, and SQL** are the **most in-demand**, even though they offer relatively **lower median salaries**, generally under **$100K**.

 This indicates a clear distinction between **high-paying niche skills** and **high-demand foundational tools**. Mastery of in-demand tools ensures employability, while expertise in specialized platforms can lead to **significantly higher compensation**.

 To **maximize career growth**, aspiring Data Analysts should focus on a **hybrid skill set**—building a strong base with tools like SQL and Power BI while also gaining **depth in specialized technologies** like PySpark, GitLab, or PostgreSQL.

**4. What are the most optimal skills to learn for Data Analysts?**

To identify the most optimal skills to learn ( the ones that are the highest paid and highest in demand) I calculated the percent of skill demand and the median salary of these skills. To easily identify which are the most optimal skills to learn.

**Visualize Data**

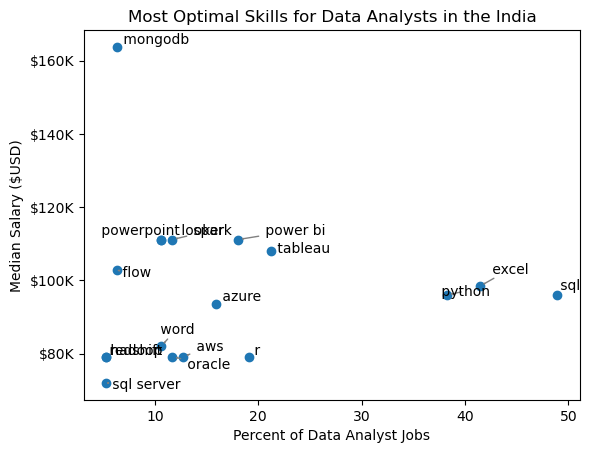
from adjustText import adjust\_text

import matplotlib.pyplot as plt

plt.scatter(df\_DA\_skills\_high\_demand['skill\_percent'], df\_DA\_skills\_high\_demand['median\_salary'])

plt.show()

**Results**

  
*A scatter plot visualizing the most optimal skills (high paying & high demand) for data analysts in the India.*

**Insights:**

 **MongoDB** stands out with the **highest median salary (~$165K)** despite being required in a **small percentage of data analyst jobs**, making it a highly rewarding niche skill.

 **SQL** and **Excel** are the **most frequently required** skills, appearing in nearly **50% and 40%** of job listings respectively. However, their median salaries (~$95K–$100K) are **modest** compared to specialized tools.

 **Power BI**, **Tableau**, and **Spark** strike a **balance between demand and salary**, making them some of the most **optimal tools** for data analysts to learn. They offer both solid pay (~$110K) and moderate job presence (20–25%).

 Tools like **PowerPoint**, **Looker**, and **Flow** also offer **above-average salaries** with **lower job competition**, indicating a good return on investment for niche use cases.

 Foundational cloud and scripting tools such as **Azure**, **AWS**, **R**, and **Oracle** are moderately in demand but offer **relatively lower compensation**, suggesting they are valuable support skills rather than primary differentiators.

**Visualizing Different Techonologies**

Let's visualize the different technologies as well in the graph. We'll add color labels based on the technology (e.g., {Programming: Python})

**Visualize Data**

from matplotlib.ticker import PercentFormatter

# Create a scatter plot

scatter = sns.scatterplot(

data=df\_DA\_skills\_tech\_high\_demand,

x='skill\_percent',

y='median\_salary',

hue='technology', # Color by technology

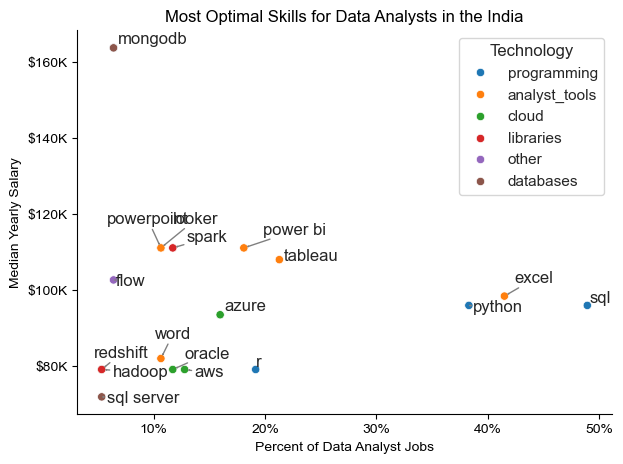
palette='bright', # Use a bright palette for distinct colors

legend='full' # Ensure the legend is shown

)

plt.show()

**Results**

  
*A scatter plot visualizing the most optimal skills (high paying & high demand) for data analysts in the India with color labels for technology.*

**Insights:**

 The scatter plot highlights that **programming skills** (blue), such as **Python** and **SQL**, are the most **widely required** in job listings and provide **stable salary levels**, making them foundational for anyone entering the data analytics field.

 **Database technologies** (brown), like **MongoDB**, stand out with the **highest salary potential (~$165K)** despite lower demand, indicating that **specialized data handling skills** are highly valued in niche roles.

 **Analyst tools** (orange), including **Power BI** and **Tableau**, offer a **strong balance** of **moderate-to-high demand** and **competitive median salaries**, emphasizing their importance in transforming data into actionable insights.

 **Libraries** and **specialized tools** (red), such as **Spark** and **Looker**, while less common in postings, tend to offer **above-average compensation**, signaling their value in **advanced analytics** or **big data environments**.

 **Cloud skills** (green), such as **AWS** and **Azure**, though not the highest in salary, appear consistently across postings, reflecting their **growing relevance** in data infrastructure and deployment.

 Less common tools like **Flow** and **Redshift** (purple/red) still command decent pay, showing that **niche or domain-specific tools** can offer good returns for specialized job roles.

**What I Learned**

Throughout this project, I deepened my understanding of the data analyst job market and enhanced my technical skills in Python, especially in data manipulation and visualization. Here are a few specific things I learned:

* **Advanced Python Usage**: Utilizing libraries such as Pandas for data manipulation, Seaborn and Matplotlib for data visualization, and other libraries helped me perform complex data analysis tasks more efficiently.
* **Data Cleaning Importance**: I learned that thorough data cleaning and preparation are crucial before any analysis can be conducted, ensuring the accuracy of insights derived from the data.
* **Strategic Skill Analysis**: The project emphasized the importance of aligning one's skills with market demand. Understanding the relationship between skill demand, salary, and job availability allows for more strategic career planning in the tech industry.

**Insights**

This project provided several general insights into the data job market for analysts:

* **Skill Demand and Salary Correlation**: There is a clear correlation between the demand for specific skills and the salaries these skills command. Advanced and specialized skills like Python and Oracle often lead to higher salaries.
* **Market Trends**: There are changing trends in skill demand, highlighting the dynamic nature of the data job market. Keeping up with these trends is essential for career growth in data analytics.
* **Economic Value of Skills**: Understanding which skills are both in-demand and well-compensated can guide data analysts in prioritizing learning to maximize their economic returns.

**Challenges I Faced**

This project was not without its challenges, but it provided good learning opportunities:

* **Data Inconsistencies**: Handling missing or inconsistent data entries requires careful consideration and thorough data-cleaning techniques to ensure the integrity of the analysis.
* **Complex Data Visualization**: Designing effective visual representations of complex datasets was challenging but critical for conveying insights clearly and compellingly.
* **Balancing Breadth and Depth**: Deciding how deeply to dive into each analysis while maintaining a broad overview of the data landscape required constant balancing to ensure comprehensive coverage without getting lost in details.

**Conclusion**

This exploration into the data analyst job market has been incredibly informative, highlighting the critical skills and trends that shape this evolving field. The insights I got enhance my understanding and provide actionable guidance for anyone looking to advance their career in data analytics. As the market continues to change, ongoing analysis will be essential to stay ahead in data analytics. This project is a good foundation for future explorations and underscores the importance of continuous learning and adaptation in the data field.